SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 1–30, January 1996

INFINITE-HORIZON LINEAR-QUADRATIC REGULATOR PROBLEMS FOR NONAUTONOMOUS PARABOLIC SYSTEMS WITH BOUNDARY CONTROL*

PAOLO ACQUISTAPACE† AND BRUNELLO TERRENI‡

Abstract. This paper concerns the classical linear-quadratic regulator problem for general nonautonomous parabolic systems with boundary control over infinite time horizon from the point of view of semigroup theory. Under appropriate assumptions we prove existence and uniqueness of the optimal pair, as well as existence, uniqueness, and further properties of the solution of the associated Riccati equation. Several examples are discussed in detail.

Key words. optimal control, parabolic systems, boundary control, infinite horizon, Riccati equation

AMS subject classifications. 49J20, 49N10, 49L20, 34G20

Introduction. This paper concerns the classical linear-quadratic regulator (LQR) problem for general nonautonomous parabolic systems with boundary control over infinite time horizon. Under appropriate assumptions we prove here existence and uniqueness of the optimal pair as well as existence, uniqueness, and further properties of the solution of the associated Riccati equation. Such results generalize the similar ones known in the autonomous case [F3], [LT2], [BDDM] and those of [DI3] relative to nonautonomous problems with distributed control; they also constitute a development of the theory of [AFT] concerning the case of finite time horizon.

Our assumptions are, generally speaking, not uniform with respect to t, with few exceptions concerning the spectra of the elliptic operators $A(t)$ appearing in the state equation and the regularity of the Green maps $G(t)$ associated with them: see Hypotheses 1.1 and 1.3 below. In particular, we do not assume any global exponential estimate for the evolution operator $U(t,s)$ or any boundedness for $G(t)$ and the operators appearing in the cost functional.

On the other hand, some uniform requirements arise in the study of certain features of the Riccati equation. Thus, in order to construct a minimal solution $P_\infty(t)$ of such an equation and to solve the synthesis, we need the "finite cost condition" (Hypothesis 2.2), which is necessary and sufficient; moreover a uniform version of this condition (Hypothesis 3.1) is necessary and sufficient for the existence of a bounded solution of the Riccati equation. Further uniform assumptions (Hypotheses 3.4, 3.5, 3.6 and 3.9) guarantee other properties, such as stability of the optimal state and uniqueness of $P_\infty(t)$. The periodic case is also analyzed. All these results seem to be new even in the case of distributed control of [DI3].

We now list some notations. If X is a Hilbert space, we denote the inner product and the norm of X by $(\cdot\,,\cdot)_X$ and $\|\cdot\|_X$. If Y is another Hilbert space, $\mathcal{L}(X,Y)$ is the Banach space of bounded linear operators from X into Y, and $|\cdot|_{\mathcal{L}(X,Y)}$ denotes its usual norm; we write $\mathcal{L}(X)$ instead of $\mathcal{L}(X,X)$.

If $A : D(A) \subseteq X \to Y$ is a closed linear operator with dense domain, the adjoint operator $A^* : D(A^*) \subseteq Y^* \to X^*$ is defined in the usual way. In particular we denote by $\Sigma(X)$ the set of operators $A \in \mathcal{L}(X)$ such that $A = A^*$, and we set

$$\Sigma^+(X) := \{A \in \Sigma(X) : (Ax,x)_X \geq 0 \ \forall x \in X\},$$
$$\Sigma^{++}(X) := \{A \in \Sigma^+(X) : \exists \nu > 0 : (Ax,x)_X \geq \nu\|x\|_X \ \forall x \in X\}.$$

If $I \subseteq \mathbb{R}$ is an interval, we will use the spaces $L^p(I,X) := \{f : I \to X : f$ is strongly

* Received by the editors February 10, 1993; accepted for publication (in revised form) August 9, 1994.
† Dipartimento di Matematica, Università di Pisa, Via F. Buonarroti 2, 56127 Pisa, Italy.
‡ Dipartimento di Matematica Pura ed Applicata, Università dell'Aquila, Via Vetoio, 67010 Coppito (L'Aquila), Italy.

measurable and $\int_I \|f(t)\|_X^p\, dt < \infty\}$ $(1 \le p < \infty)$, and $L^\infty(I,X)$, $C(I,X)$, whose definitions are similar. Finally we will also use the spaces

$$L^p_{\mathrm{loc}}(I,X) := \bigcap_{J \subset\subset I} L^p(J,X) \quad (1 \le p \le \infty).$$

1. Problem formulation and hypotheses.

1.1. Abstract formulation of a parabolic differential system. Let $\Omega \subset R^d$ be a bounded open domain with boundary $\partial\Omega$ of class C^2, and consider the following parabolic system:

$$\begin{cases} y_t(t,x) = \mathcal{A}(t,x,D)y(t,x) & \text{in } [0,\infty[\times \Omega, \\ \mathcal{B}(t,x,D)y(t,x) = u(t,x) & \text{in } [0,\infty[\times \partial\Omega, \\ y(0,x) = y_0(x) & \text{in } \Omega. \end{cases} \tag{1.1}$$

Here the strongly elliptic differential operators $\{\mathcal{A}(t,\cdot\,,D)\}_{t\ge 0}$ and the boundary operators $\{\mathcal{B}(t,\cdot\,,D)\}_{t\ge 0}$ are assumed to be such that the abstract hypotheses listed in the next subsection are satisfied (see, for instance, the conditions of [AFT, §2.2]).

For each $t \ge 0$ we define $A(t)$ as the realization in $L^2(\Omega)$ of the operator $\mathcal{A}(t,\cdot\,,D)$ with homogeneous boundary conditions determined by $\mathcal{B}(t,\cdot\,,D)$, i.e.,

$$\begin{aligned} &D_{A(t)} := \{y \in L^2(\Omega) \mid \mathcal{A}(t,\cdot\,,D)y \in L^2(\Omega) \text{ and } \mathcal{B}(t,\cdot\,,D)y = 0 \text{ on } \partial\Omega\}, \\ &A(t)y := \mathcal{A}(t,\cdot\,,D)y \quad \forall\, y \in D_{A(t)}. \end{aligned}$$

If we choose $\lambda_0 \in \mathbb{R}$ large enough, we can define simultaneously (for $t \ge 0$) the fractional powers $(\lambda_0 - A(t))^\alpha$ with $0 < \alpha < 1$. We also require that for each $u \in L^2(\partial\Omega)$ we can solve (in the sense of [AFT, §2.4]) simultaneously for $t \ge 0$ the following elliptic problems:

$$\begin{cases} \lambda_0\Phi - \mathcal{A}(t,\cdot\,,D)\Phi = 0 & \text{in } \Omega, \\ \mathcal{B}(t,\cdot\,,D)\Phi = u & \text{on } \partial\Omega; \end{cases} \tag{1.2}$$

in other words, we can define the map $G : [0,\infty) \times L^2(\partial\Omega) \to L^2(\Omega)$ as $G(t)u := \Phi$, where Φ is the unique solution of problem (1.2). In the next section we shall need certain regularity properties for $G(t)$: for instance, we shall assume that

$$t \to (\lambda_0 - A(t))^\alpha G(t) \in L^\infty_{\mathrm{loc}}([0,\infty[; \mathcal{L}(L^2(\partial\Omega), L^2(\Omega)))$$

for some $\alpha \in]0,1[$. It is shown in [AFT] that systems of type (1.1) fulfill this condition.

We remark also that the map G depends on the initial choice of λ_0.

Let

$$z(t) := e^{-\lambda_0 t} y(t),$$

where y solves problem (1.1); then z solves the following problem:

$$\begin{cases} z_t = (\mathcal{A}(t,\cdot\,,D) - \lambda_0)z & \text{in } [0,\infty[\times \Omega, \\ \mathcal{B}(t,\cdot\,,D)z = e^{-\lambda_0 t}u(t) & \text{on } [0,\infty[\times \partial\Omega, \\ z(0) = y_0 & \text{in } \Omega, \end{cases}$$

so using the representation formula proved in [AFT, §2.5] we have

$$z(t) = U_{\lambda_0}(t,0)y_0 + \int_0^t U_{\lambda_0}(t,s)(\lambda_0 - A(s))G(s)u(s)e^{-\lambda_0 s}\, ds. \tag{1.3}$$

Here $U_{\lambda_0}(t,s) := e^{-\lambda_0(t-s)}U(t,s)$ where $U(t,s)$ is the evolution operator associated with $\{A(t)\}_{t\geq 0}$. Thus

$$y(t) = U(t,0)y_0 + \int_0^t U(t,s)(\lambda_0 - A(s))G(s)u(s)\,ds. \tag{1.4}$$

We must recall that formula (1.4) is very useful for our calculations, but we understand that its exact form is

$$y(t) = U(t,0)y_0 + \int_0^t [(\lambda_0 - A(s)^*)^{1-\alpha}U(t,s)^*]^*((\lambda_0 - A(s))^\alpha G(s)u(s)\,ds; \tag{1.5}$$

for more details we refer to [AFT, (2.80) – (2.74)].

Throughout this paper, equation (1.4) or (1.5) will be considered the state equation for our abstract control problem.

1.2. Standing assumptions. In the following discussion we will consider three Hilbert spaces: H (state space), U (control space), and V (space of observations), and we will study the optimal control of equation (1.4) over an unbounded time interval I (which could be $[T_0,\infty[$ for some $T_0 \in \mathbb{R}$, or even $\mathbb{R}$, as for instance in the periodic case). We will consider equation (1.4) as an abstract evolution equation in the Hilbert space H, subject to the abstract assumptions listed below. Thus, equation (1.4) can also cover concrete problems different from those explicitly described above and in the examples in §4. We assume the following hypotheses.

Hypothesis 1.1. $\{A(t)\}_{t\in I}$ is a family of infinitesimal generators of analytic semigroups in H; the spectrum of $A(t)$ is such that the fractional powers $(\lambda_0 - A(t))^\alpha$ are well defined for any $\alpha > 0$, simultaneously with respect to $t \in I$, for some fixed $\lambda_0 \in \mathbb{R}$.

Hypothesis 1.2. The assumptions of [AFT] hold locally, i.e., over every bounded interval $J \subset\subset I$, possibly with constants depending on the interval. More precisely, given $J = [a,b] \subset I$, we assume

(i) the evolution operators $U(t,s)$ and $U(t,s)^*$ are strongly continuous in $\bar{\Delta}$, where $\Delta := \{(t,s) \in [a,b]^2 : t > s\}$, and there exists $M_0 > 0$ such that

$$|U(t,s)|_{\mathcal{L}(H)} + |U(t,s)^*|_{\mathcal{L}(H)} \leq M_0 \quad \forall\,(t,s) \in \bar{\Delta};$$

(ii) for every $\beta,\ \mu \in [-1,1]$ and $(t,s) \in \Delta$ the operators

$$(\lambda_0 - A(t))^\beta U(t,s)(\lambda_0 - A(s))^{-\mu}, \qquad (\lambda_0 - A(s)^*)^\beta U(t,s)^*(\lambda_0 - A(t)^*)^{-\mu}$$

have continuous extensions to H, the maps

$$(t,s) \to (\lambda_0 - A(t))^\beta U(t,s)(\lambda_0 - A(s))^{-\mu}, \quad (t,s) \to (\lambda_0 - A(s)^*)^\beta U(t,s)^*(\lambda_0 - A(t)^*)^{-\mu}$$

are strongly continuous, and there exists $M_{\beta,\mu} > 0$ such that

$$\begin{aligned} &|(\lambda_0 - A(t))^\beta U(t,s)(\lambda_0 - A(s))^{-\mu}|_{\mathcal{L}(H)} + |(\lambda_0 - A(s)^*)^\beta U(t,s)^*(\lambda_0 - A(t)^*)^{-\mu}|_{\mathcal{L}(H)} \\ &\qquad \leq M_{\beta,\mu}[(t-s)^{\mu-\beta} + 1] \quad \forall\,(t,s) \in \Delta. \end{aligned} \tag{1.6}$$

Hypothesis 1.3. There exists $\alpha \in\,]0,1]$ such that, for each $t \in I$, $G(t)$ maps U into the domain of $(\lambda_0 - A(t))^\alpha$, and $(\lambda_0 - A(\cdot))^\alpha G(\cdot)$ is strongly measurable and bounded over each $[a,b] \subset I$.

Remark 1.4. (i) The only novelty with respect to [AFT] is the uniformity with respect to t in Hypotheses 1.1 and 1.3 for the choices of λ_0 and α, respectively. In particular, we do not assume any global exponential estimate for $U(t,s)$ and $(\lambda_0 - A(t))^\beta U(t,s)(\lambda_0 - A(s))^{-\mu}$ or uniform boundedness for $(\lambda_0 - A(t)^*)^\alpha G(t)$, $C(t)$, $N(t)$, and $N(t)^{-1}$ (defined below). Under these assumptions, the representation formulas (1.4), (1.5) can be studied as in [AFT] over any bounded time interval $[a,b]$.

(ii) In §2 the sentence "c depends on $[a,b]$" will mean that c in fact depends, besides $[a,b]$ itself, on all constants involved in estimating functions and operators defined in $[a,b]$. In §3 some questions of stability are treated, and there we shall assume the necessary uniformities and point out the independence of the constants.

(iii) We formulated Hypothesis 1.2 in terms of the evolution operators $U(t,s)$ and $U(t,s)^*$, rather than of the family $\{A(t)\}_{t\geq 0}$; this choice is motivated by the existence in the literature of many independent sets of assumptions on the family $\{A(t)\}_{t\geq 0}$, each of which implies Hypothesis 1.2. In [AT1] and [A1] one can find a review of these assumptions.

1.3. Formulation of the infinite-horizon LQR problem. Given $t_0 \in I$ and $y_0 \in H$, we will consider the problem of minimizing the cost functional

$$J_{t_0,\infty}(u) := \int_{t_0}^{\infty} [\|C(t)y(t)\|_V^2 + (N(t)u(t), u(t))_U]\, dt \tag{1.7}$$

over all $u \in L^2_{\text{loc}}(t_0, \infty; U)$, subject to the state equation (1.5).

Concerning $C(\cdot)$ and $N(\cdot)$ we assume the following hypothesis.

Hypothesis 1.5. For every interval $[a,b] \subset I$,

$$C(\cdot) \in L^\infty([a,b]; \mathcal{L}(H,V)) \quad \text{and} \quad N(\cdot) \in L^\infty([a,b]; \Sigma^{++}(U)).$$

(This means that there exists $\nu > 0$, possibly depending on $[a,b]$, such that $N(t) \geq \nu\, \forall\, t \in [a,b]$, i.e., $(N(t)u,u)_U \geq \nu \|u\|_U^2\, \forall\, u \in U$, $\forall\, t \in [a,b]$.)

2. Solutions of the Riccati equation and related questions. The aim of this section is to solve the integral Riccati equation

$$\begin{aligned} Q(s) &= U(t,s)^* Q(t) U(t,s) \\ &\quad + \int_s^t U(r,s)^* [C(r)^* C(r) \\ &\qquad - Q(r)(\lambda_0 - A(r)) G(r) N(r)^{-1} G(r)^* (\lambda_0 - A(r)^*) Q(r)] U(r,s)\, dr, \end{aligned} \tag{2.1}$$

with $s,\ t \in I,\ s \leq t$, and to prove some further results related to this equation. We follow here along the lines of [F3], *mutatis mutandis*; in particular we shall need a new local existence result (see Theorem 2.5 below).

We recall that a solution $Q(\cdot)$ of equation (2.1) is an operator-valued function $Q \in C_s(I; \Sigma^+(H))$ such that (i) $t \to (\lambda_0 - A(t)^*)^{1-\alpha} Q(t)$ is well defined and strongly continuous from I into $\mathcal{L}(H)$, and (ii) $Q(\cdot)$ satisfies the following meaningful version of (2.1):

$$\begin{aligned} Q(s) &= U(t,s)^* Q(t) U(t,s) \\ &\quad + \int_s^t U(r,t)^* [C(r)^* C(r) \\ &\qquad - [(\lambda_0 - A(r)^*)^{1-\alpha} Q(r)]^* K(r) [(\lambda_0 - A(r)^*)^{1-\alpha} Q(r)]] U(r,s)\, dr \end{aligned}$$

with $s,\ t \in I,\ s \le t$, and where

$$K(r) := [(\lambda_0 - A(r))^{\alpha} G(r)] N(r)^{-1} [(\lambda_0 - A(r))^{\alpha} G(r)]^*;$$

but in order to simplify our calculations and for sake of clearness we will always use equation (2.1).

2.1. Definition of the operator $P_\infty(t)$. Let us denote by $P_T(t)$ the solution, defined for every $t \le T$ (and $t \in I$; we will not repeat this detail in the following discussion), of Riccati equation (2.1) with value 0 at $t = T$:

$$P_T(t) = \int_t^T U(r,t)^* [C(r)^* C(r) - [(\lambda_0 - A(r)^*)^{1-\alpha} P_T(r)]^* K(r) [(\lambda_0 - A(r)^*)^{1-\alpha} P_T(r)]] U(r,s)\, dr. \tag{2.2}$$

This equation was solved in [AFT, Thm. 3.13]. We recall that

$$(P_T(t_0) y_0, y_0)_H = \min_u J_{t_0,T}(u) \tag{2.3}$$

under the condition $y(t_0) = y_0$, where for each $u \in L^2_{\mathrm{loc}}(t_0, \infty; U)$ the functional $J_{t_0,T}(u)$ is defined as

$$J_{t_0,T}(u) = \int_{t_0}^T [\|C(t) y(t)\|_V^2 + (N(t) u(t), u(t))_U]\, dt$$

and y satisfies the state equation (1.4).

Let us introduce the following important condition:

$$(2.4_{t_0}) \quad \begin{cases} \text{there exists } c = c(t_0) > 0 \text{ such that to each } y_0 \in H \text{ there corresponds} \\ \text{a control } u = u(y_0) \in L^2_{\mathrm{loc}}(t_0, \infty; U) \text{ for which } J_{t_0,\infty}(u) \le c \|y_0\|_H^2; \end{cases}$$

in other words condition (2.4_{t_0}) requires the existence of an admissible control with respect to a given $t_0 \in I$ for each initial state $y_0 \in H$.

The following lemma holds.

LEMMA 2.1. *Under Hypotheses* 1.1 – 1.3 *and* 1.5, *we have*

(i) $P_{T_1}(t) \le P_{T_2}(t)$ *for each* $t \le T_1 \le T_2$;

(ii) *condition* (2.4_{t_1}) *implies condition* (2.4_t) *for each* $t \le t_1$;

(iii) *if condition* (2.4_{t_0}) *holds, then* $\sup_{T \ge t_0} |P_T(t_0)|_{\mathcal{L}(H)} < \infty$ *and there exists* $P_\infty(t_0) \in \Sigma^+(H)$ *such that* $P_T(t_0) \uparrow P_\infty(t_0)$ *strongly as* $T \uparrow \infty$.

(iv) *if condition* (2.4_{t_0}) *holds, then for each fixed* $\tau_0 < t_0$ *we have*

$$\sup_{\tau_0 \le t \le t_0, T > t_0} |P_T(t)|_{\mathcal{L}(H)} < \infty; \tag{2.5}$$

moreover $P_\infty(t)$ *is well defined for each* $t \le t_0$, *and*

$$P_\infty(s) \le U(t,s)^* P_\infty(t) U(t,s) + \int_s^t U(r,s)^* C(r)^* C(r) U(r,s)\, dr; \tag{2.6}$$

in particular, for each fixed $\tau_0 < t_0$ *we have*

$$\sup_{\tau_0 \le t \le t_0} |P_\infty(t)|_{\mathcal{L}(H)} < \infty. \tag{2.7}$$

Proof. It is standard and follows from [F1], [AFT]. □

Our main assumption in order to obtain a solution of equation (2.1) is the following hypothesis.

Hypothesis 2.2. Condition (2.4_t) is satisfied for every $t \in I$.

By the preceding lemma, if Hypothesis 2.2 holds, we can define $P_\infty(t)$ for each $t \in I$; this operator-valued function is the candidate solution of the Riccati equation.

Remark 2.3. As we will see, Hypothesis 2.2 is necessary and sufficient to construct $P_\infty(t)$ and to solve the synthesis; hence it is important to know when it is satisfied in concrete examples. Two general remarks in this direction are the following:

(i) if system (1.4) is exactly controllable at 0 in finite time, starting from any time t_0 and initial position y_0, then Hypothesis 2.2 holds;

(ii) if (1.4) is exponentially stabilizable, starting from any t_0, and if $C(t)$ and $N(t)$ are uniformly bounded, then Hypothesis 2.2 holds.

The analysis of these properties in concrete cases is under investigation; however, for certain classes of systems property (ii), hence Hypothesis 2.2, has been proved to hold true (see, for instance, Example 4.2 below). Note also that in the case of periodic systems it is sufficient to show that (2.4_t) is satisfied for some $t \in \mathbb{R}$, because this implies that (2.4_s) holds for each $s < t$ and thus for each s by periodicity.

2.2. A priori bound on $(\lambda_0 - A(t)^*)^\mu P_T(t)$. The following result plays a basic role in solving the Riccati equation (2.1). The same result was proved in [F1] in the case of autonomous parabolic systems, but the proof given in [F1] cannot be extended (at least in an obvious way) to the present case; thus, the proof given here is new. See also [F2], where a similar proof provides the a priori bound needed to get a global solution over a finite time horizon.

LEMMA 2.4. *Assume Hypotheses* 1.1 – 1.3, 1.5, *and* 2.2. *Then for each* $\beta \in]0, 1/2[$ *and each interval* $[a, b] \subset I$, *we have*

$$\sup_{a \le t \le b+1, T > b+2} |(\lambda_0 - A(t)^*)^\beta P_T(t)(\lambda_0 - A(t))^\beta|_{\mathcal{L}(H)} =: c_1(\beta, [a, b]) < +\infty, \tag{2.8}$$

$$\sup_{a \le t \le b+1} |(\lambda_0 - A(t)^*)^\beta P_\infty(t)(\lambda_0 - A(t))^\beta|_{\mathcal{L}(H)} < +\infty. \tag{2.9}$$

Proof. Let us fix an interval $[a, b] \subset I$. Fix $t \in [a, b+1]$, $x \in D((\lambda_0 - A(t))^\beta)$ and set $y_1 := U(b+2, t)(\lambda_0 - A(t))^\beta x$. By Hypothesis 2.2 there exists a control $\bar{u}$ belonging to $L^2_{\mathrm{loc}}(b+2, \infty; U)$ such that $J_{b+2,\infty}(\bar{u}) \le c\|y_1\|_H^2$; hence by Hypothesis 1.2(ii) we also have $J_{b+2,\infty}(\bar{u}) \le c\|x\|_H^2$. Consider the control $\tilde{u} \in L^2_{\mathrm{loc}}(t, \infty; U)$ defined by

$$\tilde{u}(s) = \begin{cases} 0 & \text{for } t \le s < b+2, \\ \bar{u}(s) & \text{for } b+2 \le s < \infty. \end{cases}$$

Using (2.3) we have

$$\begin{aligned} &(P_T(t)(\lambda_0 - A(t))^\beta x, (\lambda_0 - A(t))^\beta x)_H \\ &\quad \le J_{t,T}(\tilde{u}) \\ &\quad = \int_t^{b+2} \|C(s)U(s,t)(\lambda_0 - A(t))^\beta x\|_V^2 \, ds + J_{b+2,\infty}(\bar{u}) \\ &\quad \le \left[c \int_t^{b+2} |U(s,t)(\lambda_0 - A(t))^\beta|^2_{\mathcal{L}(H)} \, ds + c \right] \|x\|_H^2 \le c\|x\|_H^2. \end{aligned} \tag{2.10}$$

Recalling that $P_T(t) \geq 0$, by (2.10) we easily obtain (2.8); moreover, as $P_\infty(t) \geq 0$ too, (2.9) follows by letting $T \to \infty$ in (2.10). □

We state now the following local existence theorem, proved in the Appendix. Its proof is the nonautonomous version of that of [F2, Lem. 2.1].

THEOREM 2.5. *Assume Hypotheses* 1.1 – 1.3, 1.5, *and* 2.2; *fix* $\beta \in]1/2 - \alpha, 1/2[$, $t \in I$, $r_0 > 0$, *and let* $Q_t \in \Sigma^+(H)$ *be such that* $|(\lambda_0 - A(t)^*)^\beta Q_t (\lambda_0 - A(t))^\beta|_{\mathcal{L}(H)} \leq r_0$. *Then there exists* $\tau_0 = \tau_0(r_0, \beta) > 0$ *such that the Riccati equation*

$$
\begin{aligned}
Q(s) = {} & U(t,s)^* Q_t U(t,s) \\
& + \int_s^t U(r,s)^* [C(r)^* C(r) \\
& \qquad - Q(r)(\lambda_0 - A(r)) G(r) N(r)^{-1} G(r)^* (\lambda_0 - A(r)^*) Q(r)] U(r,s)\, dr
\end{aligned}
\tag{2.11}
$$

has a unique solution $Q(\cdot)$ *in* $[t - \tau_0, t[$ *such that* $Q(s) \in \Sigma^+(H)$ *for each* $s \in [t - \tau_0, t[$ *and*

$$
|(\lambda_0 - A(s)^*)^{1-\alpha} Q(s)|_{\mathcal{L}(H)} \leq c(\beta, \alpha, r_0)(t-s)^{\beta+\alpha-1} \quad \forall\, s \in [t - \tau_0, t[. \tag{2.12}
$$

LEMMA 2.6. *Assume Hypotheses* 1.1 – 1.3, 1.5 *and* 2.2. *Then for each* $\mu \in]0, 1[$ *and each interval* $[a, b] \subset I$, *we have*

$$
\sup_{a \leq s \leq b, T > b+2} |(\lambda_0 - A(s)^*)^\mu P_T(s)|_{\mathcal{L}(H)} := c_2(\mu, [a,b]) < \infty. \tag{2.13}
$$

Proof. Using Theorem 2.5 and estimate (2.8) (having fixed any $\beta \in]0, 1/2[$), we find two constants $\tau \in]0, 1]$ and $c' > 0$, depending only on $[a, b]$ and on the constant c_1 of (2.8), such that for each $s, t \in [a, b+1]$ with $s < t$ and $t - s \leq \tau$ we have

$$
(t-s)^{1-\alpha-\beta} |(\lambda_0 - A(s)^*)^{1-\alpha} P_T(s)|_{\mathcal{L}(H)} \leq c'. \tag{2.14}
$$

Fix $s \in [a, b]$ and choose $t_1 := s + \tau/2 < t_2 := s + \tau \leq b + 1$. As $P_T(s)$, in particular, solves (2.2) for $s \leq t_1$ with final datum $P_T(t_1)$, we deduce

$$
\begin{aligned}
P_T(s) = {} & U(t_1, s)^* P_T(t_1) U(t_1, s) + \int_s^{t_1} U(r,s)^* C(r)^* C(r) U(r,s)\, ds \\
& - \int_s^{t_1} U(r,s)^* [(\lambda_0 - A(r)^*)^{1-\alpha} P_T(r)]^* K(r) [(\lambda_0 - A(r)^*)^{1-\alpha} P_T(r)] U(r,s)\, dr.
\end{aligned}
$$

Applying the operator $(\lambda_0 - A(s)^*)^\mu$ to both sides and using Hypothesis 1.2(ii) (and the estimate (2.12) with $t = t_2$), we get

$$
\begin{aligned}
|(\lambda_0 - A(s)^*)^\mu P_T(s)|_{\mathcal{L}(H)} \leq {} & \frac{c}{(t_1 - s)^\mu} |P_T(t_1)|_{\mathcal{L}(H)} + \int_s^{t_1} \frac{c}{(r-s)^\mu}\, ds \sup_{r \in [a,b]} |C(r)|^2_{\mathcal{L}(H,V)} \\
& + c \int_s^{t_1} \frac{c_2^2}{(r-s)^\mu (t_2 - r)^{2(1-\alpha-\beta)}}\, dr \leq c(\tau, c', \beta, [a,b], \mu),
\end{aligned}
$$

and the proof is complete. □

2.3. Existence of solutions of the Riccati equation. We prove here the following result.

THEOREM 2.7. *Assume Hypotheses* 1.1 – 1.3, 1.5, *and* 2.2. *Then*

(i) *for each* $\mu \in]0,1[$ *and* $t \in I$ *the operator* $P_\infty(t)$, *defined in Lemma* 2.1, *maps* H *into the domain of* $(\lambda_0 - A(t)^*)^\mu$ *and*

$$(\lambda_0 - A(t)^*)^\mu P_T(t) \to (\lambda_0 - A(t)^*)^\mu P_\infty(t) \quad \textit{strongly as } T \to \infty; \tag{2.15}$$

(ii) *the operator* $P_\infty(\cdot)$ *is a solution of equation* (2.1); *i.e., for each* t, $s \in I$ *with* $s \le t$ *we have*

$$\begin{aligned} P_\infty(s) = {} & U(t,s)^* P_\infty(t) U(t,s) \\ & + \int_s^t U(r,s)^* [C(r)^* C(r) \\ & \qquad - P_\infty(r)(\lambda_0 - A(r))G(r)N(r)^{-1}G(r)^*(\lambda_0 - A(r)^*)P_\infty(r)] \\ & \times U(r,s)\, dr. \end{aligned}$$

Remark 2.8. (i) It is possible to show that the convergence in (2.15) is uniform in t over bounded intervals. However, we omit the proof because we do not need this result in what follows.

(ii) In the special case where the resolvent of $A(t)$ is compact, the proof of Theorem 2.7 is very simple (see Theorem 2 in [F1]). However, we prefer to deal here with the general case, where the proof is considerably more difficult, since it is easy to construct examples with lack of compactness (see the remark at the end of Example 4.1).

The proof of Theorem 2.7 is based on the following lemma, which has also other applications (see, for instance, [F3]). Consider the following Riccati equation for $s \in [t_0, t[$, with fixed $t_0, t \in I$:

$$\begin{aligned} Q(s) = {} & U(t,s)^* \bar{Q}_t U(t,s) \\ & + \int_s^t U(r,s)^* [C(r)^* C(r) \\ & \qquad - Q(r)(\lambda_0 - A(r)^*)G(r)N(r)^{-1}G(r)^*(\lambda_0 - A(r)^*)Q(r)]U(r,s)dr \end{aligned} \tag{2.16}$$

under the assumption that $\bar{Q}_t \in \Sigma^+(H)$ and that the operator $(\lambda_0 - A(t)^*)^\beta \bar{Q}_t(t)(\lambda_0 - A(t))^\beta$ belongs to $\mathcal{L}(H)$; denote by $Q(s; \bar{Q}_t)$ its solution in a suitable interval $[t - r_0, t[$, given by Theorem 2.5.

LEMMA 2.9. *Assume Hypotheses* 1.1 – 1.3, 1.5, *and* 2.2, *and fix* $\beta \in]0, 1/2[$. *Let* $\{Q_{t,n}\}_{n > n_0}$ *be a family of operators in* $\Sigma^+(H)$ *such that*

(i) *there exists a constant* $c_3 > 0$ *such that*

$$|(\lambda_0 - A(t)^*)^\beta Q_{t,n}(\lambda_0 - A(t))^\beta|_{\mathcal{L}(H)} \le c_3 \quad \forall\, n > n_0;$$

(ii) $Q_{t,n}$ *converges strongly as* $n \to \infty$ *to an operator* $Q_t \in \Sigma^+(H)$ *for which the operator* $(\lambda_0 - A(t)^*)^\beta Q_t (\lambda_0 - A(t))^\beta$ *also belongs to* $\mathcal{L}(H)$.

Then for each $s \in [t - r_0, t[$

$$(\lambda_0 - A(s)^*)^{1-\alpha} Q(s; Q_{t,n}) \to (\lambda_0 - A(s)^*)^{1-\alpha} Q(s; Q_t) \quad \textit{strongly as } n \to \infty.$$

Proof of Lemma 2.9. This proof is adapted from [F3]. We denote by $[\Gamma(Q)](s)$ and $[\Gamma_n(Q)](s)$ the right-hand side of (2.16) when we consider the final data Q_t and $Q_{t,n}$, respectively. Thus equation (2.16) can be rewritten as

$$Q(s) = [\Gamma(Q)](s), \qquad s \in [t_0, t], \tag{2.17}$$

if the final datum is Q_t, and

$$Q(s) = [\Gamma_n(Q)](s), \qquad s \in [t_0, t], \tag{2.17$_n$}$$

if the final datum is $Q_{t,n}$.

Next set $\gamma := \min\{1-\alpha-\beta, \beta\}$ and consider the following space:

$$\begin{aligned} X(t-r_0,t) := \{\, & Q : [t-r_0,t] \to \Sigma^+(H) :\ Q(s) \text{ maps } H \text{ into the domain of} \\ & (\lambda_0 - A(s)^*)^{1-\alpha} \text{for each } s \in [t-r_0,t[, \text{ and both } (\lambda_0 - A(\cdot)^*)^{1-\alpha}Q(\cdot) \\ & \text{and its adjoint are strongly continuous in } [t-r_0,t[; \text{ moreover} \\ & |(\lambda_0 - A(s)^*)^{1-\alpha}Q(s)|_{\mathcal{L}(H)} \\ & \quad \le c(Q)[1+(t-s)^{-(1-\alpha-\beta)}] \quad \forall\, s \in [t-r_0,t[, \\ & |(\lambda_0 - A(s)^*)^{1-\alpha}Q(s)U(s,r)(\lambda_0 - A(s))^{\beta}|_{\mathcal{L}(H)} \\ & \quad \le c(Q)(t-r)^{\gamma}[1+(t-s)^{-(1-\alpha-\beta)}](s-r)^{-\beta} \\ & \quad\ \forall\, s \in [t-r_0,t[,\ \forall\, r \in [0,s[\}, \end{aligned}$$

endowed with the norm

$$|Q|_X := \max\{A, B\},$$

where

$$A := \sup_{s\in[t-r_0,t]} (t-s)^{1-\alpha-\beta}|(\lambda_0 - A(s)^*)^{1-\alpha}Q(s)|_{\mathcal{L}(H)},$$

$$\begin{aligned} B := & \sup_{t-r_0\le r<s<t} \frac{(s-r)^{\beta}[1+(t-s)^{1-\alpha-\beta}]}{(t-r)^{\gamma}} \\ & \times|(\lambda_0 - A(s)^*)^{1-\alpha}Q(s)U(s,r)(\lambda_0 - A(r))^{\beta}|_{\mathcal{L}(H)}. \end{aligned}$$

It can be easily shown, arguing as in the Appendix below, that the maps Γ_n and Γ are equicontractions on any sufficiently large ball of $X(t_0,t)$, provided that r_0 is suitably small. Hence, possibly replacing the constant c_3 in (i) by a larger one, we may say that Γ and Γ_n are equicontractions on the ball

$$B(t-r_0,t;c_3) := \{Q \in X(t-r_0,t) : |Q|_X \le c_3\}. \tag{2.18}$$

Now we apply the contraction principle to equations (2.17$_n$) in the ball $B(t-r_0,t;c_3)$ uniformly with respect to n. Namely, let $Q_0(\cdot)$ be the initial iteration point in $B(t-r_0,t;c_3)$ for Γ and Γ_n; then, remarking that

$$\lim_{k\to\infty} |Q(\cdot\,, Q_{t,n}) - [(\Gamma_n)^k(Q)]|_X = 0 \quad \text{uniformly with respect to } n, \tag{2.19}$$

$$\lim_{k\to\infty} |Q(\cdot\,, Q_t) - [(\Gamma)^k(Q)]|_X = 0, \tag{2.20}$$

we deduce for each $k \in \mathbb{N}^+$, $s \in [t-r_0,t]$, and $x \in H$

$$\begin{aligned} & \|(\lambda_0 - A(s)^*)^{1-\alpha}Q(s;Q_{t,n})x - (\lambda_0 - A(s)^*)^{1-\alpha}Q(s;Q_t)x\|_H \\ & \le |Q(\cdot;Q_{t,n}) - [(\Gamma_n)^k(Q_0)]|_X\, \|x\|_H \\ & \quad + \|(\lambda_0 - A(s)^*)^{1-\alpha}[(\Gamma_n)^k(Q_0)](s)x - (\lambda_0 - A(s)^*)^{1-\alpha}[(\Gamma)^k(Q_0)](s)x\|_H \\ & \quad + |[(\Gamma)^n(Q_0)] - Q(\cdot\,;Q_t)|_X\, \|x\|_H. \end{aligned} \tag{2.21}$$

Thus we just need to show that for each $k \in \mathbb{N}^+$ and $s \in [t - r_0, t]$

$$(\lambda_0 - A(s)^*)^{1-\alpha}[(\Gamma_n)^k(Q_0)](s) \to (\lambda_0 - A(s)^*)^{1-\alpha}[(\Gamma)^k(Q_0)](s) \quad \text{strongly as } n \to \infty.$$

This result is obviously true when $k = 1$, since for each $s \in [t - r_0, t]$ and $x \in H$

$$\begin{aligned}(\lambda_0 - A(s)^*)^{1-\alpha}&[[\Gamma_n(Q_0)](s) - [\Gamma(Q_0)](s)]x \\ &= (\lambda_0 - A(s)^*)^{1-\alpha}U(t,s)^*(Q_{t,n} - Q_t)U(t,s)x;\end{aligned} \tag{2.22}$$

on the other hand, if the result is true for the integer $k - 1$, then we have

$$\begin{aligned}&(\lambda_0 - A(s)^*)^{1-\alpha}[[(\Gamma_n)^k(Q_0)](s) - [(\Gamma)^k(Q_0)](s)]x \\ &\quad = (\lambda_0 - A(s)^*)^{1-\alpha}U(t,s)^*(Q_{t,n} - Q_t)U(t,s)x \\ &\qquad - \int_s^t (\lambda_0 - A(s)^*)^{1-\alpha}U(r,s)^*[(\lambda_0 - A(r)^*)^{1-\alpha}(\Gamma_n)^{k-1}(Q_0)(r)]^* \\ &\qquad \times K(r)(\lambda_0 - A(r)^*)^{1-\alpha}[(\Gamma_n)^{k-1}(Q_0)(r) - (\Gamma)^{k-1}(Q_0)(r)]U(r,s)x\,dr \\ &\qquad - \int_s^t (\lambda_0 - A(s)^*)^{1-\alpha}U(r,s)^*[(\lambda_0 - A(r)^*)^{1-\alpha}[(\Gamma_n)^{k-1}(Q_0)(r) - (\Gamma)^{k-1}(Q_0)(r)]]^* \\ &\qquad \times K(r)(\lambda_0 - A(r)^*)^{1-\alpha}(\Gamma)^{k-1}(Q_0)(r)U(r,s)x\,dr,\end{aligned} \tag{2.23}$$

and remarking that $(\Gamma_n)^{k-1}(Q_0)$ and $(\Gamma)^{k-1}(Q_0)$ belong to $B(t - r_0, t; c_3)$ by the induction hypothesis we get the result for the integer k. This proves Lemma 2.9. □

Proof of Theorem 2.7. Fix $t, t_0 \in I$ with $t > t_0$. We have to show that

$$P_\infty(s) = Q(s, P_\infty(t)) \quad \forall\, s \in [t_0, t[. \tag{2.24}$$

We apply Lemma 2.9 with $Q_t = P_\infty(t)$, $Q_{t,n} = P_n(t)$ (i.e., the solution of equation (2.2) with final time $T = n$); this is allowed by Lemmas 2.4 and 2.1. As a consequence we get

$$(\lambda_0 - A(s)^*)^{1-\alpha}[Q(s, P_n(t)) - Q(s, P_\infty(t))] \to 0 \quad \text{strongly in } [t - r_0, t[\text{ as } n \to \infty.$$

On the other hand we have

$$Q(s, P_n(t)) \equiv P_n(s) \to P_\infty(s) \quad \text{strongly in } [t - r_0, t] \text{ as } n \to \infty,$$

and (2.24) follows for each $s \in [t - r_0, t[$. The same result for all $s \in [t_0, t[$ follows now by standard uniqueness arguments. □

2.4. Minimality property of P_∞. Let $\hat{P} \in C_s(I, \Sigma^+(H))$ be any solution of equation (2.1), and consider the evolution operator $\hat{\Phi}(t,r)$ corresponding to $\hat{P}(\cdot)$, i.e., the operator-valued function defined for r, $t \in I$, $r \le t$, by the following equation:

$$\begin{aligned}\hat{\Phi}(t,r) &= U(t,r) \\ &\quad - \int_r^t U(t,s)(\lambda_0 - A(s))G(s)N(s)^{-1}G(s)^*(\lambda_0 - A(s)^*)\hat{P}(s)\hat{\Phi}(s,r)\,ds.\end{aligned} \tag{2.25}$$

We know (see, e.g., [G], [LT1]) that for each $s,\ t \in I$ with $s \le t$ the following identities hold:

$$\hat{P}(s) = \hat{\Phi}(t,s)^* P(t)\hat{\Phi}(t,s) + \int_s^t \Phi(v,s)^* \times [C(v)^*C(v) + \hat{P}(v)(\lambda_0 - A(v))G(v)N(v)^{-1}G(v)(\lambda_0 - A(v)^*)\hat{P}(v)]\hat{\Phi}(v,s)\,dv, \tag{2.26}$$

$$\hat{P}(s) = U(t,s)^*\hat{P}(t)\hat{\Phi}(t,s) + \int_s^t U(v,s)^*C(v)^*C(v)\hat{\Phi}(v,s)\,dv. \tag{2.27}$$

We have the following proposition.

PROPOSITION 2.10. *Assume Hypotheses* 1.1 – 1.3, *and* 1.5. *Then*

(i) *equation* (2.1) *has a solution if and only if Hypothesis* 2.2 *holds*;

(ii) *if this is the case, the function* $P_\infty(\cdot)$ *defined in Lemma* 2.1 *is the minimal solution of equation* (2.1); *i.e., for any solution* $\hat{P}(\cdot)$ *of equation* (2.1) *we have*

$$P_\infty(t) \le \hat{P}(t) \quad \forall\, t \in I.$$

Proof. (i) Theorem 2.7 shows the if part of the proposition. Conversely, if $\hat{P}(\cdot)$ is a solution of (2.1) and $y_0 \in H,\ t_1 \in I$ are given, we consider the control

$$\hat{u}(t) = -N(t)^{-1}G(t)^*(\lambda_0 - A(t)^*)\hat{P}(t)\hat{\Phi}(t,t_1)y_0, \qquad t \ge t_1.$$

A simple calculation shows that the corresponding state is $\hat{y}(t) = \hat{\Phi}(t,t_1)y_0$. Using equation (2.26) we easily obtain

$$(\hat{P}(t_1)y_0, y_0)_H = (\hat{P}(t)\hat{y}(t), \hat{y}(t))_H + \int_{t_1}^t [\|C(v)\hat{y}(v)\|_V^2 + (N(v)\hat{u}(v), \hat{u}(v))_U]\,dv.$$

Hence for each $t \ge t_1$ we have

$$\int_{t_1}^t [\|C(v)\hat{y}(v)\|_V^2 + (N(v)\hat{u}(v), \hat{u}(v))_U]\,dv \le (\hat{P}(t_1)y_0, y_0)_H;$$

consequently

$$J_{t_1,\infty}(\hat{u}) \le (\hat{P}(t_1)y_0, y_0)_H \le |\hat{P}(t_1)|_{\mathcal{L}(H)}\|y_0\|_H^2. \tag{2.28}$$

By the local coercivity of $N(\cdot)$ (Hypothesis 1.5) we then get $\hat{u} \in L^2_{\text{loc}}(t_1,\infty;U)$, so condition (2.4_{t_1}) holds.

(ii) Using (2.3) we obtain

$$(P_T(t_1)y_0, y_0)_H \le J_{t_1,T}(u) \quad \forall\, t_1 \in I,\ \forall\, T > t_1,\ \forall\, y_0 \in H,\ \forall u \in L^2_{\text{loc}}(t_1,\infty;U),$$

and consequently we have

$$(P_T(t_1)y_0, y_0)_H \le J_{t_1,\infty}(u) \quad \forall\, t_1 \in I,\ \forall T > t_1,\ \forall\, y_0 \in H,\ \forall\, u \in L^2_{\text{loc}}(t_1,\infty;U);$$

letting $T \to \infty$ we obtain

$$(P_\infty(t_1)y_0, y_0)_H \le J_{t_1,\infty}(u) \quad \forall\, t_1 \in I,\ \forall\, y_0 \in H,\ \forall\, u \in L^2_{\text{loc}}(t_1,\infty;U) \tag{2.29}$$

so that, in particular, by (2.28)

$$(P_\infty(t_1)y_0, y_0)_H \le J_{t_1,\infty}(\hat{u}) \le (\hat{P}(t_1)y_0, y_0)_H \quad \forall\, t_1 \in I,\ \forall\, y_0 \in H, \tag{2.30}$$

and the result follows. □

2.5. Synthesis of the infinite-horizon LQR problem. We use the properties of the operator $P_\infty(\cdot)$ to solve the problem of the synthesis. We have the following theorem.

THEOREM 2.11. *Assume Hypotheses* 1.1 – 1.3, 1.5, *and* 2.2. *Let* $t_0 \in I$ *and* $y_0 \in H$ *be given. Then*

(i) *there exists a unique optimal control* $u^* \in L^2_{\text{loc}}(t_0, \infty; U)$ *for problem* (1.7);

(ii) *if* (u^*, y^*) *is the optimal pair and* $P_\infty(\cdot)$ *is defined by Lemma* 2.1, *then*

$$u^*(t) = -N(t)^{-1}G(t)^*(\lambda_0 - A(t)^*)P_\infty(t)y^*(t) \quad \forall\, t \geq t_0;$$

(iii) *the optimal cost is*

$$J_{t_0,\infty}(u^*) = (P_\infty(t_0)y_0, y_0)_H;$$

(iv) *the optimal state is given by*

$$y^*(t) = \Phi_\infty(t, t_0)y_0,$$

where $\Phi_\infty(t, s)$ *is the evolution operator defined by equation* (2.25) *with* $P_\infty(\cdot)$ *in place of* $\hat{P}(t)$.

Proof. Given $t_0 \in I$ and $y_0 \in H$, set

$$u^*(t) := -N(t)^{-1}G(t)^*(\lambda_0 - A(t)^*)P_\infty(t)\Phi_\infty(t, t_0)y_0, \qquad t \geq t_0;$$

by the same arguments in the proof of Proposition 2.10 we easily see that $u^* \in L^2_{\text{loc}}(t_0, \infty; U)$ is an admissible control with respect to t_0, whereas $y^*(t) := \Phi_\infty(t, t_0)y_0$ is the state corresponding to u^*. By (2.29) and (2.30), with $\hat{P}$ replaced by P_∞, $\hat{u}$ by u^*, and t_1 by t_0, we obtain

$$(P_\infty(t_0)y_0, y_0)_H = \min_u J_{t_0,\infty}(u) = J_{t_0,\infty}(u^*);$$

i.e., u^* is an optimal control.

Finally it is clear that Hypothesis 1.5 on $N(\cdot)$ implies the strict convexity of $J_{t_0,\infty}$, so the optimal control is unique. □

3. Further properties of solutions of the Riccati equation.

3.1. Bounded solutions. In many cases it is important to know whether some bounded solution of Riccati equation (2.1) exists. In order to obtain boundedness we have to assume some uniformity in Hypothesis 2.2. Thus, following [DI3], we introduce a stronger version of that assumption.

Hypothesis 3.1. There exists a constant $\bar{c} > 0$ such that for each $t \in I$ and $y_0 \in H$ there exists a control $u \in L^2_{\text{loc}}(t, \infty; U)$ such that

$$J_{t,\infty}(u) < \bar{c}\,\|y_0\|^2_H.$$

We have the following proposition.

PROPOSITION 3.2. *Assume Hypotheses* 1.1 – 1.3, *and* 1.5. *Then there exists a bounded solution of Riccati equation* (2.1) *if and only if Hypothesis* 3.1 *holds.*

Proof. If Hypothesis 3.1 holds, then in particular Hypothesis 2.2 holds too, so that by the results of §2 the function $P_\infty(\cdot)$, defined in Lemma 2.1, is a solution of equation (2.1). In addition we have for each $t \in I$, $T > t$, and $y \in H$

$$(P_T(t)y, y)_H \leq \min_u J_{t,T}(u) \leq \bar{c}\,\|y\|^2_H;$$

thus letting $T \to \infty$ we get

$$(P_\infty(t)y, y)_H \le \bar{c}\,\|y\|_H^2 \quad \forall t \in I, \forall y \in H,$$

and recalling that $P_\infty(t) \ge 0$ we obtain

$$\sup_{t\in I} |P_\infty(t)|_{\mathcal{L}(H)} \le \bar{c}.$$

Conversely, if there exists a bounded solution $\hat{P}(\cdot)$ of (2.1), then we can repeat the argument of Proposition 2.10(i), and (2.28) shows that Hypothesis 3.1 holds with

$$\bar{c} = \sup_{t\in I} |P_\infty(t)|_{\mathcal{L}(H)}.$$

Remark 3.3. Hypothesis 3.1 is fulfilled in several cases.

(i) In the periodic case of [L1], [F], [DI2] (see §3.4), if condition (2.4_t) is satisfied for some $t \in \mathbb{R}$, then Hypothesis 3.1 holds.

(ii) If system (1.4) is stabilizable, i.e., there exists $K \in L^\infty(I, \mathcal{L}(H,U))$ such that the evolution operator associated with the family $\{[A-(\lambda_0-A)GK](t)\}$ is stable, and in addition the operators $C(\cdot), N(\cdot)$ are bounded, then Hypothesis 3.1 holds (compare with the comments after Hypothesis (H3) in [DI3]).

(iii) In Example 4.2 below, Hypothesis 3.1 holds naturally.

3.2. Stability of the perturbed evolution operator. Theorem 2.11 shows, under suitable assumptions, the existence of a unique optimal pair (u^*, y^*) for problem (1.7) with $t_0 = 0$; we also know that

$$y^* = \Phi_\infty(\cdot\,, 0)y_0, \qquad u^* = -[N^{-1}G^*(\lambda_0 - A^*)P_\infty y^*].$$

(From now on we will drop the indication of the variable t if unnecessary.) Here $P_\infty(t)$ is the minimal solution of Riccati equation (2.1) and $\Phi_\infty(t,s)$ is the evolution operator associated to the closed-loop operator family

$$\{A - G(\lambda_0 - A)N^{-1}G^*(\lambda_0 - A^*)P_\infty\}$$

by the integral equation (2.25); in other words, $\Phi_\infty(t,s)$ is the solution, for $t, s \in I$, $t \ge s$, of

$$\begin{aligned}\Phi_\infty(t,s) &= U(t,s)\\ &\quad - \int_s^t U(t,r)(\lambda_0 - A(r))G(r)N(r)^{-1}G(r)^*(\lambda_0 - A(r)^*)P_\infty(r)\Phi_\infty(r,s)\,dr.\end{aligned} \tag{3.1}$$

In this subsection, following the ideas of [DI1], [DI2] and [BDDM, Chap. IV.2, §3.2], we will prove a stability result for $y^*(t) = \Phi_\infty(t,0)$ as $t \to \infty$. In order to do this we have to assume that Hypotheses 1.2, 1.3, and 1.5 hold uniformly over the time interval I. More precisely we formulate the following hypothesis.

Hypothesis 3.4. (i) The evolution operators $U(t,s)$ and $U(t,s)^*$ are strongly continuous in $\bar{\Delta}_I$, where $\Delta_I := \{(t,s) \in I^2 : t > s\}$ and there exist $M_0 > 0$ and $\omega \in \mathbb{R}$ such that

$$|U(t,s)|_{\mathcal{L}(H)} + |U(t,s)^*|_{\mathcal{L}(H)} \le M_0 \exp(\omega_0(t-s)) \quad \forall (t,s) \in \Delta_I;$$

(ii) for each β, $\mu \in [-1,1]$ and $(t,s) \in \Delta_I$, the operators

$$(\lambda_0 - A(t))^\beta U(t,s)(\lambda_0 - A(s))^{-\mu}, \qquad (\lambda_0 - A(s)^*)^\beta U(t,s)^*(\lambda_0 - A(t)^*)^{-\mu}$$

have continuous extensions to H, the maps

$$(t,s) \to (\lambda_0 - A(t))^\beta U(t,s)(\lambda_0 - A(s))^{-\mu}, \quad (t,s) \to (\lambda_0 - A(s)^*)^\beta U(t,s)^*(\lambda_0 - A(t)^*)^{-\mu}$$

are strongly continuous, and there exists $M_{\beta,\mu} > 0$ such that

$$|(\lambda_0 - A(t))^\beta U(t,s)(\lambda_0 - A(s))^{-\mu}|_{\mathcal{L}(H)} + |(\lambda_0 - A(s)^*)^\beta U(t,s)^*(\lambda_0 - A(t)^*)^{-\mu}|_{\mathcal{L}(H)} \leq M_{\beta,\mu}[(t-s)^{\mu-\beta} + 1]\exp(\omega_0(t-s)) \quad \forall\, (t,s) \in \Delta_I.$$

Hypothesis 3.5. There exists $\alpha \in]0,1]$ such that, for each $t \in I$, $G(t)$ maps U into the domain of $(\lambda_0 - A(t))^\alpha$, and $(\lambda_0 - A(\cdot))^\alpha G(\cdot)$ is strongly measurable and bounded over I.

Hypothesis 3.6. We have

$$C(\cdot) \in L^\infty(I; \mathcal{L}(H,V)), \qquad N(\cdot) \in L^\infty(I; \Sigma^{++}(U)).$$

(This means that there exists $\nu > 0$ such that $N(t) \geq \nu$, $\forall\, t \in I$.)

Under the assumption listed above we can revisit the proof of Lemma 2.6, and we get the following lemma.

LEMMA 3.7. *Assume Hypotheses* 1.1, 3.4–3.6, *and* 3.1. *Then for each* $\mu \in]0,1[$ *we have*

$$\sup_{s \in I} |(\lambda_0 - A(s)^*)^\mu P_\infty(s)|_{\mathcal{L}(H)} < \infty.$$

Proof. Fix $t \in I$, $0 \leq \beta < 1/2$, $x \in D[(\lambda_0 - A(t))^\beta]$, and set $y_1 := U(t+1,t)(\lambda_0 - A(t))^\beta x$. By Hypothesis 3.1 there exists a control $\bar{u} \in L^2_{\text{loc}}(t+1, \infty; U)$ such that

$$J_{t+1,\infty}(\bar{u}) \leq \bar{c}\, \|y_1\|_H^2,$$

and by Hypothesis 3.4(ii) we also have

$$J_{t+1,\infty}(\bar{u}) \leq c\, \|x\|_H^2. \tag{3.2}$$

Consider the control $\tilde{u} \in L^2_{\text{loc}}(t, \infty; U)$ defined by

$$\tilde{u}(s) = \begin{cases} 0 & \text{if } t \leq s < t+1, \\ \bar{u}(s) & \text{if } t+1 \leq s < \infty. \end{cases}$$

Using Theorem 2.11(iii) and (3.2) we have

$$\begin{aligned}
(P_\infty(t)(\lambda_0 - A(t))^\beta x, (\lambda_0 - A(t))^\beta x)_H &\leq J_{t,\infty}(\tilde{u}) \\
&= \int_t^{t+1} \|C(s)U(s,t)(\lambda_0 - A(t))^\beta x\|_V^2\, ds + J_{t+1,\infty}(\bar{u}) \\
&\leq \left(\|C\|^2_{L^\infty(I;\mathcal{L}(V,H))} \int_t^{t+1} |U(s,t)(\lambda_0 - A(t))^\beta|^2_{\mathcal{L}(H)}\, ds + c \right) \|x\|_H^2 \\
&\leq \left(|C|^2_{L^\infty(I;\mathcal{L}(V,H))} M^2_{\beta,0} \exp(\omega_0) + c \right) \|x\|_H^2 \leq c \|x\|_H^2.
\end{aligned}$$

Recalling that $P_\infty(t) \geq 0$, the above estimate shows that

$$\sup_{t \in I} |(\lambda_0 - A(t)^*)^\beta P_\infty(t)(\lambda_0 - A(t))^\beta|_{\mathcal{L}(H)} =: L < +\infty.$$

We now repeat the argument of the proof of Lemma 2.6; invoking Theorem 2.5 and noting that our assumptions are uniform in t now, we find two constants $\tau \in]0, 1]$ and $c' > 0$, depending on L and β but independent of $t \in I$, for which the analogue of (2.14) holds, i.e.,

$$(t-s)^{1-\alpha-\beta}|(\lambda_0 - A(s)^*)^{1-\alpha}P_\infty(s)|_{\mathcal{L}(H)} \le c' \quad \forall\, s \in I \cap [t-\tau, t[.$$

Arguing as in the proof of Lemma 2.6, one arrives easily at the estimate

$$\sup_{s\in I} |(\lambda_0 - A(s)^*)^{\mu}P_\infty(s)|_{\mathcal{L}(H)} \le c(\tau, c', \beta, L, \mu),$$

which concludes the proof. □

Using the result of Lemma 3.7 it is easy to show that the evolution operator $\Phi_\infty(t, s)$ has an exponential growth. Namely, we have the following lemma.

LEMMA 3.8. *Under Hypotheses* 1.1, 3.4–3.6, *and* 3.1, *there exist* $M_1 > 0$ *and* $\omega_1 > \omega_0$ *such that*

$$|\Phi_\infty(t, s)|_{\mathcal{L}(H)} \le M_1 \exp(\omega_1(t-s)) \quad \forall\, (t, s) \in \bar{\Delta}_I.$$

Proof. The result follows easily by equation (3.1), using our assumptions and the result of Lemma 3.7. □

It is important, in some applications, to give conditions under which the evolution operator Φ_∞ is exponentially stable, i.e., there exist $M > 0$ and $\gamma > 0$ such that

$$\|\Phi_\infty(t, 0)\|_{\mathcal{L}(H)} \le M \exp(-\gamma t).$$

A simple situation where this occurs is when the operator $C(s)$ is invertible for each $s \in I$, and C^{-1} belongs to $L^\infty(I; \mathcal{L}(V, H))$ (see [BDDM]); indeed, if this is the case, we fix $x \in H$ and argue as in the proof of Proposition 2.10, replacing t_1 with t_0, $\hat{\Phi}$ with Φ_∞, $\hat{P}$ with P_∞, $\hat{y}$ with $y^* := \Phi(\cdot\,, 0)x$, and $\hat{u}$ with u^*. Then we obtain, for each $t \in I$,

$$\int_{t_0}^{t} [\|Cy^*\|_V^2 + \|N^{-1/2}G^*(\lambda_0 - A^*)P_\infty y^*\|_H^2]\, ds \le (P_\infty(0)x, x)_H \le \bar{c}\,\|x\|_H^2.$$

So we have

$$C(\cdot)\Phi_\infty(\cdot\,, 0)x = Cy^* \in L^2(t_0, \infty; V), \tag{3.3}$$

$$N^{-1/2}G^*(\lambda_0 - A^*)P_\infty y^* \in L^2(t_0, \infty; H), \tag{3.4}$$

and by (3.3) we deduce that $\Phi_\infty(\cdot\,, 0)x \in L^2(I; H)$; thus by the classical results of Datko [D], we obtain the exponential stability of Φ_∞.

A sufficient condition yielding the same property, even if C is not invertible, is given by the following detectability condition [F1], [DI1], [DI3].

Hypothesis 3.9. The family $\{(A, C)\}$ is detectable; this means that there exists a mapping $K : I \to \mathcal{L}(V, H)$, strongly measurable and bounded, such that the evolution operator $U_{A-KC}(t, s)$ associated with$\{A - KC\}$ is stable; i.e., there exist two constants $M_2 > 0$, $\omega_2 > 0$ such that

$$|U_{A-KC}(t, s)|_{\mathcal{L}(H)} \le M_2 \exp(-\omega_2(t-s)) \quad \forall\, (t, s) \in \Delta_I. \tag{3.5}$$

LEMMA 3.10. *Assume Hypotheses* 3.4, 3.6, *and* 3.9. *Then there exists a constant* $c > 0$ *such that for each* $(t, s) \in \Delta_I$ *the operator* $U_{A-KC}(t, s)(\lambda_0 - A(s))^{1-\alpha}$ *has a continuous extension to* H *and*

$$|U_{A-KC}(t, s)(\lambda_0 - A(s))^{1-\alpha}|_{\mathcal{L}(H)} \le c(t-s)^{\alpha-1} \exp(-\omega_2(t-s)) \quad \forall\, (t, s) \in \Delta_I.$$

Proof. By Hypotheses 3.6 and 3.9 it follows that $KC \in L^\infty(I;\mathcal{L}(H))$ and the construction of $U_{A-KC}(t,s)$ is standard. Next, for each $s \in I$ set

$$V(t,s)x := U_{A-KC}(t,s)(\lambda_0 - A(s))^{1-\alpha}x, \qquad t \geq s,\ x \in D((\lambda_0 - A(s))^{1-\alpha});$$

then it is immediately seen that

$$V(t,s)x = U(t,s)(\lambda_0 - A(s))^{1-\alpha}x + \int_s^t U(t,r)K(r)C(r)V(r,s)x\,dr, \qquad (t,s) \in \Delta_I.$$

Using Hypothesis 3.4 we easily get

$$\|V(t,s)x\|_H \leq c(t-s)^{\alpha-1}\|x\|_H \quad \forall\,(t,s) \in \Delta_I \text{ with } t-s \leq 1;$$

hence, taking (3.3) into account we easily get the result. □

As a simple consequence of the above lemma we have the following theorem.

THEOREM 3.11. *Assume Hypotheses* 1.1, 3.1, 3.4–3.6, *and* 3.9. *Then* $\Phi_\infty(\cdot\,,0)$ *is exponentially stable.*

Proof. We have

$$U_{A-KC}(t,s) = U(t,s) + \int_s^t U(t,r)K(r)C(r)U_{A-KC}(r,s)\,dr, \quad (t,s) \in I. \tag{3.6}$$

Comparing with (3.1) we easily obtain for each $(t,s) \in I$

$$\Phi_\infty(t,s) = U_{A-KC}(t,s) - \int_s^t U_{A-KC}(t,r)[KC - (\lambda_0 - A)GN^{-1}G^*(\lambda_0 - A^*)P_\infty](r)\Phi_\infty(r,s)\,dr.$$

Now using (3.5), Lemma 3.10, Hypothesis 3.6, and the boundedness of K, by Young's inequality we deduce that $\Phi_\infty(\cdot\,,0) \in L^2(I;H)$, and finally the exponential stability is a consequence of the results of Datko [D]. □

3.3. Uniqueness of the solution of the Riccati equation. By Proposition 2.10 it is clear that if a bounded solution $\hat{P}$ of equation (2.1) exists, then P_∞ also is bounded. Under suitable assumptions on the LQR system we are able to show uniqueness of bounded solutions.

We have the following result, which generalizes [F1, Thm. 4].

THEOREM 3.12. *Assume Hypotheses* 1.1–1.3, 1.5, *and* 3.1; *in addition, assume that the optimal trajectory* $y^*(\cdot)$ *is stable. Then the only bounded solution of equation* (2.1) *is* P_∞.

Proof. By Proposition 3.2 we know that P_∞ is a bounded solution of (2.1). Now let $\hat{P}$ be another bounded solution of (2.1); by Proposition 2.10 we know that

$$P_\infty(t) \leq \hat{P}(t) \quad \forall\, t \in I,$$

so it is sufficient to prove the converse inequality.

Fix $y_0 \in H$ and $t_1 \in I$; by [AFT, Thm. 3.14] we deduce that

$$(\hat{P}(t_1)y_0, y_0)_H \leq J_{t_1,t}(u) + (\hat{P}(t)y(t), y(t))_H \quad \forall\, u \in L^2(t_1,t;U),\ \forall\, t > t_1, \tag{3.7}$$

where $y(\cdot)$ satisfies the state equation (1.4).

We apply (3.7), using the optimal control $u^* = -N^{-1}G^*(\lambda_0 - A^*)P_\infty y^*$; we recall that the optimal trajectory is given by $y^*(t) = \Phi_\infty(t, t_1)y_0$. We obtain

$$\begin{aligned}(\hat{P}(t_1)y_0, y_0)_H &\le \int_{t_1}^{t} [\|Cy^*\|_V^2 + (Nu^*, u^*)_U]\, dv + (\hat{P}(t)y^*(t), y^*(t))_H \\ &\le J_{t_1,\infty}(u^*) + (\hat{P}(t)y^*(t), y^*(t))_H \\ &= (P_\infty(t_1)y_0, y_0)_H + (\hat{P}(t)y^*(t), y^*(t))_H.\end{aligned}$$

By assumption we have $y^*(t) \to 0$ as $t \to \infty$, whereas $\hat{P}$ is bounded; hence, as $t \to \infty$ we obtain

$$(\hat{P}(t_1)y_0, y_0)_H \le (P_\infty(t_1)y_0, y_0)_H.$$

This shows that $\hat{P} \le P_\infty$. □

3.4. Periodic case and autonomous case. We consider now two special cases of equation (2.1): the periodic case and the time-invariant case. We assume the following hypothesis.

Hypothesis 3.13. There exists $\vartheta > 0$ such that $A(t + \vartheta) = A(t)$, $G(t + \vartheta) = G(t)$, $C(t + \vartheta) = C(t)$, and $N(t + \vartheta) = N(t)$ for all $t \in \mathbb{R}$. If this is the case we say that the system is ϑ-periodic.

Remark 3.14. If the system is ϑ-periodic, then

(i) evidently all assumptions concerning the uniform behaviour of the operators follow from the local assumptions listed in §1;

(ii) if $\hat{P}(t)$ is a bounded solution of (2.1), then $\hat{P}_\vartheta(t) := \hat{P}(t + \vartheta)$, $t \in \mathbb{R}$, is also a bounded solution;

(iii) some stabilizability results for equation (1.4) can be found in [L2].

As in [DI3, Prop. 3.4] we have the following proposition.

PROPOSITION 3.15. *Assume Hypotheses* 1.1 – 1.3, 1.5, 2.2, *and* 3.13. *Then the minimal solution* P_∞ *of* (2.1) *is* ϑ*-periodic. If Hypothesis* 3.9 *holds too, then* P_∞ *is the unique nonnegative* ϑ*-periodic solution of* (2.1) *and the corresponding optimal trajectory for problem* (1.7) *is exponentially stable.*

Proof. The periodicity of P_∞ follows from the same argument as in [DI3, Prop. 3.4]; the stability of the optimal trajectory is a direct consequence of Theorem 3.11. □

Finally assume that A, G, C, and N are independent of t. Then our assumptions correspond to those assumed by Flandoli [F1], and the corresponding result is the following proposition.

PROPOSITION 3.16. *Suppose that* A *is the infinitesimal generator of an analytic semigroup* e^{tA}, *and let* $G \in \mathcal{L}(U, D((\lambda_0 - A)^\alpha))$, $C \in \mathcal{L}(U, V)$, $N, N^{-1} \in \Sigma^+(U)$. *In addition assume that condition* (2.4_0) *holds* (*i.e., there exists an admissible control*). *Then* $P_\infty(t) \equiv P_\infty$ *is independent of* t, *and it is the minimal solution of the algebraic Riccati equation*

$$A^*Q + QA + C^*C - Q(\lambda_0 - A)^{1-\alpha}[(\lambda_0 - A)^\alpha G]N^{-1}[(\lambda_0 - A)^\alpha G]^*(\lambda_0 - A^*)^{1-\alpha}Q = 0. \tag{3.8}$$

Furthermore, if (A, C) *is detectable, then* P_∞ *is the unique nonnegative solution of* (3.8). □

4. Examples.

4.1. A finite-dimensional example. Consider the family of 2×2 matrices $\{A(t)\} = \{(1+t)A_1\}_{t\geq 0}$, where

$$A_1 = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$

In $\mathbb{R}^2$ consider the state equation

$$y'(t) = A(t)y(t) + B(t)u(t), \qquad t \geq 0, \quad y(0) = (x_1, x_2) \in \mathbb{R}^2, \tag{4.1}$$

where $u \in L^2_{\mathrm{loc}}(\mathbb{R}^+, \mathbb{R}^2)$ is the control and $B(t) = b(t)I$ (I is the identity matrix), with $b(\cdot)$ a nonzero continuous function with polynomial growth as $t \to \infty$. We want to minimize the quadratic cost functional given by

$$J_{0,\infty}(u) = \int_0^\infty [\|C(t)y(t)\|^2_{\mathbb{R}^2} + (N(t)u(t), u(t))_{\mathbb{R}^2}]\, dt,$$

with y, u subject to equation (4.1); here $C(t) = (2\sqrt{2(1+t)^2 - 1})I$, $N(t) = b(t)^2 I$.

In this situation the eigenvalues of the matrix $A(t)$ are $(1+t, \pm(1+t)i)$ and

$$U(t,s) = \exp([(t^2 - s^2)/2 + (t-s)]A_1).$$

For a given $t_0 \geq 0$ an admissible feedback control relative to t_0 is easily found by choosing $\hat{u}(t) = K(t)\hat{y}(t)$, with

$$K(t) = b(t)^{-1}(t+1)\begin{pmatrix} -2 & -1 \\ 1 & -2 \end{pmatrix},$$

so all our assumptions hold locally over the time interval $[0, +\infty[$.

The Riccati equation (2.1) becomes, in this simple situation,

$$P'(t) + (1+t)[A_1^* P(t) + P(t)A_1] + 4[2(1+t)^2 - 1]I - P(t)^2 = 0, \qquad t \geq 0. \tag{4.2}$$

The nonnegative symmetric solution $P_T(\cdot)$ of equation (4.2) over the interval $[0, T]$, with final datum $P(T) = 0$, is given by

$$P_T(t) = 4(1+t)I - v_T(t)^{-1} I,$$

where

$$v_T(t) = \frac{\exp[3(T^2 - t^2) + 6(T-t)]}{4(1+T)} - \int_t^T \exp[3(s^2 - t^2) + 6(s-t)]\, ds.$$

It is easily seen that $v_T(t) > 0$ for each $t \in [0, T]$ and that $\lim_{T\to\infty} v_T(t) = +\infty$; hence for each $t \geq 0$

$$P_T(t) \uparrow P_\infty(t) := 4(1+t)I \quad \text{as } T \uparrow \infty.$$

The optimal control u^* is given by $u^*(t) = -4b(t)^{-1}(1+t)Iy^*(t)$, and the optimal trajectory y^* is the solution of the closed-loop system

$$y'(t) = (1+t)A_2 y, \quad t \geq 0, \quad y(0) = (x_1, x_2),$$

where

$$A_2 = \begin{pmatrix} -3 & 1 \\ -1 & -3 \end{pmatrix}.$$

We remark that the optimal trajectory is stable. Furthermore it can be seen that $P_\infty(t) = 4(1+t)I$ $(t \geq 0)$ is the only positive solution of the Riccati equation (4.2).

An infinite-dimensional example can be easily obtained by the above example, by just adding to it, as a direct sum, a control problem with unbounded time-invariant operators. In a similar way one can easily arrange things in such a way that the resolvent operator of $A(t)$ is not compact for any $t \geq 0$, using, for instance, multiplicative operators in infinite-dimensional spaces (compare with Remark 2.8(ii)).

4.2. Parabolic equations in noncylindrical domains. Let Ω_0 be a bounded open set of $\mathbb{R}^n$ with smooth boundary Γ_0. Following [DZ], [A2] we consider the family of mappings $\{T_t(\cdot) : \mathbb{R}^n \to \mathbb{R}^n,\ t \geq 0\}$ associated with a family of regular vector fields $\{V(t,\cdot) : \mathbb{R}^n \to \mathbb{R}^n,\ t \geq 0\}$ by the dynamic system

$$\frac{\partial}{\partial t} T_t(x) = V(t, T_t(x)), \quad T_0(x) = x, \quad t \geq 0,\ x \in \mathbb{R}^n.$$

Consider the sets $\Omega_t := T_t(\Omega_0)$ with boundary $\Gamma_t := T_t(\Gamma)$, and the evolution domain $Q = \cup_{t>0}\{t\} \times \Omega_t$ with boundary $\Sigma = \cup_{t>0}\{t\} \times \Gamma_t$.

We want to apply the results of the preceding sections to the following problem: minimize among all $u \in L^2_{\text{loc}}(\Sigma)$ the functional

$$J(u) = \int_0^\infty \int_{\Omega_t} |y(t,\xi)|^2 d\xi\, dt + \int_\Sigma |u(t,\xi)|^2\, d\Sigma(t,\xi), \tag{4.3}$$

where y is the solution of the parabolic boundary problem

$$\begin{cases} y_t(t,\xi) = \Delta y(t,\xi), & t \geq 0,\ \xi \in \Omega_t, \\ y(t,\xi) = u(t,\xi) \quad \text{or} \quad \dfrac{\partial y}{\partial \nu_t}(t,\xi) = u(t,\xi), & t \geq 0,\ \xi \in \Gamma_t, \\ y(0,\xi) = y_0(\xi), & \xi \in \Omega_0. \end{cases} \tag{4.4}$$

(ν_t is the outward normal to Γ_t.)

Denote by DT_t the Jacobian matrix of T_t and by J_t its determinant; then the change of variable

$$z(t,x) = y(t, T_t(x)), \quad t \geq 0, \quad x \in \Omega_0, \quad v(t,x) = u(t, T_t(x)), \quad t \geq 0, \quad x \in \Gamma_0,$$

transforms problem (4.3), (4.4) into the following one: minimize among all $v \in L^2_{\text{loc}}(\partial\Omega \times \mathbb{R}^+)$ the functional

$$\begin{aligned} J_0(v) = &\int_0^\infty \int_{\Omega_0} |z(t,x)|^2 J_t(x)\, dx\, dt \\ &+ \int_0^\infty \int_{\partial\Omega_0} |v(t,\xi)|^2 J_t(x) B(t,x)\, dH_{n-1}(x)\, dt, \end{aligned} \tag{4.5}$$

where z is the solution of the parabolic boundary problem

$$\begin{cases} z_t(t,x) = \mathcal{A}(t,x,D)y, & t \geq 0,\ x \in \Omega_0, \\ z(t,x) = v(t,x) \quad \text{or} \quad \dfrac{\partial z(t,x)}{\partial \nu_{A(t)}} = v(t,x)\beta(t,x), & t \geq 0,\ x \in \Gamma_0, \\ z(0,x) = y_0(x), & x \in \Omega_0, \end{cases} \tag{4.6}$$

with

$$\begin{aligned} B(t,x) &:= \sqrt{1+[(V(t,x),n_t(x))_{\mathbb{R}^n}]^2}, \\ \mathcal{A}(t,x,D)w &:= -J_t(x)^{-1}\,\mathrm{div}\,((DT_t(x)^{-1})(DT_t(x)^{-1})^*\cdot Dw(x)) \\ &\quad +((DT_t(x)^{-1})\cdot Dw(x),V(t,x))_{\mathbb{R}^n}, \\ \beta(t,x) &:= J_t(x)|(DT_t(x)^{-1})^*\cdot n_0(x)|, \\ \nu_{A(t)} &:= (DT_t(x)^{-1})(DT_t(x)^{-1})^*\cdot \nu_0(x). \end{aligned}$$

Problem (4.5) – (4.6) can be studied with the methods of this paper. It is shown in [DZ] (in the case of Dirichlet boundary control) and in [A2] (in the case of Neumann boundary control) that an admissible control exists; more precisely, in both cases Hypothesis 3.1 holds true.

4.3. Strongly damped wave equation. Let $\Omega\subset\mathbb{R}^n$ be a bounded open set with smooth boundary $\partial\Omega$. Consider the Dirichlet or Neumann boundary control problem for the damped wave equation in $]0,\infty[\times\Omega$:

$$\text{(4.7)}\qquad \begin{cases} y_{tt}(t,x)=\Delta y(t,x)+\rho(t)\Delta y_t(t,x), & t>0,\ x\in\Omega, \\ y(0,x)=y_0(x), \quad y_t(0,x)=w_0(x), & x\in\Omega, \\ By(t,x)=u(t,x), \quad t>0,\ x\in\partial\Omega & \left[B=I \text{ or } B=\dfrac{\partial}{\partial\nu}\right]; \end{cases}$$

here ρ is a scalar function belonging to $C^{\varepsilon+1/2}([0,\infty[)$; the data y_0, w_0 belong to $H^1(\Omega)$ and $L^2(\Omega)$, respectively; and Δ is the Laplace operator. The cost functional

$$\text{(4.8)}\qquad \begin{aligned} J(u)=\int_0^\infty \{ &\|C_1(t)y(t,\cdot)\|^2_{L^2(\Omega)}+\|C_2(t)y_t(t,\cdot)\|^2_{L^2(\Omega)}+\|C_3(t)Dy(t,\cdot)\|^2_{L^2(\Omega)} \\ &+(N_1(t)u(t,\cdot),u(t,\cdot))_{L^2(\partial\Omega)}+(N_2(t)u_t(t,\cdot),u_t(t,\cdot))_{L^2(\partial\Omega)}\}\,dt \end{aligned}$$

has to be minimized among all $u\in W^{1,2}_{\mathrm{loc}}(0,\infty;L^2(\partial\Omega))$, with y subject to (4.7); the operators C_1, C_2, C_3 and N_1, N_2 belong to $L^\infty_{\mathrm{loc}}(0,\infty;\mathcal{L}(L^2(\Omega)))$ and $L^\infty_{\mathrm{loc}}(0,\infty;\Sigma^{++}(L^2(\partial\Omega)))$, respectively.

In order to apply the results of this paper we rewrite problem (4.7) – (4.8) in abstract form. Define

$$\text{(4.9)}\qquad \begin{cases} D_A=\{y\in H^2(\Omega)\colon By=0 \text{ on } \partial\Omega\}, \\ Ay=\Delta y, \end{cases}$$

$$\text{(4.10)}\qquad G: L^2(\partial\Omega)\to L^2(\Omega), \qquad Gu=z \Leftrightarrow \begin{cases} \Delta z=z & \text{in } \Omega, \\ Bz=u & \text{in } \partial\Omega. \end{cases}$$

Then, following [B2], it is easy to see that if $u\in W^{2,2}_{\mathrm{loc}}(0,\infty;L^2(\partial\Omega))$ then the function

$$Z:=\begin{pmatrix} y \\ y_t \end{pmatrix}-G\begin{pmatrix} u \\ u_t \end{pmatrix}$$

solves

$$\text{(4.11)}\qquad \begin{cases} Z_t=\begin{pmatrix} 0 & 1 \\ \Delta & \rho(t)\Delta \end{pmatrix}Z+\begin{pmatrix} 0 \\ Gu+\rho(t)Gu_t-Gu_{tt} \end{pmatrix}, & t>0,\ x\in\Omega, \\ Z(0)=\begin{pmatrix} y_0-Gu(0) \\ w_0-Gu_t(0) \end{pmatrix}, \quad x\in\Omega; \quad BZ=0, \quad t>0,\ x\in\partial\Omega. \end{cases}$$

Now set $H := H^1(\Omega) \times L^2(\Omega)$, $U := L^2(\partial\Omega)$, $V := H$, and

$$\begin{cases} D_{\mathcal{A}(t)} = \left\{ \begin{pmatrix} y \\ w \end{pmatrix} \in H : y + \rho(t)w \in D_A \right\}, \\ \mathcal{A}(t) \begin{pmatrix} y \\ w \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \Delta & \rho(t)\Delta \end{pmatrix} \begin{pmatrix} y \\ w \end{pmatrix}; \end{cases}$$

then it is shown in [L3] that (i) the operators $\mathcal{A}(t)$ generate analytic semigroups in H, and (ii) they satisfy the assumptions of [AT1], [AT2] and [AT3,§6]; this in turn yields that the evolution operator $U_{\mathcal{A}}(t,s)$ associated with $\{\mathcal{A}(t)\}$ exists and fulfills the hypotheses of this paper. (We remark that by choosing $H = L^2(\Omega) \times L^2(\Omega)$, as done in [B1], [B2] in the autonomous case, we still would have (i), but (ii) would no longer be true.)

Consequently we can write the mild form of (4.11), and after integrating by parts we get

$$\begin{aligned} Z(t) = U_{\mathcal{A}}(t,0) \begin{pmatrix} y_0 \\ w_0 \end{pmatrix} - \begin{pmatrix} 0 \\ Gu'(t) \end{pmatrix} + \int_0^t U_{\mathcal{A}}(t,s) \begin{pmatrix} 0 \\ Gu(s) + \rho(s)u'(s) \end{pmatrix} ds \\ - \int_0^t U_{\mathcal{A}}(t,s)\mathcal{A}(s)G \begin{pmatrix} 0 \\ u'(s) \end{pmatrix} ds, \qquad t \geq 0 \end{aligned}$$

(where the last term has to be interpreted as in (1.4)–(1.5)); by density we see that this formula holds for all $u \in W^{1,2}_{\mathrm{loc}}(0,\infty; L^2(\partial\Omega))$. Hence, setting $L := \binom{G}{0}$, $M := \binom{0}{G}$ for the sake of simplicity, we obtain, for $Y := \binom{y}{y_t} \in L^2_{\mathrm{loc}}(0,\infty; H)$,

$$\begin{aligned} Y(t) = U_{\mathcal{A}}(t,0) \begin{pmatrix} y_0 \\ w_0 \end{pmatrix} + Lu(t) + \int_0^t U_{\mathcal{A}}(t,s)Mu(s)\,ds \\ + \int_0^t U_{\mathcal{A}}(t,s)[\rho(s) - \mathcal{A}(s)]Mu'(s)\,ds. \end{aligned}$$

Now, as in [B1], [B2], we regard the control u as an auxiliary component of the state and define u' as a new control; namely, we set $v := u'$, $X := (Y,u)$, $\bar{H} := H \times U$, $\bar{U} := U$, and $\bar{V} := \bar{H}$ and look for the state equation satisfied by X. As shown in [B1,§2], $X(t)$ is the mild solution of

$$\begin{cases} X'(t) = \mathcal{B}(t)X(t) + Q(t)v(t), \qquad t > 0, \\ X(0) = X_0, \end{cases}$$

where

$$\text{(4.12)} \qquad \begin{cases} D_{\mathcal{B}(t)} = \left\{ \begin{pmatrix} Y \\ u \end{pmatrix} \in \bar{H} : Y - Lu \in D_{\mathcal{A}(t)} \right\} \\ \mathcal{B}(t) = \begin{pmatrix} \mathcal{A}(t) & M - \mathcal{A}(t)L \\ 0 & 0 \end{pmatrix}, \end{cases}$$

$$Q(t) = \begin{pmatrix} L + [\rho(t) - \mathcal{A}(t)]M \\ 1 \end{pmatrix}, \qquad X_0 = \begin{pmatrix} Y(0) \\ u(0) \end{pmatrix}.$$

It is easy to see that the operators $\mathcal{B}(t)$ possess in $\bar{H}$ the same properties enjoyed by the operators $\mathcal{A}(t)$ in H; in particular $\rho(\mathcal{B}(t)) = \rho(\mathcal{A}(t))$ and

$$\begin{aligned} &[\lambda - \mathcal{B}(t)]^{-1} \\ &= \begin{pmatrix} [\lambda - \mathcal{A}(t)]^{-1} & [\lambda - \mathcal{A}(t)]^{-1}M - \mathcal{A}(t)[\lambda - \mathcal{A}(t)]^{-1}L \\ 0 & 1 \end{pmatrix} \quad \forall\, \lambda \in \rho(\mathcal{B}(t)). \end{aligned}$$

In addition the evolution operator $U_{\mathcal{B}}(t,s)$ associated with $\{\mathcal{B}(t)\}$ exists; it fulfills the hypotheses of this paper and has the following explicit representation:

$$U_{\mathcal{B}}(t,s) = \begin{pmatrix} U_{\mathcal{A}}(t,s) & \int_s^t U_{\mathcal{A}}(t,\sigma)M\,d\sigma + [1 - U_{\mathcal{A}}(t,s)]L \\ 0 & 1 \end{pmatrix}.$$

The operator $Q(t)$ may be also written (improperly but usually) as $Q(t) = [1 - \mathcal{B}(t)]\mathcal{G}(t)$, where $\mathcal{G}(t) = [1 - \mathcal{B}(t)]^{-1}Q(t)$ is given, after some manipulations, by

$$\mathcal{G}(t) = \begin{pmatrix} L + M + \rho(t)[1 - \mathcal{A}(t)]^{-1}M \\ 1 \end{pmatrix}.$$

Hence the state $X(t)$ solves the equation

$$X(t) = U_{\mathcal{B}}(t,0)X_0 + \int_0^t U_{\mathcal{B}}(t,s)[1 - \mathcal{B}(s)]\mathcal{G}(s)v(s)\,ds, \qquad t \geq 0. \tag{4.13}$$

We remark that, conversely, if $X(t)$ is given by this formula, then, setting $X(t) = (Y(t), u(t))$, the second component of X' gives $u' = v$, and from the first component it is easy to go back to (4.11) and hence to the solution y of the original problem.

Concerning the cost functional $J(u)$, we can rewrite it as $\bar{J}(v)$, where

$$\bar{J}(v) = \int_0^\infty \left\{ \|C(t)X(t)\|_{\bar{H}}^2 + (N(t)v(t), v(t))_{\bar{U}} \right\} dt, \tag{4.14}$$

with $C(t)$ and $N(t)$ given by

$$C(t)\begin{pmatrix} Y \\ u \end{pmatrix} \equiv C(t)\begin{pmatrix} Y_1 \\ Y_2 \\ u \end{pmatrix} = C_1(t)Y_1 + C_2(t)Y_2 + C_3(t)DY_1 + [N_1(t)]^{1/2}u,$$
$$N(t)v = N_2(t)v.$$

Thus the original control problem (4.7)–(4.8) is equivalent to minimizing $\bar{J}(v)$ among all $v \in L^2_{\mathrm{loc}}(0,\infty;\bar{U})$, with X subject to equation (4.13).

In order to apply the theory of this paper we still need to verify Hypothesis 1.3 for $\mathcal{G}(t)$ (and this follows by the results of [B1], [B2]) and the finite cost condition (Hypothesis 2.2). Concerning the latter, in the case of Dirichlet boundary control it is satisfied by choosing $u = 0$, as the following proposition shows.

PROPOSITION 4.1. *Let $J(u)$ be given by* (4.8), *where y satisfies* (4.7) *with $B = I$, and assume that $\rho_1 \geq \rho(t) \geq \rho_0 > 0$ for each $t > 0$. Then we have $J(0) < \infty$.*

Proof. Multiply the partial differential equation (PDE) in (4.7) by y_t and integrate over Ω; then

$$\frac{1}{2}\frac{d}{dt}\int_\Omega |y_t(t,x)|^2\,dx = -\frac{1}{2}\frac{d}{dt}\int_\Omega |Dy(t,x)|^2\,dx - \rho(t)\int_\Omega |Dy_t(t,x)|^2\,dx.$$

Integrating over $]0,T[$ we get

$$\int_\Omega |y_t(T,x)|^2\,dx + \int_\Omega |Dy(T,x)|^2\,dx$$
$$= \int_\Omega |w_0|^2\,dx + \int_\Omega |Dy_0|^2\,dx - 2\int_0^T \rho(t)\int_\Omega |Dy_t(t,x)|^2\,dx\,dt;$$

this implies that

$$\sup_{T>0}\left[\int_\Omega |y_t(T,x)|^2\,dx+\int_\Omega |Dy(T,x)|^2\,dx\right]+\int_0^\infty \rho(t)\int_\Omega |Dy_t(t,x)|^2\,dx\,dt \le c\left[\int_\Omega |w_0|^2\,dx+\int_\Omega |Dy_0|^2\,dx\right]. \tag{4.15}$$

As $\rho(t)\ge\rho_0$, by the Poincaré inequality we also get

$$\int_0^\infty\int_\Omega |y_t(t,x)|^2\,dx\,dt \le c\left[\int_\Omega |w_0|^2\,dx+\int_\Omega |Dy_0|^2\,dx\right]. \tag{4.16}$$

On the other hand, multiplying the PDE in (4.7) by y and integrating over $]0,T[\times\Omega$, we obtain after some integrations by parts

$$\begin{aligned}&\int_\Omega y_t(T,x)y(T,x)\,dx-\int_\Omega w_0y_0\,dx\\ &\quad=\int_0^T\int_\Omega |y_t(t,x)|^2\,dx\,dt-\int_0^T\int_\Omega |Dy(t,x)|^2\,dx\,dt\\ &\quad\quad-\int_0^T\rho(t)\int_\Omega Dy_t(t,x)\cdot Dy(t,x)\,dx\,dt,\end{aligned}$$

which implies

$$\begin{aligned}&\int_0^T\int_\Omega |Dy(t,x)|^2\,dx\,dt\\ &\quad\le\int_\Omega w_0y_0\,dx+\frac12\int_\Omega |y(T,x)|^2\,dx+\frac12\int_\Omega |y_t(T,x)|^2\,dx\\ &\quad\quad+\frac{1}{\eta}\rho_1\int_0^T\int_\Omega |Dy_t|^2\,dx\,dt+\eta\rho_1\int_0^T\int_\Omega |Dy|^2\,dx\,dt;\end{aligned}$$

hence if η is sufficiently small, again using the Poincaré inequality we get

$$\begin{aligned}&\int_0^\infty\int_\Omega |y(t,x)|^2\,dx\,dt+\int_0^\infty\int_\Omega |Dy(t,x)|^2\,dx\,dt\\ &\quad\le c\left[\int_\Omega w_0y_0\,dx+\int_\Omega |y(T,x)|^2\,dx+\int_\Omega |y_t(T,x)|^2\,dx+\int_0^T\int_\Omega |Dy_t|^2\,dx\,dt\right],\end{aligned}$$

and by (4.15) we finally obtain

$$\int_0^\infty\int_\Omega |y|^2\,dx\,dt+\int_0^\infty\int_\Omega |Dy|^2\,dx\,dt\le c\left[\int_\Omega |w_0|^2\,dx+\int_\Omega |y_0|^2\,dx+\int_\Omega |Dy_0|^2\,dx\right]. \tag{4.17}$$

The result now follows by (4.8), (4.17), and (4.16). □

Remark 4.2. The above proposition and the results of [D] imply that the evolution operator $U_{\mathcal{A}}(t,s)$ associated with $\{\mathcal{A}(t)\}$ is exponentially stable; i.e., it satisfies

$$\|U_{\mathcal{A}}(t,s)\|_{\mathcal{L}(H)}\le c\,e^{-\beta(t-s)}\quad\forall\,0\le s\le t$$

for some $\beta>0$.

In the case of Neumann boundary control, the finite cost condition is fulfilled too; indeed, we have the following proposition.

PROPOSITION 4.3. *Let $J(u)$ be given by* (4.8), *where y satisfies* (4.7) *with $B = \partial/\partial\nu$, and assume that $\rho_1 \geq \rho(t) \geq \rho_0 > 0$ for each $t > 0$. Then there exists $u \in L^2(0,\infty; L^2(\partial\Omega))$ such that $J(u) < \infty$.*

Proof. Multiply the PDE in (4.7) by y_t and integrate over Ω; then

$$\frac{1}{2}\frac{d}{dt}\int_\Omega |y_t(t,x)|^2\,dx = \int_{\partial\Omega} u(t,x)y_t(t,x)\,d\sigma_x + \rho(t)\int_{\partial\Omega} u_t(t,x)y_t(t,x)\,d\sigma_x$$
$$-\frac{1}{2}\frac{d}{dt}\int_\Omega |Dy(t,x)|^2\,dx - \rho(t)\int_\Omega |Dy_t(t,x)|^2\,dx.$$

Choose the feedback control $u = -y|_{\partial\Omega}$; we then have

$$\frac{1}{2}\frac{d}{dt}\int_\Omega |y_t(t,x)|^2\,dx + \frac{1}{2}\frac{d}{dt}\int_\Omega |Dy(t,x)|^2\,dx + \frac{1}{2}\frac{d}{dt}\int_{\partial\Omega} |y(t,x)|^2\,d\sigma_x$$
$$= -\rho(t)\int_{\partial\Omega} |y_t(t,x)|^2\,d\sigma_x - \rho(t)\int_\Omega |Dy_t(t,x)|^2\,dx$$

so that integrating over $]0,T[$ we get

$$\int_\Omega |y_t(T,x)|^2\,dx + \int_\Omega |Dy(T,x)|^2\,dx + \int_{\partial\Omega} |y(T,x)|^2\,d\sigma_x$$
$$= \int_\Omega |w_0|^2\,dx + \int_\Omega |Dy_0|^2\,dx + \int_{\partial\Omega} |y_0|^2\,d\sigma_x$$
$$-2\int_0^T \rho(t)\int_{\partial\Omega} |y_t(t,x)|^2\,d\sigma_x\,dt - 2\int_0^T \rho(t)\int_\Omega |Dy_t(t,x)|^2\,dx\,dt.$$

This implies that

$$\sup_{T>0}\left[\int_\Omega |y_t(T,x)|^2\,dx + \int_\Omega |Dy(T,x)|^2\,dx + \int_{\partial\Omega} |y(T,x)|^2\,d\sigma_x\right]$$
$$+\int_0^\infty \rho(t)\int_\Omega |Dy_t(t,x)|^2\,dx\,dt + \int_0^\infty \rho(t)\int_{\partial\Omega} |y_t(t,x)|^2\,d\sigma_x\,dt \tag{4.18}$$
$$\leq c\left[\int_\Omega |w_0|^2\,dx + \int_\Omega |Dy_0|^2\,dx + \int_{\partial\Omega} |y_0|^2\,d\sigma_x\right].$$

On the other hand, multiplying the PDE in (4.7) by y and integrating over $]0,T[\times\Omega$, we obtain after some integrations by parts

$$\int_\Omega y_t(T,x)y(T,x)\,dx - \int_\Omega w_0y_0\,dx + \int_0^T\int_{\partial\Omega} |y(t,x)|^2\,d\sigma_x\,dt$$
$$= \int_0^T\int_\Omega |y_t(t,x)|^2\,dx\,dt - \int_0^T\int_\Omega |Dy(t,x)|^2\,dx\,dt$$
$$-\int_0^T \rho(t)\int_{\partial\Omega} y_t(t,x)y(t,x)\,d\sigma_x\,dt - \int_0^T \rho(t)\int_\Omega Dy_t(t,x)\cdot Dy(t,x)\,dx\,dt;$$

hence by (4.18) one easily finds that

$$\begin{aligned}\frac{1}{2}\int_0^\infty\int_\Omega |Dy(t,x)|^2\,dx\,dt + \frac{1}{2}\int_0^\infty\int_{\partial\Omega} |y(t,x)|^2\,d\sigma_x\,dt \\ \leq \int_\Omega |y(T,x)|^2\,dx + c\left[\int_\Omega |w_0|^2\,dx + \int_\Omega |Dy_0|^2\,dx + \int_\Omega |y_0|^2\,dx\right. \\ \left. + \int_{\partial\Omega} |y_0|^2\,d\sigma_x + \int_0^\infty\int_\Omega |y_t(t,x)|^2\,dx\,dt\right].\end{aligned} \tag{4.19}$$

Now we have the following lemma.

LEMMA 4.4. *There exists $c > 0$ such that*

$$\|f\|_{L^2(\Omega)} \leq c\left[\|Df\|_{L^2(\Omega)} + \|f\|_{L^2(\partial\Omega)}\right] \quad \forall f \in H^1(\Omega).$$

Proof. The proof is by contradiction; otherwise there should exist a sequence $\{f_k\}$ in $H^1(\Omega)$ such that

$$\|f_k\|_{L^2(\Omega)} > k\left[\|Df_k\|_{L^2(\Omega)} + \|f_k\|_{L^2(\partial\Omega)}\right] \quad \forall k \in \mathbb{N}^+.$$

In particular, the right member is not zero (since in that case $f_k \equiv 0$). Then if we set

$$g_k(x) = \frac{f_k(x)}{\|Df_k\|_{L^2(\Omega)} + \|f_k\|_{L^2(\partial\Omega)}},$$

we have

$$\|g_k\|_{L^2(\Omega)} > k, \quad \|Dg_k\|_{L^2(\Omega)} + \|g_k\|_{L^2(\partial\Omega)} = 1 \quad \forall k \in \mathbb{N}^+.$$

Hence for a suitable subsequence we get $Dg_k \to z$ weakly in $L^2(\Omega)$ and $g_k \to w$ weakly in $L^2(\partial\Omega)$.

Now let $\varphi \in L^2(\Omega)$ and take the solution $\Phi \in H^2(\Omega) \cap H_0^1(\Omega)$ of $\Delta\Phi = \varphi$ in Ω. Then as $k \to \infty$ we have

$$\begin{aligned}\int_\Omega g_k\varphi\,dx &= \int_\Omega g_k\,\Delta\Phi\,dx \\ &= \int_{\partial\Omega} g_k\frac{\partial\Phi}{\partial\nu}d\sigma_x - \int_\Omega Dg_k\cdot D\Phi\,dx \to \int_{\partial\Omega} w\frac{\partial\Phi}{\partial\nu}d\sigma_x - \int_\Omega z\cdot D\Phi\,dx,\end{aligned}$$

which implies that $\{g_k\}$ is weakly convergent in $L^2(\Omega)$, but this is impossible since $\{g_k\}$ is not bounded in $L^2(\Omega)$.

Let us return to (4.19); by Lemma 4.4 and (4.18) we obtain

$$\begin{aligned}&\int_0^\infty\int_\Omega |Dy(t,x)|^2\,dx\,dt + \int_0^\infty\int_{\partial\Omega} |y(t,x)|^2\,d\sigma_x\,dt \\ &\quad + \int_0^\infty\int_\Omega |Dy_t(t,x)|^2\,dx\,dt + \int_0^\infty\int_{\partial\Omega} |y_t(t,x)|^2\,d\sigma_x\,dt \\ &\quad \leq c\left[\int_\Omega |w_0|^2\,dx + \int_\Omega |Dy_0|^2\,dx + \int_\Omega |y_0|^2\,dx + \int_{\partial\Omega} |y_0|^2\,d\sigma x\right],\end{aligned}$$

and consequently the choice $u = y|_{\partial\Omega}$ implies $J(u) < \infty$. □

Remark 4.4. We have in fact verified that Hypothesis 3.1 holds too.

Remark 4.5. A more general approach to problem (4.7)–(4.8) in the autonomous case, which allows one to take controls $u \in L^2(0,\infty;L^2(\partial\Omega))$, can be found in [LLP], [T].

4.4. Structurally damped plate equation. Let $\Omega \subset \mathbb{R}^n$ be a bounded open set with smooth boundary $\partial\Omega$. Consider the following Dirichlet or Neumann boundary control problem for the structurally damped plate equation in $]0,\infty[\times \Omega$:

$$\text{(4.20)} \quad \begin{cases} y_{tt}(t,x) = -\Delta^2 y(t,x) + \rho(t)\Delta y_t(t,x), & t>0,\ x\in\Omega, \\ y(0,x) = y_0(x), \quad y_t(0,x) = w_0(x), & x\in\Omega, \\ By(t,x) = 0, \quad By_t(t,x) = u(t,x), & t>0,\ x\in\partial\Omega \end{cases} \left[B = I \text{ or } B = \frac{\partial}{\partial\nu}\right].$$

Here ρ is a scalar function belonging to $C^\varepsilon([0,\infty[)$; the data y_0, w_0 belong to $H^2(\Omega)$ and $L^2(\Omega)$, respectively. The cost functional

$$\text{(4.21)} \quad \begin{aligned} J(u) = \int_0^\infty \{ &\|y(t,\cdot)\|^2_{L^2(\Omega)} + \|y_t(t,\cdot)\|^2_{L^2(\Omega)} + \|Dy(t,\cdot)\|^2_{L^2(\Omega)} \\ &+ \|D^2 y(t,\cdot)\|^2_{L^2(\Omega)} \\ &+ \|u(t,\cdot)\|^2_{L^2(\partial\Omega)} + \|u_t(t,\cdot)\|^2_{L^2(\partial\Omega)}\}\, dt \end{aligned}$$

has to be minimized among all $u \in W^{1,2}_{\text{loc}}(0,\infty; L^2(\partial\Omega))$, with y subject to (4.20). Following the same method of the preceding example, we define A, G as in (4.9), (4.10), set $H = D_A \times L^2(\Omega)$, $U = L^2(\partial\Omega)$, and finally rewrite the problem in abstract form. It turns out that if $u \in W^{2,2}_{\text{loc}}(0,\infty; L^2(\partial\Omega))$, the function

$$Z := \begin{pmatrix} y \\ y_t \end{pmatrix} - G\begin{pmatrix} u \\ u_t \end{pmatrix}$$

solves

$$\begin{cases} Z'(t) = \mathcal{A}(t)Z(t) + F(t), & t>0, \\ Z(0) = Z_0, \end{cases}$$

where

$$\mathcal{A}(t) := \begin{pmatrix} 0 & 1 \\ -\Delta^2 & \rho(t)\Delta \end{pmatrix}, \qquad Z_0 := \begin{pmatrix} y_0 + (1-A)^{-1}Gu(0) \\ w_0 + (1-A)^{-1}Gu'(0) \end{pmatrix},$$

$$F(t) := \begin{pmatrix} 0 \\ -2Gu(t) + \rho(t)Gu'(t) + (1-A)^{-1}[Gu''(t) - \rho(t)Gu'(t) + Gu(t)] \end{pmatrix}.$$

As $\mathcal{A}(t)$ fulfills the assumptions of [AT1], [AT2], and [AT3, §6], there exists its evolution operator $U_{\mathcal{A}}(t,s)$; hence setting $Y := \binom{y}{y_t}$ and integrating by parts we get

$$\begin{aligned} Y(t) &= Z(t) - \begin{pmatrix} (1-A)^{-1}Gu(t) \\ (1-A)^{-1}Gu'(t) \end{pmatrix} \\ &= U_{\mathcal{A}}(t,0)\begin{pmatrix} y_0 + (1-A)^{-1}Gu(0) \\ w_0 \end{pmatrix} - \begin{pmatrix} (1-A)^{-1}Gu(t) \\ 0 \end{pmatrix} \\ &\quad + \int_0^t U_{\mathcal{A}}(t,s)\begin{pmatrix} (1-A)^{-1}Gu'(s) \\ (1-A)^{-1}Gu(s) - 2Gu(s) \end{pmatrix} ds, \end{aligned}$$

i.e., defining

$$L := \begin{pmatrix} (1-A)^{-1}G \\ 0 \end{pmatrix}, \quad M := \begin{pmatrix} 0 \\ 2G - (1-A)^{-1}G \end{pmatrix}, \quad Y_0 := \begin{pmatrix} y_0 + (1-A)^{-1}Gu(0) \\ w_0 \end{pmatrix},$$

$$Y(t) = U_{\mathcal{A}}(t,0)(Y_0 - Lu(0)) - Lu(t) + \int_0^t U_{\mathcal{A}}(t,s)(Lu'(s) + Mu(s))\,ds.$$

This formula holds for $u \in W^{1,2}_{\mathrm{loc}}(0,\infty; L^2(\partial\Omega))$ as well.

Now, as in the preceding example, we set $\bar{H} := H \times U$, $\bar{U} := U$, and $v(t) := u'(t)$, $X(t) := (Y(t), u(t))$. The state X satisfies

$$\begin{cases} X'(t) = \mathcal{B}(t)X(t) + Qv(t), & t > 0, \\ X(0) = X_0, \end{cases}$$

where $\mathcal{B}(t)$ is defined as in (4.12) and

$$Q := \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad X_0 := \begin{pmatrix} Y_0 \\ u(0) \end{pmatrix}.$$

Arguing as in the preceding example we arrive again to the state equation (4.13) for $X(t)$, where now

$$\mathcal{G}(t) := \begin{pmatrix} (1-\mathcal{A}(t))^{-1}M - \mathcal{A}(t)(1-\mathcal{A}(t))^{-1}L \\ 1 \end{pmatrix}.$$

Note that $\mathcal{G}(t)$ is uniformly bounded in $]0,\infty[$ as an element of $\mathcal{L}(\bar{U}, D_{\mathcal{A}(t)})$. The cost functional (4.21) transforms into

$$\bar{J}(v) = \int_0^\infty \{\|X(t)\|^2_{\bar{H}} + \|v(t)\|^2_{\bar{U}}\}\,dt, \tag{4.22}$$

and our abstract theory applies to the control problem (4.13), (4.22), provided that we verify the finite cost condition (Hypothesis 2.2). Now it turns out that in the case of Dirichlet boundary conditions ($B = I$) one can choose the control $u = 0$, whereas in the case of Neumann boundary conditions $\left(B = \frac{\partial}{\partial\nu}\right)$ one can choose the feedback control $u = y + y_t$. The proof that the cost is finite can be done by adapting the arguments of Propositions 4.1 and 4.3.

Remark 4.6. A more general approach to problem (4.20) – (4.21) in the autonomous case, which allows one to take controls $u \in L^2(0,\infty; L^2(\partial\Omega))$, can be found in [LLP], [T].

Appendix: Proof of Theorem 2.5. We are going to use the contraction principle on a suitable Banach space. For fixed $T_0, T \in \mathbb{R}$ with $T_0 < T$ we set, as in the proof of Lemma 2.9,

$$\begin{aligned} X(T_0,T) := \{\ & P : [T_0,T] \to \Sigma(H) \text{ such that} \\ & \text{(i) } (\lambda_0 - A^*(\cdot))^{1-\alpha}P(\cdot) \in C_s([T_0,T], \mathcal{L}(H)); \\ & \text{(ii) } |[\lambda_0 - A(t)^*]^{1-\alpha}P(t)|_{\mathcal{L}(H)} \le c[1 + (T-t)^{\beta+\alpha-1}]\ \forall t \in [T_0,T[; \\ & \text{(iii) } |[\lambda_0 - A(t)^*]^{1-\alpha}P(t)U(t,s)[\lambda_0 - A(s)]^{\beta}|_{\mathcal{L}(H)} \\ & \quad \le c(T-s)^{\gamma}[1 + (T-t)^{\beta+\alpha-1}](t-s)^{-\beta}\ \forall T_0 \le s < t < T\} \end{aligned} \tag{A.1}$$

with $\gamma := \min\{1-\alpha-\beta, \beta\}$. We endow $X(T_0,T)$ by its natural norm, i.e.,

$$\|P\|_{X(T_0,T)} := \max\{A, B\},$$

where

$$A := \sup_{t\in[T_0,T[} [1+(T-t)^{1-\alpha-\beta}]|[\lambda_0 - A(t)^*]^{1-\alpha}P(t)|_{\mathcal{L}(H)},$$

$$B := \sup_{T_0\le s<t<T} \frac{[1+(T-t)^{1-\alpha-\beta}](t-s)^{\beta}}{(T-s)^{\gamma}} |[\lambda_0 - A(t)^*]^{1-\alpha}P(t)U(t,s)[\lambda_0 - A(s)]^{\beta}|_{\mathcal{L}(H)}.$$

We also set

$$B(\rho) := \{P \in X(T_0,T) : \|P\|_{X(T_0,T)} \le \rho\}.$$

Theorem 2.5 will be a consequence of the following lemma.

LEMMA A.1. *For each $\rho_0 > 0$ there exist $T_0 < T$ and $\rho > 0$ such that for any P_T satisfying*

$$|[\lambda_0 - A(T)^*]^{\beta}P_T[\lambda_0 - A(T)]^{\beta}|_{\mathcal{L}(H)} \le \rho_0$$

the Riccati equation

$$P(t) = U(T,t)^*P_TU(T,t) + \int_t^T U(r,t)^* \times [C(r)^*C(r) - P(r)(\lambda_0 - A(r))G(r)N(r)^{-1}G(r)^*(\lambda_0 - A(r)^*)P(r)]U(r,t)\,dr, \quad t \in [T_0,T[, \tag{A.2}$$

has a unique solution $P(\cdot)$ in $B(\rho)$.

Proof. First set

$$Q_T = (\lambda_0 - A(T)^*)^{\beta}P_T(\lambda_0 - A(T))^{\beta}. \tag{A.3}$$

Now fix $\rho_0 > 0$ and let P_T be such that $|Q_T|_{\mathcal{L}(H)} \le \rho_0$. Consider the map Γ defined on $B(\rho)$ in the following way:

$$[\Gamma(P)](t) = U(T,t)^*P_TU(T,t) + \int_t^T U(r,t)^* \times [C(r)^*C(r) - [(\lambda_0 - A(r)^*)^{1-\alpha}P(r)]^*K(r)[(\lambda_0 - A(r)^*)^{1-\alpha}P(r)]]U(r,t)\,dr, \tag{A.4}$$

where $t \in [T_0,T[$ and

$$K(r) := [(\lambda_0 - A(r))^{\alpha}G(r)]^*N(r)^{-1}[[(\lambda_0 - A(r))^{\alpha}G(r)]^*. \tag{A.5}$$

We remark that $K(\cdot) \in L^{\infty}(]T_0,T[,\mathcal{L}(H))$ by Hypothesis 1.3.

We will show that for suitable T_0 and ρ (independent of the choice of P_T) the map Γ is a contraction in $B(\rho)$.

We start with the following estimate, which is true for $t < r < T$ and follows by (A.1(iii)) and (1.6):

$$\begin{aligned} &|(\lambda_0 - A(r)^*)^{1-\alpha}P(r)U(r,t)(\lambda_0 - A(t))^{1-\alpha}|_{\mathcal{L}(H)} \\ &\le \left|(\lambda_0 - A(r)^*)^{1-\alpha}P(r)U\left(r,\frac{r+t}{2}\right)\left(\lambda_0 - A\left(\frac{r+t}{2}\right)\right)^{\beta}\right|_{\mathcal{L}(H)} \\ &\quad \times \left|\left(\lambda_0 - A\left(\frac{r+t}{2}\right)\right)^{-\beta}U\left(\frac{r+t}{2},t\right)(\lambda_0 - A(t))^{1-\alpha}\right|_{\mathcal{L}(H)} \\ &\le c\,\rho\,(T-t)^{\gamma}[1+(T-r)^{\alpha+\beta-1}](r-t)^{\alpha-1}. \end{aligned} \tag{A.6}$$

By (1.6), (A.5), and (A.4) we deduce (with Q_T and $K(r)$ given by (A.3) and (A.5))

$$
\begin{aligned}
&|(\lambda_0 - A(t)^*)^{1-\alpha}\Gamma(P)(t)|_{\mathcal{L}(H)} \\
&\quad \le |(\lambda_0 - A(t)^*)^{1-\alpha}U(T,t)^*(\lambda_0 - A(T)^*)^{-\beta}Q_T(\lambda_0 - A(T))^{-\beta}U(T,t)|_{\mathcal{L}(H)} \\
&\qquad + \left| \int_t^T (\lambda_0 - A(t)^*)^{1-\alpha}U(r,t)^*C(r)^*C(r)U(r,t)\,dr \right|_{\mathcal{L}(H)} \\
&\qquad + \left| \int_t^T [(\lambda_0 - A(t)^*)^{1-\alpha}U(r,t)^*P(r)(\lambda_0 - A(r))^{1-\alpha}] \right. \\
&\qquad\qquad \left. \times K(r)[(\lambda_0 - A(r)^*)^{1-\alpha}P(r)]U(r,t)\,dr \right|_{\mathcal{L}(H)} \\
&\quad \le c\,\rho_0[1 + (T-t)^{\beta+\alpha-1}] + c(T-t)^{\alpha} \\
&\qquad + c\,\rho^2 \int_t^T (T-t)^{\gamma}[1 + (T-r)^{(\alpha+\beta-1)}]^2(r-t)^{\alpha-1}\,dr \\
&\quad \le c[\rho_0 + 1 + \rho^2(T-t)^{\min\{\gamma+\alpha,\gamma+\beta+2\alpha-1\}}][1 + (T-t)^{\alpha+\beta-1}] \quad \forall t \in [T_0, T[.
\end{aligned}
$$

On the other hand, for $T_0 \le s < t < T$ we have by (A.4), (1.6), and (A.1(ii))–(A.1(iii))

$$
\begin{aligned}
&|(\lambda_0 - A(t)^*)^{1-\alpha}[\Gamma(P)](t)U(t,s)(\lambda_0 - A(s))^{\beta}|_{\mathcal{L}(H)} \\
&\quad \le |(\lambda_0 - A(t)^*)^{1-\alpha}U(T,t)^*(\lambda_0 - A(T)^*)^{-\beta} \\
&\qquad \times Q_T(\lambda_0 - A(T))^{-\beta}U(T,s)(\lambda_0 - A(s))^{\beta}|_{\mathcal{L}(H)} \\
&\qquad + \left| \int_t^T (\lambda_0 - A(t)^*)^{1-\alpha}U(r,t)^*C(r)^*C(r)U(r,s)(\lambda_0 - A(s))^{\beta}\,dr \right|_{\mathcal{L}(H)} \\
&\qquad + \left| \int_t^T [(\lambda_0 - A(t)^*)^{1-\alpha}U(r,t)^*P(r)(\lambda_0 - A(r))^{1-\alpha}] \right. \\
&\qquad\qquad \left. \times K(r)[(\lambda_0 - A(r)^*)^{1-\alpha}P(r)]U(r,s)(\lambda_0 - A(s))^{\beta}\,dr \right|_{\mathcal{L}(H)} \\
&\quad \le c\,\rho_0[1 + (T-t)^{\beta+\alpha-1}] + c(T-t)^{\alpha}(t-s)^{-\beta} \\
&\qquad + c\,\rho^2 \int_t^T (T-t)^{\gamma}[1 + (T-r)^{\beta+\alpha-1}]^2(r-t)^{\alpha-1}(T-s)^{\gamma}(r-s)^{-\beta}\,dr \\
&\quad \le c[\rho_0[1 + (T-t)^{\beta+\alpha-1}] + (T-t)^{\alpha}(t-s)^{-\beta} \\
&\qquad + \rho^2[(T-t)^{\gamma+\alpha} + (T-t)^{\gamma+2\beta+3\alpha-2}](T-s)^{\gamma}(t-s)^{-\beta}] \\
&\quad \le c(T-s)^{\gamma}[1 + (T-t)^{\beta+\alpha-1}](t-s)^{-\beta}\,[\rho_0 + 1 + \rho^2(T-t)^{\min\{\gamma+\alpha,\gamma+\beta+2\alpha-1\}}].
\end{aligned}
$$

The above estimates show that

$$
\|\Gamma(P)\|_{X(T_0,T)} \le c[\rho_0 + 1 + \rho^2(T-t)^{\min\{\gamma+\alpha,\gamma+\beta+2\alpha-1\}}].
$$

As $\gamma = \min\{\beta, 1-\alpha-\beta\}$ and $\beta > 1/2 - \alpha$, we have in any case $\gamma + \beta + 2\alpha - 1 > 0$. Hence we can find a large ρ and a T_0 sufficiently close to T such that

$$
\Gamma(P) \in B(\rho) \quad \forall\, P \in B(\rho). \tag{A.7}
$$

Now we have to prove that the map Γ is a contraction in $B(\rho)$. Indeed, if $P, Q \in B(\rho)$ we can estimate the $X(T_0,T)$-norm of $\Gamma(P) - \Gamma(Q)$ exactly as before (and the calculation is even simpler); the result is

$$
\|\Gamma(P) - \Gamma(Q)\|_{X(T_0,T)} \le c\,\rho\,\|P - Q\|_{X(T_0,T)}(T - T_0)^{\min\{\gamma+\alpha,\gamma+\beta+2\alpha-1\}}. \tag{A.8}
$$

Hence we can find a large ρ and a T_0 sufficiently close to T such that both (A.7) and (A.8) hold, and the result follows by the contraction principle. □

REFERENCES

[A1] P. Acquistapace, *Abstract linear non-autonomous parabolic equations: A survey*, in Differential Equations in Banach Spaces, Proceedings Bologna 1991, G. Dore, A. Favini, E. Obrecht, and A. Venni, eds., Lecture Notes in Pure and Appl. Math. 148, M. Dekker, New York, 1993, pp. 1 – 19.

[A2] ———, *Boundary control for parabolic problems in non-cylindrical domains*, in Boundary Control and Variation, Proceedings Sophia-Antipolis 1992, J. P. Zolésio, ed., Lecture Notes in Pure and Appl. Math. 163, M. Dekker, New York, 1994, pp. 1 – 12.

[AT1] P. Acquistapace and B. Terreni, *A unified approach to abstract linear non-autonomous parabolic equations*, Rend. Sem. Mat. Univ. Padova, 78 (1987), pp. 47 – 107.

[AT2] ———, *On fundamental solutions for abstract parabolic equations*, in Differential Equations in Banach Spaces, Proceedings Bologna 1985, A. Favini and E. Obrecht, eds., Lecture Notes in Math. 1223, Springer-Verlag, Berlin, 1986, pp. 1 – 11.

[AT3] ———, *Regularity properties of the evolution operator for abstract linear parabolic equations*, Differential Integral Equations, 5 (1992), pp. 1151 – 1184.

[AFT] P. Acquistapace, F. Flandoli, and B. Terreni, *Initial boundary value problems and optimal control for non-autonomous parabolic systems*, SIAM J. Control Optim., 29 (1991), pp. 89 – 118.

[BDDM] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite Dimensional Systems*, Vol. II, Birkhäuser-Verlag, Basel, 1993.

[B1] F. Bucci, *A Dirichlet boundary control problem for the strongly damped wave equation*, SIAM J. Control Optim., 30 (1992), pp. 1092 – 1100.

[B2] ———, *A Boundary Control Problem with Infinite Horizon for the Strongly Damped Wave Equation*, preprint, Dip. Mat., Univ. Pisa, 1990.

[D] R. Datko, *Uniform asymptotic stability of evolutionary process in Banach spaces*, SIAM J. Math. Anal., 3 (1972), pp. 428 – 445.

[DI1] G. Da Prato and A. Ichikawa, *Bounded solutions on the real line to non-autonomous Riccati equations*, Rend. Accad. Naz. Lincei Cl. Sci. Fis. Mat. Nat. (8), 79 (1985), pp. 107 – 112.

[DI2] ———, *Quadratic control for linear periodic systems*, Appl. Math. Optim., 18 (1988), pp. 39 – 66.

[DI3] ———, *Quadratic control for linear time-varying systems*, SIAM J. Control Optim., 28 (1990), pp. 359 – 381.

[DZ] G. Da Prato and J. P. Zolésio, *A boundary control problem for a parabolic equation in noncylindrical domain*, in Stability of Flexible Structures, Proceedings Montpellier 1987, A. V. Balakrishnan and J. P. Zolésio, eds., Optimization Software Inc. Publications Division, New York, 1988, pp. 52 – 61.

[F1] F. Flandoli, *Algebraic Riccati equation arising in boundary control problems*, SIAM J. Control Optim., 25 (1987), pp. 612 – 636.

[F2] ———, *A new proof of an a priori estimate arising in boundary control theory*, Appl. Math. Lett., 2 (1989), pp. 341 – 343.

[F3] ———, *On the direct solutions of Riccati equations arising in boundary control theory*, Ann. Mat. Pura Appl. (4), 163 (1993), pp. 93 – 131.

[F] M. Fuhrman, *Bounded solutions for abstract time-periodic parabolic equations with nonconstant domains*, Differential Integral Equations, 4 (1991), pp. 493 – 518.

[G] J. S. Gibson, *The Riccati integral equations for optimal control problems on Hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537 – 565.

[LT1] I. Lasiecka and R. Triggiani, *Dirichlet boundary control problem for parabolic equations with quadratic cost: Analyticity and Riccati's feedback synthesis*, SIAM J. Control Optim., 21 (1983), pp. 41 – 67.

[LT2] ———, *Differential and Algebraic Riccati Equations with Application to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*. Lecture Notes in Control and Inform. Sci. 164, Springer-Verlag, Berlin Heidelberg, 1991.

[LLP] I. Lasiecka, D. Lukes, and L. Pandolfi, *Input dynamics and nonstandard Riccati equations with applications to boundary control of damped wave and plate equations*, preprint, 1993, J. Optim. Theory Appl., 84 (1995), pp. 549 – 574.

[L1] A. Lunardi, *Bounded solutions of linear periodic abstract parabolic equations*, Proc. Roy. Soc. Edinburgh Sect. A, 110 (1988), pp. 135 – 159.

[L2] ———, *Stabilizability of time periodic parabolic equations*, SIAM J. Control Optim., 29 (1991), pp. 810 – 828.

[L3] ———, *Neumann boundary stabilization of structurally damped time periodic wave and plate equations*, in Differential Equations with Applications in Biology, Physics and Engineering, Proceedings Leibnitz 1990, J. A. Goldstein, F. Kappel, and W. Schappacher, eds., Lecture Notes in Pure and Appl. Math. 133, M. Dekker, New York, 1991, pp. 241 – 257.

[T] R. Triggiani, *Optimal Quadratic Boundary Control Problem for Wave- and Plate-like Equations with High Internal Damping: An Abstract Approach*, in Control of Partial Differential Equations, Proceedings Trento 1993, G. Da Prato and L. Tubaro, eds., Lecture Notes in Pure and Appl. Math. 165, M. Dekker, New York, 1994, pp. 215 – 263.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 31–61, January 1996

002

ON THE AVERAGED STOCHASTIC APPROXIMATION FOR LINEAR REGRESSION*

LÁSZLÓ GYÖRFI† AND HARRO WALK‡

Abstract. For a linear regression function the average of stochastic approximation with constant gain is considered. In case of ergodic observations almost sure convergence is proved, where the limit is biased with small bias for small gain. For independent and identically distributed observations and also under martingale and mixing assumptions, asymptotic normality with $(n^{-1/2})$-convergence order is obtained. In the martingale case the asymptotic covariance matrix is close to the optimum one if the gain is small.

Key words. averaged stochastic approximation, constant gains, linear regression, adaptive filtering, ergodicity, mixing, martingales, almost sure convergence, asymptotic normality

AMS subject classifications. 62L20, 62J05, 93E12

1. Introduction.

According to one of the most important recent results on stochastic approximation the optimal convergence order $n^{-1/2}$, together with an optimal (with respect to the trace) asymptotic covariance matrix, can be achieved if the output of the conventional stochastic approximation is averaged, where the corresponding stochastic approximation has a gain sequence decreasing to zero slower than the usual choice constant/n [21], [24].

In the following discussion $(\cdot,\cdot)$ denotes the inner product in the Euclidean space $\mathbf{R}^d$ and $\|\cdot\|$ stands for the norm of either a d-dimensional vector or a linear operator on $\mathbf{R}^d$. Let I denote the $d\times d$ identity matrix and $\lambda(B)$ and $\Lambda(B)$ denote the smallest and the largest eigenvalues of a symmetric matrix B, respectively; χ is used for denoting an indicator function.

The general assumption is that a sequence of random symmetric positive semidefinite $d \times d$ matrices A_n and a sequence of random d-dimensional vectors V_n, $n = 0, \pm 1, \pm 2, \dots$, are given, where $((A_n, V_n))$ is stationary and ergodic with $\mathbb{E}\|A_n\| < \infty$, $\mathbb{E}\|V_n\| < \infty$. Let $A := \mathbb{E}A_n$, $V := \mathbb{E}V_n$. Assume that A^{-1} exists. The aim is to estimate

$$\vartheta := A^{-1}V$$

on the basis of observations of A_n, V_n for $n \geq 1$.

For this purpose as a primary algorithm a stochastic approximation in $\mathbf{R}^d$ with constant gain $\alpha > 0$ is introduced:

$$X_{n+1} = X_n - \alpha(A_{n+1}X_n - V_{n+1}), \quad n \geq 0, \tag{1.1}$$

with an arbitrary X_0. It is followed by an averaging,

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{1.2}$$

* Received by the editors February 24, 1992; accepted for publication (in revised form) July 19, 1994.

† Department of Mathematics, Technical University of Budapest, Stoczek u. 2, H-1521 Budapest, Hungary.

‡ Mathematisches Institut A, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany.

For reasons of simplicity X_0 is assumed to be deterministic. (In view of assertions on mean convergence one would have to impose integrability assumptions on X_0.)

In the case of independent and identically distributed (i.i.d.) observations with $A_n = A$ almost surely (a.s.), for sufficiently small gain $\alpha > 0$ the rate of convergence of (Y_n) and the asymptotic covariance matrix do not depend on α and are optimal [21], [22]. Unfortunately this property does not hold in general if the sequence (A_n) is random.

The usual motivation of using a constant gain algorithm is to deal with nonstationary cases where the algorithm (1.1) is expected to have some tracking abilities for the time-varying parameters. In this paper we deal only with stationary cases. Here consistency does not hold for the process X_n, but for the averages Y_n. Without averaging it is applied in practice if the computational complexity of the algorithm is a question of interest. Although the main purpose of this paper is to study the properties of the averaging rule (1.2), as a by-product we get some interesting features of (1.1) too, under the very general condition that the observation sequence is ergodic; therefore, the results may be interesting for readers who apply constant gain rule without averaging.

The paper is organized as follows. Let

$$A_{n,k}(\alpha) := (I - \alpha A_n) \dots (I - \alpha A_k) \quad \text{for} \quad k \le n,$$
$$A_{n,n+1}(\alpha) := I.$$

In §2 it will be shown that under the above general conditions for sufficiently small $\alpha > 0$ the almost sure limit δ_α of $(Y_n - \vartheta)$ exists, if

$$U_\alpha := \alpha \sum_{n=1}^{\infty} \|A_{n,1}(\alpha)(V_0 - A_0\vartheta)\| \quad (< \infty \text{ a.s. })$$

is integrable; for the asymptotic bias δ_α one has

$$\delta_\alpha = EX_0^* - \vartheta,$$

where $(X_n^*)_{n \ge 0}$ defined by

$$X_n^* := \vartheta + \alpha \sum_{i=-\infty}^{0} A_{n,n+i+1}(\alpha)(V_{n+i} - A_{n+i}\vartheta)$$

is a stationary and ergodic sequence satisfying recursion (1.1) (Theorem 2.1). In §2, by an averaging argument we also study assumptions (uniform integrability of X_0^* with respect to α, fulfilled under φ- or α-mixing together with uniform boundedness or under M-dependence together with moment conditions) under which $\delta_\alpha \to 0$ $(\alpha \to 0)$ (Theorem 2.7); mixing conditions are often assumed in adaptive filtering, and boundedness or moment conditions prevent the constant gain algorithm from exploding. Under more restrictive assumptions on the dependence of the observations (independence and also martingale case) it is proved in §3 that $\delta_\alpha = 0$ for sufficiently small α (Theorems 3.1 and 3.2). It should be mentioned that in the case of a nonlinear regression function even under i.i.d. observations there is an asymptotic bias. Section 4 concerns asymptotic normality, with $(n^{-1/2})$-convergence order, in the case of not necessarily vanishing δ_α (Theorem 4.1). The more special situation of §3 is studied in more detail in §5, where the $(n^{-1/2})$-convergence order of $Y_n - \vartheta$ to 0, together with

an asymptotic covariance matrix differing from the optimal one by a term of order α, is obtained (Theorems 5.3 and 5.6).

Without averaging the constant gain stochastic approximation has been mainly applied for adaptive filtering, when based on an observed random d-dimensional vector R one has to construct a linear estimate (x, R) of the unobserved real random variable Z such that the vector x^* minimizes the mean square error

$$\mathbb{E}((x, R) - Z)^2.$$

If a training sequence $(R_n, Z_n), n = 0, \pm 1, \pm 2, \ldots$, is given such that (R_n, Z_n) has the same distribution as (R, Z), then (1.1) can be applied with

$$A_n = R_n R_n^T$$

and

$$V_n = Z_n R_n.$$

In this case (1.1) can be written as

$$X_{n+1} = X_n - \alpha((R_{n+1}, X_n) - Z_{n+1})R_{n+1},$$

where there are no matrix operations, and if α is a negative integer power of 2, then one can save some multiplications, which is very important for high-speed communication and signal processing applications. (References on the possible application can be seen, for example, in [12] and [6].)

Remark 1.1. It is easy to construct a 3-dependent sequence for (A_n, V_n) for which $\delta_\alpha \neq 0$, i.e., Y_n is asymptotically biased: let $\{W_i\}$ be i.i.d., ± 1 valued, $EW_i = 0$, and one wants to predict $W_i + W_{i+1}$ from $W_{i-1} + W_i$; this is a one-step prediction problem for a moving average process. Then $d = 1$ and

$$\begin{aligned} A_i &= (W_{i-1} + W_i)^2 = 2(1 + W_{i-1}W_i), \\ V_i &= (W_{i-1} + W_i)(W_i + W_{i+1}) = 1 + W_{i-1}W_i + W_iW_{i+1} + W_{i-1}W_{i+1}, \end{aligned}$$

and $\vartheta = 1/2$. Obviously the observation sequences are 3-dependent. In this example, with the notation

$$d_n = \alpha\mathbb{E}(A_{n,2}(\alpha)(V_1 - A_1\vartheta)), \quad n \geq 2,$$

one has

$$\begin{aligned} &\mathbb{E}(A_{n,2}(\alpha)(V_1 - A_1\vartheta) \mid W_0, \ldots, W_{n-1}) \\ &\quad = \mathbb{E}(1 - \alpha A_n \mid W_0, \ldots, W_{n-1})A_{n-1,2}(\alpha)(V_1 - A_1\vartheta) \\ &\quad = (1 - 2\alpha)A_{n-1,2}(\alpha)(V_1 - A_1\vartheta), \quad n \geq 3, \\ d_2 &= \alpha\mathbb{E}\left((1 - \alpha A_2)(V_1 - A_1\vartheta)\right) = -\alpha^2\mathbb{E}\left(A_2(V_1 - A_1\vartheta)\right) = -2\alpha^2, \end{aligned}$$

and thus

$$\delta_\alpha = \sum_{n=2}^{\infty} d_n = \sum_{n=2}^{\infty}(-2\alpha^2)(1 - 2\alpha)^{n-2} = -\alpha;$$

therefore, Y_n is asymptotically biased, but the bias is small if α is small. It will be shown that this is true for much more general conditions on dependence, too.

The following list of assumptions will be explained in their interdependence and used later.

Assumption A1. There is some $\alpha' > 0$ such that U_α is integrable for $0 < \alpha < \alpha'$.

Assumption A2. A_1 is bounded a.s.; there is some $\alpha'' > 0$ such that X_0^* is uniformly integrable with respect to $0 < \alpha < \alpha''$.

Assumption B1. $E\|A_1\|^q < \infty$, $E\|V_1\|^q < \infty$ for all $q > 0$; there is some $M \in \mathbf{N}$ such that the sequence $((A_n, V_n))$ is M-dependent, i.e., $\sigma(A_0, V_0, A_{-1}, V_{-1}, \ldots)$ and $\sigma(A_{M+1}, V_{M+1}, A_{M+2}, V_{M+2}, \ldots)$ are independent.

Assumption B2a. A_1 and V_1 are bounded a.s.; the sequence $((A_n, V_n))$ is α-mixing, i.e.,

$$\begin{aligned}\alpha_n := \sup\{&|P(B \cap D) - P(B)P(D)|;\ \ B \in \sigma(A_0, V_0, A_{-1}, V_{-1}, \ldots),\\ &D \in \sigma(A_n, V_n, A_{n+1}, V_{n+1}, \ldots)\}\\ &\to 0 \ \ (n \to \infty),\end{aligned}$$

with $\sum \alpha_n < \infty$.

Assumption B2b. A_1 and V_1 are bounded a.s.; the sequence $((A_n, V_n))$ is φ-mixing, i.e.,

$$\begin{aligned}\varphi_n := \sup\{&|P(B \cap D) - P(B)P(D)|P(B)^{-1};\\ &B \in \sigma(A_0, V_0, A_{-1}, V_{-1}, \ldots),\ \ D \in \sigma(A_n, V_n, A_{n+1}, V_{n+1}, \ldots)\}\\ &\to 0 \ \ (n \to \infty).\end{aligned}$$

Assumption B2c. A_1 and V_1 are bounded a.s.; the sequence $((A_n, V_n))$ is φ-mixing (and thus α-mixing) with $\alpha_n = O(n^{-\tau})$ for some $\tau > 1$ or is α-mixing with $\alpha_n = O(n^{-\tau})$ for some $\tau > 2$.

2. Ergodic observations. We note the above-mentioned general assumptions that $((A_n, V_n),\ n = 0, \pm 1, \pm 2, \ldots)$ is stationary and ergodic with existence of the $d \times d$ matrix $\mathbb{E}A_1 =: A$ and of the d-dimensional vector $\mathbb{E}V_1 =: V$ and that the random matrices A_n are symmetric, positive semidefinite, A^{-1} exists. For ergodic observations the standard stochastic approximation works; namely if

$$\alpha_n = \frac{\text{constant}}{n},$$

then

$$X_n \to \vartheta \quad \text{a.s.},$$

which was proved by [3], [5], [15], [18], and [27] for $\|A_1\| < 1$ a.s., $\mathbb{E}\|A_1\|^2 < \infty$, and $\mathbb{E}\|A_1\| < \infty$, respectively. For constant gain α, X_n generally does not converge, but Y_n is a.s. convergent, as the following theorem states. Different choices of X_0 lead to different versions of the process (X_n). The first assertion of Theorem 2.1 tells that there exists a stationary and ergodic version of the process and that each of these versions is asymptotically close to it.

THEOREM 2.1. a) *There exists an* $\alpha''' > 0$ *such that for all* $0 < \alpha < \alpha'''$

$$U_\alpha := \alpha \sum_{n=1}^{\infty} \|A_{n,1}(\alpha)(V_0 - A_0\vartheta)\| \tag{2.1}$$

and

$$\alpha \sum_{i=-\infty}^{0} \|A_{0,i+1}(\alpha)(V_i - A_i\vartheta)\|$$

are a.s. finite. The random elements

$$X_n^* := \vartheta + \alpha \sum_{i=-\infty}^{0} A_{n,n+i+1}(\alpha)(V_{n+i} - A_{n+i}\vartheta), \quad n \geq 0, \tag{2.2}$$

are a.s. defined. (X_n^*) *satisfies recursion* (1.1) *and is stationary and ergodic; further*

$$X_n - X_n^* \to 0 \quad (n \to \infty) \text{ a.s.} \tag{2.3}$$

b) *Suppose that the random element* X_0^* *is integrable; then*

$$Y_n \to \mathbb{E}X_0^* =: \vartheta + \delta_\alpha \quad (n \to \infty) \text{ a.s.} \tag{2.4}$$

Remark 2.2. a) Under assumption A1, for $0 < \alpha < \alpha' (\leq \alpha''')$ the random element X_0^* is integrable. This follows from the relation

$$\alpha \sum_{i=-\infty}^{0} \mathbb{E}\|A_{0,i+1}(\alpha)(V_i - A_i\vartheta)\| = \mathbb{E}U_\alpha$$

obtained by stationarity of $((A_n, V_n), \ n = 0, \pm 1, \pm 2, \ldots)$.

b) Assumption A1 is satisfied under each of the conditions B1, B2a, and B2b.

c) Obviously the integrability of X_0^* is a necessary condition for almost sure convergence of (Y_n).

The proof of Theorem 2.1 requires Lemmas 2.3 and 2.4, which will be proved first. Remark 2.2b) is immediately proved by Lemma 2.6 and the moment or boundedness conditions on A_1 and V_1; Lemma 2.6 will be stated and proved after Remark 2.5.

LEMMA 2.3. *For all* $\alpha > 0$

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E} \log \|A_{n,1}(\alpha)\| = E \tag{2.5}$$

exists and $E < \infty$; *moreover*

$$\lim_{n\to\infty} \frac{1}{n} \log \|A_{n,1}(\alpha)\| = E \quad \text{a.s.} \tag{2.6}$$

Proof. Introduce the notations

$$M_n = I - \alpha A_n,$$
$$E_n = \frac{1}{n} \mathbb{E} \log \|M_n M_{n-1} \ldots M_1\|.$$

According to Furstenberg and Kesten [4], for a stationary and ergodic sequence of square matrices M_n with $\mathbb{E}\{(\log \|M_1\|)^+\} < \infty$ the limit

$$E = \lim_{n\to\infty} E_n \tag{2.7}$$

exists and $E < \infty$; moreover

$$\lim_{n\to\infty} \frac{1}{n} \log \|M_n M_{n-1} \dots M_1\| = E \quad \text{a.s.} \tag{2.8}$$

This result can be applied here since

$$\mathbb{E}\{(\log \|M_1\|)^+\} \le \mathbb{E}\{\log(1 + \alpha\|A_1\|)\} \le \mathbb{E}\{\alpha\|A_1\|\} < \infty. \qquad \square$$

LEMMA 2.4. *There exists an $\alpha''' > 0$ such that for all $0 < \alpha < \alpha'''$ the relation $E < 0$ holds. Consequently*

$$\|A_{n,1}(\alpha)\| \le e^{-n\varepsilon} \quad \text{a.s.} \tag{2.9}$$

for large n, with $\varepsilon > 0$ depending on $\alpha \in (0, \alpha''')$.

Proof. First, one shows

$$\lim_{n\to\infty} E_n = \inf_n E_n. \tag{2.10}$$

Obviously

$$\lim_{n\to\infty} E_n \ge \inf_n E_n.$$

In view of the reverse inequality one obtains for all integers n, N

$$E_{nN} = \frac{1}{nN}\mathbb{E}\log\|M_{nN}M_{nN-1}\dots M_1\| \le \frac{1}{nN}\mathbb{E}\sum_{k=1}^{n}\log\|M_{kN}M_{kN-1}\dots M_{(k-1)N+1}\|$$

$$= \frac{1}{N}\mathbb{E}\log\|M_N M_{N-1}\dots M_1\| = E_N;$$

therefore,

$$\lim_n E_{nN} \le E_N,$$

which implies

$$\lim_n E_n \le \inf_n E_n.$$

Because of (2.10) it is enough to find an integer N^* such that

$$E_{N^*} < 0.$$

By Jensen's inequality

$$E_n \le \log \mathbb{E}\|A_{n,1}(\alpha)\|^{1/n};$$

thus it is sufficient to show that there is an integer N^* such that

$$\mathbb{E}\|A_{N^*,1}(\alpha)\|^{1/N^*} < 1 \tag{2.11}$$

for α sufficiently small. (A_n) is ergodic; therefore

$$\left\| \frac{1}{n}\sum_{i=1}^{n} A_i - A \right\| \to 0 \text{ a.s.}$$

where A is positive definite. Because of Fatou's lemma there is an integer N^* such that

$$\lambda_n := \mathbb{E}\lambda\left(\frac{1}{n}\sum_{i=1}^{n} A_i\right) > 0 \text{ for all } n \geq N^*. \tag{2.12}$$

One shows (2.11) for this N^*. In view of this one proves that

$$\overline{\lim_{\alpha \downarrow 0}} \frac{\mathbb{E}\|A_{N^*,1}(\alpha)\|^{1/N^*} - 1}{\alpha} \leq -\lambda_{N^*}. \tag{2.13}$$

Obviously

$$\begin{aligned}\|A_{N^*,1}(\alpha)\|^{1/N^*} &\leq (\|I - \alpha A_{N^*}\| \, \|I - \alpha A_{N^*-1}\| \ldots \|I - \alpha A_1\|)^{1/N^*} \\ &\leq \left(\prod_{i=1}^{N^*}(1 + \alpha\|A_i\|)\right)^{1/N^*} \\ &\leq 1 + \alpha\frac{1}{N^*}\sum_{i=1}^{N^*}\|A_i\|;\end{aligned}$$

therefore,

$$\frac{\|A_{N^*,1}(\alpha)\|^{1/N^*} - 1}{\alpha} \leq \frac{1}{N^*}\sum_{i=1}^{N^*}\|A_i\|,$$

and because

$$\mathbb{E}\left(\frac{1}{N^*}\sum_{i=1}^{N^*}\|A_i\|\right) = \mathbb{E}\|A_1\| < \infty$$

one can apply Fatou's lemma:

$$\overline{\lim_{\alpha \downarrow 0}} \frac{\mathbb{E}\|A_{N^*,1}(\alpha)\|^{1/N^*} - 1}{\alpha} \leq \mathbb{E}\,\overline{\lim_{\alpha \downarrow 0}} \frac{\|A_{N^*,1}(\alpha)\|^{1/N^*} - 1}{\alpha}.$$

Further, one notices

$$\left\|A_{N^*,1}(\alpha) - \left[I - \alpha\sum_{i=1}^{N^*} A_i\right]\right\| \leq \alpha^2 \prod_{i=1}^{N^*}(1 + \|A_i\|) \text{ for } 0 < \alpha \leq 1, \tag{2.14}$$

$$\left\|I - \alpha\sum_{i=1}^{N^*} A_i\right\| = 1 - \alpha\lambda\left(\sum_{i=1}^{N^*} A_i\right) \text{ for } \alpha \text{ sufficiently small} \tag{2.15}$$

by positive semidefiniteness of $\sum_{i=1}^{N^*} A_i$, and thus obtains (2.13). □

Proof of Theorem 2.1. a) The first assertion immediately follows from Lemmas 2.3 and 2.4. For the second assertion, after a transposition of the matrices one argues in the same way, noting also the a.s. ergodic theorem for $(\|V_i - A_i\vartheta\|)$; the further argument is based on the reverse extension of (1.1). Thus the random element

$$X_0^* := \vartheta + \alpha\sum_{i=-\infty}^{0} A_{0,i+1}(\alpha)(V_i - A_i\vartheta)$$

is a.s. defined. Analogously one obtains for general $n = 0, 1, \ldots$ that X_n^* in (2.2) is a.s. defined. By induction with respect to n, noting

$$(I - \alpha A_{n+1}) \sum_{i=-\infty}^{0} A_{n,n+i+1}(\alpha)(V_{n+i} - A_{n+i}\vartheta)$$
$$= \sum_{i=-\infty}^{-1} A_{n+1,n+i+2}(\alpha)(V_{n+1+i} - A_{n+1+i}\vartheta),$$

one shows that (X_n^*) satisfies recursion (1.1). By (2.2), (X_n^*) is a time-invariant function of the stationary and ergodic sequence $((A_i, V_i))$; thus it is stationary and ergodic, too, where the latter follows by an application of Proposition 4.3 in [11, Chap. 1] to $((A_n, V_n))$ with reversed order of indices. Moreover

$$X_n - X_n^* = A_{n,1}(\alpha)(X_0 - X_0^*) \to 0 \qquad (n \to \infty) \quad \text{a.s.} \tag{2.16}$$

because of (2.9) in Lemma 2.4.

b) If X_0^* is integrable with $\mathbb{E}X_0^* =: \vartheta + \delta_\alpha$, by the almost sure ergodic theorem

$$\frac{1}{n}\sum_{i=1}^{n} X_i^* \to \vartheta + \delta_\alpha \qquad (n \to \infty) \quad \text{a.s.}$$

This and (2.16) yield

$$Y_n \to \vartheta + \delta_\alpha \qquad (n \to \infty) \quad \text{a.s.} \qquad \square$$

Remark 2.5. Unfortunately Lemma 2.4 proves only the existence of α'''. There is a special case of A_n where α''' can be given explicitly. Consider the case of $\|A_1\| < C$ a.s. Choose $\alpha''' = \min\{1/C, 1\}$. Then a.s.

$$\|A_{n,1}(\alpha)\| \le \|1 - \alpha A_n\| \, \|1 - \alpha A_{n-1}\| \ldots \|1 - \alpha A_1\| \le 1 \text{ for } \alpha < \alpha''';$$

therefore, (2.11) is proved if

$$\mathbb{P}\{\|A_{n,1}(\alpha)\| < 1\} > 0.$$

Let $N(A_i)$ be the null-space of A_i; then

$$\|A_{n,1}(\alpha)\| = 1$$

iff

$$\bigcap_{i=1}^{n} N(A_i) \neq 0.$$

A_1 is positive semidefinite; therefore

$$\bigcap_{i=1}^{n} N(A_i) = N\left(\sum_{i=1}^{n} A_i\right) = N\left(\frac{1}{n}\sum_{i=1}^{n} A_i\right).$$

Because of ergodicity

$$\frac{1}{n}\sum_{i=1}^{n} A_i \to A \quad \text{a.s.},$$

and A is invertible; therefore, there is an n such that

$$\mathbb{P}\left\{N\left(\frac{1}{n}\sum_{i=1}^{n}A_i\right)=0\right\}=\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}A_i \text{ is invertible}\right\}>0.$$

LEMMA 2.6. a) *Under Assumption* B1 *or* B2b, *for each* $q\in\mathbf{N}$ *there exist real numbers* $K>0$, $\alpha^*>0$, $\rho>0$ *such that*

$$\mathbb{E}\|A_{n,1}(\alpha)\|^q\le Ke^{-\alpha\rho n} \text{ for all } n\in\mathbf{N} \text{ and } 0<\alpha<\alpha^*.$$

If the second version of Assumption B2c *holds, then for each* $q\in\mathbf{N}$ *and* $p\in\mathbf{N}$ *there exist real numbers* $K>0$, $\alpha^*>0$ *such that*

$$\mathbb{E}\|A_{n,1}(\alpha)\|^q\le Kn^{-p} \text{ for all } n\in\mathbf{N} \text{ and } 0<\alpha<\alpha^*.$$

b) *Under Assumption* B2a

$$\sum_n \mathbb{E}\|A_{n,1}(\alpha)\|<\infty \text{ for sufficiently small } \alpha>0.$$

Proof. a) Choose N^* according to (2.12). First assume B1. Let $k^*:=\max\{N^*,M,2\}$. Because of the moment condition and by the Cauchy–Schwarz inequality, it suffices to show a corresponding assertion for $\|A_{nk^*,1}(\alpha)\|^{2q}$, $n\in\{2,3,\dots\}$. One obtains

$$\begin{aligned}
&\mathbb{E}\|A_{nk^*,1}(\alpha)\|^{2q}\\
&=\mathbb{E}\|A_{nk^*,(n-1)k^*+1}(\alpha)A_{(n-1)k^*,(n-2)k^*+1}(\alpha)\dots A_{k^*,1}(\alpha)\|^{2q}\\
&\le\mathbb{E}\left(\prod_{j=0}^{[\frac{n-1}{2}]}\|A_{(2j+1)k^*,2jk^*+1}(\alpha)\|^{2q}\prod_{j=1}^{[\frac{n}{2}]}\|A_{2jk^*,(2j-1)k^*+1}(\alpha)\|^{2q}\right)\\
&\le\left(\prod_{j=0}^{[\frac{n-1}{2}]}\mathbb{E}\|A_{(2j+1)k^*,2jk^*+1}(\alpha)\|^{4q}\right)^{\frac12}\left(\prod_{j=1}^{[\frac{n}{2}]}\mathbb{E}\|A_{2jk^*,(2j-1)k^*+1}(\alpha)\|^{4q}\right)^{\frac12}\\
&=\left(\mathbb{E}\|A_{k^*,1}(\alpha)\|^{4q}\right)^{\frac{n}{2}},
\end{aligned}$$

where for the second inequality one uses the Cauchy–Schwarz inequality once more and M-dependence. (2.14) for k^* instead of N^* yields

$$\begin{aligned}
&\|A_{k^*,1}(\alpha)\|\\
&\le\left\|I-\alpha\sum_{i=1}^{k^*}A_i\right\|+\alpha^2\prod_{i=1}^{k^*}(1+\|A_i\|)\\
&\le 1-\alpha\lambda\left(\sum_{i=1}^{k^*}A_i\right)+2\alpha\Lambda\left(\sum_{i=1}^{k^*}A_i\right)\chi_{\{\alpha\lambda(\sum_{i=1}^{k^*}A_i)\ge1\}}+\alpha^2\prod_{i=1}^{k^*}(1+\|A_i\|)\\
&\le 1-\alpha\lambda\left(\sum_{i=1}^{k^*}A_i\right)+2\alpha^2\Lambda\left(\sum_{i=1}^{k^*}A_i\right)^2+\alpha^2\prod_{i=1}^{k^*}(1+\|A_i\|).
\end{aligned}$$

Each of the terms

$$\lambda\left(\sum_{i=1}^{k^*} A_i\right),\ \Lambda\left(\sum_{i=1}^{k^*} A_i\right)^2,\ \prod_{i=1}^{k^*}(1+\|A_i\|)$$

is majorized by

$$2^{k^*+1}\max\left\{1, \|A_1\|^{k^*}, \ldots, \|A_{k^*}\|^{k^*}\right\}.$$

Now by the trivial inequality

$$(1+x)^n \le 1 + nx + 2^n \max\{x^2, |x|^n\} \quad (x \in \mathbf{R},\ n \in \mathbf{N})$$

one obtains

$$\begin{aligned}&\|A_{k^*,1}(\alpha)\|^{4q}\\ &\le 1 - 4q\alpha\lambda\left(\sum_{i=1}^{k^*} A_i\right) + \alpha^2 2^{4q(k^*+4)+1}\max\left\{1, \|A_1\|^{4qk^*}, \ldots, \|A_{k^*}\|^{4qk^*}\right\}\end{aligned}$$

and thus

$$\begin{aligned}&\mathbb{E}\|A_{k^*,1}(\alpha)\|^{4q}\\ &\le 1 - 4q\alpha\mathbb{E}\lambda\left(\sum_{i=1}^{k^*} A_i\right) + \alpha^2 2^{4q(k^*+4)+1}\left(1 + k^*\mathbb{E}\|A_1\|^{4qk^*}\right)\end{aligned}$$

for $0 < \alpha \le 1$, which together with (2.12) and the moment condition yields the assertion.

Now assume B2b. Because of almost sure boundedness of $\|A_1\|$ and stationarity of (A_n),

$$A^* := \operatorname{ess\,sup} \|A_n\| < \infty$$

is independent of n. Neglecting a set of $\mathbb{P}$-measure zero, one has

$$\sup_{n,\omega} \|A_n(\omega)\| \le A^*.$$

In view of the assertion it suffices to show a corresponding assertion for $\|A_{nN^*,1}\|^q$, $n \in \mathbf{N}$. According to (2.14) and (2.15) one has—for $0 < \alpha < \min\left\{1, \frac{1}{N^*A^*}\right\}$—the representation

$$\|A_{nN^*,(n-1)N^*+1}(\alpha)\| = 1 - \alpha T_n(\alpha),\quad n \in \mathbf{N},$$

with

$$T_n(\alpha) \ge 0,\ |T_n(\alpha) - \lambda(A_{(n-1)N^*+1}(\alpha) + \cdots + A_{nN^*}(\alpha))| \le \alpha(1+A^*)^{N^*}.$$

Let $\delta^* := \frac{1}{2}\mathbb{E}\lambda(A_1 + \ldots + A_{N^*}) > 0$. For $0 < \alpha < \min\left\{1, \frac{1}{N^*A^*}, \frac{\delta^*}{2(1+A^*)^{N^*}}\right\}$ one now obtains

$$\|A_{nN^*,1}(\alpha)\|^q$$

$$
\begin{aligned}
&\leq \|A_{nN^*,(n-1)N^*+1}(\alpha)\|^q \dots \|A_{N^*,1}(\alpha)\|^q \\
&= (1-\alpha T_n(\alpha))^q \dots (1-\alpha T_1(\alpha))^q \\
&\leq e^{-\alpha q(T_1(\alpha)+\dots+T_n(\alpha))} \\
&\leq e^{-\alpha q(T_1(\alpha)+\dots+T_n(\alpha))} \chi_{\left\{\left|\frac{T_1(\alpha)+\dots+T_n(\alpha)}{n} - E\lambda(A_1+\dots+A_{N^*})\right| \leq \delta^*\right\}} \\
&\quad + \chi_{\left\{\left|\frac{T_1(\alpha)+\dots+T_n(\alpha)}{n} - E\lambda(A_1+\dots+A_{N^*})\right| > \delta^*\right\}} \\
&\leq e^{-\alpha q \delta^* n} + \chi_{B_n}
\end{aligned}
$$

with

$$B_n = \left\{\left|\frac{\xi_1 + \dots + \xi_n}{n}\right| > \frac{\delta^*}{2}\right\},$$

where

$$\xi_n = \lambda\left(A_{(n-1)N^*+1} + \dots + A_{nN^*}\right) - \mathbb{E}\lambda(A_{(n-1)N^*+1} + \dots + A_{nN^*}), \quad n \in \mathbf{N}.$$

The sequence (ξ_n) is bounded and φ-mixing. Thus, according to an inequality of Collomb [2] (see the details in Györfi et al. [7, pp. 19, 20]) a constant $c_0 > 0$ exists with

$$P(B_n) \leq e^{-c_0 n}, \quad n \in \mathbf{N}.$$

Therefore,

$$\mathbb{E}\|A_{nN^*,1}\|^q \leq e^{-\min\{\alpha q \delta^*, c_0\} n}, \quad n \in \mathbf{N}.$$

If the second version of B2c is assumed, one argues as before, noting

$$
\begin{aligned}
P(B_n) &\leq (n\delta^*/2)^{-2p} \mathbb{E}\|\xi_1 + \dots + \xi_n\|^p \\
&\leq c_1 n^{-p}, \quad n \in \mathbf{N},
\end{aligned}
$$

with a suitable constant $c_1 < \infty$, where for the latter inequality Theorem 2 in [10] is used instead of Collomb's inequality.

b) We use the notations in the second part of a). As there, one obtains

$$
\begin{aligned}
\sum_{n=1}^{\infty} \mathbb{E}\|A_{n,1}(\alpha)\| &= \sum_{n=1}^{\infty} \sum_{k=(n-1)N^*+1}^{nN^*} \mathbb{E}\|A_{k,1}(\alpha)\| \\
&\leq N^* \sum_{n=1}^{\infty} \mathbb{E}\|A_{(n-1)N^*,1}(\alpha)\| \\
&\leq N^* \left(\sum_{n=0}^{\infty} e^{-\alpha\delta^* n} + \sum_{n=1}^{\infty} \mathbb{P}(B_n)\right)
\end{aligned}
$$

for sufficiently small $\alpha > 0$. (ξ_n) is α-mixing where the corresponding mixing coefficients α'_n fulfill $\sum \alpha'_n < \infty$. According to Ibragimov and Linnik [9, Lem. 18.5.2 and its proof] one has, with a suitable constant $c < \infty$,

$$\sum_{n=1}^{\infty} \mathbb{P}(B_n) \leq \left(\frac{2}{\delta^*}\right)^4 \sum_{n=1}^{\infty} n^{-4} \mathbb{E}\left\|\sum_{j=1}^{n} \xi_j\right\|^4$$

$$\leq c \sum_{n=1}^{\infty} n^{-2} \sum_{j=1}^{n} j\alpha_j$$

$$= c \sum_{j=1}^{\infty} \left(\sum_{n=j}^{\infty} n^{-2} \right) j\alpha_j < \infty. \qquad \square$$

The theorem below concerns the asymptotic mean estimation error of (X_n) and the asymptotic bias δ_α of (Y_n) for small α. Its proof is based on an argument of Sanchez-Palencia [25] concerning the averaging method for deterministic differential equations (compare also [26, §4.2]). The first assertion in Theorem 2.7b), under an assumption similar to B2a and under Assumption B1, is due to Kushner and Shwartz [13, especially p. 180] and to Macchi and Eweda [16], respectively. A corresponding result for an algorithm with projection has been obtained by Krieger and Masry [12] under conditions concerning α-mixing and finite moments.

THEOREM 2.7. a) *Under Assumption* A2,

$$\lim_n \mathbb{E}\|X_n - \vartheta\| \to 0 \quad (\alpha \to 0)$$

holds and therefore

$$\delta_\alpha \to 0 \quad (\alpha \to 0).$$

b) *If Assumption* B2a *or* B1 *is fulfilled, then*

$$\lim_n \mathbb{E}\|X_n - \vartheta\| = O(\alpha^{1/2}), \quad \delta_\alpha = O(\alpha^{1/2}) \quad (\alpha \to 0).$$

The following remark is proved similarly to Remark 2.2b).

Remark 2.8. Assumption A2 is satisfied under each of the conditions B2a and B2b.

Proof of Theorem 2.7. Via the recursion for $(X_n - \vartheta)$, without loss of generality $V = \vartheta = 0$ may be assumed.

a) Choose (X_n^*) as in (2.2). Let A^* be a uniform bound of $\|A_n\|$ as in the proof of Lemma 2.6. Noting that A_1 is bounded and positive semidefinite, and noting stationarity of (A_n) and relation (2.9) in Lemma 2.4, one obtains for α sufficiently small

$$\mathbb{E}\|X_n - X_n^*\| = \mathbb{E}\|A_{n,1}(\alpha)(X_0 - X_0^*)\| \to 0 \quad (n \to \infty)$$

by Lebesgue's dominated convergence theorem. Therefore, without loss of generality $X_0 = X_0^*$ may be assumed. Since (X_n) is then stationary and ergodic, in view of the first assertion it suffices to prove

$$H(\alpha) := \mathbb{E}\|X_n\| = \mathbb{E}\|X_0\| \to 0 \quad (\alpha \to 0).$$

$\lambda(A) > 0$ yields

$$\|(I - \alpha A)^n\| \leq (1 - \alpha\lambda(A))^n \leq e^{-\alpha\lambda(A)n}$$

for all $0 < \alpha < \lambda(A)^{-1}$ and $n \in \mathbf{N}$. Choose Q such that $e^{-\lambda(A)Q} = \frac{1}{2}$, $D := D(\alpha) := [\frac{Q}{\alpha}] + 1$; thus $\|(1 - \alpha A)^D\| \leq \frac{1}{2}$ for all $0 < \alpha < \lambda(A)^{-1}$. Further choose an integer $T = T(\alpha)$ such that

$$\frac{1}{4} < e^{\alpha\|A\|D}\alpha(T+1)(2A^* + \alpha D A^{*^2}) < \frac{1}{2},$$

which is possible for all $0 < \alpha < \tilde{\alpha}$ with some $\tilde{\alpha} \leq \min\{\lambda(A)^{-1}, \alpha''\}$. In the following discussion, let $0 < \alpha < \tilde{\alpha}$. Further, let

$$g(n, y) := A_{n+1}y - V_{n+1},$$
$$g_T(n, y) := \frac{1}{T}\sum_{i=1}^{T} g(n+i, y), \quad n \geq 0, \quad y \in \mathbf{R}^m.$$

Thus

$$X_{n+1} = X_n - \alpha g(n, X_n) = X_{DN} - \alpha \sum_{k=DN}^{n} g(k, X_k), \quad n \geq DN, \quad N \geq 0.$$

Let

$$Y_{n+1}^N := Y_n^N - \alpha g_T(n, Y_n^N), \quad DN \leq n \leq D(N+1) - 1$$

with

$$Y_{DN}^N := X_{DN}.$$

Thus

$$Y_{n+1}^N = Y_{DN}^N - \alpha \sum_{k=DN}^{n} g_T(k, Y_k^N), \quad DN \leq n \leq D(N+1) - 1.$$

With

$$\phi(n, N) := \sum_{j=DN}^{n} g(j, X_j)$$

one has

$$X_{n+1} = X_{DN} - \alpha\phi(n, N), \quad DN \leq n,$$

and

$$\begin{aligned} & X_{n+1} - Y_{n+1}^N \\ &= -\alpha\left[\phi(n, N) - \sum_{k=DN}^{n} g_T(k, Y_k^N)\right] \\ &= -\alpha\left[(\phi(n, N) - \phi_T(n, N)) + \left(\phi_T(n, N) - \sum_{k=DN}^{n} g_T(k, X_k)\right)\right. \\ &\qquad \left. + \sum_{k=DN}^{n} \left(g_T(k, X_k) - g_T(k, Y_k^N)\right)\right] \end{aligned}$$

for $DN \leq n \leq D(N+1) - 1$, where

$$\phi_T(n, N) := \frac{1}{T}\sum_{k=1}^{T} \phi(n+k, N).$$

Now

$$\begin{aligned}&\phi(n,N)-\phi_T(n,N)\\&=\frac{1}{T}\sum_{k=1}^{T}\big(\phi(n,N)-\phi(n+k,N)\big)\\&=-\frac{1}{T}\sum_{k=1}^{T}\sum_{j=n+1}^{n+k}(A_{j+1}X_j-V_{j+1})\end{aligned}$$

for $DN\le n\le D(N+1)-1$. Further

$$\begin{aligned}&\phi_T(n,N)\\&=\frac{1}{T}\sum_{k=1}^{T}\sum_{j=DN+k}^{n+k}g(j,X_j)+\frac{1}{T}\sum_{k=1}^{T}\sum_{j=DN}^{DN+k-1}g(j,X_j)\\&=\sum_{j=DN}^{n}g_T(j,X_j)+\sum_{j=DN}^{n}\frac{1}{T}\sum_{k=1}^{T}\big[g(j+k,X_{j+k})-g(j+k,X_j)\big]\\&\quad+\frac{1}{T}\sum_{k=1}^{T}\sum_{j=DN}^{DN+k-1}g(j,X_j)\\&=\sum_{j=DN}^{n}g_T(j,X_j)+\sum_{j=DN}^{n}\frac{1}{T}\sum_{k=1}^{T}A_{j+k+1}\left(\alpha\sum_{l=j}^{j+k-1}(A_{l+1}X_l-V_{l+1})\right)\\&\quad+\frac{1}{T}\sum_{k=1}^{T}\sum_{j=DN}^{DN+k-1}(A_{j+1}X_j-V_{j+1}),\\&\quad\sum_{k=DN}^{n}\big(g_T(k,X_k)-g_T(k,Y_k^N)\big)\\&=\sum_{k=DN}^{n}\frac{1}{T}\sum_{i=1}^{T}A_{k+i+1}\big(X_k-Y_k^N\big)\end{aligned}$$

for $DN\le n\le (D+1)N-1$. In the next step, set

$$Z_{n+1}^N:=Z_n^N-\alpha AZ_n^N,\quad DN\le n\le D(N+1)-1,$$

with

$$Z_{DN}^N:=X_{DN}.$$

Then

$$\begin{aligned}&Y_{n+1}^N-Z_{n+1}^N\\&=-\alpha\sum_{k=DN}^{n}g_T\big(k,Y_k^N\big)+\alpha\sum_{k=DN}^{n}AZ_k^N\\&=-\alpha A\sum_{k=DN}^{n}\big(Y_k^N-Z_k^N\big)-\alpha\sum_{k=DN}^{n}\left[\frac{1}{T}\sum_{j=k+1}^{k+T}(A_{j+1}-A)Y_k^N\right]\end{aligned}$$

$$+\alpha\frac{1}{T}\sum_{k=DN}^{n}\sum_{j=k+1}^{k+T}V_{j+1}.$$

Now

$$\begin{aligned}(2.17)\qquad & X_{n+1}-Z_{n+1}^N\\ &=-\alpha A\sum_{k=DN}^{n}\left(X_k-Z_k^N\right)-\alpha\sum_{k=DN}^{n}\frac{1}{T}\sum_{i=1}^{T}(A_{k+i+1}-A)X_k\\ &\quad+\alpha\frac{1}{T}\sum_{k=1}^{T}\left(\sum_{j=n+1}^{n+k}A_{j+1}X_j-\sum_{j=DN}^{DN+k-1}A_{j+1}X_j\right)\\ &\quad-\alpha^2\sum_{j=DN}^{n}\frac{1}{T}\sum_{k=1}^{T}A_{j+k+1}\sum_{l=j}^{j+k-1}A_{l+1}X_l\\ &\quad-\alpha\frac{1}{T}\sum_{k=1}^{T}\left(\sum_{j=n+1}^{n+k}V_{j+1}-\sum_{j=DN}^{DN+k-1}V_{j+1}\right)\\ &\quad+\alpha^2\sum_{j=DN}^{n}\frac{1}{T}\sum_{k=1}^{T}A_{j+k+1}\sum_{l=j}^{j+k-1}V_{l+1}\\ &\quad+\alpha\frac{1}{T}\sum_{k=DN}^{n}\sum_{j=k+1}^{k+T}V_{j+1}\end{aligned}$$

for $DN\le n\le D(N+1)-1$. Thus

$$\begin{aligned}&\mathbb{E}\|X_{n+1}-Z_{n+1}^N\|\\ &\le\alpha\|A\|\sum_{k=DN}^{n}\mathbb{E}\|X_k-Z_k^N\|+\kappa(\alpha,D,T),\quad DN\le n\le D(N+1)-1,\end{aligned}$$

where

$$\begin{aligned}\kappa(\alpha,D,T)&=\alpha D\mathbb{E}\left\|\frac{1}{T}\sum_{i=1}^{T}(A_{i+1}-A)X_0\right\|+\alpha(T+1)A^*\mathbb{E}\|X_0\|\\ &\quad+\alpha^2D\frac{T+1}{2}A^{*^2}\mathbb{E}\|X_0\|+2\alpha\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left\|\sum_{j=1}^{k}V_j\right\|\\ &\quad+\alpha^2DA^*\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left\|\sum_{l=1}^{k}V_l\right\|+\alpha D\frac{1}{T}\mathbb{E}\left\|\sum_{j=1}^{T}V_j\right\|.\end{aligned}$$

By induction one obtains

$$\begin{aligned}&\mathbb{E}\|X_n-Z_n^N\|\\ &\le e^{\alpha\|A\|(n-DN)}\kappa(\alpha,D,T)\\ &\le e^{\alpha\|A\|D}\kappa(\alpha,D,T)=:\sigma(\alpha)\end{aligned}$$

for $DN \le n \le D(N+1)$. Let

$$Z_{n+1} := Z_n - \alpha A Z_n, \quad n \ge 0,$$

with $Z_0 := X_0$. Noting

$$\left\| Z_{D(N+1)} - Z^N_{D(N+1)} \right\| \le \frac{1}{2} \left\| Z_{DN} - Z^N_{DN} \right\| = \frac{1}{2} \left\| Z_{DN} - X_{DN} \right\|$$

one obtains

$$\mathbb{E}\|X_{D(N+1)} - Z_{D(N+1)}\| \le \sigma(\alpha) + \frac{1}{2}\mathbb{E}\|Z_{DN} - X_{DN}\|$$

and thus by induction

$$\mathbb{E}\|X_{DN} - Z_{DN}\| \le 2\sigma(\alpha), \quad N \ge 0.$$

Because

$$\mathbb{E}\|Z_{DN}\| \le \|(I - \alpha A)^{DN}\| \, \mathbb{E}\|X_0\| \to 0 \quad (N \to \infty)$$

one has

$$H(\alpha) \le 2\sigma(\alpha).$$

Moreover

$$H(\alpha) \le 4e^{\alpha\|A\|D} \left[\alpha D \mathbb{E} \left\| \frac{1}{T} \sum_{i=1}^{T} (A_{i+1} - A) X_0 \right\| \right.$$
$$\left. + (2 + \alpha D A^*) \alpha \frac{1}{T} \sum_{k=1}^{T} \mathbb{E} \left\| \sum_{j=1}^{k} V_j \right\| + \alpha D \frac{1}{T} \mathbb{E} \left\| \sum_{j=1}^{T} V_j \right\| \right].$$

Noting $D(\alpha) = O(\frac{1}{\alpha})$, $T(\alpha) = O(\frac{1}{\alpha})$, $T(\alpha) \to \infty$ $(\alpha \to 0)$,

$$\sup_{0<\alpha<\tilde{\alpha}} \mathbb{E}\left(\|X_0\| \chi_{[\|X_0\| > c]} \right) \to 0 \quad (c \to \infty)$$

(uniform integrability of X_0), boundedness of A_1, and

$$\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} (A_i - A) \right\| \to 0, \quad \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} V_i \right\| \to 0 \quad (n \to \infty)$$

(mean ergodic theorem), one obtains $H(\alpha) \to 0$ $(\alpha \to 0)$. $\delta_\alpha \to 0$ $(\alpha \to 0)$ now follows from

$$\|\delta_\alpha\| = \lim \|\mathbb{E} Y_n\| \le \underline{\lim}_n \mathbb{E}\|Y_n\| \le \underline{\lim}_n \mathbb{E}\|X_n\| = H(\alpha).$$

b) First assume B2a. Consider α sufficiently small and choose $D = D(\alpha)$ and $T = T(\alpha)$ as in a). Without loss of generality $X_0 = X_0^*$ may be assumed. According to the proof of Lemma 2.6 one obtains

$$\mathbb{E} \left\| \frac{1}{T} \sum_{i=1}^{T} (A_{i+1} - A) X_0 \right\|$$
$$\le C \mathbb{E} \left\| \frac{1}{T} \sum_{i=1}^{T} (A_{i+1} - A) \right\| + 2A^* C \alpha \sum_{n=1}^{\infty} \mathbb{P}(B_n)$$

with a suitable constant $C < \infty$ and $\sum \mathbb{P}(B_n) < \infty$. Because

$$\mathbb{E}\left\|\sum_{i=1}^{n}(A_i - A)\right\| = O(n^{\frac{1}{2}}), \quad \mathbb{E}\left\|\sum_{i=1}^{n} V_i\right\| = O(n^{\frac{1}{2}})$$

according to Ibragimov and Linnik [9, proof of Thm. 18.5.4], one has

$$H(\alpha) = O\left(T(\alpha)^{-\frac{1}{2}} + \alpha T(\alpha)^{\frac{1}{2}}\right) = O\left(\alpha^{\frac{1}{2}}\right), \quad \alpha \to 0;$$

further, $\delta_\alpha = O(\alpha^{1/2})$ $(\alpha \to 0)$.

Now assume B1. One uses the stationary sequence (X_n^*) defined by (2.2) for sufficiently small $\alpha > 0$ and notes

$$\mathbb{E}\|X_0^*\|^2 = O(1) \quad (\alpha \to 0).$$

This relation follows by an argument similar to that in the proof of Lemma 2.6b. In fact, it suffices to prove

$$\begin{aligned}
&\sum_{n=1}^{\infty} \mathbb{E}\|A_{nk^*,(n-1)k^*+1}(\alpha) \dots A_{k^*,1}(\alpha)\|^4 \\
&+ \sum_{1 \le m < n} \mathbb{E}(\|A_{nk^*,(n-1)k^*+1}(\alpha) \dots A_{(m+1)k^*,mk^*+1}(\alpha)\|^2 \\
&\qquad \cdot \|A_{mk^*,(m-1)k^*+1}(\alpha) \dots A_{k^*,1}(\alpha)\|^4) \\
&\le c\alpha^{-2}
\end{aligned}$$

with $k^* = \max\{N^*, M\}$ (N^* according to (2.12)) for sufficiently small $\alpha > 0$ with a suitable constant c. Because of M-dependence the left side is majorized by

$$\sum_{n=1}^{\infty} n \left(\max\left\{\mathbb{E}\|A_{k^*,1}(\alpha)\|^4, \mathbb{E}\|A_{k^*,1}(\alpha)\|^8\right\}\right)^n$$

and thus, via Lemma 2.6a), by

$$\sum_{n=1}^{\infty} n(1 - \rho\alpha)^n = O\left(\frac{1}{(\rho\alpha)^2}\right), \quad \alpha \to 0,$$

with a suitable constant $\rho > 0$. Square integrability of X_0^* and Lemma 2.6a) with $q = 4$, together with M-dependence, yield

$$\mathbb{E}\|X_n - X_n^*\| = \mathbb{E}\|A_{n,1}(\alpha)(X_0 - X_0^*)\| \to 0 \quad (n \to \infty).$$

Thus, as in a), without loss of generality $X_0 = X_0^*$ may be assumed. It suffices to prove

$$\mathbb{E}\|X_0\| = O(\alpha^{\frac{1}{2}}), \quad \alpha \to 0.$$

One uses a modification of the argument in a). Q and $D = D(\alpha)$ are chosen as in a); the integer $T = T(\alpha)$ is chosen in such a way that

$$\frac{1}{4} < e^{\alpha\|A\|D}\alpha(T+1)\|A\|(2 + \alpha D\mathbb{E}\|A_1\|) < \frac{1}{2}.$$

First a refined treatment of the third and the fourth terms in the right side of (2.17) will be given. By partial summation and use of (1.1) one obtains

$$\sum_{j=1}^{k}(A_{j+1}-A)X_j$$
$$=\sum_{j=1}^{k}(A_{j+1}-A)X_k+\alpha\sum_{j=1}^{k-1}\left(\sum_{i=1}^{j}(A_{i+1}-A)\right)(A_{j+1}X_j-V_{j+1})$$

for $k=1,2,\dots$. Thus, by stationarity, one has

$$\alpha\mathbb{E}\frac{1}{T}\left\|\sum_{k=1}^{T}\left(\sum_{j=n+1}^{n+k}A_{j+1}X_j-\sum_{j=DN}^{DN+k-1}A_{j+1}X_j\right)\right\|$$
$$\le\alpha(T+1)\|A\|\mathbb{E}\|X_0\|+2\alpha\frac{1}{T}\mu(\alpha,T)$$

for $DN\le n\le D(N+1)-1$ with

$$\mu(\alpha,T)=\sum_{k=1}^{T}\mathbb{E}\left(\left\|\sum_{j=1}^{k}(A_{j+1}-A)\right\|\,\|X_k\|\right)$$
$$+\alpha\sum_{k=1}^{T}\sum_{j=1}^{k-1}\mathbb{E}\left(\left\|\sum_{i=1}^{j}(A_{i+1}-A)\right\|\,\|A_{j+1}\|\,\|X_j\|\right)$$
$$+\alpha\sum_{k=1}^{T}\sum_{j=1}^{k-1}\mathbb{E}\left(\left\|\sum_{i=1}^{j}(A_{i+1}-A)\right\|\,\|V_{j+1}\|\right).$$

Noting $\mathbb{E}\|X_0\|^2=O(1)$, $\alpha\to 0$, the moment conditions on A_1 and the assumption of M-dependence, one further obtains

$$\alpha^2\mathbb{E}\left\|\sum_{j=DN}^{n}\frac{1}{T}\sum_{k=1}^{T}A_{j+k+1}\sum_{l=j}^{j+k-1}A_{l+1}X_l\right\|$$
$$\le\alpha^2D\|A\|\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left(\|A_{k+2}\|\sum_{l=1}^{k}\|X_l\|\right)$$
$$+\alpha^2D\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left(\|A_{k+2}\|\left\|\sum_{l=1}^{k}(A_{l+1}-A)X_l\right\|\right)$$
$$\le\alpha^2D\|A\|\frac{1}{T}\sum_{k=M}^{T}\mathbb{E}\left(\|A_{k+2}\|\sum_{l=1}^{k-M+1}\|X_l\|\right)$$
$$+\alpha^2D\|A\|\frac{1}{T}\sum_{k=M}^{T}\mathbb{E}\left(\|A_{k+2}\|\left\|\sum_{l=1}^{k-M+1}(A_{l+1}-A)X_l\right\|\right)$$
$$+c\alpha^2D$$
$$\le\frac{1}{2}\alpha^2D\|A\|(T+1)\mathbb{E}\|A_1\|\,\mathbb{E}\,\|X_0\|$$
$$+\alpha^2D\|A\|\,\mathbb{E}\,\|A_1\|\frac{1}{T}\mu(\alpha,T)+c\alpha^2D$$

for $DN \le n \le D(N+1)-1$ with a suitable $c \in \mathbf{R}_+$. Set

$$\begin{aligned}
&\kappa^*(\alpha, D, T)\\
&:= \alpha D \mathbb{E}\left\| \frac{1}{T}\sum_{i=1}^{T}(A_{i+1}-A)X_0 \right\| + \alpha(T+1)\|A\|\,\mathbb{E}\,\|X_0\|\\
&\quad + \frac{1}{2}\alpha^2 D\|A\|(T+1)\mathbb{E}\|A_1\|\,\mathbb{E}\,\|X_0\|\\
&\quad + (2\alpha + \alpha^2 D\|A\|\,\mathbb{E}\,\|A_1\|)\frac{1}{T}\mu(\alpha, T) + c\alpha^2 D\\
&\quad + 2\alpha\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left\|\sum_{j=1}^{k}V_j\right\| + \alpha^2 D\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left(\|A_{k+1}\|\left\|\sum_{l=1}^{k}V_l\right\|\right)\\
&\quad + \alpha D\frac{1}{T}\mathbb{E}\left\|\sum_{j=1}^{T}V_j\right\|,\\
&\sigma^*(\alpha) := e^{\alpha\|A\|D}\kappa^*(\alpha, D, T).
\end{aligned}$$

As in a) one obtains

$$H(\alpha) \le 2\sigma^*(\alpha);$$

moreover

$$\begin{aligned}
H(\alpha) \le 4e^{\alpha\|A\|D}\Bigg[&\alpha D\mathbb{E}\left\|\frac{1}{T}\sum_{i=1}^{T}(A_{i+1}-A)X_0\right\|\\
&+(2\alpha + \alpha^2 D\|A\|E\|A_1\|)\frac{1}{T}\mu(\alpha, T) + c\alpha^2 D\\
&+2\alpha\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left\|\sum_{j=1}^{k}V_j\right\| + \alpha^2 D\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left(\|A_{k+1}\|\left\|\sum_{l=1}^{k}V_l\right\|\right)\\
&+\alpha D\frac{1}{T}\mathbb{E}\left\|\sum_{j=1}^{T}V_j\right\|\Bigg].
\end{aligned}$$

The right-hand side is $O(\alpha^{1/2})$, $\alpha \to 0$. To show this one uses the Cauchy–Schwarz inequality several times, notes $D(\alpha) = O(1/\alpha)$, $T(\alpha) = O(1/\alpha)$, $T(\alpha)^{-1} = O(\alpha)$, and the moment conditions; further one uses

$$\mathbb{E}\left\|\sum_{k=1}^{n}(A_k - A)\right\|^2 = O(n), \quad \mathbb{E}\left\|\sum_{k=1}^{n}V_k\right\|^2 = O(n)$$

according to Billingsley [1, §20, Lem. 3] and Ibragimov and Linnik [9, Thm. 18.5.2] (compare also Peligrad [19, Thm. 1.1]), and

$$\mathbb{E}\|X_0\|^2 = O(1), \quad \alpha \to 0.$$

Here in view of $\mu(\alpha, T)$ one notes, by M-dependence,

$$E\left(\left\|\sum_{i=1}^{j-M-1}(A_{i+1}-A)\right\|\|A_j\|\right)^2 = O(j). \qquad \square$$

3. Consistency under independence or martingale assumptions. In this section it will be shown that in the case of independence and in the martingale case the assertion of Theorem 2.1b) on (Y_n) holds with $\delta_\alpha = 0$. This means that Y_n is strongly consistent and asymptotically unbiased.

THEOREM 3.1. *If the random elements (A_n, V_n), $n = 0, \pm 1, \pm 2, \ldots$, are i.i.d., then there exists an α^* with $0 < \alpha^* \leq \alpha'''$ such that for all $0 < \alpha < \alpha^*$*

$$Y_n \to \vartheta \quad (n \to \infty) \text{ a.s. and in the first mean.} \tag{3.1}$$

Proof. One chooses N^* according to (2.12) and argues similarly to the last part of the proof of Lemma 2.4. By the independence assumption, $E\|A_1\| < \infty$, and Lebesgue's dominated convergence theorem for $\alpha \to 0$, one obtains

$$\mathbb{E}\|A_{N^*,1}(\alpha)\| \leq 1 - \frac{1}{2}\alpha \mathbb{E}\lambda\left(\sum_{i=1}^{N^*} A_i\right)$$

for sufficiently small α, and then $\mathbb{E}\|U_\alpha\| < \infty$. Now Theorem 2.1b) yields the first part of the assertion, where obviously $\delta_\alpha = 0$. Further one obtains $E\|X_n - X_n^*\| \to 0$ for X_n^* in (2.2) and then, by the mean ergodic theorem for (X_n^*), together with $\mathbb{E}X_0^* = \vartheta$, the second assertion. □

For the special case of nonrandom $A_n = A$ almost sure convergence in (3.1) was proved by Pflug [20] and Polyak and Juditsky [22]. In what follows we extend Theorem 3.1 for martingale differences that are not necessarily stationary and ergodic.

THEOREM 3.2. *Assume that $((A_n - A, V_n - V), \mathcal{F}_{n-1})$ is a martingale difference sequence such that*

$$c^2 := \sup_{n,\omega} \mathbb{E}\left(\|A_n - A\|^2 \mid \mathcal{F}_{n-1}\right)(\omega) < \infty \tag{3.2}$$

and

$$\sup_n \mathbb{E}\|V_n\|^2 < \infty. \tag{3.3}$$

Then for all $0 < \alpha < 2/\Lambda(A)$ satisfying

$$\frac{\alpha^2 c^2}{1 - \|1 - \alpha A\|^2} < 1 \tag{3.4}$$

the relations

$$\sup_n \mathbb{E}\|X_n\|^2 < \infty, \quad \overline{\lim_n} \mathbb{E}\|X_n - \vartheta\|^2 = O(\alpha) \quad (\alpha \to 0)$$

hold and

$$Y_n \to \vartheta \quad (n \to \infty) \text{ a.s. and in the second mean.} \tag{3.5}$$

Proof. Noting the recursion for $(X_n - \vartheta)$ and the equivalence of (3.3) and

$$\sup_n \mathbb{E}\|V_n - A_n\vartheta\|^2 < \infty$$

(because of (3.2)), one may assume $V = \vartheta = 0$ without loss of generality. (1.1) can be written in the form

$$X_{n+1} = X_n - \alpha A X_n + \alpha W_{n+1}, \quad n \geq 0,$$

where $W_{n+1} = -(A_{n+1} - A)X_n + V_{n+1}$. Obviously $\mathbb{E}(W_{n+1}, X_n) = 0$; therefore, for all sufficiently small $\epsilon > 0$

$$\begin{aligned} &\mathbb{E}\|X_{n+1}\|^2 \\ &\leq \|1 - \alpha A\|^2 \mathbb{E}\|X_n\|^2 + \alpha^2 \mathbb{E}\|W_{n+1}\|^2 \\ &\leq \rho \mathbb{E}\|X_n\|^2 + \alpha^2 \left(1 + \frac{1}{\epsilon}\right) \mathbb{E}\|V_{n+1}\|^2 \end{aligned}$$

with $\rho = \|1 - \alpha A\|^2 + \alpha^2(1+\epsilon)c^2 < 1$, because of (3.2) and (3.4). This, together with (3.3), yields the assertions on (X_n). In the special case $V_n = 0$, here with $\epsilon = 0$, one obtains

$$X_n = A_{n,1}(\alpha) X_0,$$

$$\mathbb{E}\|X_{n+1}\|^2 \leq \rho \mathbb{E}\|X_n\|^2,$$

and thus

$$\mathbb{E}\|A_{n,1}(\alpha) X_0\|^2 = O(\rho^n).$$

The further specialization to unit vectors X_0 yields

$$\mathbb{E}\|A_{n,1}(\alpha)\|^2 = O(\rho^n) \tag{3.6}$$

and then

$$\|A_{n,1}(\alpha)\| \to 0 \quad (n \to \infty) \text{ a.s.} \tag{3.7}$$

Therefore, in view of (3.5), without loss of generality $X_0 = 0$ may be assumed. It is easy to verify

$$Y_{n+1} = Y_n - \alpha A Y_n - \frac{I - \alpha A}{n+1} Y_n + \frac{\alpha}{n+1} \sum_{k=1}^{n+1} W_k.$$

Moreover

$$\frac{1}{n+1} \sum_{k=1}^{n+1} (A_k - A) X_{k-1} \to 0 \quad \text{a.s. and in the second mean}$$

because

$$\mathbb{E}((A_k - A) X_{k-1} \mid \mathcal{F}_{k-1}) = 0$$

and

$$\mathbb{E}\|(A_k - A) X_{k-1}\|^2 \leq c^2 \mathbb{E}\|X_{k-1}\|^2 = O(1).$$

From this and from

$$\frac{1}{n+1} \sum_{k=1}^{n+1} V_k \to 0 \text{ a.s.},$$

by $\|I - \alpha A\| < 1$ one obtains relation (3.5) (compare [28, Lem. 2b]). □

4. Asymptotic normality. In this section it will be shown that the sequence (Y_n) of arithmetic means under rather weak assumptions has convergence order $n^{-1/2}$; there holds a (functional) central limit theorem.

THEOREM 4.1. *If Assumption* B1 *or* B2c *holds, then for α sufficiently small, $\sqrt{n}(Y_n-\vartheta-\delta_\alpha)$, i.e., $n^{-1/2}\sum_{k=1}^n(X_k-\vartheta-\delta_\alpha)$, converges in distribution to a Gaussian random vector with zero expectation.*

Remark 4.2. Theorem 4.1 may be generalized to a weak invariance principle of Donsker type.

Proof of Theorem 4.1. Without loss of generality $(X_n) = (X_n^*)$, with X_n^* as in (2.2), may be assumed, because for the corresponding arithmetic means one has

$$\begin{aligned}\sqrt{n}\|Y_n - Y_n^*\| &\le n^{-\frac{1}{2}} \sum_{k=1}^{n} \|A_{k,1}\| \, \|X_0 - X_0^*\| \\ &\le n^{-\frac{1}{2}} \sum_{k=1}^{\infty} \|A_{k,1}\| \, \|X_0 - X_0^*\| \to 0 \quad \text{a.s.}\end{aligned}$$

by Theorem 2.1a). Let

$$\tilde{X}_n := X_n - \vartheta - \delta_\alpha = \alpha \sum_{i=-\infty}^{0} A_{n,n+i-1}(\alpha)(V_{n+i} - A_{n+i}\vartheta) - \delta_\alpha,$$

$$\tilde{X}_n^{(l)} := \alpha \sum_{i=-l}^{0} A_{n,n+i+1}(\alpha)(V_{n+i} - A_{n+i}\vartheta) - \delta_\alpha, \quad n \in \mathbf{N}, \quad l \in \mathbf{N}.$$

One has

$$\mathbb{E}\tilde{X}_n = 0,$$

and further (compare [1, §21, especially p. 183], [8])

$$\begin{aligned}&[\mathbb{E}\|\tilde{X}_0^{(l)} - E(\tilde{X}_0|A_l, V_l, \ldots, A_0, V_0, \ldots, A_{-l}, V_{-l})\|^2]^{\frac{1}{2}} \\ &\le [\mathbb{E}\{\mathbb{E}(\|\tilde{X}_0^{(l)} - \tilde{X}_0\|^2|A_l, V_l, \ldots, A_0, V_0, \ldots, A_{-l}, V_{-l})\}]^{\frac{1}{2}} \\ &= [\mathbb{E}\|\tilde{X}_0^{(l)} - \tilde{X}_0\|^2]^{\frac{1}{2}} \\ &= \alpha \left[\mathbb{E}\left\|\sum_{i=-\infty}^{-l-1} A_{0,i+1}(\alpha)(V_i - A_i\vartheta)\right\|^2\right]^{\frac{1}{2}} \\ &\le \alpha \left[\mathbb{E}\left(\sum_{i=-\infty}^{-l-1} \|A_{0,i+1}(\alpha)(V_i - A_i\vartheta)\|\right)^2\right]^{\frac{1}{2}} \\ &= \alpha \left[\mathbb{E} \sum_{i,j\in\{-l-1,-l-2,\ldots\}} \|A_{0,i+1}(\alpha)(V_i - A_i\vartheta)\| \, \|A_{0,j+1}(\alpha)(V_j - A_j\vartheta)\|\right]^{\frac{1}{2}} \\ &\le \alpha \left[\sum_{i,j\in\{-l-1,-l-2,\ldots\}} \left(\mathbb{E}(\|A_{0,i+1}(\alpha)\|^2\|V_i - A_i\vartheta\|^2)\right)^{\frac{1}{2}}\right.\end{aligned}$$

$$\left. \left(\mathbb{E}(\|A_{0,j+1}(\alpha)\|^2 \|V_j - A_j\vartheta\|^2) \right)^{\frac{1}{2}} \right]^{\frac{1}{2}}$$

$$= \alpha \sum_{i \in \{-l-1, -l-2, \ldots\}} \left(\mathbb{E}(\|A_{0,i+1}\|^2 \|V_i - A_i\vartheta\|^2) \right)^{\frac{1}{2}}$$

$$= \alpha \sum_{n=l+1}^{\infty} \left(\mathbb{E}(\|A_{n,1}(\alpha)\|^2 \|V_0 - A_0\vartheta\|^2) \right)^{\frac{1}{2}}.$$

By Lemma 2.6a) one obtains

$$\sum_{l=1}^{\infty} \left(\mathbb{E}\|\tilde{X}_0^{(l)} - E(\tilde{X}_0 | A_l, V_l, \ldots, A_0, V_0, \ldots, A_{-l}, V_{-l})\|^2 \right)^{\frac{1}{2}} < \infty;$$

further, in a similar way,

$$\mathbb{E}\|\tilde{X}_0\|^q < \infty \text{ for each } q \in \mathbf{N}.$$

One uses Theorem 18.6.2 (and Remark 18.6.1) in [9], or Theorem 4.2 in [17], for which the statement after Corollary (3.9) there concerning $\sigma^2 = 0$ also holds, together with Definition (2.4) and the remarks in §§2 and 3 there, and obtains then the assertion by the Cramér–Wold device. □

5. Asymptotic normality and covariance. In this section we assume that the observations either are independent or form a martingale difference sequence. Polyak and Juditsky [22] considered asymptotic normality in the martingale case too.

THEOREM 5.1 (see [22, Thm. 1], [21, Thm. 1]). *Consider the iteration*

$$X_0 \quad \textit{arbitrary},$$

$$X_{n+1} = X_n - \alpha(AX_n - V - W_{n+1}), \quad n \geq 0, \tag{5.1}$$

where $0 < \alpha < 2/\Lambda(A)$, *under the assumption that* $(W_n, \mathcal{F}_{n-1})$ *is a martingale difference sequence.*

a) *If*

$$\sup_n \mathbb{E}(\|W_n\|^2 \mid \mathcal{F}_{n-1}) < \infty \quad a.s., \tag{5.2}$$

$$\lim_{C \to \infty} \overline{\lim_{n \to \infty}} \mathbb{E}(\|W_n\|^2 \chi_{[\|W_n\| > C]} \mid \mathcal{F}_{n-1}) = 0 \quad \textit{in probability, and} \tag{5.3}$$

$$\lim_{n \to \infty} \mathbb{E}(W_n W_n^T \mid \mathcal{F}_{n-1}) = S \quad \textit{in probability}, \tag{5.4}$$

where S *is positive semidefinite, then*

$$n^{\frac{1}{2}}(Y_n - \vartheta) \to \mathcal{N}(0, A^{-1}SA^{-1}) \quad \textit{in distribution.} \tag{5.5}$$

b) *If* X_0 *is square integrable and*

$$\lim_{n \to \infty} \mathbb{E}(W_n W_n^T) = S, \tag{5.6}$$

then

$$\lim_{n\to\infty} n\mathbb{E}\left((Y_n-\vartheta)(Y_n-\vartheta)^T\right) = A^{-1}SA^{-1}. \tag{5.7}$$

We need a slight modification of Theorem 5.1a.

LEMMA 5.2. *Consider iteration* (5.1), *where* $0<\alpha<2/\Lambda(A)$ *and* $(W_n,\mathcal{F}_{n-1})$ *is a stationary and ergodic martingale difference sequence such that*

$$\mathbb{E}(W_1W_1^T) = S$$

with S *positive semidefinite. Then*

$$n^{\frac{1}{2}}(Y_n-\vartheta) \to \mathcal{N}(0, A^{-1}SA^{-1}) \text{ in distribution.}$$

Proof. Without loss of generality we may assume $V=\vartheta=0$ and, according to the proof of Theorem 3.2 with $\rho = \|I-\alpha A\|^2$, also $X_0=0$. One obtains

$$X_k = \alpha\sum_{l=1}^{k}\left(I-\alpha A\right)^{k-l}W_l;$$

therefore,

$$\begin{aligned} Y_n &= \alpha\frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{k}\left(I-\alpha A\right)^{k-l}W_l = \alpha\frac{1}{n}\sum_{k=1}^{n}\left(\sum_{l=0}^{n-k}\left(I-\alpha A\right)^{l}\right)W_k \\ &= \alpha\frac{1}{n}\sum_{k=1}^{n}\left((I-(I-\alpha A)^{n-k+1})(\alpha A)^{-1}\right)W_k \\ &= \frac{1}{n}\sum_{k=1}^{n}A^{-1}W_k - \frac{1}{n}\sum_{k=1}^{n}\left(I-\alpha A\right)^{n-k+1}A^{-1}W_k. \end{aligned}$$

On the one hand

$$n^{\frac{1}{2}}\frac{1}{n}\sum_{k=1}^{n}A^{-1}W_k \to \mathcal{N}(0, A^{-1}SA^{-1}) \text{ in distribution}$$

according to Billingsley [1, Thm. 23.1] and the Cramér–Wold device, and on the other hand

$$n^{\frac{1}{2}}\frac{1}{n}\sum_{k=1}^{n}\left(I-\alpha A\right)^{n-k+1}A^{-1}W_k \to 0$$

in probability; therefore, the proof is complete. □

Although $A_n = A$ is a special case, (5.5) and (5.7) are surprises: the asymptotic covariance does not depend on α and is the best possible covariance. Unfortunately it is not true for random A_n; the asymptotic covariance differs from the best possible covariance by a term $O(\alpha)$. In the following X_0 is considered as a random vector.

THEOREM 5.3. *Assume that* $(A_n,V_n), n=0,\pm1,\pm2,\dots,$ *are i.i.d.,* $\mathbb{E}\|A_1\|^2<\infty$, *and* $\mathbb{E}\|V_1\|^2<\infty$.

a) *Let X_0 be square integrable. Then there is an $\alpha^* > 0$ such that for all $0 < \alpha < \alpha^*$ the limits*

$$\lim_{n\to\infty} \mathbb{E}((X_n - \vartheta)(X_n - \vartheta)^T) = \Sigma_\alpha \tag{5.8}$$

and

$$\lim_{n\to\infty} n\mathbb{E}((Y_n - \vartheta)(Y_n - \vartheta)^T) = \Sigma \tag{5.9}$$

exist, and

$$\Sigma = A^{-1}SA^{-1} + \mathbb{E}((A^{-1}A_0 - I)\Sigma_\alpha(A^{-1}A_0 - I)^T), \tag{5.10}$$

where

$$S = \mathbb{E}((V_1 - A_1\vartheta)(V_1 - A_1\vartheta)^T). \tag{5.11}$$

b) *Furthermore, with X_0 not necessarily square integrable, for $0 < \alpha < \alpha^*$ with suitable α^*,*

$$n^{\frac{1}{2}}(Y_n - \vartheta) \to \mathcal{N}(0, \Sigma) \text{ in distribution.} \tag{5.12}$$

Remark 5.4. It is well known [13], [20] that $\Sigma_\alpha = O(\alpha)$; therefore,

$$\Sigma = A^{-1}SA^{-1} + O(\alpha).$$

Proof of Theorem 5.3. a) Without loss of generality one may assume $V = \vartheta = 0$ and $X_0 = 0$, the latter because of (3.6) for sufficiently small $\alpha > 0$ in the proof of Theorem 3.2. Then

$$X_n = \sum_{i=1}^{n} B_{ni}V_i, \tag{5.13}$$

where

$$B_{ni} = \alpha A_{n,i+1}(\alpha),$$

and

$$\begin{aligned} &\mathbb{E}(X_nX_n^T) \\ &= \mathbb{E}\left(\sum_{i=1}^{n} B_{ni}V_i\left(\sum_{j=1}^{n} B_{nj}V_j\right)^T\right) = \mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} B_{ni}V_i(B_{nj}V_j)^T\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}\left(B_{ni}V_i(B_{nj}V_j)^T\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}\left(B_{ni}V_iV_j^TB_{nj}^T\right) \\ &= \sum_{i=1}^{n} \mathbb{E}\left(B_{ni}V_iV_i^TB_{ni}^T\right) = \sum_{i=1}^{n} \mathbb{E}\left(B_{ni}SB_{ni}^T\right) =: \Sigma_{\alpha,n}. \end{aligned} \tag{5.14}$$

Using (3.6) once more one obtains that $\mathbb{E}\|A_1\|^2 < \infty$ implies the existence of an $\alpha^* > 0$ such that for all $0 < \alpha < \alpha^*$

$$\sum_{n=1}^{\infty} \mathbb{E}\|B_{n1}\|^2 < \infty;$$

consequently

$$\begin{aligned} \lim_{n\to\infty} \Sigma_{\alpha,n} &= \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}\left(B_{ni} S B_{ni}^T\right) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}\left(B_{i1} S B_{i1}^T\right) = \sum_{i=1}^{\infty} \mathbb{E}\left(B_{i1} S B_{i1}^T\right) = \Sigma_\alpha \end{aligned} \tag{5.15}$$

exists and is finite. From (5.13) and (5.15) one obtains

$$\Sigma_\alpha - \mathbb{E}((I - \alpha A_0)\Sigma_\alpha (I - \alpha A_0)) = \alpha^2 S$$

or equivalently

$$A\Sigma_\alpha + \Sigma_\alpha A - \alpha \mathbb{E}(A_0 \Sigma_\alpha A_0) = \alpha S;$$

thus

$$\Sigma_\alpha A^{-1} + A^{-1}\Sigma_\alpha - \alpha A^{-1}\mathbb{E}(A_0 \Sigma_\alpha A_0)A^{-1} = \alpha A^{-1} S A^{-1}. \tag{5.16}$$

Moreover

$$n\mathbb{E}(Y_n Y_n^T) = \frac{1}{n}\mathbb{E}\left(\sum_{k=1}^{n} X_k X_k^T + \sum_{k=1}^{n-1}\sum_{l=k+1}^{n} X_k X_l^T + \sum_{l=1}^{n-1}\sum_{k=l+1}^{n} X_k X_l^T\right). \tag{5.17}$$

Then

$$\frac{1}{n}\mathbb{E}\left(\sum_{k=1}^{n} X_k X_k^T\right) = \frac{1}{n}\sum_{k=1}^{n} \Sigma_{\alpha,k} \to \Sigma_\alpha. \tag{5.18}$$

For $k < l$

$$\begin{aligned} \mathbb{E}(X_k X_l^T) &= \mathbb{E}\left(\sum_{i=1}^{k} B_{ki} V_i \left(\sum_{j=1}^{l} B_{lj} V_j\right)^T\right) = \mathbb{E}\left(\sum_{i=1}^{k}\sum_{j=1}^{l} B_{ki} V_i (B_{lj} V_j)^T\right) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{l} \mathbb{E}\left(B_{ki} V_i (B_{lj} V_j)^T\right) = \sum_{i=1}^{k}\sum_{j=1}^{l} \mathbb{E}\left(B_{ki} V_i V_j^T B_{lj}^T\right) \\ &= \sum_{i=1}^{k} \mathbb{E}\left(B_{ki} V_i V_i^T B_{li}^T\right) = \sum_{i=1}^{k} \mathbb{E}\left(B_{ki} S B_{li}^T\right) \\ &= \sum_{i=1}^{k} \mathbb{E}\left(B_{ki} S B_{ki}^T\right)(I - \alpha A)^{l-k} = \Sigma_{\alpha,k}(I - \alpha A)^{l-k}; \end{aligned}$$

therefore,

$$\begin{aligned}(5.19)\qquad &\frac{1}{n}\mathbb{E}\left(\sum_{k=1}^{n-1}\sum_{l=k+1}^{n} X_k X_l^T\right)\\ &= \frac{1}{n}\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}\Sigma_{\alpha,k}\,(I-\alpha A)^{l-k} = \frac{1}{n}\sum_{k=1}^{n-1}\Sigma_{\alpha,k}\sum_{l=k+1}^{n}(I-\alpha A)^{l-k}\\ &= \frac{1}{n}\sum_{k=1}^{n-1}\Sigma_{\alpha,k}\left(I-(I-\alpha A)^{n-k}\right)(\alpha A)^{-1}(I-\alpha A)\\ &\to \Sigma_\alpha(\alpha A)^{-1}(I-\alpha A) = \Sigma_\alpha(\alpha A)^{-1} - \Sigma_\alpha.\end{aligned}$$

In the same way one obtains

$$(5.20)\qquad \frac{1}{n}\mathbb{E}\left(\sum_{l=1}^{n-1}\sum_{k=l+1}^{n} X_k X_l^T\right) \to (\alpha A)^{-1}\Sigma_\alpha - \Sigma_\alpha.$$

By (5.17)–(5.20) one obtains

$$(5.21)\qquad \Sigma = (\Sigma_\alpha A^{-1} + A^{-1}\Sigma_\alpha)/\alpha - \Sigma_\alpha.$$

(5.16) and (5.21) imply (5.9).

b) Without loss of generality $V = \vartheta = 0$ is assumed. For sufficiently small $\alpha > 0$ the random variable U_α defined by (2.1) is integrable. Let (X_n^*) be defined by (2.2) and

$$Y_n^* := \frac{1}{n}\sum_{k=1}^{n} X_k^*;$$

then

$$Y_n^* - Y_n = \frac{1}{n}\sum_{k=1}^{n}\left(X_k^* - X_k\right) = \frac{1}{n}\sum_{k=1}^{n} A_{k,1}(\alpha)(X_0^* - X_0),$$

and thus because of (2.9) in Lemma 2.4

$$\begin{aligned}n^{\frac{1}{2}}\|Y_n^* - Y_n\| &\le n^{-\frac{1}{2}}\left(\sum_{k=1}^{n}\|A_{k,1}(\alpha)\|\right)\|X_0^* - X_0\|\\ &\le n^{-\frac{1}{2}}\left(\sum_{k=1}^{\infty}\|A_{k,1}(\alpha)\|\right)\|X_0^* - X_0\| \to 0 \quad \text{a.s.}\end{aligned}$$

Therefore, it suffices to consider only X_n^*. Now one applies Lemma 5.2 for $W_{n+1} = -(A_{n+1} - A)X_n^* + V_{n+1}$, noting square integrability of X_0^* (see Lemma 5.5 below) with $\mathbb{E}(X_0^* X_0^{*T}) = \Sigma_\alpha$ by a) and (5.9), (5.10). □

In order to treat the convergence behaviour in the martingale case the following lemma will be used.

LEMMA 5.5. *Assume that* $((A_n - A, V_n - V), \mathcal{F}_{n-1})$ *is a stationary and ergodic martingale difference sequence such that*

$$E(\|A_1\|^2 \mid \mathcal{F}_0) \le c^2 < \infty \quad a.s.$$

and

$$E\|V_1\|^2 < \infty.$$

Assume X_0 square integrable and let $\alpha > 0$ be sufficiently small.

a) *U_α defined by* (2.1) *is integrable.*

Let (X_n^*) be defined by (2.2). *Then*

b) $E\|X_0^*\|^2 < \infty$, *even* $E\|X_0^* - \vartheta\|^2 = O(\alpha)$;

c) $E\|X_n - X_n^*\|^2 \to 0\,(n \to \infty)$.

Proof. a) The proof is established by (3.6).

b) Because (X_n^*) is stationary and ergodic (by Theorem 2.1), for arbitrary fixed $M \in \mathbf{N}$ the sequence $(Z_{k,M})_{k\in\mathbf{N}}$ with

$$Z_{k,M} = \begin{cases} \|X_k^*\|^2 & \text{if } \|X_k^*\|^2 \le M, \\ M & \text{otherwise} \end{cases}$$

is stationary and ergodic, thus

$$\frac{1}{n}\sum_{k=1}^{n} Z_{k,M} \to \mathbb{E}Z_{1,M} \quad (n \to \infty) \text{ a.s.},$$

$$\mathbb{E}Z_{1,M} = \mathbb{E}\lim_n \frac{1}{n}\sum_{k=1}^{n} Z_{k,M} \le \mathbb{E}\varliminf_n \frac{1}{n}\sum_{k=1}^{n} \|X_k^*\|^2, \qquad M \in \mathbf{N},$$

$$\mathbb{E}\|X_0^*\|^2 \le \mathbb{E}\varliminf_n \frac{1}{n}\sum_{k=1}^{n} \|X_k^*\|^2.$$

Because of $\|X_n - X_n^*\| \to 0$ a.s. (by (3.7)) and

$$\frac{1}{n}\sum_{k=1}^{n} \|X_k + X_k^*\| \le \frac{1}{n}\sum_{k=1}^{n} \|X_k - X_k^*\| + 2\frac{1}{n}\sum_{k=1}^{n} \|X_k^*\| \to 2\mathbb{E}\|X_0^*\| < \infty \quad \text{a.s.}$$

(by a), Remark 2.2a), and stationarity and ergodicity of (X_n^*)), one obtains

$$\begin{aligned} &\frac{1}{n}\left|\sum_{k=1}^{n}\left(\|X_k^*\|^2 - \|X_k\|^2\right)\right| \\ &= \frac{1}{n}\left|\sum_{k=1}^{n}(X_k^* + X_k,\, X_k^* - X_k)\right| \\ &\le \frac{1}{n}\sum_{k=1}^{n} \|X_k^* + X_k\|\,\|X_k^* - X_k\| \to 0 \quad \text{a.s.} \end{aligned}$$

and thus

$$\varliminf_n \frac{1}{n}\sum_{k=1}^{n} \|X_k^*\|^2 = \varliminf_n \frac{1}{n}\sum_{k=1}^{n} \|X_k\|^2.$$

Therefore, by Fatou's lemma and Theorem 3.2,

$$\mathbb{E}\|X_0^*\|^2 \le \mathbb{E}\varliminf_n \frac{1}{n}\sum_{k=1}^{n} \|X_k\|^2$$

$$\leq \underline{\lim_n} \mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n}\|X_k\|^2\right)$$
$$\leq \sup_n \mathbb{E}\|X_n\|^2 < \infty\,.$$

$\mathbb{E}\|X_0^* - \vartheta\|^2 = O(\alpha)$ will follow from c), stationarity of (X_n^*), and $\overline{\lim_n}\,\mathbb{E}\|X_n - \vartheta\|^2 = O(\alpha)$ $(\alpha \to 0)$ (by Theorem 3.2).

c) One uses square integrability of X_0 and X_0^* and employs (3.6). □

THEOREM 5.6. *Under the conditions of Lemma* 5.5 *there is an* $\alpha^* > 0$ *such that for all* $0 < \alpha < \alpha^*$

$$\lim_{n\to\infty} \mathbb{E}\dot{(}(X_n - \vartheta)(X_n - \vartheta)^T) = \Sigma_\alpha^*, \tag{5.22}$$

where

$$\Sigma_\alpha^* = \mathbb{E}((X_0^* - \vartheta)(X_0^* - \vartheta)^T),$$

and

$$\lim_{n\to\infty} n\mathbb{E}((Y_n - \vartheta)(Y_n - \vartheta)^T) = \Sigma^*, \tag{5.23}$$

where

$$\Sigma^* = \mathbb{E}(BB^T)$$

with

$$B = (A^{-1}A_1 - I)(X_0^* - \vartheta) - (A^{-1}V_1 - A^{-1}A_1\vartheta) = (A^{-1}A_1 - I)X_0^* + \vartheta - A^{-1}V_1.$$

It holds that

$$\Sigma^* - A^{-1}SA^{-1} = \begin{cases} O(\alpha) & \text{if } \mathbb{E}(A_1(X_0^* - \vartheta)V_1^T) = 0, \\ O(\alpha^{\frac{1}{2}}) & \text{otherwise.} \end{cases} \tag{5.24}$$

Furthermore, with X_0 *not necessarily square integrable,*

$$n^{\frac{1}{2}}(Y_n - \vartheta) \to \mathcal{N}(0, \Sigma^*) \text{ in distribution.} \tag{5.25}$$

Proof. Without loss of generality $V = \vartheta = 0$ and, according to (3.6) in the proof of Theorem 3.2, $X_0 = 0$ may be assumed. Let $\alpha > 0$ be sufficiently small. By Lemma 5.5 and the auxiliary formula $\|CC^T - DD^T\| \leq \|C - D\|(\|C\| + \|D\|)$ for d-dimensional vectors or $d \times d$ matrices C, D one obtains

$$\|\mathbb{E}(X_n X_n^T) - \mathbb{E}(X_0^* X_0^{*T})\|$$
$$= \|\mathbb{E}(X_n X_n^T) - \mathbb{E}(X_n^* X_n^{*T})\| \leq \mathbb{E}(\|X_n - X_n^*\|(\|X_n\| + \|X_n^*\|))$$
$$\leq (\mathbb{E}\|X_n - X_n^*\|^2)^{\frac{1}{2}}[(\mathbb{E}\|X_n\|^2)^{\frac{1}{2}} + (\mathbb{E}\|X_0^*\|^2)^{\frac{1}{2}}] \to 0 \quad (n \to \infty),$$

which proves (5.22). In order to prove (5.23) one applies Theorem 5.1b) for $W_{n+1} = -(A_{n+1} - A)X_n + V_{n+1}$, which form a martingale difference sequence. It is sufficient to verify (5.6), i.e.,

$$\mathbb{E}(W_n W_n^T) \to \mathbb{E}\left(((A_1 - A)X_0^* - V_1)((A_1 - A)X_0^* - V_1)^T\right).$$

As before, as an upper bound for the difference one obtains, with a suitable constant c,

$$\mathbb{E}(\|A_{n+1} - A\|\|X_n - X_n^*\|[\|A_{n+1} - A\|(\|X_n\| + \|X_n^*\|) + 2\|V_{n+1}\|])$$

$$\leq c(\mathbb{E}\|X_n - X_n^*\|^2)^{\frac{1}{2}}[(\mathbb{E}\|X_n\|^2)^{\frac{1}{2}} + (\mathbb{E}\|X_n^*\|^2)^{\frac{1}{2}} + (\mathbb{E}\|V_{n+1}\|^2)^{\frac{1}{2}}],$$

which tends to 0 because of Lemma 5.5. (5.24) also follows from this lemma. (5.25) concerns asymptotic normality, which can be proved as that of Theorem 5.3b). □

6. Conclusion. The method treated above is of low computational complexity with the following disadvantages: (i) for general (dependent) observations it is usually asymptotically biased, (ii) for weakly dependent observations it is asymptotically normal with $(n^{-1/2})$-convergence order but with a nonoptimal covariance matrix, even in the i.i.d. case. In order to avoid these disadvantages one can use decreasing gains, e.g., $a_n = \alpha n^{-\gamma}$ $(\alpha > 0, \quad 3/4 < \gamma < 1)$. Under assumptions close to ergodicity (see (10) and (11) in [28] and (2.15a), (2.15b) in [15]) Ljung [15] proved that X_n is a.s. convergent to ϑ (asymptotically unbiased) and therefore Y_n also is. Under the assumption that a functional central limit theorem for $(V_k - A_k x)$ holds, $x \in \mathbf{R}^d$ (see [17] for sufficient conditions) with asymptotic covariance matrix S in the case $x = \vartheta$, and under the assumptions

$$\mathbb{E}\left\|\sum_{k=1}^{n}(A_k - A)\right\|^2 = O(n), \qquad \mathbb{E}\left\|\sum_{k=1}^{n}(V_k - V)\right\|^2 = O(n)$$

we proved for the above gains, besides almost sure convergence, asymptotic normality with the best possible covariance matrix $A^{-1}SA^{-1}$. This problem has been considered by Polyak and Juditsky [22] under i.i.d. observations and by Yin [29] under φ-mixing and bounded observations. The window averaging version of the decreasing gain algorithm has been investigated by Kushner and Yang [14].

Acknowledgment. The authors are grateful to the referees for making useful suggestions and raising relevant questions which led to a considerable improvement and extension of this paper.

Note added in proof. As to Theorem 2.7b), considering $\|\mathbb{E}X_n\|$ instead of $\mathbb{E}\|X_n\|$ in its proof, one can show

$$\delta_\alpha = O(\alpha) \qquad (\alpha \to 0),$$

if Assumption B1 or B2c is fulfilled. The auxiliary result that

$$\sum_{i=1}^{n}\|\mathbb{E}(A_{i+1} - A)X_0\| \quad \text{and} \quad \sum_{i=1}^{n}\|\mathbb{E}(A_i - A)(V_1 - AX_0)\|$$

are uniformly bounded with respect to n and α is obtained by use of Lemma 3.5 $(p = 1)$ in [17], Collomb's [2] inequality, and Theorem 2 in [10].

The authors thank L. Gerencsér for a stimulating discussion.

REFERENCES

[1] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.
[2] G. Collomb, *Propriétés de convergence presque complète du prédicteur à noyau*, Z. Wahrscheinlichkeitstheorie verw. Gebiete, 66 (1984), pp. 441–460.

[3] J. FRITZ, *Learning from ergodic training sequence*, in Limit Theorems of Probability Theory, P. Révész, ed., North-Holland, Amsterdam, 1974, pp. 79–91.
[4] H. FURSTENBERG AND H. KESTEN, *Products of random matrices*, Ann. Math. Stat., 31 (1960), pp. 457–469.
[5] L. GYÖRFI, *Stochastic approximation from ergodic sample for linear regression*, Z. Wahrscheinlichkeitstheorie verw. Gebiete, 54 (1980), pp. 47–55.
[6] ———, *Adaptive linear procedures under general conditions*, IEEE Trans. Inform. Theory, IT-30 (1984), pp. 262–267.
[7] L. GYÖRFI, W. HÄRDLE, P. SARDA, AND PH. VIEU, *Nonparametric Curve Estimation from Time Series*, Springer-Verlag, Berlin, 1989.
[8] I. A. IBRAGIMOV, *Some limit theorems for stationary processes*, Theory Probab. Appl., 7 (1962), pp. 349–382.
[9] I. A. IBRAGIMOV AND YU. V. LINNIK, *Independent and Stationary Sequences of Random Variables*, Wolters–Noordhoff, Groningen, 1971.
[10] T. Y. KIM, *A note on moment bounds for strong mixing sequences*, Statist. Probab. Lett., 16 (1993), pp. 163–168.
[11] U. KRENGEL, *Ergodic Theorems*, de Gruyter, Berlin, New York, 1985.
[12] A. KRIEGER AND E. MASRY, *Convergence analysis of adaptive linear estimation for dependent stationary processes*, IEEE Trans. Inform. Theory, IT-34 (1988), pp. 642–654.
[13] H. J. KUSHNER AND A. SHWARTZ, *Weak convergence and asymptotic properties of adaptive filters with constant gains*, IEEE Trans. Inform. Theory, IT-30 (1984), pp. 177–182.
[14] H. J. KUSHNER AND J. YANG, *Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes*, SIAM J. Control Optim. 31 (1993), pp. 1045–1062.
[15] L. LJUNG, *Analysis of stochastic gradient algorithms for linear regression problems*, IEEE Trans. Inform. Theory, IT-30 (1984), pp. 151–160.
[16] O. MACCHI AND E. EWEDA, *Second-order convergence analysis of stochastic adaptive linear filtering*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 76–85.
[17] D. L. MCLEISH, *Invariance principles for dependent variables*, Z. Wahrscheinlichkeitstheorie verw. Gebiete, 32 (1975), pp. 165–178.
[18] M. MÉTIVIER AND P. PRIOURET, *Applications of a Kushner and Clark lemma to general classes of stochastic algorithms*, IEEE Trans. Informat. Theory, IT-30 (1984), pp. 140–151.
[19] M. PELIGRAD, *The convergence of moments in the central limit theorem for ρ-mixing sequences of random variables*, Proc. Amer. Math. Soc., 101 (1987), pp. 142–148.
[20] G. CH. PFLUG, *Non-asymptotic confidence bounds for stochastic approximation algorithms with constant step size*, Monatsh. Math., 110 (1990), pp. 297–314.
[21] B. T. POLYAK, *New stochastic approximation type procedures*, Avtomat. i Telemekh., 51 (1990), pp. 98–107 (in Russian), translated in Automat. Remote Control, 51 (1990), pp. 937–946.
[22] B. T. POLYAK AND A. B. JUDITSKY, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.
[23] H. ROBBINS AND D. SIEGMUND, *A convergence theorem for nonnegative almost supermartingales and some applications*, in Optimizing Methods in Statistics, J. S. Rustagi, ed., Academic Press, New York, 1971, pp. 233–257.
[24] D. RUPPERT, *Efficient estimators from a slowly convergent Robbins–Monro process*, Technical report no. 781, School of Oper. Res. and Ind. Eng., Cornell University, Ithaca, N.Y., 1988 (see §2.8. of D. Ruppert, *Stochastic Approximation*, in Handbook of Sequential Analysis, B. K. Ghosh and P. K. Sen, eds., Marcel Dekker, New York, 1991, pp. 503–529).
[25] E. SANCHEZ-PALENCIA, *Méthode de centrage-estimation de l'erreur et comportement des trajectoires dans l'espace des phases*, Internat. J. Non-Linear Mechanics, 11 (1976), pp. 251–263.
[26] J. A. SANDERS AND F. VERHULST, *Averaging Methods in Nonlinear Dynamical Systems*, Springer-Verlag, New York, Berlin, 1985.
[27] H. WALK, *Almost sure convergence of stochastic approximation processes*, Statist. Decisions, supplement issue 2 (1985), pp. 137–141.
[28] H. WALK AND L. ZSIDÓ, *Convergence of the Robbins–Monro method for linear problems in a Banach space*, J. Math. Anal. Appl., 139 (1989), pp. 152–177.
[29] G. YIN, *On extensions of Polyak's averaging approach to stochastic approximation*, Stochastics Stochastics Rep., 36 (1991), pp. 245–264.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 62–73, January 1996

003

ON A CERTAIN PARAMETER OF THE DISCRETIZED EXTENDED LINEAR-QUADRATIC PROBLEM OF OPTIMAL CONTROL*

CIYOU ZHU†

Abstract. The number $\gamma := \|\hat{Q}^{-\frac{1}{2}}\hat{R}\hat{P}^{-\frac{1}{2}}\|$ is an important parameter for the extended linear-quadratic programming (ELQP) problem associated with the Lagrangian $L(\hat{u},\hat{v}) = \hat{p}\cdot\hat{u} + \frac{1}{2}\hat{u}\cdot\hat{P}\hat{u} + \hat{q}\cdot\hat{v} - \frac{1}{2}\hat{v}\cdot\hat{Q}\hat{v} - \hat{v}\cdot\hat{R}\hat{u}$ over polyhedral sets $\hat{U}\times\hat{V}$. Some fundamental properties of the problem, as well as the convergence rates of certain newly developed algorithms for large-scale ELQP, are all related to γ.

In this paper, we derive an estimate of γ for the ELQP problems resulting from discretization of an optimal control problem. We prove that the parameter γ of the discretized problem is bounded independently of the number of subintervals in the discretization.

Key words. extended linear-quadratic programming, minimax problem, optimal control, primal-dual projected gradient algorithm

AMS subject classifications. 65K05, 65K10, 90C20

1. Introduction.

The extended linear-quadratic programming (ELQP) problem, in its standard minimax form, is to find a saddle point of the Lagrangian

$$L(\hat{u},\hat{v}) = \hat{p}\cdot\hat{u} + \tfrac{1}{2}\hat{u}\cdot\hat{P}\hat{u} + \hat{q}\cdot\hat{v} - \tfrac{1}{2}\hat{v}\cdot\hat{Q}\hat{v} - \hat{v}\cdot\hat{R}\hat{u} \quad \text{over} \quad \hat{U}\times\hat{V}, \tag{1.1}$$

where $\hat{U}$ and $\hat{V}$ are polyhedral sets in $\mathbb{R}^{\hat{k}}$ and $\mathbb{R}^{\hat{l}}$, respectively, and $\hat{P}\in\mathbb{R}^{\hat{k}\times\hat{k}}$ and $\hat{Q}\in\mathbb{R}^{\hat{l}\times\hat{l}}$ are symmetric positive semidefinite matrices [1]. The associated primal and dual problems are

$$(\hat{\mathcal{P}}) \qquad \text{minimize } f(\hat{u}) \text{ over all } \hat{u}\in\hat{U}, \text{ where } f(\hat{u}) := \sup_{\hat{v}\in\hat{V}} L(\hat{u},\hat{v}),$$

$$(\hat{\mathcal{Q}}) \qquad \text{maximize } g(\hat{v}) \text{ over all } \hat{v}\in\hat{V}, \text{ where } g(\hat{v}) := \inf_{\hat{u}\in\hat{U}} L(\hat{u},\hat{v}).$$

The problem is called *fully quadratic* if both P and Q are positive definite.

The number

$$\gamma := \|\hat{Q}^{-\frac{1}{2}}\hat{R}\hat{P}^{-\frac{1}{2}}\| \tag{1.2}$$

introduced by Rockafellar [2] is an important parameter for the problem in the fully quadratic case. (We use the Euclidean norm for vectors and the associated operator norm for matrices unless otherwise specified.) It serves as a Lipschitz constant for the mappings $F:\mathbb{R}^{\hat{k}}\to\mathbb{R}^{\hat{l}}$ and $G:\mathbb{R}^{\hat{l}}\to\mathbb{R}^{\hat{k}}$ defined as

$$F(\hat{u}) = \operatorname*{argmax}_{\hat{v}\in\hat{V}} L(\hat{u},\hat{v}) \quad \text{and} \quad G(\hat{v}) = \operatorname*{argmin}_{\hat{u}\in\hat{U}} L(\hat{u},\hat{v}),$$

*Received by the editors July 29, 1993; accepted for publication (in revised form) July 26, 1994. This research was supported by the Office of Scientific Computing of the Department of Energy under contract W-31-109-Eng-38 and by the National Science Foundation under contract ASC-9213149.

†Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439. Present address: EECS Department, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208.

respectively [2]. It also plays a central role in the convergence results of several newly developed algorithms for large-scale ELQP problems.

Let $\varepsilon_\nu = f(u^\nu) - g(v^\nu)$ be the νth duality gap related to the primal-dual pair of iterates (u^ν, v^ν). Rockafellar proved that the sequence $\{(u^\nu, v^\nu)\}$ generated by the finite-envelope algorithm [2] satisfies

$$\frac{\varepsilon_{\nu+1}}{\varepsilon_\nu} \leq 1 - \frac{1}{4(1+\gamma^2)}. \tag{1.3}$$

In [10], Zhu proved that the sequence $\{(u^\nu, v^\nu)\}$ generated by certain variants of the primal-dual steepest descent algorithm developed in [9] satisfies

$$\frac{\varepsilon_{\nu+1}}{\varepsilon_\nu} \leq \begin{cases} \frac{\gamma^2}{2-\gamma^2} & \text{if } 0 \leq \gamma^2 < \frac{1}{2}, \\ 1 - \frac{1}{2\gamma^2+0.5} & \text{if } \gamma^2 \geq \frac{1}{2}. \end{cases} \tag{1.4}$$

One of the variants uses a "fixed step length" strategy [10], where the step lengths are also related to γ. In [6], [7], S. J. Wright described interior point algorithms for linear complementarity problems (LCPs). If the ELQP problem is formulated as an LCP and if the standard conjugate gradient algorithm is used to solve the resulting linear equations, the convergence rate for the "inner iterations" [8] will be

$$\frac{\varepsilon_{\nu+1}}{\varepsilon_\nu} \leq 1 - \frac{2}{1 + (1+\gamma^2)^{\frac{1}{2}}}. \tag{1.5}$$

The right-hand sides of (1.3)–(1.5) all depend on the parameter γ of the problem. The smaller the value of γ, the faster the convergence for these algorithms.

In this paper, we consider the ELQP problem resulting from discretizing a continuous-time optimal control problem with time-independent data. As we show in §2, the matrix $\hat{R}$ for such problem consists of a large number of nonzero blocks, each of which is a product of infinite series in terms of the matrices in the original continuous-time problem. It is usually impractical to compute γ from the definition (1.2). Actually, a primary goal in algorithm design is to avoid computations involving the $\hat{R}$ matrix, because of its size and density. All three of the above-mentioned algorithms for the discretized ELQP problem could be implemented in such a way that this goal is reached by taking advantage of the discretized system dynamics in their computations [3], [4], [8].

Mathematically, however, an unanswered question is the dependence of γ on the data of the original continuous-time problem and on the number of subintervals used in the discretization. Zhu and Rockafellar [9] observe that the number of iterations needed for their algorithms to converge remains essentially unchanged as the discretization is refined. This observation suggests strongly that the value of γ approaches a constant, or is at least bounded above, as the number of subintervals increases. In this paper, we will prove this conjecture on γ. In §2, we derive expressions for the matrices $\hat{P}$, $\hat{Q}$, and $\hat{R}$ in the Lagrangian (1.1) for the discretized problem. In §3, we give an estimate of γ in terms of the matrices in the original continuous-time extended linear-quadratic problem of optimal control, an estimate that is independent of the mesh width.

2. Data matrices in the Lagrangian for the discretized problem. The continuous-time extended linear-quadratic problem of optimal control (with time-independent data and normalized time interval) is

$(\mathcal{P}^{\text{cont}})$ minimize $\mathcal{F}(u_e, u)$

$$= \int_0^1 [p{\cdot}u(t) + \tfrac{1}{2}u(t){\cdot}Pu(t) - c{\cdot}x(t)]dt + [p_e{\cdot}u_e + \tfrac{1}{2}u_e{\cdot}P_eu_e - c_e{\cdot}x(1)]$$
$$+ \int_0^1 \rho_{V,Q}\big(q - Cx(t) - Du(t)\big)dt + \rho_{V_e,Q_e}\big(q_e - C_ex(1) - D_eu\big)$$

over the state trajectory

$$\dot{x}(t) = Ax(t) + Bu(t) + b \quad \text{a.e.}, \quad x(0) = B_eu_e + b_e \quad (x(t) \in \mathbb{R}^m)$$

with the control space

$$\mathcal{U} = \{(u_e, u) \in \mathbb{R}^{k_e} \times \mathcal{L}_k^\infty[0,1] \,\big|\, u_e \in U_e, u(t) \in U \text{ a.e. }\}$$

(Rockafellar [1]). Here U, U_e, V, and V_e are polyhedral convex sets, and P, P_e, Q, and Q_e are symmetric positive semidefinite matrices. Each ρ term, defined as

$$\rho_{V,Q}(s) = \sup_{v \in V}\{s{\cdot}v - \tfrac{1}{2}v{\cdot}Qv\},$$

is a lower semicontinuous convex piecewise linear-quadratic function [1, Prop. 2.3]. The dual problem is

$(\mathcal{Q}^{\text{cont}})$ maximize $\mathcal{G}(v, v_e)$

$$= \int_0^1 [q{\cdot}v(t) - \tfrac{1}{2}v(t){\cdot}Qv(t) - b{\cdot}y(t)]dt + [q_e{\cdot}v_e + \tfrac{1}{2}v_e{\cdot}Q_ev_e - b_e{\cdot}y(0)]$$
$$- \int_0^1 \rho_{U,P}\big(B^Ty(t) + D^Tv(t) - p\big)dt - \rho_{U_e,P_e}\big(B_e^Ty(0) + D_e^Tv - p_e\big)$$

over the state trajectory

$$-\dot{y}(t) = A^Ty(t) + C^Tv(t) + c \quad \text{a.e.}, \quad y(1) = C_e^Tv_e + c_e \quad (y(t) \in \mathbb{R}^m)$$

with the control space

$$\mathcal{V} = \{(v, v_e) \in \mathcal{L}_l^\infty[0,1] \times \mathbb{R}^{l_e} \,\big|\, v(t) \in V \text{ a.e., } v_e \in V_e\},$$

where

$$\rho_{U,P}(r) = \sup_{u \in U}\{r{\cdot}u - \tfrac{1}{2}u{\cdot}Pu\}.$$

Problems $(\mathcal{P}^{\text{cont}})$ and $(\mathcal{Q}^{\text{cont}})$ differ from the conventional linear-quadratic models in optimal control in that they allow for piecewise linear-quadratic penalty terms in the objective functionals, as well as constraints on the controls. See Rockafellar [1] for a detailed presentation.

Problems $(\mathcal{P}^{\text{cont}})$ and $(\mathcal{Q}^{\text{cont}})$ are equivalent to a saddle point problem under certain *finiteness conditions* (which will be satisfied if, for example, the matrices P, Q, P_e, and Q_e, are positive definite) [1, Thm. 6.1 and Coro. 6.4]. The saddle point problem is

$$(\mathcal{S}^{\text{cont}}) \qquad \underset{(u_e,u)\in\mathcal{U},(v,v_e)\in\mathcal{V}}{\text{minimax}} \mathcal{J}(u_e,u;v,v_e),$$

where

$$\mathcal{J}(u_e,u;v,v_e)=\int_0^1 J(u(t),v(t))dt+J_e(u_e,v_e)-\langle(u_e,u);(v,v_e)\rangle$$

with

$$J(u,v)=p\cdot u+\tfrac{1}{2}u\cdot Pu+q\cdot v-\tfrac{1}{2}v\cdot Qv-v\cdot Du \text{ for } u\in\mathbb{R}^k, v\in\mathbb{R}^l,$$
$$J_e(u_e,v_e)=p_e\cdot u_e+\tfrac{1}{2}u_e\cdot P_e u_e+q_e\cdot v_e-\tfrac{1}{2}v_e\cdot Q_e v_e \text{ for } u_e\in\mathbb{R}^{k_e}, v\in\mathbb{R}^{l_e},$$

and

$$\begin{aligned}\langle(u_e,u);(v,v_e)\rangle&=\int_0^1 x(t)\cdot(C^Tv(t)+c)dt+x(1)\cdot(C_e^Tv_e+c_e)\\&=\int_0^1 y(t)\cdot(Bu(t)+b)dt+y(0)\cdot(B_eu_e+b_e).\end{aligned}$$

A numerical solution of $(\mathcal{S}^{\text{cont}})$ can be obtained by discretizing the problem into the following approximate version [4], [5]:

$$(\mathcal{S}_{\text{n}}^{\text{cont}}) \qquad \underset{\mathcal{U}_n\times\mathcal{V}_n}{\text{minimax}} \mathcal{J}(u_e,u;v,v_e),$$

where

$$\mathcal{U}_n=\left\{(u_e,u)\in\mathcal{U}\,\middle|\,u(t) \text{ is constant on } \left(\frac{\tau-1}{n},\frac{\tau}{n}\right),\ \tau=1,\dots,n\right\},$$
$$\mathcal{V}_n=\left\{(v,v_e)\in\mathcal{V}\,\middle|\,v(t) \text{ is constant on } \left(\frac{\tau-1}{n},\frac{\tau}{n}\right),\ \tau=1,\dots,n\right\}.$$

If we denote the constant values of $u(t)$ and $v(t)$ on $\left(\frac{\tau-1}{n},\frac{\tau}{n}\right)$ by u_τ and v_τ, respectively, for $\tau=1,\dots,n$, problem $(\mathcal{S}_{\text{n}}^{\text{cont}})$ can be written in the form of a finite-dimensional discrete-time saddle point problem as

$$(\mathcal{S}_{\text{n}}^{\text{disc}}) \qquad \underset{U_n\times V_n}{\text{minimax}} \mathcal{J}_n(u_e,u_1,\dots,u_n;\ v_1,\dots,v_n,v_e),$$

where

$$\begin{aligned}&\mathcal{J}_n(u_e,u_1,\dots,u_n;\ v_1,\dots,v_n,v_e)\\&\quad=\sum_{\tau=1}^{n}J_n(u_\tau,v_\tau)+J_e(u_e,v_e)-\langle(u_e,u_1,\dots,u_n);\ (v_1,\dots,v_n,v_e)\rangle_n\end{aligned}\tag{2.1}$$

with

$$\begin{aligned}J_n(u_\tau,v_\tau)&=p_n\cdot u_\tau+q_n\cdot v_\tau+\tfrac{1}{2}u_\tau\cdot P_nu_\tau-\tfrac{1}{2}v_\tau\cdot Q_nv_\tau-v_\tau\cdot D_nu_\tau+d_n,\\J_e(u_e,v_e)&=p_e\cdot u_e+q_e\cdot v_e+\tfrac{1}{2}u_e\cdot P_eu_e-\tfrac{1}{2}v_e\cdot Q_ev_e,\end{aligned}\tag{2.2}$$

and

$$\begin{aligned}&\langle (u_e, u_1, \ldots, u_n); (v_1, \ldots, v_n, v_e)\rangle_n \\ &= \sum_{\tau=1}^{n} y_{\tau+1}\cdot(B_n u_\tau + b_n) + y_1\cdot(B_e u_e + b_e) \\ &= \sum_{\tau=1}^{n} x_{\tau-1}\cdot(C_n^T v_\tau + c_n) + x_n\cdot(C_e^T v_e + c_e).\end{aligned} \tag{2.3}$$

The trajectories are given by the discretized system dynamics

(2.4)
$$\begin{aligned}&x_\tau = A_n x_{\tau-1} + B_n u_\tau + b_n \text{ for } \tau = 1, \ldots, n, \quad x_0 = B_e u_e + b_e \quad (x_\tau \in \mathbb{R}^m), \\ &y_\tau = A_n^T y_{\tau+1} + C_n^T v_\tau + c_n \text{ for } \tau = 1, \ldots, n, \quad y_{n+1} = C_e^T v_e + c_e \quad (y_\tau \in \mathbb{R}^m),\end{aligned}$$

where we impose

$$\begin{aligned}(u_e, u_1, \ldots, u_n) \in U_n :=& U_e \times (U)^n \subseteq \mathbb{R}^{k_e} \times (\mathbb{R}^k)^n, \\ (v_1, \ldots, v_n, v_e) \in V_n :=& (V)^n \times V_e \subseteq (\mathbb{R}^l)^n \times \mathbb{R}^{l_e}.\end{aligned}$$

The transformation of the data is

$$A_n = I + M_n A, \tag{2.5a}$$

$$B_n = M_n B, \qquad b_n = M_n b, \tag{2.5b}$$

$$C_n = C M_n, \qquad c_n = M_n^T c, \tag{2.5c}$$

$$D_n = \frac{1}{n} D + C S_n B, \qquad d_n = -c\cdot S_n b, \tag{2.5d}$$

$$P_n = \frac{1}{n} P, \qquad p_n = \frac{1}{n} p - B^T S_n^T c, \tag{2.5p}$$

$$Q_n = \frac{1}{n} Q, \qquad q_n = \frac{1}{n} q - C S_n b, \tag{2.5q}$$

where

$$S_n = \sum_{i=2}^{\infty} \frac{1}{i!} \left(\frac{1}{n}\right)^i A^{i-2}, \qquad M_n = \frac{1}{n} I + A S_n. \tag{2.6}$$

(Wright [4], [5]). The associated primal and dual problems are

$(\mathcal{P}_{\mathrm{n}}^{\mathrm{disc}})$ minimize $f(u)$ over $u \in U_n$, where
$$f(u_e, u_1, \ldots, u_n) := \max_{v \in V_n} \mathcal{J}_n(u_e, u_1, \ldots, u_n;\, v_1, \ldots, v_n, v_e)$$

and

$(\mathcal{Q}_{\mathrm{n}}^{\mathrm{disc}})$ maximize $g(v)$ over $v \in V_n$, where
$$g(v_1, \ldots, v_n, v_e) := \min_{u \in U_n} \mathcal{J}_n(u_e, u_1, \ldots, u_n;\, v_1, \ldots, v_n, v_e).$$

Problems $(\mathcal{P}_{\mathrm{n}}^{\mathrm{disc}})$ and $(\mathcal{Q}_{\mathrm{n}}^{\mathrm{disc}})$ are ELQP in the multistage format, which could be solved directly by the techniques mentioned in §1 without forming the huge $\hat{R}$ matrix in the Lagrangian (1.1) of its standard form. However, in order to get an expression for γ in terms of the matrices A, B, C, D, P, and Q from the continuous-time problem, we

eliminate the state variables x_τ and y_τ in the expression of $\mathcal{J}_n$. From the discretized system dynamics (2.4), we obtain

$$\begin{bmatrix} x_e \\ x_1 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} B_e & & & & \\ A_nB_e & B_n & & & \\ \cdot & \cdot & \cdots & & \\ \cdot & \cdot & \cdots & & \\ A_n^nB_e & A_n^{n-1}B_n & \cdots & A_nB_n & B_n \end{bmatrix} \begin{bmatrix} u_e \\ u_1 \\ \cdot \\ \cdot \\ u_n \end{bmatrix} + \begin{bmatrix} b_e \\ A_nb_e + b_n \\ \cdot \\ \cdot \\ A_n^nb_e + A_n^{n-1}b_n + \ldots + b_n \end{bmatrix}$$

$$= \begin{bmatrix} I & & & & \\ A_n & I & & & \\ \cdot & \cdot & \cdots & & \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ A_n^n & A_n^{n-1} & \cdots & A_n & I \end{bmatrix} \left(\begin{bmatrix} B_eu_e \\ B_nu_1 \\ \cdot \\ \cdot \\ B_nu_n \end{bmatrix} + \begin{bmatrix} b_e \\ b_n \\ \cdot \\ \cdot \\ b_n \end{bmatrix} \right). \tag{2.7}$$

By substituting (2.7) in the second expression of (2.3), we obtain

$$\begin{aligned} &\langle (u_e, u_1, \ldots, u_n); (v_1, \ldots, v_n, v_e) \rangle_n \\ &= \begin{bmatrix} c_n \\ \cdot \\ \cdot \\ c_n \\ c_e \end{bmatrix} \cdot \begin{bmatrix} x_e \\ x_1 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} + \begin{bmatrix} v_1 \\ \cdot \\ \cdot \\ v_n \\ v_e \end{bmatrix} \cdot \begin{bmatrix} C_n & & & \\ & \cdots & & \\ & & C_n & \\ & & & C_e \end{bmatrix} \begin{bmatrix} x_e \\ x_1 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \\ &= \begin{bmatrix} c_n \\ \cdot \\ \cdot \\ c_n \\ c_e \end{bmatrix} \cdot \begin{bmatrix} B_e & & & & \\ A_nB_e & B_n & & & \\ \cdot & \cdot & \cdots & & \\ \cdot & \cdot & \cdots & \cdot & \\ A_n^nB_e & A_n^{n-1}B_n & \cdots & A_nB_n & B_n \end{bmatrix} \begin{bmatrix} u_e \\ u_1 \\ \cdot \\ \cdot \\ u_n \end{bmatrix} \\ &\quad + \begin{bmatrix} v_1 \\ \cdot \\ \cdot \\ v_n \\ v_e \end{bmatrix} \cdot \begin{bmatrix} C_nB_e & & & & \\ C_nA_nB_e & C_nB_n & & & \\ \cdot & \cdot & \cdots & & \\ \cdot & \cdot & \cdots & \cdot & \\ C_eA_n^nB_e & C_eA_n^{n-1}B_n & \cdots & C_eA_nB_n & C_eB_n \end{bmatrix} \begin{bmatrix} u_e \\ u_1 \\ \cdot \\ \cdot \\ u_n \end{bmatrix} \\ &\quad + \begin{bmatrix} v_1 \\ \cdot \\ \cdot \\ v_n \\ v_e \end{bmatrix} \cdot \begin{bmatrix} C_n & & & & \\ C_nA_n & C_n & & & \\ \cdot & \cdot & \cdots & & \\ \cdot & \cdot & \cdots & \cdot & \\ C_eA_n^n & C_eA_n^{n-1} & \cdots & C_eA_n & C_e \end{bmatrix} \begin{bmatrix} b_e \\ b_n \\ \cdot \\ \cdot \\ b_n \end{bmatrix} \\ &\quad + \begin{bmatrix} c_n \\ \cdot \\ \cdot \\ c_n \\ c_e \end{bmatrix} \cdot \begin{bmatrix} I & & & & \\ A_n & I & & & \\ \cdot & \cdot & \cdots & & \\ \cdot & \cdot & \cdots & \cdot & \\ A_n^n & A_n^{n-1} & \cdots & A_n & I \end{bmatrix} \begin{bmatrix} b_e \\ b_n \\ \cdot \\ \cdot \\ b_n \end{bmatrix}. \end{aligned} \tag{2.8}$$

Similarly, we have from (2.2) that

$$\begin{aligned} &\sum_{\tau=1}^{n} J_n(u_\tau, v_\tau) + J_e(u_e, v_e) \\ &= \begin{bmatrix} p_e \\ p_n \\ \cdot \\ \cdot \\ p_n \end{bmatrix} \cdot \begin{bmatrix} u_e \\ u_1 \\ \cdot \\ \cdot \\ u_n \end{bmatrix} + \begin{bmatrix} q_n \\ \cdot \\ \cdot \\ q_n \\ q_e \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ \cdot \\ \cdot \\ v_n \\ v_e \end{bmatrix} + \frac{1}{2} \begin{bmatrix} u_e \\ u_1 \\ \cdot \\ \cdot \\ u_n \end{bmatrix} \cdot \begin{bmatrix} P_e & & & \\ & P_n & & \\ & & \cdots & \\ & & & P_n \end{bmatrix} \begin{bmatrix} u_e \\ u_1 \\ \cdot \\ \cdot \\ u_n \end{bmatrix} \end{aligned}$$

$$-\frac{1}{2}\begin{bmatrix} v_1 \\ \cdot \\ \cdot \\ v_n \\ v_e \end{bmatrix} \cdot \begin{bmatrix} Q_n & & \cdots & & \\ & Q_n & & & \\ & & & Q_e & \end{bmatrix} \begin{bmatrix} v_1 \\ \cdot \\ \cdot \\ v_n \\ v_e \end{bmatrix}$$

$$-\begin{bmatrix} v_1 \\ \cdot \\ \cdot \\ v_n \\ v_e \end{bmatrix} \cdot \begin{bmatrix} 0 & D_n & & & \\ & \cdots & \cdots & & \\ & & 0 & D_n & \\ & & & 0 & \end{bmatrix} \begin{bmatrix} u_e \\ u_1 \\ \cdot \\ \cdot \\ u_n \end{bmatrix} \tag{2.9}$$

By substituting (2.8) and (2.9) in (2.1), we obtain equality of $\mathcal{J}_n$ with the Lagrangian $L(\hat{u}, \hat{v})$ in (1.1) by noting the identities

$$\hat{P} = \operatorname{diag}[P_e, P_n, \cdots, P_n], \qquad \hat{Q} = \operatorname{diag}[Q_n, \cdots, Q_n, Q_e], \tag{2.10}$$

$$\hat{R} = \begin{bmatrix} C_nIB_e & D_n & & & & & \\ C_nA_nB_e & C_nIB_n & D_n & & & & \\ C_nA_n^2B_e & C_nA_nB_n & C_nIB_n & D_n & & & \\ \cdot & \cdot & \cdot & \cdot & \cdots & & \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \\ C_nA_n^{n-1}B_e & C_nA_n^{n-2}B_n & \cdot & \cdot & \cdots & C_nIB_n & D_n \\ C_eA_n^nB_e & C_eA_n^{n-1}B_n & \cdot & \cdot & \cdots & C_eA_nB_n & C_eIB_n \end{bmatrix}, \tag{2.11}$$

$$\hat{p} = \begin{bmatrix} p_e \\ p_n \\ \cdot \\ \cdot \\ p_n \end{bmatrix} - \begin{bmatrix} B_e^Tc_n + \ldots + B_e^T(A_n^T)^{n-1}c_n + B_e^T(A_n^T)^nc_e \\ B_n^Tc_n + \ldots + B_n^T(A_n^T)^{n-2}c_n + B_n^T(A_n^T)^{n-1}c_e \\ \cdot \\ \cdot \\ B_n^Tc_n + B_n^TA_n^Tc_e \\ B_n^Tc_e \end{bmatrix},$$

$$\hat{q} = \begin{bmatrix} q_n \\ \cdot \\ \cdot \\ q_n \\ q_e \end{bmatrix} - \begin{bmatrix} C_nb_e \\ C_nA_nb_e + C_nb_n \\ \cdot \\ \cdot \\ C_nA_n^{n-1}b_e + C_nA_n^{n-2}b_n + \ldots + C_nb_n \\ C_eA_n^nb_e + C_eA_n^{n-1}b_n + \ldots + C_eb_n \end{bmatrix}$$

with

$$\hat{u} = (u_e, u_1, \cdots, u_n), \qquad \hat{v} = (v_1, \cdots, v_n, v_e).$$

(The additive constants in $\mathcal{J}_n$ are dropped since they play no role in the problem.)

3. Estimation of the parameter γ of the discretized problem. In this section, we prove the following estimate of the parameter γ of the discretized problem in terms of the continuous-time problem data.

THEOREM 3.1. *Suppose the matrices P, Q, P_e, and Q_e in the continuous-time extended linear-quadratic problems $(\mathcal{P}^{\text{cont}})$ and $(\mathcal{Q}^{\text{cont}})$ of optimal control are positive definite. Then the parameter γ_n of the discretized versions $(\mathcal{P}_n^{\text{disc}})$ and $(\mathcal{Q}_n^{\text{disc}})$ satisfies*

$$\begin{aligned} \gamma_n \le\ & \|Q^{-\frac{1}{2}}C\|\|B_eP_e^{-\frac{1}{2}}\|e^{(n-1)\|A\|/n} \\ & + \|Q_e^{-\frac{1}{2}}C_e\|\|BP^{-\frac{1}{2}}\|e^{(n-1)\|A\|/n} + \|Q_e^{-\frac{1}{2}}C_e\|\|B_eP_e^{-\frac{1}{2}}\|e^{\|A\|} \\ & + \|Q^{-\frac{1}{2}}DP^{-\frac{1}{2}}\| + \|Q^{-\frac{1}{2}}C\|\|BP^{-\frac{1}{2}}\|\frac{1-e^{(n-1)\|A\|/n}}{n(1-e^{\|A\|/n})} + O(n^{-1}), \end{aligned} \tag{3.1}$$

where n is the number of equal-length subintervals used in the discretization. Moreover,

$$\begin{aligned}\limsup_{n\to\infty}\gamma_n \le\ & \|Q^{-\frac{1}{2}}C\|\|B_eP_e^{-\frac{1}{2}}\|e^{\|A\|} \\ & + \|Q_e^{-\frac{1}{2}}C_e\|\|BP^{-\frac{1}{2}}\|e^{\|A\|} + \|Q_e^{-\frac{1}{2}}C_e\|\|B_eP_e^{-\frac{1}{2}}\|e^{\|A\|} \\ & + \|Q^{-\frac{1}{2}}DP^{-\frac{1}{2}}\| + \|Q^{-\frac{1}{2}}C\|\|BP^{-\frac{1}{2}}\|\frac{e^{\|A\|}-1}{\|A\|}.\end{aligned} \tag{3.2}$$

Two conclusions follow immediately from Theorem 3.1:

(i) The parameter γ_n of the discretized problem, as a function of n, is bounded above when $n \to \infty$. Hence for the algorithms with their convergence rates having an upper bound determined solely by γ_n, the number of iterations needed for convergence should remain essentially the same as n increases. If, in addition, the algorithm has only $O(n)$ operations in each iteration, then the total central processing unit time needed for convergence should be proportional to n. These results are consistent with the observations of Zhu and Rockafellar [9].

(ii) The only part of the original data that has an exponential contribution to γ_n is the norm of the matrix A in the system dynamics as the coefficient of the state variables. The norms of all the other matrices contribute linearly.

Before proving the theorem, we first state two simple propositions. The proofs of these propositions are elementary and therefore skipped.

PROPOSITION 3.2. *Suppose matrix E can be partitioned as*

$$E = \begin{bmatrix} E_{11} & E_{12} & \cdots & E_{1s} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ E_{r1} & E_{r2} & \cdots & E_{rs} \end{bmatrix}.$$

Then

$$\|E\| \le \sum_{i=1}^{r}\sum_{j=1}^{s}\|E_{ij}\|.$$

PROPOSITION 3.3. *Suppose matrix E is of block diagonal form*

$$E = \operatorname{diag}[E_1, E_2, \ldots, E_r].$$

Then

$$\|E\| = \max_{1\le i\le r}\{\|E_i\|\}.$$

Proof of the theorem. Let $\Gamma := \hat{Q}^{-\frac{1}{2}}\hat{R}\hat{P}^{-\frac{1}{2}}$. Then $\gamma_n = \|\Gamma\|$. Observe that the matrix $\hat{R}$ in (2.11) is block lower Hessenberg, while the matrices $\hat{P}$ and $\hat{Q}$ in (2.10) are block diagonal with the corresponding block structure. Hence the matrix $\Gamma = \hat{Q}^{-\frac{1}{2}}\hat{R}\hat{P}^{-\frac{1}{2}}$ is also of block lower Hessenberg with the same block structure as that of $\hat{R}$. Partition Γ as

$$\Gamma = \begin{bmatrix} \Gamma_{ne} & \Gamma_{nn} \\ \Gamma_{ee} & \Gamma_{en} \end{bmatrix}, \tag{3.3}$$

where

$$\Gamma_{ne} = \begin{bmatrix} Q_n^{-\frac{1}{2}} C_n I B_e P_e^{-\frac{1}{2}} \\ Q_n^{-\frac{1}{2}} C_n A_n B_e P_e^{-\frac{1}{2}} \\ \cdot \\ \cdot \\ \cdot \\ Q_n^{-\frac{1}{2}} C_n A_n^{n-1} B_e P_e^{-\frac{1}{2}} \end{bmatrix}, \quad \Gamma_{ee} = \left[Q_e^{-\frac{1}{2}} C_e A_n^n B_e P_e^{-\frac{1}{2}} \right], \tag{3.4}$$

$$\Gamma_{en} = \left[Q_e^{-\frac{1}{2}} C_e A_n^{n-1} B_n P_n^{-\frac{1}{2}} \quad \cdots \quad Q_e^{-\frac{1}{2}} C_e A_n B_n P_n^{-\frac{1}{2}} \quad Q_e^{-\frac{1}{2}} C_e I B_n P_n^{-\frac{1}{2}} \right], \tag{3.5}$$

and

(3.6)

$$\Gamma_{nn} = \begin{bmatrix} Q_n^{-\frac{1}{2}} D_n P_n^{-\frac{1}{2}} & & & \\ Q_n^{-\frac{1}{2}} C_n I B_n P_n^{-\frac{1}{2}} & Q_n^{-\frac{1}{2}} D_n P_n^{-\frac{1}{2}} & & \\ \cdot & \cdot & \cdots & \\ \cdot & \cdot & \cdots & \cdot \\ Q_n^{-\frac{1}{2}} C_n A_n^{n-2} B_n P^{-\frac{1}{2}} & Q_n^{-\frac{1}{2}} C_n A_n^{n-3} B_n P^{-\frac{1}{2}} & \cdots & Q_n^{-\frac{1}{2}} C_n I B_n P_n^{-\frac{1}{2}} \quad Q_n^{-\frac{1}{2}} D_n P_n^{-\frac{1}{2}} \end{bmatrix}.$$

It follows from (2.6) that

$$S_n = \sum_{i=2}^{\infty} \frac{1}{i!} \left(\frac{1}{n}\right)^i A^{i-2} = \frac{1}{2n^2} I + O(n^{-3}), \tag{3.7a}$$

$$M_n = \frac{1}{n} I + A S_n = \frac{1}{n} I + O(n^{-2}). \tag{3.7b}$$

Hence, by equations (2.5), we have the following first-order approximations for the matrices in the discretized problem:

$$A_n = I + M_n A = \sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{A}{n}\right)^i = e^{A/n}, \tag{3.8a}$$

$$B_n = M_n B = \frac{1}{n} B + O(n^{-2}), \tag{3.8b}$$

$$C_n = C M_n = \frac{1}{n} C + O(n^{-2}), \tag{3.8c}$$

$$D_n = \frac{1}{n} D + C S_n B = \frac{1}{n} D + O(n^{-2}), \tag{3.8d}$$

$$P_n = \frac{1}{n} P, \tag{3.8p}$$

$$Q_n = \frac{1}{n} Q, \tag{3.8q}$$

Applying Proposition 3.2 to the partitioned form of Γ in (3.3), we have

$$\|\Gamma\| \leq \|\Gamma_{ne}\| + \|\Gamma_{ee}\| + \|\Gamma_{en}\| + \|\Gamma_{nn}\|. \tag{3.9}$$

Now we estimate the norms on the right-hand side of (3.9). The matrix Γ_{ne} in (3.4) can be written as

(3.10)
$$\Gamma_{ne} = \operatorname{diag}[Q_n^{-\frac{1}{2}} C_n B_e P_e^{-\frac{1}{2}}, Q_n^{-\frac{1}{2}} C_n A_n B_e P_e^{-\frac{1}{2}}, \ldots, Q_n^{-\frac{1}{2}} C_n A_n^{n-1} B_e P_e^{-\frac{1}{2}}][I \ \cdots \ I]^T.$$

However,

$$\begin{aligned}
&\| \operatorname{diag}[Q_n^{-\frac{1}{2}} C_n B_e P_e^{-\frac{1}{2}}, Q_n^{-\frac{1}{2}} C_n A_n B_e P_e^{-\frac{1}{2}}, \ldots, Q_n^{-\frac{1}{2}} C_n A_n^{n-1} B_e P_e^{-\frac{1}{2}}] \| \\
&\quad = \max_{0 \le i \le n-1} \{ \| Q_n^{-\frac{1}{2}} C_n A_n^i B_e P_e^{-\frac{1}{2}} \| \} \\
&\quad \le \max_{0 \le i \le n-1} \{ \| Q_n^{-\frac{1}{2}} C_n \| \| A_n^i \| \| B_e P_e^{-\frac{1}{2}} \| \} \\
&\quad \le \| Q_n^{-\frac{1}{2}} C_n \| \| B_e P_e^{-\frac{1}{2}} \| \max_{0 \le i \le n-1} \| A_n \|^i \} \\
&\quad \le \| Q_n^{-\frac{1}{2}} C_n \| \| B_e P_e^{-\frac{1}{2}} \| \max\{1, \| A_n \|^{n-1}\} \qquad (3.11)
\end{aligned}$$

by Proposition 3.3. It follows from (3.8a) and (3.8c) that

$$\| A_n \|^{n-1} \le e^{(n-1)\|A\|/n}, \qquad (3.12)$$

$$\| Q_n^{-\frac{1}{2}} C_n \| = n^{-\frac{1}{2}} \| Q^{-\frac{1}{2}} C \| + O(n^{-\frac{3}{2}}). \qquad (3.13)$$

Substituting (3.12) and (3.13) in (3.11), and using $\| [I \ \cdots \ I]^T \| = n^{\frac{1}{2}}$, we obtain

$$\begin{aligned}
\| \Gamma_{ne} \| &\le \big(\| Q^{-\frac{1}{2}} C \| + O(n^{-1}) \big) \| B_e P_e^{-\frac{1}{2}} \| \max\{1, e^{(n-1)\|A\|/n}\}, \\
&\le \| Q^{-\frac{1}{2}} C \| \| B_e P_e^{-\frac{1}{2}} \| e^{(n-1)\|A\|/n} + O(n^{-1}). \qquad (3.14)
\end{aligned}$$

We can show in a similar way that

$$\| \Gamma_{en} \| \le \| Q_e^{-\frac{1}{2}} C_e \| \| B P^{-\frac{1}{2}} \| e^{(n-1)\|A\|/n} + O(n^{-1}). \qquad (3.15)$$

For the matrix Γ_{ee}, it is obvious by (3.8a) that

$$\| \Gamma_{ee} \| = \| Q_e^{-\frac{1}{2}} C_e e^A B_e P_e^{-\frac{1}{2}} \| \le \| Q_e^{-\frac{1}{2}} C_e \| \| B_e P_e^{-\frac{1}{2}} \| e^{\|A\|}. \qquad (3.16)$$

Next, we estimate $\|\Gamma_{nn}\|$. Let $\Gamma_{nn}^{(0)}$ be the matrix obtained by zeroing out all the blocks of Γ_{nn} except the diagonal blocks. Let $\Gamma_{nn}^{(i)}$, $i = 1, \ldots, n-1$, be the matrix obtained by zeroing out all the blocks of Γ_{nn} except the blocks on the ith subdiagonal. Then by Propositions 3.2 and 3.3, we have

$$\| \Gamma_{nn}^{(0)} \| = \| Q_n^{-\frac{1}{2}} D_n P_n^{-\frac{1}{2}} \| = \| Q^{-\frac{1}{2}} D P^{-\frac{1}{2}} \| + O(n^{-1})$$

and

$$\| \Gamma_{nn}^{(i)} \| = \| Q_n^{-\frac{1}{2}} C_n A_n^{i-1} B_n P_n^{-\frac{1}{2}} \| = \frac{1}{n} \| Q^{-\frac{1}{2}} C e^{(i-1)A/n} B P^{-\frac{1}{2}} \| + O(n^{-2})$$

for $i = 1, \ldots, n-1$, where the first-order approximations in (3.8) are used to get the right-hand sides. Hence

$$\begin{aligned}\|\Gamma_{nn}\| &\le \|\Gamma_{nn}^{(0)}\| + \sum_{i=1}^{n-1} \|\Gamma_{nn}^{(i)}\| \\ &= \|Q^{-\frac{1}{2}} D P^{-\frac{1}{2}}\| + \frac{1}{n} \sum_{i=1}^{n-1} \|Q^{-\frac{1}{2}} C e^{(i-1)A/n} B P^{-\frac{1}{2}}\| + O(n^{-1}) \\ &\le \|Q^{-\frac{1}{2}} D P^{-\frac{1}{2}}\| + \frac{1}{n} \|Q^{-\frac{1}{2}} C\| \|B P^{-\frac{1}{2}}\| \sum_{i=1}^{n-1} \|e^{(i-1)A/n}\| + O(n^{-1}).\end{aligned}$$

But

$$\sum_{i=1}^{n-1} \|e^{(i-1)A/n}\| \le \sum_{i=1}^{n-1} e^{(i-1)\|A\|/n} = \frac{1 - e^{(n-1)\|A\|/n}}{1 - e^{\|A\|/n}}.$$

Therefore

$$\|\Gamma_{nn}\| \le \|Q^{-\frac{1}{2}} D P^{-\frac{1}{2}}\| + \|Q^{-\frac{1}{2}} C\| \|B P^{-\frac{1}{2}}\| \frac{1 - e^{(n-1)\|A\|/n}}{n(1 - e^{\|A\|/n})} + O(n^{-1}). \tag{3.17}$$

Substituting (3.14), (3.15), (3.16), and (3.17) in (3.9), we get the inequality (3.1) in Theorem 3.1. Taking limsup on both sides of (3.1), we obtain (3.2). □

In all the above discussions, we assume time-independent data in the optimal control problem. As a final remark, we point out that it would not be difficult to derive a similar bound for the ELQP problem arising from the Euler difference scheme when the optimal control data is time dependent. Actually the proof of Theorem 3.1, with minor adaptations, still works in this latter case if the data elements in

$$A(t), B(t), C(t), D(t), b(t), c(t), P(t), Q(t), p(t), q(t), U(t), V(t)$$

are all Lipschitzian in t. A detailed exposition for this kind of time-dependent case will be presented elsewhere.

Acknowledgments. The author is indebted to S. J. Wright and the associate editor for their helpful comments and suggestions and to an anonymous referee for remarks on the time-dependent case. The final remark in the last section was due to this referee.

REFERENCES

[1] R.T. ROCKAFELLAR, *Linear-quadratic programming and optimal control*, SIAM J. Control Optim. 25 (1987), pp. 781–814.

[2] ———, *Computational schemes for solving large-scale problems in extended linear-quadratic programming*, Math. Programming, 48 (1990), pp. 447–474.

[3] ———, *Large-scale extended linear-quadratic programming and multistage optimization*, in Proc. Fifth Mexico-U.S. Workshop on Numerical Analysis, S. Gomez, J.-P. Hennart, and R. Tapia, eds., Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[4] S. E. WRIGHT, (WITH INTRODUCTION BY R.T. ROCKAFELLAR), *DYNFGM: Dynamic Finite Generation Method*, report, Department of Mathematics, University of Washington, 1989.

[5] ———, *Consistency of primal-dual approximations of convex optimal control problems*, SIAM J. Control Optim., 33 (1995), pp. 1489–1509.

[6] S. J. WRIGHT, *An infeasible-interior-point algorithm for linear complementarity problems,* Math. Programming, Series A, 67 (1994), pp. 29–52.

[7] ———, *A path-following infeasible-interior-point algorithm for linear complementarity problems,* Optimization Methods and Software, 2 (1993), pp. 79-106.

[8] ———, unpublished information, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1992.

[9] C. ZHU AND R. T. ROCKAFELLAR, *Primal-dual projected gradient algorithms for extended linear-quadratic programming,* SIAM J. Optim., 3 (1993), pp. 751–783.

[10] C. ZHU, *On the primal-dual steepest descent algorithm for extended linear-quadratic programming,* SIAM J. Optim., 5 (1995), pp. 114–128.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 74–101, January 1996

004

BELLMAN EQUATIONS OF RISK-SENSITIVE CONTROL*

H. NAGAI†

Dedicated to Professor M. Fukushima on the occasion of his sixtieth birthday

Abstract. Risk-sensitive control problems are considered. Existence of a nonnegative solution to the Bellman equation of risk-sensitive control is shown. The result is applied to prove that no breaking down occurs. Asymptotic behaviour of the nonnegative solution is studied in relation to ergodic control problems and the relationship between the asymptotics and the large deviation principle is noted.

Key words. risk-sensitive control, Bellman equation, ergodic control, breaking down, asymptotic behaviour, large deviation

AMS subject classifications. 93E20, 49L20, 35K55, 60F10, 60H30

Introduction. Let us consider the following stochastic control problem minimizing

$$I(T,x;z,\theta)=\frac{1}{\theta}\log E\left[e^{\theta\int_0^T\{V(X_s)+\phi(X_s,z_s)\}ds}\right],\quad \theta\in R\backslash\{0\},\tag{0.1}$$

subject to controlled processes governed by the stochastic differential equation

$$\begin{cases} dX_t=\sigma(X_t)dB_t+b(X_t)dt+c(X_t,z_t)dt,\\ X_0=x, \end{cases}$$

where B_t is a standard Brownian motion process defined on a probability space $(\Omega,\mathcal{F},P)$ and z_t is a control process assuming its value on a control region $Z\subset R^{N_1}$. The constant θ is called a risk-sensitive parameter and its meaning is realized by considering the asymptotics as $\theta\to 0$:

$$I(T,x;z,\theta)=E[\Phi_T]+\theta\mathrm{Var}[\Phi_T]+O(\theta^2),$$

where $\Phi_T=\int_0^T\{V(X_s)+\phi(X_s,z_s)\}ds$. The case where $\theta>0$ is called risk averse and $\theta<0$, risk seeking. We assume that V and ϕ are nonnegative functions such that $V(x)\to\infty$ as $|x|\to\infty$ and $\phi(x,z)\to\infty$ as $|z|\to\infty$, and therefore it may occur in risk-averse cases that (0.1) never has finite value for any control process z_t. We then say that the control problem breaks down. Thus we are led to the problem of finding the conditions where no breaking down occurs.

It is natural to relate the problem with the existence of the solution to the Bellman equation from the control theoretical point of view since the value function should satisfy the equation if it has finite value and sufficient regularity. Actually we shall first study the existence of a nonnegative solution to the Bellman equation:

$$\begin{cases} \frac{\partial u}{\partial t}=\frac{1}{2}a^{ij}D_{ij}u+b^iD_iu+Q_0(x,\nabla u)+V(x),\\ u(0,x)=0, \end{cases}\tag{0.2}$$

* Received by the editors September 9, 1993; accepted for publication (in revised form) July 26, 1994.

† Department of Mathematical Science, Faculty of Engineering Science, Osaka University, Toyonaka, Osaka, Japan.

with

$$Q_0(x,p) = \frac{\theta}{2}a^{ij}p_ip_j + \inf_{z\in Z}\{c^i(x,z)p_i + \phi(x,z)\}, \quad p \in R^N$$

(cf. Theorem 1.1) and then prove in §2 that no breaking down occurs under the assumptions of Theorem 1.1. To obtain the existence theorem the estimate (1.28) plays a key role and the assumption (1.16) is essential to deriving the estimate. Moreover the assumption indicates the bound of the size of the risk-sensitive parameter θ ensuring finiteness of the value function for any terminal time T in the risk-averse case. In fact, Theorem 2.3 combined with Theorem 1.1 shows that no breaking down occurs under the assumption besides other conditions of Theorem 1.1 (cf. §2.2 and Remark 1.2). The linear exponential quadratic Gaussian (LEQG) case is covered by our assumptions. More general examples are illustrated in §1.5. LEQG refers to the case where V (resp., ϕ) is a quadratic function of x (resp., z), b (resp., c) a linear function of x (resp., z), and σ a constant matrix, and it has been studied from various points of view (cf. Jacobson [15], Whittle [25], Bensoussan and Van Schuppen [3]).

The next problem is to see how the value function $I^*(T,x;\theta) = \inf I(T,x;z,\theta)$ behaves as $T \to \infty$. Related to this problem we shall study the asymptotic behaviour of the nonnegative solution $u(t,x)$ to the Bellman equation (0.2) as $t \to \infty$. We shall prove in a specialized case of $Q_0(x,p) = a^{ij}p_ip_j$ that $u(t,x) - u(t,0)$ converges to a function v and $\frac{\partial u}{\partial t}$ to a constant χ, characterized by the following Bellman equation of ergodic type:

$$\text{(0.3)} \qquad \begin{cases} \chi = \frac{1}{2}a^{ij}D_{ij}v + b^iD_iv + Q_0(x,\nabla v) + V(x), & x \in R^N, \\ \chi \equiv \text{constant}, \quad v \in C^2(R^N). \end{cases}$$

We find as its corollary that

$$\text{(0.4.)} \qquad \lim_{T\to\infty}\frac{u(T,x)}{T} = \lim_{T\to\infty}\frac{\partial u}{\partial t}(T,x) = \chi.$$

Several authors so far (cf. [1], [4], [21], [11], [19]) have deduced an equation such as (0.3) from the Bellman equation of discounted type as the discounted factor tends to 0, with the most far-reaching study on the matter by Bensoussan and Frehse [2]. We further mention the studies on the asymptotic behaviour of the solution u in the case of LEQG done by several authors, e.g., by Whittle [25], Glover and Doyle [12], Runolfsson [23], and in other cases by Fleming and McEneaney [9].

We note the relationship between the asymptotics and large deviation principle due to Donsker and Varadhan [7], which has been noticed by Runolfsson [23]. We illustrate a typical example in §3.3 indicating the relationship, where χ is realized as the constant relating to the principal eigenvalue of a Schrödinger operator.

1. Existence.

1.1. Statement of Theorem 1.1. Let $(\Omega, \mathcal{F}, P)$ be a probability space with filtration $\mathcal{F}_t$, $t \geq 0$, B_t a standard N-dimensional $\mathcal{F}_t$- Brownian motion process, and z_t a progressively measurable process with the value on a Borel subset Z of R^{N_1}. We consider the following stochastic differential equation (SDE):

$$\text{(1.1)} \qquad \begin{cases} dX_s^i = \sigma_j^i(X_s)dB_s^j + b^i(X_s)ds + c^i(X_s,z_s)ds, & i = 1,\dots,N, \\ X_0 = x \end{cases}$$

and introduce the value functions J_+^* and J_-^* as follows:

$$J_+^*(t,x;T,\theta) = \inf_{\mathcal{A}_{T-t}} E_x[e^{\theta\Phi_{T-t}}], \quad \theta > 0, \tag{1.2}$$

$$J_-^*(t,x;T,\theta) = \sup_{\mathcal{A}_{T-t}} E_x[e^{\theta\Phi_{T-t}}], \quad \theta < 0, \tag{1.3}$$

where

$$\Phi_t = \int_0^t \{V(X_s) + \phi(X_s, z_s)\}ds$$

and $\mathcal{A}_{T-t}$ is the totality of $(\Omega, \mathcal{F}, \mathcal{F}_s, P, B_s, z_s)_{0\le s<T-t}$ such that (1.1) has a unique solution for $0 \le s < T-t$. In this paper we employ the summation convention that in cases where the same indices appear in a term twice, the symbol of summation is omitted. We assume the following conditions:

(1.4) σ, b, c, V, and ϕ are smooth,

(1.5) $\|\sigma(x) - \sigma(y)\| \le M|x-y|$, $|b(x) - b(y)| \le M|x-y|$, $\exists M > 0$,

(1.6) all derivatives of σ, b, and $V$are dominated by $M(1+|x|)^m$, $\exists m > 0$,

(1.7) $|c(x,z)| \le c_0(z)$ for some locally bounded function $c_0(z)$,

(1.8) $V(x) \ge 0$ and $\lim_{|x|\to\infty} V(x) = \infty$,

(1.9) $\phi(x,z) \ge 0$ and

$$\lim_{|z|\to\infty} \phi(x,z) = \infty, \quad \lim_{|z|\to\infty} \frac{|c(x,z)|}{\phi(x,z)} = 0 \text{ uniformly in } x,$$

(1.10) $a^{ij}\xi_j\xi_j \ge \nu|\xi|^2$, $\xi \in R^N$ $\exists \nu > 0$,

where $a^{ij} = (\sigma\sigma^*)^{ij}$. In the case where $\theta > 0$ the Bellman equation for the stochastic control problem (1.2) is formally written as

$$\begin{cases} \mathcal{L}J_+ + \inf_{z\in Z}\{c^i(x,z)D_iJ_+ + \theta(V(x)+\phi(x,z))J_+\} = 0, & [0,T)\times R^N, \\ J_+(T,x) = 1, \end{cases} \tag{1.11}$$

where

$$\mathcal{L}J_+ = \frac{\partial J_+}{\partial t} + \frac{1}{2}a^{ij}D_{ij}J_+ + b^iD_iJ_+$$

and

$$D_{ij} = \frac{\partial^2}{\partial x_i \partial x_j}, \quad D_i = \frac{\partial}{\partial x_j}.$$

For $\theta < 0$, replace inf by sup in equation (1.11) and denote it as (1.11′). We put the solution to the equation J_-. In both cases, taking a transformation $e^{\theta w(t,x)} = J_\pm(t,x)$, we obtain the equation

$$\begin{cases} \mathcal{L}w + Q_0(x,\nabla w) + V(x) = 0, & [0,T)\times R^N, \\ w(T,x) = 0, \end{cases} \tag{1.12}$$

where

$$Q_0(x,p) = \frac{\theta}{2}a^{ij}p_ip_j + \inf_{z\in Z}\{c^i(x,z)p_i + \phi(x,z)\}, \quad p \in R^N. \tag{1.13}$$

Let us set

$$u(t,x) = w(T-t,x), \quad 0 \le t \le T. \tag{1.14}$$

Then we have the equation

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{1}{2}a^{ij}D_{ij}u + b^iD_iu + Q_0(x,\nabla u) + V(x), \\ u(0,x) = 0 \end{cases} \tag{1.15}$$

on $[0,T]\times R^N$. If we have the solution u to (1.15) on $[0,\infty)\times R^N$, taking $T>0$ and setting $w(t,x) = u(T-t,x)$, we obtain the solution w to (1.12), and accordingly the solution $J_\pm = e^{\theta w(t,x)}$ to (1.11) or (1.11′), respectively. Now we are going to consider the existence of the solution to (1.15). For that we further assume that

$$-\frac{k_1}{2}a^{ij}p_ip_j \le Q_0(x,p) \le -\frac{k_2}{2}a^{ij}p_ip_j, \quad \exists k_i, k_2 > 0 \tag{1.16}$$

and that $Q_0(x,p)$ is a smooth function such that

$$\left|\frac{\partial Q_0(x,p)}{\partial p}\right| \le M_1|p| + M_2, \quad \left|\frac{\partial Q_0(x,p)}{\partial x}\right| \le M_1|p|^2 + M_2 \tag{1.17}$$

for some locally bounded functions M_1 and M_2. Then we have the following theorem.

THEOREM 1.1. *Under the assumptions* (1.4)–(1.10), (1.16), *and* (1.17), *equation* (1.15) *has a nonnegative solution* $u \in C^{1+\frac{\alpha}{2},2+\alpha}((0,\infty)\times R^N)\cap C([0,\infty)\times R^N)$. *Moreover it satisfies the following estimates*:

$$\frac{\partial u}{\partial t} \ge 0, \tag{1.18}$$

$$t\left(|\nabla u|^2 + \gamma\frac{\partial u}{\partial t}\right) \le tK_{r,\gamma} + L_{r,\gamma}, \quad (0,\infty)\times B_r \tag{1.19}$$

for $\gamma > \frac{2}{k_2\nu}$, *where* $K_{r,\gamma}$ *and* $L_{r,\gamma}$ *are the constants independent of* t.

Remark 1.1. Examples satisfying assumptions (1.16) and (1.17) are illustrated in §1.5. Example 1 deals with our main concerns. More general cases where c grows like z and ϕ like $|z|^2$ for large and small z can be also covered (cf. Examples 4 and 5). However, we don't assume that V is quadratic but admit any function V growing to infinity as $|x|\to\infty$ with at most polynomial growth rate.

Remark 1.2. Assumption (1.16) indicates the bound of the size of the risk-sensitive parameter θ ensuring the existence of a solution of the Bellman equation (1.15). We will see in Theorem 2.3 that it implies the finiteness of the value function (1.2). Thus we will see that (1.16) gives a condition of θ under which no breaking down occurs. The condition is more clearly seen in the examples in §1.5.

1.2. Dirichlet problem. Let us consider the following SDE:

$$\begin{cases} dY_s^i = \sigma_j^i(Y_s)dB_s^j + b^i(Y_s)ds, \quad i = 1,\dots,N, \\ Y_0 = x \end{cases} \tag{1.20}$$

and set

$$\psi_l(t,x) = E_x[e^{-k_l\int_0^t V(Y_s)ds}], \quad l = 1,2. \tag{1.21}$$

Owing to our assumptions, ψ_l is a smooth function on $[0,\infty) \times R^N$ and satisfies the equation

$$\text{(1.22)} \qquad \begin{cases} \frac{\partial \psi_l}{\partial t} = \frac{1}{2} a^{ij} D_{ij} \psi_l + b^i D_i \psi_l - k_l V, \\ \psi_l(0,x) = 1 \end{cases}$$

for each $l = 1, 2$ (cf. [14], [16]). Put

$$u_l(t,x) = -\frac{1}{k_l} \log \psi_l(t,x), \quad l = 1,2,$$

which turns out to be the solution to the equation

$$\begin{cases} \frac{\partial u_l}{\partial t} = \frac{1}{2} a^{ij} D_{ij} u_l + b^i D_i u_l - \frac{k_l}{2} a^{ij} D_i u_l D_j u_l + V(x), \\ u_l(0,x) = 0 \end{cases}$$

for each $l = 1, 2$. Therefore we see that u_1 (resp., u_2) is a nonnegative subsolution (resp., supersolution) to the equation (1.15) because of assumption (1.16). We furthermore see that $u_1(t,x) \leq u_2(t,x)$ by using Hölder's inequality in Kac's representation (1.21). We shall find a nonnegative solution u to (1.15) such that $u_1 \leq u \leq u_2$. For that we first consider the following Dirichlet problem:

$$\text{(1.23)} \qquad \begin{cases} \frac{\partial u}{\partial t} = \frac{1}{2} a^{ij} D_{ij} u + Q(x, \nabla u) + V(x), \quad (0,T] \times B_R \equiv U_{T,R}, \\ u(t,x) = u_1(t,x), \quad \partial' U_{T,R}, \end{cases}$$

where $Q(x,p) = Q_0(x,p) + b^i p_i$, $\partial' U_{T,R} = \{(0,x); x \in \overline{B_R}\} \cup \{(t,x);\ 0 \leq t \leq T, x \in \partial B_R\}$, and $B_R = \{x \in R^N; |x| \leq R\}$. Owing to Theorem 6.1 in §6 of [17, Chap. V] we have the solution to (1.23). In fact, we consider the family of linear problems

$$\text{(1.23')} \qquad \begin{cases} Lv = (1-\tau) L u_1 + \tau(Q(x, \nabla \eta) + V), \quad x \in U_{T,R}, \\ v(t,x) = u_1(t,x), \quad x \in \partial' U_{T,R} \end{cases}$$

for $\eta \in C^{\frac{1}{2}+\frac{\alpha}{2}, 1+\alpha}$, where

$$Lv = \frac{\partial v}{\partial t} - \frac{1}{2} a^{ij} D_{ij} v.$$

These linear problems define an operator $\Phi(\eta; \tau)$ which associates each function $\eta \in C^{\frac{1}{2}+\frac{\alpha}{2}, 1+\alpha}$ with a solution v of (1.23'). The fixed points of ϕ for $\tau = 1$ are solutions of the problem (1.23). Let u^τ be one of the fixed points of the transformation $\Phi(\eta; \tau)$: $u^\tau = \Phi(u^\tau; \tau)$. Then u^τ is a solution of the nonlinear problem

$$\begin{cases} Lu^\tau = (1-\tau) L u_1 + \tau(Q(x, Du^\tau) + V), \quad x \in U_{T,R}, \\ u^\tau = u_1(t,x), \quad x \in \partial' U_{T,R}. \end{cases}$$

It can be seen that

$$\sup_{U_{T,R}} |u^\tau(t,x)| \leq K_1, \quad \tau \in [0,1]$$

by Theorem 2.9 in [17, Chap. I] and then we can obtain the estimate

$$\sup_{U_{T,R}} |\nabla u^\tau(t,x)| \leq K_2, \quad \tau \in [0,1]$$

in a way similar to the preceding discussion of Theorem 6.1 in [17, Chap. V], where K_2 is a constant depending only on K_1, ν, $\sup_{U_{T,R}} |a^{ij}(x)|$, $\sup_{B_R} |M_i(x)|$, $i = 1, 2$, k_1, k_2, $\sup_{B_R} |V(x)|$, $\sup_{B_R} |\nabla V|$, and $\sup_{\partial B_R} |\nabla u_1(x)|$. By (1.17), $Q(x,p)$ is uniformly Hölder continuous on $B_R \times \{|p| \leq K_2\}$, $\frac{\partial Q}{\partial p}$ is bounded on the set, and all other

assumptions in Theorem 6.1 in [17, Chap. 17] are satisfied. Thus we can obtain the following lemma.

LEMMA 1.2. *There exists a unique solution* $u \in C^{1+\frac{\alpha}{2},2+\alpha}(\overline{U}_{T,R})$ *to the Dirichlet problem* (1.23).

Now we give a probabilistic representation of the solution. Let $u_R(t,x)$ be the solution to (1.23) and set $J_R(t,x) = e^{\theta u(T-t,x)}$, $0 \le t \le T$. Then J_R satisfies the following equation on $[0,T) \times B_R \equiv \hat{U}_{T,R}$:

(1.24)
$$\begin{cases} \mathcal{L}J_R + \inf_{z\in Z}\{c^i(x,z)D_iJ_R + \theta(\phi(x,z)+V(x))J_R\} = 0, & (t,x) \in \hat{U}_{T,R}, \\ J_R(t,x) = e^{\theta u_1(T-t,x)}, & (t,x) \in \partial'\hat{U}_{T,R} \end{cases}$$

in the case that $\theta > 0$. If $\theta < 0$, replacing inf by sup in (1.24), we obtain the corresponding equation. By standard arguments using Ito's formula we obtain the following lemma.

LEMMA 1.3. *If* $\theta > 0$ *we have*

$$J_R(t,x) = \inf_{\tilde{\mathcal{A}}_{T-t}} E_x[e^{\theta\Phi_{(T-t)\wedge\tau}}\hat{\psi}_1(t+(T-t)\wedge\tau, X_{(T-t)\wedge\tau})], \tag{1.25}$$

where $\hat{\psi}_1(t,x) = e^{\theta u_1(T-t,x)}$, $\tau = \inf\{t; X_t \in \partial B_R\}$ *and* $\tilde{\mathcal{A}}_{T-t}$ *is the subset of* $\mathcal{A}_{T-t}$ *such that* z_s *is bounded.*

Proof. The inequality $\le$ can be seen by applying Ito's formula to $J_+(t+\cdot,\cdot)$ on the solution to the SDE (1.1). The converse inequality is proven by taking the ϵ-optimal Markovian strategy $z^\epsilon(t+\cdot,\cdot)$. It is possible since $Z \subset R^{N_1}$ is a separable metric space and c and ϕ are nice functions (cf. [16]). □

If $\theta < 0$, (1.25) holds by replacing inf by sup in the right-hand side.

Remark 1.3. From Lemma 1.3 it follows that $J_R(t,x) \ge J_R(t',x)$, $t < t'$ for $\theta > 0$ and accordingly

$$u_R(t,x) \le u_R(t',x), \quad t < t'. \tag{1.26}$$

In fact, we see that $\hat{\psi}_1(t+(T-t)\wedge\tau, X_{(T-t)\wedge\tau}) \ge \hat{\psi}_1(t'+(T-t')\wedge\tau, X_{(T-t)\wedge\tau})$ for $t < t'$. It is obvious that $e^{\theta\Phi_{(T-t)\wedge\tau}} \ge e^{\theta\Phi_{(T-t')\wedge\tau}}$ and we obtain $J_R(t,x) \ge J_R(t',x)$, $t < t'$. In the case where $\theta < 0$, $J_R(t,x) \le J_R(t',x)$, $t < t'$ holds and implies (1.26).

1.3. Gradient estimate. We first prepare a lemma on linear algebra.

LEMMA 1.4. *Let* A *be a symmetric nonnegative definite* $N \times N$*-matrix with the maximum eigenvalue* λ *and* B *a symmetric matrix. Then we have*

$$\{\mathrm{tr}(AB)\}^2 \le N\lambda\,\mathrm{tr}(AB^2).$$

Proof. By taking a orthogonal matrix T, diagonalize the matrix A

$$TAT^* = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_N \end{pmatrix} \equiv \Lambda$$

with $0 \le \lambda_1 \le \lambda_2, \dots, \lambda_N \le \lambda$. Here T^* stands for the transposed matrix of T. We then have $\mathrm{tr}(AB) = \sum_{i=1}^N \lambda_i b_{ii}$, where $b_{ij} = (TBT^*)_{ij}$. Therefore

$$\{\mathrm{tr}(AB)\}^2 = \sum_{i=1}^N \lambda_i^2 b_{ii}^2 + \sum_{i,j=1, i\neq j}^N \lambda_1\lambda_j b_{ii}b_{jj}. \tag{1.27}$$

On the other hand

$$\operatorname{tr}((AB)^2) = \sum_{i=1}^{N} \lambda_1^2 b_{ii}^2 + \sum_{i \neq j} \lambda_i \lambda_j b_{ij}^2,$$

since TBT^* is symmetric. Thus we obtain

$$\begin{aligned} N \operatorname{tr}((AB)^2) - \{\operatorname{tr}(AB)\}^2 &\geq (N-1) \sum_{i=1}^{N} \lambda_1^2 b_{ii}^2 - \sum_{i \neq j} \lambda_i \lambda_j b_{ii} b_{jj} \\ &= \frac{1}{2} \sum_{i \neq j} (\lambda_i b_{ii} - \lambda_j b_{jj})^2 \geq 0. \end{aligned}$$

Hence we conclude that

$$\{\operatorname{tr}(AB)\}^2 \leq N \operatorname{tr}((AB)^2) \leq N\lambda \sum_{i,j=1}^{N} \lambda_i b_{ij}^2 = N\lambda \operatorname{tr}(AB^2). \qquad \square$$

The following lemma plays an essential role in the following arguments. Similar kinds of estimates have been studied by Li and Yau [18] to obtain parabolic Harnack's inequalities for positive solutions to heat equations (cf. also Davies [5], Nagai [22]).

LEMMA 1.5. *Let u_R be the solution to* (1.23). *Then we have the following estimate*:

$$t\left(|\nabla u_R|^2 + \gamma \frac{\partial u_R}{\partial t}\right) \leq tK_{r,\gamma} + L_{r,\gamma}, \quad [0,T] \times B_r \tag{1.28}$$

for $r < \frac{1}{2}R$, $\gamma > \frac{2}{k_2 \nu}$, where $K_{r,\gamma}$ and $L_{r,\gamma}$ are the constants independent of t, R, and T.

Proof. Let us take a point $\alpha \in B_r$, $0 < r < \frac{1}{2}R$ and a cut-off function $\tau(x)$ such that

$$\begin{aligned} &0 \leq \tau(x) \leq \tau(\alpha) = 1, \quad a^{ij} D_i \tau D_j \tau \leq \frac{c_1 \tau}{r^2}, \quad a^{ij} D_{ij} \tau \geq -\frac{2c_2}{r^2}, \\ &\tau(x) = 0, \quad x \notin B_r(\alpha) = \{x; |x - \alpha| < r\}. \end{aligned} \tag{1.29}$$

In fact, set

$$\tau(x) = \begin{cases} (\frac{|x-\alpha|^2}{r^2} - 1)^2, & |x - \alpha| \leq r, \\ 0, & |x - \alpha| > 0. \end{cases} \tag{1.30}$$

Then it satisfies (1.29). Set

$$F(t,x) = t\left(|\nabla u|^2 + \gamma \frac{\partial u}{\partial t}\right), \quad u = u_R. \tag{1.31}$$

Then $F(t,x) \geq 0$ since $\frac{\partial u}{\partial t} \geq 0$ by (1.26). Let us take a maximum point (s,x) of the function τF in $[0,t] \times B_r(\alpha)$. Then it is obvious that $s > 0$. To see (1.28) it suffices to prove that

$$(\tau F)(s,x) \leq sK_{r,\gamma} + L_{r,\gamma} \tag{1.32}$$

for some constants $K_{r,\gamma}$ and $L_{r,\gamma}$ independent of t, T, and R. In fact, from (1.32) it follows that

$$\begin{aligned} F(t,\alpha) = \tau(\alpha) F(t,\alpha) &\leq \tau(x) F(s,x) \\ &\leq sK_{r,\gamma} + L_{r,\gamma} \leq tK_{r,\gamma} + L_{r,\gamma} \end{aligned}$$

for each $(t,\alpha) \in [0,T] \times B_r$. We note that at the point (s,x)

$$\nabla(\tau F) = 0, \quad a^{ij} D_{ij}(\tau F) \le 0, \quad \frac{\partial F}{\partial s} \ge 0. \tag{1.33}$$

Therefore we have

$$\begin{aligned} \frac{\tau F}{s} &\ge \frac{1}{2} a^{ij} D_{ij}(\tau F) + \frac{\partial Q}{\partial p_i}(x, \nabla u) D_i(\tau F) - \tau \frac{\partial F}{\partial s} + \frac{\tau F}{s} \\ &= \tau \left(\frac{1}{2} a^{ij} D_{ij} F + \frac{\partial Q}{\partial p_i}(x, \nabla u) D_i F - \frac{\partial F}{\partial s} + \frac{F}{s} \right) \\ &\quad \frac{1}{2}(a^{ij} D_{ij} \tau) F - a^{ij} \frac{D_i \tau D_j \tau}{\tau} F + \left(\frac{\partial Q}{\partial p_i}(x, \nabla u) D_i \tau \right) F. \end{aligned} \tag{1.34}$$

Here we have used the equality $\nabla F = -F \frac{\nabla \tau}{\tau}$, which follows from (1.33). Put

$$\Gamma(F) = \frac{1}{2} a^{ij} D_{ij} F + \frac{\partial Q}{\partial p_i}(x, \nabla u) D_i F - \frac{\partial F}{\partial s} + \frac{F}{s}. \tag{1.35}$$

Then from (1.29) and (1.34) it follows that

$$\frac{\tau F}{s} \ge \tau \Gamma(F) - \frac{c_2}{r^2} F - \frac{c_1}{r^2} F - \frac{c_3 \sqrt{\tau}}{r} \left| \frac{\partial Q}{\partial p}(x, \nabla u) \right| F. \tag{1.36}$$

Set $G = |\nabla u|^2$ and $H = s \frac{\partial u}{\partial s}$. Then we have

$$\Gamma(H) = 0. \tag{1.37}$$

In fact,

$$\begin{aligned} \frac{\partial H}{\partial s} &= \frac{\partial u}{\partial s} + s \frac{\partial^2 u}{\partial s^2} \\ &= \frac{H}{s} + \frac{1}{2} a^{ij} D_{ij} H + \frac{\partial Q}{\partial p_i}(x, \nabla u) D_i H. \end{aligned}$$

Thus we obtain

$$\begin{aligned} \Gamma(F) &= \frac{s}{2} a^{ij} D_{ij} G + s \frac{\partial Q}{\partial p_i}(x, \nabla u) D_i G - s \frac{\partial G}{\partial s} \\ &= s a^{ij} D_{ki} u D_{kj} u + s a^{ij} D_k D_{ijk} u \\ &\quad + 2s \frac{\partial Q}{\partial p_j}(x, \nabla u) D_i u D_{ij} u - 2s D_i u D_{is} u. \end{aligned} \tag{1.38}$$

By taking a derivative in equation (1.23) with respect to x_k, we have

$$\begin{aligned} a^{ij} D_{ijk} u &= 2 D_{ks} u - D_k a^{ij} D_{ij} u - 2 \frac{\partial Q}{\partial x_k}(x, \nabla u) \\ &\quad - 2 \frac{\partial Q}{\partial p_j}(x, \nabla u) D_{kj} u - 2 D_k V. \end{aligned} \tag{1.39}$$

Therefore

$$\begin{aligned} \Gamma(F) &= s a^{ij} D_{ki} u D_{kj} u - s D_k u D_k a^{ij} D_{ij} u \\ &\quad - 2s D_k u \frac{\partial Q}{\partial x_k}(x, \nabla u) - 2s D_k u D_k V. \end{aligned} \tag{1.40}$$

Since

$$D_k u(D_k a^{ij})D_{ij}u \le \frac{\epsilon}{2}(D_{ij}u)^2 + \frac{1}{2\epsilon}(D_k u)^2(D_k a^{ij})^2$$

for each i, j, k and $\epsilon > 0$, we obtain from (1.40)

$$\begin{aligned}\Gamma(F) \ge{}& s a^{ij} D_{ki}u D_{kj}u - \frac{Ns\epsilon}{2}\sum_{ij}(D_{ij}u)^2 \\ &- \frac{1}{2\epsilon}\sum_{ij}(D_k u)^2(D_k a^{ij})^2 - 2s|\nabla u|\left|\frac{\partial Q}{\partial x}(x,\nabla u)\right| - 2s|\nabla u||\nabla V|.\end{aligned} \tag{1.41}$$

Taking ϵ such that $\nu \ge N\epsilon$ and utilizing Lemma 1.4, we have

$$\begin{aligned}\Gamma(F) \ge{}& \frac{s}{2N\lambda_r}(a^{ij}D_{ij}u)^2 - \frac{N^2K_r}{2\epsilon}|\nabla u|^2 \\ &- 2s|\nabla u|\left|\frac{\partial Q}{\partial x}(x,\nabla u)\right| - 2s|\nabla u||\nabla V|,\end{aligned} \tag{1.42}$$

where λ_r is the upper bound of the largest eigenvalues of the matrix $a^{ij}(x)$ in B_r and K_r is the constant such that $|\nabla a^{ij}| \le K_r$. From (1.23) and (1.42) it follows that

$$\begin{aligned}\Gamma(F) \ge{}& \frac{2s}{N\lambda_r}\left(\frac{\partial u}{\partial s} - Q(x,\nabla u) - V\right)^2 \\ &- \frac{sN^2K_r}{2\epsilon}|\nabla u|^2 - 2s|\nabla u|\left|\frac{\partial Q}{\partial x}(x,\nabla u)\right| - 2s|\nabla u||\nabla V|.\end{aligned} \tag{1.43}$$

Note that from (1.16), (1.17), and (1.13) it follows that

$$-\frac{k_1'}{2}a^{ij}p_ip_j - K(x) \le Q(x,p) \le -\frac{k_2'}{2}a^{ij}p_ip_j + K(x), \tag{1.44}$$

$$\left|\frac{\partial Q}{\partial x}(x,p)\right| \le K_1|p|^2 + K_2, \tag{1.45}$$

$$\left|\frac{\partial Q}{\partial p}(x,p)\right| \le K_1|p| + K_2, \tag{1.46}$$

for some locally bounded functions K, K_1, and K_2 and some positive constants k_1' and k_2'. Since $\frac{\partial u}{\partial s} = \frac{F}{s\gamma} - \frac{G}{\gamma}$, we have

$$\begin{aligned}-\frac{k_1'}{2}a^{ij}D_iuD_ju - 2K - \frac{F}{s\gamma} + \frac{G}{\gamma} &\le Q(x,\nabla u) - \frac{\partial u}{\partial s} - K \\ &\le -\frac{k_2'}{2}a^{ij}D_iuD_ju - \frac{F}{s\gamma} + \frac{G}{\gamma}.\end{aligned}$$

Note that $\frac{k_2\nu}{2} - \frac{1}{\gamma} > 0$ and k_2' can be taken as $\frac{k_2'\nu}{2} - \frac{1}{\gamma} > 0$. Therefore, setting $G = \beta F$, we obtain

$$\left(Q - \frac{\partial u}{\partial s} - K\right)^2 \ge \left(\left(k - \frac{1}{\gamma}\right)\beta + \frac{1}{s\gamma}\right)^2 F^2, \quad k = \frac{k_2'\nu}{2}. \tag{1.47}$$

Thus from (1.36), (1.43), and (1.47) we obtain

$$\frac{\tau F}{s} \geq \frac{2s\tau}{N\lambda_r}\left\{\left(\frac{1}{s\gamma}+\left(k-\frac{1}{\gamma}\right)\beta\right)^2 F^2 - 2(K+V)\left(\frac{1}{s\gamma}+\left(k'-\frac{1}{\gamma}\right)\beta\right)F - K'\right\}$$

(1.48)

$$\begin{aligned}&- \frac{s\tau N^2 K_r}{\epsilon}|\nabla u|^2 - 2s\tau|\nabla u|\left|\frac{\partial Q}{\partial x}(x,\nabla u)\right| - 2s\tau|\nabla u||\nabla V| \\ &- \frac{c_1+c_2}{r^2}F - \frac{c_3\sqrt{\tau}}{r}\left|\frac{\partial Q}{\partial p}(x,\nabla u)\right|F,\end{aligned}$$

where $k' = \frac{k_1\lambda_r}{2}$ and $K' = 4K(K+V)$. Therefore, using (1.45) and (1.46), we have

$$\begin{aligned}\frac{\tau F}{s} &\geq \frac{2s\tau}{N\lambda_r}\left\{\left(\frac{1}{s\gamma}+\left(k-\frac{1}{\gamma}\right)\beta\right)^2 F^2 - 2(K+V)\left(\frac{1}{s\gamma}+\left(k'-\frac{1}{\gamma}\right)\beta\right)F - K'\right\} \\ &- \frac{s\tau N^2 K_r}{\epsilon}\beta F - 2s\tau K_1\beta^{\frac{3}{2}}F^{\frac{3}{2}} - 2s\tau K_2\beta^{\frac{1}{2}}F^{\frac{1}{2}} - 2s\tau|\nabla V|\beta^{\frac{1}{2}}F^{\frac{1}{2}} \\ &- \frac{c_1+c_2}{r^2}F - \frac{c_2\sqrt{\tau}}{r}K_1\beta^{\frac{1}{2}}F^{\frac{3}{2}} - \frac{c_3\sqrt{\tau}}{r}K_2 F.\end{aligned}$$

Now we can assume that $F(s,x) \geq K'(x)s$ and $F(s,x) \geq s(K_2(x)+|\nabla V|)$. In fact, otherwise (1.32) holds already. Therefore

$$\begin{aligned}\frac{F}{s} \geq \frac{\tau F}{s} &\geq \frac{2s\tau}{N\lambda_r}\left(\frac{1}{s\gamma}+\left(k-\frac{1}{\gamma}\right)\beta\right)^2 F^2 - \frac{4s\tau}{N\lambda_r}(K+V)\left(\frac{1}{s\gamma}+\left(k'-\frac{1}{\gamma}\right)\beta\right)F \\ &- \frac{2\tau F}{N\lambda_r} - \frac{s\tau N^2 K_r}{\epsilon}\beta F - 2s\tau K_1\beta^{\frac{3}{2}}F^{\frac{3}{2}} - 2\beta^{\frac{1}{2}}\tau F^{\frac{3}{2}} \\ &- \frac{c_1+c_2}{r^2}F - \frac{c_3\sqrt{\tau}}{r}K_1\beta^{\frac{1}{2}}F^{\frac{3}{2}} - \frac{c_3\sqrt{\tau}}{r}K_2 F.\end{aligned}$$

Multiplying by $\frac{s}{F}$, we obtain

$$\begin{aligned}&1 + \frac{4s^2}{N\lambda_r}(K+V)\left(\frac{1}{s\gamma}+\left(k'-\frac{1}{\gamma}\right)\beta\right) + \frac{2s}{N\lambda_r} + \frac{N^2K_r\beta}{\epsilon}s^2 \\ &+ \frac{c_1+c_2}{r^2}s + \frac{c_3K_2}{r}s \geq \frac{2s^2}{N\lambda_r}\left(\frac{1}{s\gamma}+\left(k-\frac{1}{\gamma}\right)\beta\right)^2\tau F \\ &- 2s^2K_1\beta^{\frac{3}{2}}\tau^{\frac{1}{2}}F^{\frac{1}{2}} - \left(2+\frac{c_3K_1}{r}\right)\beta^{\frac{1}{2}}s\tau^{\frac{1}{2}}F^{\frac{1}{2}}.\end{aligned} \tag{1.49}$$

Let us set $X = \tau^{\frac{1}{2}}F^{\frac{1}{2}}$ and $k\gamma - 1 = \delta$. Then the right-hand side of (1.49)

$$\begin{aligned}&= \frac{2}{N\lambda_r\gamma^2}(1+\delta\beta s)^2X^2 - \beta^{\frac{1}{2}}s(2K_1\beta s + c_r)X, \qquad c_r = 2 + \frac{c_3K_1}{r} \\ &= \frac{2}{N\lambda_r\gamma^2}(1+\delta\beta s)\left(1+\frac{\delta\beta s}{2}\right)X^2 + \frac{\delta\beta(1+\delta\beta s)}{N\lambda_r\gamma^2}X^2 - \beta^{\frac{1}{2}}s(2K_1\beta s + c_r)X \\ &\geq \frac{2}{N\lambda_r\gamma^2}(1+\delta\beta s)\left(1+\frac{\delta\beta s}{2}\right)X^2 - \frac{sN\lambda_r\gamma^2}{(1+\delta\beta s)\delta}(2K_1\beta s + c_r)^2,\end{aligned}$$

where $c_r = 2 + \frac{c_3 K_1}{r}$. Hence we obtain

$$X^2 \le \frac{N\lambda_r\gamma^2}{2(1+\delta\beta s)(1+\frac{\delta\beta s}{2})} + \frac{(N\lambda_r\gamma^2)^2}{2\delta}\frac{(2K_1\beta s + c_r)^2}{(1+\delta\beta s)(1+\frac{\delta\beta s}{2})}s$$
$$+ 2(K+V)\frac{(1+(k'\gamma-1)\beta s)}{(1+\delta\beta s)(1+\frac{\delta\beta s}{2})}s + \frac{\gamma^2 s}{(1+\delta\beta s)(1+\frac{\delta\beta s}{2})}$$
$$+ \frac{N^3\lambda_r\gamma^2 K_r}{2\epsilon}\frac{\beta s}{(1+\delta\beta s)(1+\frac{\delta\beta s}{2})}s + \frac{N\lambda_r\gamma^2(c_1+c_2+c_3K_2r)}{2r^2(1+\delta\beta s)(1+\frac{\delta\beta s}{2})}s.$$

Since $\frac{1}{(1+\delta\beta s)(1+\frac{\delta\beta s}{2})}$, $\frac{(c_r+2K_1\beta s)^2}{(1+\delta\beta s)(1+\frac{\delta\beta s}{2})}$, $\frac{1+(k'\gamma-1)\beta s}{(1+\delta\beta s)(1+\frac{\delta\beta s}{2})}$, and $\frac{\beta s}{(1+\delta\beta s)(1+\frac{\delta\beta s}{2})}$ are dominated by a constant independent of β and s, we conclude that

$$X^2 \le sK_{r,\gamma} + L_{r,\gamma}$$

for some constants $L_{r,\gamma}$ and $L_{r,\gamma}$. □

1.4. Proof of Theorem 1.1. According to [17] we formulate a generalized solution to the equation

$$\frac{\partial u}{\partial t} = \frac{1}{2}a^{ij}D_{ij}u + Q(x,\nabla u) + V(x), \quad (0,T)\times R^N. \tag{1.50}$$

Take ϵ and T_1 such that $0 < \epsilon < T_1 < T$ and $r > 0$ and define $W_2^{1,1}(U^\epsilon_{T_1,r})$ as the Hilbert space with an inner product

$$(u,v) = \iint_{U^\epsilon_{T_1,r}} \left(uv + \nabla u\cdot\nabla v + \frac{\partial u}{\partial t}\frac{\partial v}{\partial t}\right) dxdt, \tag{1.51}$$

where $U^\epsilon_{T_1,r} = (\epsilon, T_1)\times B_r$. Let us set

$$V_2^{1,0} = L^2(\epsilon, T_1; H^1(B_r)) \cap C([\epsilon, T_1]; L^2(B_r)).$$

The norm of the Banach space $V_2^{1,0}$ is defined by

$$\|v\| = \max_{\epsilon\le t\le T_1}\|v(t,x)\|_{2,B_r} + \|\nabla v\|_{2,U^\epsilon_{T_1,\epsilon}},$$

where $\|v\|_{2,B} = \int_B |v(t,x)|^2 dx$. We say that u is a generalized solution to (1.50) if, for each ϵ, T_1, and r, u belongs to $V_2^{1,0}(U^\epsilon_{T_1,r})$ and satisfies

$$\iint_{U^\epsilon_{T_1,r}} u\frac{\partial\eta}{\partial t}dxdt - \frac{1}{2}\iint_{U^\epsilon_{T_1,r}} a^{ij}D_iuD_j\eta dxdt - \frac{1}{2}\iint_{U^\epsilon_{T_1,r}}\frac{\partial a^{ij}}{\partial x^j}D_iu\eta dxdt$$
$$+ \iint_{U^\epsilon_{T_1,r}} Q(x,\nabla u)\eta dxdt + \iint_{U^\epsilon_{T_1,r}} V(x)\eta dxdt = 0 \tag{1.52}$$

for each $\eta \in \overset{\circ}{W}{}_2^{1,1}(U^\epsilon_{T_1,r})$ such that $\eta(\epsilon,x) = \eta(T_1,x) = 0$. Now we have the following lemma.

LEMMA 1.6. *Let $u_R = u(t,x;T,R)$ be a solution to (1.23). Then for each $r > 0$, ϵ, and T_1 such that $0 < \epsilon < T_1 < T$, there exists $\{R_n\} \subset R_+$ such that u_{R_n} converges weakly in $W_2^{1,1}(U^\epsilon_{T_1,r})$ to $u \in W_2^{1,1}(U^\epsilon_{T_1,r})$, which satisfies (1.52).*

Proof. Let u_R be a solution to (1.23) and take $r' > 0$, ϵ' and T_1' such that $r' > r$, $0 < \epsilon' < \epsilon < T_1 < T_1' < T$. Then by Lemma 1.5 and (1.26) we see that $\{u_R\}_{R>2r'}$ is a bounded subset of the Hilbert space $W_2^{1,1}(U^{\epsilon'}_{T_1',r'})$ and has a subsequence

$\{u_{R_n}\}$ converging weakly in $W_2^{1,1}(U^{\epsilon'}_{T_1',r'})$ and strongly in $L^2(U^{\epsilon'}_{T_1',r'})$ to some function $u \in W_2^{1,1}(U^{\epsilon'}_{T_1',r'})$. To see that this u satisfies (1.52) we shall prove that ∇u_{R_n} converges to ∇u strongly in $L^2(U^{\epsilon}_{T_1,r})$. Take a function $s(t,x) \in C_0^\infty(U^{\epsilon'}_{T_1',r'})$ such that $s(t,x) = 1$ on $U^{\epsilon}_{T_1,r}$. Then, by taking $s(u_n - u)$ as a test function, we have

$$
\begin{aligned}
0 = &\iint_{U^{\epsilon'}_{T_1',r'}} u_n \frac{\partial s(u_n-u)}{\partial t} dxdt - \frac{1}{2}\iint_{U^{\epsilon'}_{T_1',r'}} a^{ij} D_i u_n D_j (s(u_n-u))dxdt \\
&- \frac{1}{2}\iint_{U^{\epsilon'}_{T_1',r'}} \frac{\partial a^{ij}}{\partial x^j} D_i u_n s(u_n-u)dxdt + \iint_{U^{\epsilon'}_{T_1',r'}} Q(x,\nabla u_n)s(u_n-u)dxdt \\
&+ \iint_{U^{\epsilon'}_{T_1',r'}} V(x)s(u_n-u)dxdt,
\end{aligned} \tag{1.53}
$$

where $u_n = u_{R_n}$. Therefore we deduce that

$$
\begin{aligned}
&\frac{1}{2}\iint_{U^{\epsilon'}_{T_1',r'}} a^{ij} D_i(u_n-u) D_j(u_n-u) s dxdt \\
&= \iint_{U^{\epsilon'}_{T_1',r'}} \frac{\partial u_n}{\partial t} s(u_n-u)dxdt - \frac{1}{2}\iint_{U^{\epsilon'}_{T_1',r'}} a^{ij} D_i u D_j(u_n-u) s dxdt \\
&- \iint_{U^{\epsilon'}_{T_1',r'}} a^{ij} D_i u_n D_j s (u_n-u)dxdt - \frac{1}{2}\iint_{U^{\epsilon'}_{T_1',r'}} \frac{\partial a^{ij}}{\partial x^j} D_i u_n (u_n-u) s dxdt \\
&+ \iint_{U^{\epsilon'}_{T_1',r'}} Q(x,\nabla u_n)s(u_n-u)dxdt + \iint_{U^{\epsilon'}_{T_1',r'}} V(x)s(u_n-u)dxdt.
\end{aligned} \tag{1.54}
$$

It is easy to see that the right-hand side of (1.54) converges to 0 as $n \to \infty$, by Lemma 1.5, (1.26), and (1.44). Thus, by using uniform ellipticity of a^{ij}, we can prove that ∇u_n converges to ∇u strongly in $L^2(U^{\epsilon}_{T_1,r})$ and also almost everywhere (a.e.) by taking a subsequence, if necessary. Hence we see that u satisfies (1.52). □

Because of Lemma 1.6 we see that there exists a generalized nonnegative solution $u \in W^{1,1}_{2,loc}((0,T)\times R^N)$. Moreover regularity theorems for parabolic equations imply that $u \in C^{1+\frac{\alpha}{2},2+\alpha}((0,T)\times R^N)$, since Q satisfies (1.44) and $|\nabla u| \in L^\infty_{loc}$ (cf. Theorem 12.1 and Theorem 12.2 in §12 of [17, Chap. III]). We furthermore note that the solution u satisfies (1.18) and (1.28) for each r.

Now we prepare two lemmas to see that u is continuous on $[0,T)\times R^N$ and that $u(0,x) = 0$.

LEMMA 1.7. *Let u_1 and u_2 be functions defined by* (1.21) *and u_R the solution to* (1.23). *Then we have*

$$
u_1(t,x) \le u_R(t,x) \le u_2(t,x). \tag{1.55}
$$

Proof. These inequalities follow from the maximum principle. In fact, let us assume that $u_1(t_0,x_0) > u_R(t_0,x_0)$ for some point $(t_0,x_0) \in \overline{U}_{T,R}$. Set

$$
\rho(t,x) = e^{-\lambda t}(u_1(t,x) - u_R(t,x)), \quad \lambda > 0
$$

and take a maximum point (t_1, x_1) of $\rho(t,x)$ in $\overline{U}_{T,R}$

$$\rho(t_1,x_1) = \sup_{\overline{U}_{T,R}} \rho(t,x) > 0. \tag{1.56}$$

Then it is obvious that $t_1 > 0$ and that $x_1 \in B_R$ because $u_1 = u_R$ on $\partial' U_{T,R}$. At (t_1, x_1) we have

$$\frac{\partial \rho}{\partial t} \geq 0, \quad \nabla \rho = 0, \quad a^{ij} D_{ij} \rho \leq 0. \tag{1.57}$$

If we set $\bar{u}_1 = e^{-\lambda t} u_1$ and $\bar{u}_R = e^{-\lambda t}$ then we have

$$\frac{\partial \bar{u}_R}{\partial t} = \frac{1}{2} a^{ij} D_{ij} \bar{u}_R + e^{-\lambda t} Q(x, e^{\lambda t} \nabla \bar{u}_R) + e^{-\lambda t} V - \lambda \bar{u}_R$$

and

$$\frac{\partial \bar{u}_1}{\partial t} \leq \frac{1}{2} a^{ij} D_{ij} \bar{u}_1 + e^{-\lambda t} Q(x, e^{\lambda t} \nabla \bar{u}_1) + e^{-\lambda t} V - \lambda u_1.$$

$D\bar{u}_1(t_1,x_1) = D\bar{u}_R(t_1,x_1)$ by (1.57) and $Q(x_1, e^{\lambda t}\nabla \bar{u}_R) = Q(x_1, e^{\lambda t}\nabla \bar{u}_1)$. Therefore we obtain

$$\frac{\partial}{\partial t}(\bar{u}_1 - \bar{u}_R) \leq \frac{1}{2} a^{ij} D_{ij}(\bar{u}_1 - \bar{u}_R) - \lambda(\bar{u}_1 - \bar{u}_R).$$

Hence it follows from (1.57) that $0 \leq -\lambda \rho(t_1, x_1)$, which contradicts (1.56). The other inequality is proven in a similar way since $u_R(t,x) \leq u_2(t,x)$ on $\partial' U_{T,R}$. □

LEMMA 1.8. *Let u_R be a solution to* (1.23). *Then $\{u_R\}_{R>3r}$ are equicontinuous on $[0,T) \times B_r$ for each r.*

Proof. For each $\epsilon > 0$ there exists $\delta > 0$ such that $0 \leq u_2(t,x) < \epsilon$, $t < 2\delta$, $x \in B_r$. Take (t_1, x_1) and (t_2, x_2) such that $|x_1 - x_2| + |t_1 - t_2|^{\frac{1}{2}} < \delta_1$, $(t_i, x_i) \in U_{T,r}$, $i = 1, 2$, where $\delta_1 = \delta \wedge \epsilon^2$. In the case where $t_1 < \frac{\delta}{2}$ or $t_2 < \frac{\delta}{2}$ we have

$$|u_R(t_1,x_1) - u_R(t_2,x_2)| \leq 2\epsilon,$$

because $|t_1 - t_2|^{\frac{1}{2}} < \delta_1$ implies $t_1, t_2 < 2\delta$. If $t_1, t_2 > \frac{\delta}{2}$ it follows from Lemma 1.5 that

$$|u_R(t_1,x_1) - u_R(t_2,x_2)| \leq |u_R(t_1,x_1) - u_R(t_1,x_2)| + |u(t_1,x_2) - u(t_2,x_2)| \leq c_r \epsilon, \tag{1.58}$$

where c_r is a constant independent of R. In fact, by Lemma 1.5 and (1.26) we have

$$|\nabla u_R(t,x)|^2, \left| \frac{\partial u_R}{\partial t}(t,x) \right| \leq \frac{c_r'}{\delta}, \quad t \geq \frac{\delta}{2}, \, x \in B_r$$

and the mean value theorem implies (1.58). Thus we conclude this lemma. □

Let u_R be a solution to (1.23). Then because of Lemmas 1.7 and 1.8 there exists a subsequence $\{u_{R_n}\} \subset \{u_R\}$ such that u_{R_n} converges uniformly to a solution to (1.50) on each compact subset of $[0,T) \times R^N$ and we see that $u \in C([0,T) \times R^N)$ and $u(0,x) = 0$. Thus we find a nonnegative solution u to (1.15) on $[0,T) \times R^N$. The estimate (1.28) is independent of T and we can obtain a nonnegative solution u to (1.15) on $[0,\infty) \times R^N$ in a similar way as we did above. The solution also satisfies $\frac{\partial u}{\partial t} \geq 0$ and (1.28) on $(0,\infty) \times B_r$ for each r. Hence we complete the proof of Theorem 1.1.

1.5. Examples.

Example 1. Let $Z = R^N$, $c^i(x,z) = B^i_k(x) z^k$, and $\phi(x,z) = \frac{1}{2} S_{ij}(x) z^i z^j$, where $(S_{ij}(x))$ is a symmetric matrix such that $S_{ij}(x)\xi^i\xi^j \geq \mu|\xi|^2$, $\forall \xi \in R^N$, $\mu > 0$, and

each component is a smooth function. Then

$$
\begin{aligned}
Q_0(x,p) &= \frac{\theta}{2}a^{ij}p_ip_j + \inf_{z\in R^N}\left\{B_k^i z^k p_i + \frac{1}{2}S_{ij}z^i z^j\right\} \\
&= \frac{\theta}{2}a^{ij}p_ip_j - \frac{1}{2}(BS^{-1}B^*)p_ip_j.
\end{aligned} \tag{1.59}
$$

We assume that $B_k^i(x)$ is a bounded function for each i and k and

$$
k_2 a^{ij}p_ip_j \le ((BS^{-1}B^*) - \theta a)^{ij}p_ip_j, \quad \forall p \in R^N \tag{1.60}
$$

for some positive constant k_2. Then (1.16) and (1.17) are satisfied. Let $\lambda(x)$ be the smallest eigenvalue of the matrix $BS^{-1}B^*$ and $\bar{\nu}(x)$ the largest one of a. If

$$
\theta < \inf_x \frac{\lambda(x)}{\bar{\nu}(x)}, \tag{1.61}
$$

then (1.60) is satisfied. We moreover assume that (σ^{ij}), (b^i), and V satisfy the assumptions of Theorem 1.1. Then Theorem 1.1 applies to this example and includes the case of LEQG.

Example 2. Let us specialize the above example. We moreover assume that $BS^{-1}B^* = ka$ for some positive constant k such that

$$
\theta < k.
$$

It is then obvious that this example satisfies all assumptions of Theorem 1.1. In this case equation (1.15) is reduced to

$$
\begin{cases}
\frac{\partial u}{\partial t} = \frac{1}{2}a^{ij}D_{ij}u + b^iD_iu - \frac{k-\theta}{2}a^{ij}D_iuD_ju + V, \\
u(0,x) = 0.
\end{cases} \tag{1.62}
$$

Let Y_t be a solution to the SDE (1.20) and set

$$
u(t,x) = -\frac{1}{k-\theta}\log E[e^{-\int_0^t (k-\theta)V(Y_s)ds}]. \tag{1.63}
$$

Then $u(t,x)$ defined by (1.63) turns out to be a unique nonnegative solution to (1.62). The uniqueness follows from that of the bounded solution to the linear equation (1.22).

Example 3. We specialize Example 1 to the LEQG case, namely the case where $b^i(x) = A_j^i x^j$, $V(x) = \frac{1}{2}R_{ij}x^ix^j$, and σ_j^i, A_j^i, B_k^i, S_{ij}, and R_{ij} are constant matrices. Then (1.12) reads

$$
\begin{cases}
\frac{\partial w}{\partial t} + \frac{1}{2}a^{ij}D_{ij}w + A_j^i x^j D_i w + \left\{\frac{\theta}{2}a^{ij} - \frac{(BS^{-1}B^*)^{ij}}{2}\right\}D_iwD_jw \\
\qquad + \frac{1}{2}R_{ij}x^ix^j = 0, \\
w(T,x) = 0,
\end{cases}
$$

which has the following nonnegative solution w:

$$
w(t,x) = \frac{1}{2}P_{ij}(t)x^ix^j + G(t),
$$

where $P(t)$ and $G(t)$ satisfy the ordinary differential equations

$$
\begin{cases}
\frac{dP}{dt} + R + PA + A^*P - P(BS^{-1}B^* - \theta a)P = 0, \\
P(T) = 0
\end{cases}
$$

and

$$\begin{cases} \frac{dG}{dt} + \mathrm{tr}(Pa) = 0, \\ G(T) = 0. \end{cases}$$

Example 4. Let $Z = R^1$, $c(x,z) = z$, $\phi(x,z) = \phi(z) = c_1 z^2 + c_2 z \arctan z$, $c_1 > c_2 > 0$, and $\nu_1 \le a(x) \le \nu_2$. Then $Q_0(x,p) = \frac{\theta}{2} a p^2 + Q_1(p)$, where

$$Q_1(p) = \inf_{z \in R^1} \{zp + c_1 z^2 + c_2 z \arctan z\}.$$

Note that $\phi(z)$ is convex and $0 < 2c_1 \le \frac{d^2\phi(z)}{dz^2} \le 2(c_1 + c_2)$. Let $q(p)$ be an inverse function of $\phi'(z)$. Then we have $\frac{1}{2(c_1+c_2)} \le q'(p) \le \frac{1}{2c_1}$. We can see that

$$Q_1(p) = pq(-p) + c_1 q(-p)^2 + c_2 q(-p) \arctan q(-p)$$

and that it satisfies

$$-\frac{c_1 + 2c_2}{4c_1(c_1 + c_2)} p^2 \le Q_1(p) \le -\frac{c_1 - c_2}{4c_1(c_1 + c_2)} p^2.$$

Therefore it is easy to see that if

$$\theta < \frac{c_1 - c_2}{2c_1(c_1 + c_2)\nu_2},$$

then (1.16) and (1.17) are satisfied. We moreover assume that σ, b, and V satisfy the other assumptions of Theorem 1.1, and so it applies.

Example 5. Let $Z = R^1$, $c(x,z) = z + c_1 \arctan z$, $\phi(x,z) = c_2 z^2$, $0 < c_1 < 1$, $c_2 > 0$, and $\nu_1 \le a(x) \le \nu_2$. Then we set

$$Q_1(p) = \inf_{z \in R^1} \{zp + c_1 \arctan z + c_2 z^2\} \equiv \inf_z s(z;p).$$

Note that $s'(z) = (\frac{1+c_1+z^2}{1+z^2})p + 2c_2 z$ and that $g(z) = \frac{2c_2 z(1+z^2)}{1+c_1+z^2}$ is monotone increasing and $0 < \frac{2c_2}{1+c_1} \le g'(z) \le 2c_2 + \frac{c_1 c_2}{4(1+c_1)}$. Let $q(p)$ be an inverse function of $g(z)$. Then

$$\frac{4(1+c_1)}{8c_2 + 9c_1 c_2} \le q'(p) \le \frac{1+c_1}{2c_2}.$$

Therefore we can see that similarly,

$$Q_1(p) = pq(-p) + c_1 p \arctan q(-p) + c_2 q(-p)^2$$

and

$$-\beta_1 p^2 \le Q_1(p) \le -\beta_2 p^2$$

for some $\beta_1, \beta_2 > 0$. Thus we see that if

$$\theta < \frac{2\beta_2}{\nu_2},$$

then (1.16) and (1.17) are satisfied.

2. No breaking down. In the present section we always assume the assumptions of Theorem 1.1 and shall show that no breaking down occurs in the stochastic control problem (1.1) under the assumptions, namely, $J_+^*(t,x;T)$ has finite value for each t, T, and x.

2.1. ϵ-optimal diffusion processes. The following lemma is useful for studying ϵ-optimal diffusion processes. The proof is a modification into a parabolic case of that of Theorem 3.2 in [2].

LEMMA 2.1. *Let $u \in C^{1,2}((0,T) \times R^N)$ be a supersolution to (1.15) bounded below:*

$$\frac{\partial u}{\partial t} \geq \frac{1}{2}a^{ij}D_{ij}u + Q(x, Du) + V, \quad (0,T] \times R^N. \tag{2.1}$$

Then $\lim_{|x|\to\infty} u(t,x) = \infty$, $0 < t \leq T$.

Proof. We can assume that u is nonnegative. In fact, if $u(t,x)$ satisfies (2.1), so does $u(t,x) - c$ for each constant c. By assumption (1.16) we have

$$\frac{\partial u}{\partial t} - \frac{1}{2}a^{ij}D_{ij}u - b^iD_iu + \frac{k_1}{2}a^{ij}D_iuD_ju - V \geq 0. \tag{2.2}$$

Let us take a point x_ρ such that $|x_\rho| = \rho$, $\rho > 0$ and define a function R by

$$R(t,x) = c_\rho\left(1 - \frac{4|x - x_\rho|^2}{\rho^2} - \frac{T-t}{T}\right), \quad 0 \leq t \leq T,\ |x - x_\rho| \leq \frac{\rho}{2}, \tag{2.3}$$

where c_ρ is a constant prescribed later. Let us set $z(t,x) = u(t,x) - R(t,x)$, $D_{T,\rho} = (0,T] \times \{x; |x - x_\rho| \leq \frac{\rho}{2}\}$ and

$$\partial' D_{T,\rho} = \left\{(0,x); |x - x_\rho| \leq \frac{\rho}{2}\right\} \cup \left\{(t,x); |x - x_\rho| = \frac{\rho}{2},\ 0 \leq t \leq T\right\}.$$

Then $z(t,x) \geq u(t,x) \geq 0$ on $\partial' D_{T,\rho}$. On the other hand, by (2.2) and (2.3) we have

$$\begin{aligned}
&\frac{\partial z}{\partial t} - \frac{1}{2}a^{ij}D_{ij}z - b^iD_iz \\
&\quad \geq -\frac{k_1}{2}a^{ij}D_iuD_ju + V - \left(\frac{\partial R}{\partial t} - \frac{1}{2}a^{ij}D_{ij}R - b^iD_iR\right) \\
&\quad = -\frac{k_1}{2}a^{ij}D_i(u+R)D_j(u-R) - \frac{k_1}{2}a^{ij}D_iuD_ju + \frac{1}{2}a^{ij}D_{ij}R \\
&\quad + b^iD_iR - \frac{\partial R}{\partial t} + V
\end{aligned}$$

in $D_{T,\rho}$. Therefore

$$\begin{aligned}
\frac{\partial z}{\partial t} - \frac{1}{2}a^{ij}D_{ij}z - b^iD_iz + \frac{k_1}{2}a^{ij}D_i(u+R)D_jz \\
\geq -\frac{k_1}{2}a^{ij}D_iRD_jR + \frac{1}{2}a^{ij}D_{ij}R + b^iD_iR - \frac{c_\rho}{T} + V \\
\geq -Kc_\rho^2 - Kc_\rho - \frac{c_\rho}{T} + V
\end{aligned} \tag{2.4}$$

for some constant $K > 0$ because σ and b are at most linear growth. Since $V(x) \to \infty$ as $|x| \to \infty$ we can take c_ρ such that $c_\rho \to \infty$ as $\rho \to \infty$ and

$$-Kc_\rho^2 - Kc_\rho - \frac{c_\rho}{T} + V(x) > 0, \quad |x - x_\rho| \leq \frac{\rho}{2} \tag{2.5}$$

holds for sufficiently large ρ. Then we have $z(t,x) \geq \inf_{\partial' D_{T,\rho}} z(s,y) \geq 0$ in $D_{T,\rho}$ by the maximum principle. In particular, at $(t,x) = (t,x_\rho)$ we have

$$z(t,x_\rho) = u(t,x_\rho) - c_\rho\frac{t}{T} \geq 0. \tag{2.6}$$

Hence we see that $u(t, x_\rho) \to \infty$ as $\rho \to \infty$ for each $t > 0$. □

Let u be a nonnegative solution to (1.15) on $[0, \infty) \times R^N$, the existence of which is assured by Theorem 1.1. We take $T > 0$ and define $w(t, x) = u(T - t, x)$, $0 \le t \le T$. Then w satisfies (1.12). For each $\epsilon > 0$ there exists a Borel function $z^\epsilon(t, x)$ (cf. Krylov [16]) such that

$$\inf_{z \in Z} \{c^i(x, z) D_i w(t, x) + \phi(x, z)\} \ge c^i(x, z^\epsilon(t, x)) D_i w(t, x) + \phi(x, z^\epsilon(t, x)) - \epsilon, \tag{2.7}$$

since Z is a separable metric space and c, ∇w, and ϕ are nice functions. We further note that $z^\epsilon(t, x)$ becomes locally bounded because of our assumption (1.9). Let us consider the following SDE:

$$\begin{cases} dX_s = \sigma(X_s) dB_s + b(X_s) ds + c(X_s, z^\epsilon(t + s, X_s)) ds, & 0 \le s \le T - t, \\ X_0 = x. \end{cases} \tag{2.8}$$

Since $c(x, z^\epsilon(t + s))$ is locally bounded and σ and b satisfy the conditions in Theorem 1.1, we can see that SDE (2.8) has a unique solution if no explosion occurs. Namely, we consider the martingale problem for the SDE

$$\begin{cases} dX_s = \sigma_n(X_s) dB_s + b_n(X_s) ds + c_n(s, X_s) ds, & 0 \le s \le T - t, \\ X_0 = x, \end{cases} \tag{2.9}$$

with a_n, b_n, and c_n defined as follows: $a_n = \chi_n + (1 - \chi_n) I$, $b_n = \chi_n$, and $c_n = \chi_n \tilde{c}$, where $\tilde{c}(s, x) = c(x, z^\epsilon(t + s, x))$ and χ_n is a smooth function such that

$$\chi_n = \begin{cases} 1 & B_n, \\ 0 & B_{n+1}^c. \end{cases}$$

Let (P_x^n, X_s) be a solution to the martingale problem. If we have

$$\lim_{n \to \infty} P_x^n \left(\sup_{0 \le s \le T_0 - t} |X_s| > n \right) = 0, \quad \forall T_0 < T, \tag{2.10}$$

then (2.8) has a unique solution for $0 \le s < T - t$ (cf. Stroock and Varadhan [24]).

LEMMA 2.2. *Equation* (2.10) *holds.*

Proof. We first note that

$$\lim_{|x| \to \infty} \inf_{0 \le s \le T_0} w(s, x) = \infty, \quad \forall T_0 < T$$

holds because of Lemma 2.1 and (1.55). Let us set

$$\mathcal{G}(z) w = \frac{\partial w}{\partial t} + \frac{1}{2} a^{ij}(x) D_{ij} w + b^i(x) D_i w + c^i(x, z) D_i w \tag{2.11}$$

and

$$\mathcal{B} w = \inf_{z \in Z} \{\mathcal{G}(z) w + \phi(x, z)\}. \tag{2.12}$$

We then obtain by Ito's formula

$$\begin{aligned} w(t + s \wedge \zeta_n, X_{s \wedge \zeta_n}) - w(t, X_0) &= \int_0^{s \wedge \zeta_n} \mathcal{G}(z^\epsilon(t + \tau, X_\tau)) w(t + \tau, X_\tau) d\tau \\ &+ \int_0^{s \wedge \zeta_n} \sigma_j^i(X_\tau) D_i w(t + \tau, X_\tau) dB_\tau^j, \end{aligned}$$

where $\zeta_n = \inf\{s; |X_s| \geq n\}$. Therefore we have

$$\begin{aligned}
&E_x^n[w(t+(T_0-t)\wedge\zeta_n, X_{(T_0-t)\wedge\zeta_n})] \\
&\quad \leq w(t,x) + E_x^n\left[\int_0^{(T_0-t)\wedge\zeta_n} (\mathcal{B}w(t+\tau, X_\tau) + \epsilon)d\tau\right] \\
&\quad - E_x^n\left[\int_0^{(T_0-t)\wedge\zeta_n} \phi(X_\tau, z^\epsilon(t+\tau, X_\tau))d\tau\right] \\
&\quad \leq w(t,x) - E_x^n\left[\int_0^{(T_0-t)\wedge\zeta_n} \{V(X_\tau) + \frac{\theta}{2}a^{ij}D_iwD_jw(t+\tau, X_\tau)\}d\tau\right] \\
&\quad + \epsilon(T_0-t) - E_x^n\left[\int_0^{(T_0-t)\wedge\zeta_n} \phi(X_\tau, z^\epsilon(t+\tau, X_\tau)d\tau\right] \\
&\quad \leq w(t,x) + \epsilon(T_0-t).
\end{aligned}$$

Put $\Lambda_n = \inf\{w(t+s,x);\ 0 \leq s \leq T_0 - t, |x| = n\}$. Then $\lim_{n\to\infty}\Lambda_n = \infty$. Hence (2.10) follows from the inequalities

$$\begin{aligned}
\Lambda_n P_x^n(\zeta_n \leq T_0 - t) &\leq E_x^n[w(t+(T_0-t)\wedge\zeta_n, X_{(T_0-t)\wedge\zeta_n})] \\
&\leq w(t,x) + \epsilon(T_0-t). \qquad \square
\end{aligned}$$

Remark 2.1. We don't assume controlled processes have any stability. But we see that the above-defined ϵ-optimal diffusion process has some kind of stability. In fact, it is defined up to time T. For large T it behaves like a recurrent diffusion process with finite recurrence time since $w(t,x) \geq 0$ and

$$\frac{\partial w}{\partial t} + a^{ij}D_{ij}w + b^iD_iw + c^i(x, z^\epsilon(t,x))D_iw \leq -V(x) + \epsilon \leq -1$$

for $0 \leq t < T$, $|x| \geq R$ for sufficiently large R (cf. Theorem 7.1 in [13, p. 98]).

2.2. Finiteness of the value function. The following theorem implies the finiteness of the value function (1.2) in a risk-averse case under the assumptions of Theorem 1.1. We note the bound of θ ensuring the finiteness is determined by assumption (1.16) (cf. Remark 1.2).

THEOREM 2.3. *Let u be a nonnegative solution to* (1.15) *on* $[0,\infty)\times R^N$, *the existence of which has been shown in Theorem* 1.1. *For* $T > 0$, *set* $w(t,x) = u(T-t,x)$, $0 \leq t \leq T$, *and* $J_+(t,x;T) = e^{\theta w(t,x)}$. *Then*

$$J_+^*(t,x;T) \leq J_+(t,x;T), \quad 0 \leq t \leq T, \tag{2.13}$$

where $J_+^*(t,x;T)$ *is the value function defined by* (1.2).

Proof. For each $\epsilon > 0$ take a Borel function $z^\epsilon(t,x)$ satisfying (2.7) and consider the SDE (2.8). Then, in the same way as in the proof of Lemma 2.2, we have

$$\begin{aligned}
&w(t+s, X_s) - w(t, X_0) \\
&\qquad \leq -\int_0^s \{V(X_\tau) + \phi(X_\tau, z_\tau^\epsilon)\}d\tau + \int_0^s \sigma_j^i(X_\tau)D_iw(t+\tau, X_\tau)dB_\tau^j \\
&\qquad - \frac{\theta}{2}\int_0^s a^{ij}D_iwD_jw(t+\tau, X_\tau)d\tau + \epsilon s, \quad 0 \leq s < T-t,
\end{aligned}$$

where $z_\tau^\epsilon = z^\epsilon(t+\tau, X_\tau)$. Therefore

$$e^{\theta\int_0^s\{V(X_\tau)+\phi(X_\tau,z_\tau^\epsilon)\}d\tau+\theta w(t+s,X_s)} \le e^{\theta(\epsilon s+w(t,x))}e^{M_s-\frac{1}{2}\langle M\rangle_s},$$

where $M_s = \theta\int_0^s \sigma_j^i(X_\tau)D_i w(t+\tau, X_\tau)dB_\tau^j$. Since M_s is a local martingale and $M_0 = 0$, we have

$$E_x[e^{M_s-\frac{1}{2}\langle M\rangle_s}] \le 1.$$

Thus we obtain

$$E_x[e^{\theta\int_0^s\{V(X_\tau)+\phi(X_\tau,z_\tau^\epsilon)\}d\tau}] \le e^{\theta(\epsilon s+w(t,x))}, \quad 0 \le s < T-t,$$

since $w \ge 0$. Hence we complete the proof of our theorem as $s \to T-t$ and $\epsilon \to 0$. □

Let us consider Example 1 in §1.5. We consider the nonnegative solution u to (1.15) with (1.59) and set w as above. Let us define a feedback control

$$z(t+s,x) = -S^{-1}B^*\nabla w(t+s,x), \quad 0 \le s < T-t \tag{2.14}$$

and consider the following SDE:

$$\begin{cases} dX_s = \sigma(X_s)dB_s + b(X_s)ds + B(X_s)z(t+s,X_s)ds, & 0 \le s < T-t, \\ X_0 = x. \end{cases} \tag{2.15}$$

We then have the following proposition.

PROPOSITION 2.4. *Put $J_+(t,x) = e^{\theta w(t,x)}$. Then we have*

$$J_+(t,x) = E_x[e^{\theta\int_0^{T-t}\{V(X_s)+\frac{1}{2}S_{ij}(X_s)z_s^i z_s^j\}ds}], \tag{2.16}$$

where $z_s^i = z^i(t+s, X_s)$.

Proof. We first note that the infimum in (1.59) is attained by $z = -S^{-1}B^*p$. By Ito's formula, we then obtain in a similar way as above

$$\begin{aligned} &e^{\theta\int_0^s\{V(X_\tau)+\frac{1}{2}S_{ij}(X_\tau)z_\tau^i z_\tau^j\}d\tau+\theta w(t+s,X_s)} \\ &= e^{\theta w(t,x)}e^{\theta M_s-\frac{\theta^2}{2}\langle M\rangle_s}, \end{aligned} \tag{2.17}$$

where $M_s = \int_0^s \sigma_j^i(X_\tau)D_i w(X_\tau)dB_\tau^j$. Therefore we have

$$E_x[e^{\theta\int_0^s\{V(X_\tau)+\frac{1}{2}S_{ij}(X_\tau)z_\tau^i z_\tau^j\}d\tau}] \le e^{\theta w(t,x)}, \quad s < T-t. \tag{2.18}$$

Thus we see by Fatou's lemma that

$$E_x[e^{\theta\int_0^{T-t}\frac{1}{2}S_{ij}(X_s)z_s^i z_s^j ds}] < \infty.$$

Because of (1.60), $\theta a < BS^{-1}B^*$ as quadratic forms and

$$S_{ij}z_s^i z_s^j = (BS^{-1}B^*)^{ij}D_i w D_j w(t+s, X_s).$$

Therefore

$$E_x[e^{\frac{\theta^2}{2}\langle M\rangle_{T-t}}] < \infty. \tag{2.19}$$

Owing to Novikov's theorem (cf. Theorem 6.1 in [20]), (2.19) implies that

$$L_s = e^{\theta M_s-\frac{\theta^2}{2}\langle M\rangle_s}, \quad 0 \le s \le T-t$$

is a martingale and we have $E_x[L_{T-t}] = 1$. From (2.17) and (2.18) it follows that $\lim_{s\to T-t} e^{\theta w(t+s,X_s)}$ exists and that the limit is nothing but one because of (1.8). Hence (2.17) implies (2.16). □

3. Asymptotics.

3.1. General view of asymptotics. Let us study the asymptotic behaviour as $T \to \infty$ of a nonnegative solution to (1.15) on $[0,\infty) \times R^N$, the existence of which has been proven in Theorem 1.1. We first show the following lemma.

LEMMA 3.1. *Let us assume that the conditions of Theorem* 1.1 *and* u *are a nonnegative solution to* (1.15) *on* $[0,\infty)\times R^N$. *Then there exists a subsequence* $\{T_i\} \subset R_+$ *such that* $u(T_i,x) - u(T_i,0)$ *converges to a function* $v \in C^2(R^N)$ *uniformly on each compact set and strongly in* $W^1_{2,loc}$ *and* $\frac{\partial u}{\partial t}(T_i,x)$ *to* $\chi(x) \in C(R^N)$ *uniformly on each compact set. Moreover* $(v(x),\chi(x))$ *satisfies*

$$\chi(x) = \frac{1}{2}a^{ij}D_{ij}v(x) + Q(x,\nabla v) + V(x), \quad x \in R^N. \tag{3.1}$$

Proof. Let us set $\tilde{u}(T,x) = u(T,x) - u(T,0)$. Then $\{\tilde{u}(T,x)\}_T$ turns out to be a family of uniformly bounded and equicontinuous functions of x on each compact set because of (1.19). Therefore it has a subsequence $\{\tilde{u}(T_i,x)\}$ converging to a function $v(x) \in C(R^N)$ uniformly on each compact set. Moreover the estimates (1.18) and (1.19) imply that $\{\tilde{u}(T,\cdot)\}$ forms a bounded subset of the Hilbert space $W^1_2(B_r)$ for each r and we see that there exists a subsequence $\{\tilde{u}(T_i',\cdot)\}$ converging to $\tilde{v} \in W^1_{2,loc}$ weakly in $W^1_{2,loc}$ and strongly in L^2_{loc}. Taking a subsequence, if necessary, we can see that $\tilde{u}(T_i',x) \to \tilde{v}(x)$ a.e. and that $\tilde{v}(x) = v(x)$. We furthermore see that $\nabla\tilde{u}(T_i',x) \to \tilde{v}(x)$ strongly in L^2_{loc} in a similar way to the proof of Lemma 1.6.

Put $\xi(x) = \frac{\partial u}{\partial t}$. Then we obtain from (1.15)

$$\frac{\partial \xi}{\partial t} = \frac{1}{2}a^{ij}D_{ij}\xi + \frac{\partial Q}{\partial p_i}(x,\nabla u)D_i\xi, \quad (0,\infty)\times R^N. \tag{3.2}$$

Since ξ is bounded on $(\epsilon,\infty)\times B_r$ because of (1.18) and (1.19), the regularity theorem for parabolic equations implies that $\{\xi(T,\cdot)\}$ forms a family of Hölder equicontinuous functions on $(\epsilon,\infty)\times B_r$ for each r. Thus we have a subsequence $\{\xi(T_i'',x)\}$ converging to a function $\chi(x) \in C(R^N)$ uniformly on each compact set. In a similar way to the proof of Lemma 1.6 we obtain equation (3.1). □

Equation (3.1) with $\chi(x) \equiv$ constant is called a Bellman equation of ergodic type. We shall show the uniqueness of the solution to the equation in a similar fashion to Theorem 4.1 in Bensoussan and Frehse [2].

LEMMA 3.2. *Besides the assumptions of Theorem* 1.1 *we assume that*

$$Q_0(x,\beta p) \geq \beta^2 Q_0(x,p) - \frac{\kappa}{2}\beta(1-\beta)a^{ij}p_ip_j - \beta(1-\beta)L(x), \quad 0<\beta<1 \tag{3.3}$$

for $\kappa < k_2$ *and locally bounded function* $L(x)$ *such that*

$$V(x) - L(x) \to \infty, \quad |x| \to \infty. \tag{3.4}$$

Then the solution (v,χ) *to equation* (3.5) *such that* $v(x) \to \infty$ *as* $|x| \to \infty$ *is unique, admitting additive constants with respect to* v:

$$\begin{cases} \chi = \frac{1}{2}a^{ij}D_{ij}v + b^iD_iv + Q_0(x,\nabla v) + V(x), & R^N, \\ \chi \equiv \textit{constant}, \quad v \in C^2(R^N). \end{cases} \tag{3.5}$$

Proof. Let (v_1,χ_1) and (v_2,χ_2) be solutions to (3.5) such that $v_i(x) \to \infty$, $|x| \to \infty$, $i = 1,2$. We assume that $\chi_1 \leq \chi_2$ and take α such that $v_1(x_0) + \alpha > v_2(x_0)$ holds for some x_0. Let us set

$$z = e^{-\gamma(v_1+\alpha)} - e^{-\gamma v_2}, \quad \gamma > 0. \tag{3.6}$$

Then we see that $\inf_{x\in R^N} z(x)$ is attained by some $x_\gamma \in R^N$ since $\lim_{|x|\to\infty} z(x) = 0$ and $z(x_0) < 0$. Then

$$\frac{1}{2}a^{ij}D_{ij}z(x_\gamma) \geq 0, \quad \nabla z(x_\gamma) = 0, \quad z(x_\gamma) < 0. \tag{3.7}$$

We therefore have at x_γ

$$\begin{aligned}
0 &\leq \frac{1}{2}a^{ij}D_{ij}z + b^iD_iz \\
&= \left\{-\frac{\gamma}{2}a^{ij}D_{ij}v_1 + \frac{\gamma^2}{2}a^{ij}D_iv_1D_jv_1 - \gamma b^iD_iv_1\right\}e^{-\gamma(v_1+\alpha)} \\
&+ \left\{\frac{\gamma}{2}a^{ij}D_{ij}v_2 - \frac{\gamma^2}{2}a^{ij}D_iv_2D_jv_2 + \gamma b^iD_iv_2\right\}e^{-\gamma v_2} \\
&= -\gamma\left(\chi_1 - V - Q_0(x_\gamma, \nabla v_1) - \frac{\gamma}{2}a^{ij}D_iv_1D_jv_1\right)e^{-\gamma(v_1+\alpha)} \\
&+ \gamma\left(\chi_2 - V - Q_0(x_\gamma, \nabla v_2) - \frac{\gamma}{2}a^{ij}D_iv_2D_jv_2\right)e^{-\gamma v_2}.
\end{aligned}$$

Thus we obtain

$$\begin{aligned}
&e^{-\gamma(v_1+\alpha)}\left(Q_0(x_\gamma, \nabla v_1) + \frac{\gamma}{2}a^{ij}D_iv_1D_jv_1\right) \\
&- e^{-\gamma v_2}\left(Q_0(x_\gamma, \nabla v_2) + \frac{\gamma}{2}a^{ij}D_iv_2D_jv_2\right) \geq (\chi_1 - V)e^{-\gamma(v_1+\alpha)} - (\chi_2 - V)e^{-\gamma v_2}.
\end{aligned} \tag{3.8}$$

Because of (3.7), $\nabla v_2 = \nabla v_1 e^{\gamma v_2 - \gamma(v_1+\alpha)}$ holds at x_γ and $e^{\gamma v_2 - \gamma(v_1+\alpha)} < 1$. Therefore from (3.3) it follows that the left-hand side of (3.8)

$$\begin{aligned}
&\leq e^{-\gamma(v_1+\alpha)}\left(Q_0(x_\gamma, \nabla v_1) + \frac{\gamma}{2}a^{ij}D_iv_1D_jv_1\right) \\
&- e^{\gamma v_2 - 2\gamma(v_1+\alpha)}\left(Q_0(x_\gamma, \nabla v_1) + \frac{\gamma}{2}a^{ij}D_iv_1D_jv_1\right) \\
&+ e^{\gamma v_2 - \gamma(v_1+\alpha)}\left(e^{-\gamma v_2} - e^{-\gamma(v_1+\alpha)}\right)\left(\frac{\kappa}{2}a^{ij}D_iv_1D_jv_1 + L\right) \\
&\leq e^{\gamma v_2 - \gamma(v_1+\alpha)}\left(e^{-\gamma v_2} - e^{-\gamma(v_1+\alpha)}\right)\left\{-\left(\frac{k_2}{2} - \frac{\gamma+\kappa}{2}\right)a^{ij}D_iv_1D_jv_1 + L\right\}.
\end{aligned}$$

Taking $\gamma < k_2 - \kappa$, we have

$$\begin{aligned}
&e^{\gamma v_2 - \gamma(v_1+\alpha)}(e^{-\gamma v_2} - e^{-\gamma(v_1+\alpha)})L \\
&\qquad \geq (\chi_1 - V)e^{-\gamma(v_1+\alpha)} - (\chi_2 - V)e^{-\gamma v_2} \\
&\qquad \geq (\chi_1 - V)(e^{-\gamma(v_1+\alpha)} - e^{-\gamma v_2})
\end{aligned} \tag{3.9}$$

because $\chi \geq \chi_2$. Thus we obtain $L(x_\gamma) \geq V(x_\gamma) - \chi_1$. Therefore from (3.4) it follows that $x_\gamma \in \overline{B}_{r_0}$ for some r_0 and, taking a subsequence if necessary, $\exists \lim_{\gamma\to 0} x_\gamma = \hat{x} \in \overline{B}_{r_0}$. From (3.9) it follows that $0 \geq \chi_1 - \chi_2$ by letting γ tend to 0. Hence we conclude $\chi_1 = \chi_2$.

On the other hand we have

$$e^{-\gamma(v_1(x_\gamma)+\alpha)} - e^{-\gamma v_2(x_\gamma)} \leq e^{-\gamma(v_1(x)+\alpha)} - e^{-\gamma v_2(x)}$$

for each x since x_γ is a maximum point of z. Therefore by dividing both sides by γ and letting γ tend to 0 we obtain

$$v_2(x) - v_2(\hat{x}) - (v_1(x) - v_1(\hat{x}) \geq 0, \quad \forall x.$$

Put $\psi(x) = v_2(x) - v_2(\hat{x}) - (v_1(x) - v_1(\hat{x}))$. Then $\psi(x) \geq 0$ and $\inf_{x \in B_r(\hat{x})} \psi(x) = \psi(\hat{x})$ for each r. Since $Q_0(x,p)$ is C^1 in p, there exists a continuous R^N-valued function R such that $Q_0(x,p) - Q_0(x,p') = R \cdot (p - p')$. We therefore have

$$\frac{1}{2} a^{ij} D_{ij} \psi + b^i D_i \psi + R \cdot \nabla \psi = 0.$$

By Harnack's inequality $\sup_{x \in B_r(\hat{x})} \psi(x) \leq c_1 \inf_{x \in B_r(\hat{x})} \psi(x) = 0$ for some positive constant c_1. Hence we obtain $\psi(x) \equiv 0$ on $B_r(\hat{x})$. By repeating the same arguments on a ball centered at each point on $\partial B_r(\hat{x})$ we have $\psi(x) \equiv 0$ on $B_{2r}(\hat{x})$. By continuing this procedure we conclude that $\psi \equiv 0$. □

Remark 3.1. Condition (3.3) is rather technical but we can see that Example 1 in §1.5 satisfies the condition. A more specified case, Example 2, will be studied in the following subsection.

LEMMA 3.3. *Let us assume the conditions of Theorem* 1.1 *and* u *to be a nonnegative solution to* (1.15). *Then* $u(t,x) - u(t,0)$ *is bounded below.*

Proof. Put $\eta(t,x) = e^{u(t,0) - u(t,x)}$. Then it suffices to show that

$$\eta(t,x) \leq c_1 \tag{3.10}$$

for some constant $c_1 > 0$. Since we have $0 < \sup_{t_0 \leq t \leq T} \eta(t,x) \leq e^{u(T,0) - u(t_0,x)}$ and $\lim_{|x| \to \infty} e^{u(T,0) - u(t_0,x)} = 0$, there exists $(t_T, x_T) \in [t_0, T] \times R^N$ such that $\eta(t_t, x_T) = \sup_{[t_0,T] \times R^N} \eta(t,x)$. We can assume that $t_0 < t_T$ because otherwise we have nothing to prove. We then have at (t_T, x_T)

$$\frac{\partial \eta}{\partial t} \geq 0, \quad \nabla \eta = 0, \quad a^{ij} D_{ij} \eta \leq 0. \tag{3.11}$$

It follows from (3.11) that $\nabla u(t_T, x_T) = 0$. Therefore we obtain

$$\begin{aligned} 0 &\leq \frac{\partial \eta}{\partial t} - \frac{1}{2} a^{ij} D_{ij} \eta - b^i D_i \eta \\ &= \left(\frac{\partial u}{\partial t}(t_T, 0) - V(x_T) \right) e^{u(t_T,0) - u(t_T,x_T)}. \end{aligned}$$

Thus we have $V(x_T) \leq \frac{\partial u}{\partial t}(t_T, 0) \leq M$ for some positive constant M, which implies that $x_T \in B_R$ for some R. Hence

$$\sup_{[t_0,T] \times R^N} \eta(t,x) \leq e^{u(t_T,0) - u(t_T,x_T)} = e^{\nabla u(t_T, \beta x_T) \cdot x_T}, \quad 0 < \beta < 1$$

and (1.18) and (1.19) imply (3.10). □

Remark 3.2. Owing to Lemma 3.3, $v(x)$ in Lemma 3.1 is bounded below and we can see that $v(x) \to \infty$ as $|x| \to \infty$ in a similar way to the proof of Lemma 2.1.

3.2. The case of Example 2. In the present subsection we specialize the problem to the case of Example 2 in §1.5, namely we consider the following equation:

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{1}{2} a^{ij} D_{ij} u + b^i D_i u - \frac{\kappa}{2} a^{ij} D_i u D_j u + V, \\ u(0,x) = 0 \end{cases} \tag{3.12}$$

with $\kappa > 0$. We then have the following theorem.

THEOREM 3.4. *Let us assume* (1.4)–(1.6), (1.8), *and* (1.10), *and* u *to be a non-negative solution to* (3.12). *Then* $u(t,x)-u(t,0)$ *converges to a function* $v(x)$ *in* $W^1_{2,loc}$ *and uniformly on each compact set, and* $\frac{\partial u}{\partial t}(t,x)$ *to a constant* χ *on each compact set as* $t \to \infty$. *The pair* (v,χ) *is the unique solution to* (3.5) *with* $Q_0(x,p) = -\kappa a^{ij}p_ip_j$.

COROLLARY. *Let us set* $J_\pm(t,x;T) = e^{\theta u(T-t,x)}$. *Then*

$$\lim_{T\to\infty}\frac{1}{T}\log J_\pm(0,x;T) = \lim_{T\to\infty}\frac{\theta u(T,x)}{T} = \lim_{T\to\infty}\theta\frac{\partial u}{\partial t}(T,x) = \theta\chi.$$

Lemma 3.1 applies to the present case and it suffices to show that $\chi(x) \equiv$ constant because of Lemma 3.2, Lemma 3.3 and Remark 3.1. To show this we must prepare some auxiliary facts. Put $\xi(t,x) = \frac{\partial u}{\partial t}(t,x)$. Then it satisfies

$$\frac{\partial \xi}{\partial t} = \frac{1}{2}a^{ij}D_{ij}\xi + b^iD_i\xi - \kappa a^{ij}D_juD_i\xi, \quad (0,\infty)\times R^N \tag{3.13}$$

and the estimate

$$|\xi(t,x)| \le M_{r,\epsilon}, \quad x \in B_r, \quad t \ge \epsilon > 0 \tag{3.14}$$

(cf. Theorem 1.1). For each t we set $w(s,x) = u(t-s,x)$, $0 \le s \le t$, and consider the following SDE:

$$\begin{cases} dX_s^i = \sigma_j^i(X_s)dB_s^j + b^i(X_s)ds - \kappa a^{ij}D_jw(s,X_s)ds, & 0 \le s < t, \\ X_0 = x, \end{cases} \tag{3.15}$$

the solution of which is denoted by $(P^{t,x}, X_t)$. Let us set $\hat{\mathcal{L}}(s) = \frac{\partial}{\partial t} + \frac{1}{2}a^{ij}D_{ij} + b^iD_i - \kappa a^{ij}D_jw(s,x)D_i$. Then we have by Ito's formula

(3.16)

$$\begin{aligned} w(s,X_s) - w(0,X_0) &= \int_0^s \sigma_j^iD_iw(\tau,X_\tau)dB_\tau^j + \int_0^s \hat{\mathcal{L}}(\tau)w(\tau,X_\tau)d\tau \\ &= \int_0^s \sigma_j^iD_iw(\tau,X_\tau)dB_\tau^j - \int_0^s \left\{\frac{\kappa}{2}a^{ij}D_iwD_jw(\tau,X_\tau) + V(X_\tau)\right\}d\tau, \end{aligned}$$

since w satisfies $\hat{\mathcal{L}}(s)w + \frac{\kappa}{2}a^{ij}D_iwD_jw + V = 0$. Take a constant μ such that $\kappa > \mu > 0$. Then from (3.16) it follows that

$$\begin{aligned} &e^{\mu\int_0^s\{\frac{\kappa-\mu}{2}a^{ij}D_iwD_jw(\tau,X_\tau)+V(X_\tau)\}d\tau+\mu w(s,X_s)} \\ &= e^{\mu w(0,x)+\mu\int_0^s\sigma_j^iD_iw(\tau,X_\tau)dB_\tau^j-\frac{\mu^2}{2}\int_0^s a^{ij}D_iwD_jw(\tau,X_\tau)d\tau} \end{aligned}$$

and we obtain

$$E^{t,x}[e^{\mu\int_0^s\{\frac{\kappa-\mu}{2}a^{ij}D_iwD_jw(\tau,X_\tau)+V(X_\tau)\}d\tau+\mu w(s,X_s)}] \le e^{\mu u(t,x)}$$

for $0 \le s < t$. Therefore Jensen's inequality implies that

$$E^{t,x}\left[\int_0^s\left\{\frac{\kappa-\mu}{2}a^{ij}D_iwD_jw(\tau,X_\tau) + V(X_\tau)\right\}d\tau + w(s,X_s)\right] \le u(t,x). \tag{3.17}$$

Thus we see that $\int_0^s \sigma_j^iD_iw(\tau,X_\tau)dB_\tau^j$, $0 \le s < t$, is a martingale.

LEMMA 3.5. *For each $\epsilon > 0$, $t_0 > 0$, and a compact set K, there exist $R > 0$ and $T > 1$ such that*

$$P^{t,x}(X_s \notin B_R, t-(t_0+T) < s < t-t_0) < \epsilon, \quad t > t_0+T, \quad x \in K. \tag{3.18}$$

Proof. In the same way as above we see that

$$\begin{aligned} &w(t-t_0, X_{t-t_0}) - w(t-(t_0+T), X_{t-(t_0+T)}) \\ &= -\int_{t-(t_0+T)}^{t-t_0} \left\{ \frac{\kappa}{2} a^{ij} D_i w D_j w(\tau, X_\tau) + V(X_\tau) \right\} d\tau + \text{ martingale difference.} \end{aligned}$$

We therefore obtain

$$\begin{aligned} &E^{t,x}\left[\int_{t-(t_0+T)}^{t-t_0} \left\{ \frac{\kappa}{2} a^{ij} D_i w D_j w(\tau, X_\tau) + V(X_\tau \right\} d\tau \right] \\ &= E^{t,x}[w(t-(t_0+T), X_{t-(t_0+T)})] - E^{t,x}[w(t-t_0, X_{t-t_0})] \\ &\le E^{t,x}[u(t_0+T, X_{t-(t_0+T)})]. \end{aligned} \tag{3.19}$$

Put $\eta(t,x) = E^{t,x}[u(t_0+T, X_{t-(t_0+T)})]$. Then it satisfies

$$\begin{cases} \frac{\partial \eta}{\partial t} = \frac{1}{2} a^{ij} D_{ij} \eta + b^i D_i \eta - \kappa a^{ij} D_j u(t,x) D_i \eta, & t > t_0+T, \\ \eta(t_0+T, x) = u(t_0+T, x). \end{cases} \tag{3.20}$$

Setting $f(t,x) = e^{-\kappa u(t,x)} \eta(t,x)$, we have

$$\begin{cases} \frac{\partial f}{\partial t} = \frac{1}{2} a^{ij} D_{ij} f + b^i D_i f - \kappa V f, & t > t_0+T, \\ f(t_0+T, x) = e^{-\kappa u(t_0+T, x)} u(t_0+T, x) \end{cases} \tag{3.21}$$

and accordingly, by using Kac's representation, we obtain

$$\begin{aligned} f(t,x) &= E_x[e^{-\int_0^{t-t_1} \kappa V(Y_s) ds} e^{-\kappa u(t_1, Y_{t-t_1})} u(t_1, Y_{t-t_1})] \\ &\le K E_x[e^{-\int_0^{t-t_1} \kappa V(Y_s) ds}] = K e^{-\kappa u(t-t_1, x)}, \end{aligned}$$

where $K = (\kappa e)^{-1}$, $t_1 = t_0 + T$, and (P_x, Y_s) is a solution to (1.20). Thus we have

$$\begin{aligned} \eta(t,x) &= e^{\kappa u(t,x)} f(t,x) \le K e^{\kappa u(t,x) - \kappa u(t-t_1, x)} \\ &\le K e^{\kappa t_1 \frac{\partial u}{\partial t}(s,x)} \le K e^{t_1 M_r}, \quad t - t_1 < \exists s < t, \quad x \in B_r \end{aligned}$$

by using (1.8) and (1.9). For each $\epsilon > 0$, $r > 0$, and $t_0 > 0$, take N_1 and $T > 1$ such that

$$\frac{K e^{\kappa(t_0+T) M_r}}{N_1 T} < \epsilon$$

and R such that $B_R^c \subset \{x; V(x) \ge N_1\}$. Then from (3.19) it follows that

$$\begin{aligned} N_1 T P^{t,x}(X_s \notin B_R, t-(t_0+T) < s < t-t_0) &\le E^{t,x}\left[\int_{t-(t_0+T)}^{t-t_0} V(X_s) ds \right] \\ &\le K e^{\kappa(t_0+T) M_r}. \end{aligned}$$

Hence we obtain our lemma. □

LEMMA 3.6. *For each $\epsilon > 0$, $t_0 > 0$, and a compact set K, there exist R_1 and T such that*

$$P^{t,x}(X_{t-t_0} \in B_{R_1}) > 1 - \epsilon, \quad \forall t > t_0 + T, \quad x \in K. \tag{3.22}$$

Proof. Put $\tau_R = \inf\{s \geq t-(t_0+T), X_s \in B_R\}$ and $\sigma_R = \inf\{s \geq \tau_R, X_s \notin B_{R_1}\}$, $R < R_1$. Then from Lemma 3.5 it follows that for each $\epsilon > 0$, $t_0 > 0$, and a compact set K, there exist R and $T > 1$ such that

$$P^{t,x}(\tau_R \leq t - t_0) > 1 - \frac{\epsilon}{2}, \quad t > t_0 + T, \quad x \in K. \tag{3.23}$$

On the other hand we have by Ito's formula

$$\begin{aligned} &w((t-t_0) \wedge \sigma_{R_1}, X_{(t-t_0)\wedge\sigma_{R_1}}) - w((t-t_0) \wedge \tau_R, X_{(t-t_0)\wedge\tau_R}) \\ &= -\int_{(t-t_0)\wedge\tau_R}^{(t-t_0)\wedge\sigma_{R_1}} \left\{\frac{\kappa}{2} a^{ij} D_i w D_j w(s, X_s) + V(X_s)\right\} ds \\ &+ \int_{(t-t_0)\wedge\tau_R}^{(t-t_0)\wedge\sigma_{R_1}} \sigma_j^i D_i w(s, X_s) dB_s^j, \end{aligned}$$

from which it follows that

$$\begin{aligned} &E^{t,x}[w((t-t_0) \wedge \sigma_{R_1}, X_{(t-t_0)\wedge\sigma_{R_1}}); \{\tau_R \leq t - t_0\}] \\ &\quad \leq E^{t,x}[w((t-t_0) \wedge \tau_R, X_{(t-t_0)\wedge\tau_R}); \{\tau_R \leq t - t_0\}]. \end{aligned}$$

Thus we obtain

$$\begin{aligned} &\inf_{t_0 \leq s \leq t_0+T, x \in \partial B_{R_1}} u(s,x) P^{t,x}(\sigma_{R_1} \leq t - t_0, \tau_R \leq t - t_0) \\ &\qquad \leq \sup_{t_0 \leq t \leq t_0+T, x \in \overline{B}_R} u(s,x). \end{aligned}$$

By Lemma 2.1 $\inf_{t_0 \leq s \leq t_0+T, \partial B_{R_1}} u(s,x) \to \infty$ as $R_1 \to \infty$. Therefore for each $\epsilon > 0$ there exists R_1 such that

$$P^{t,x}(\sigma_{R_1} \leq t - t_0, \tau_R \leq t - t_0) < \frac{\epsilon}{2}. \tag{3.24}$$

Hence from (3.23) and (3.24) it follows that

$$P^{t,x}(\sigma_{R_1} > t - t_0, \tau_R \leq t - t_0) > 1 - \epsilon, \quad x \in K, \quad t > t_0 + T,$$

which implies (3.22). □

LEMMA 3.7. *For each compact set K, relative compact open set G, and $t_0 > 0$, there exists $T > 0$ such that*

$$P^{t,x}(X_{t-t_0} \in G) \geq \delta_G > 0, \quad t \geq t_0 + T, \quad x \in K. \tag{3.25}$$

Proof. By Lemma 3.5 for each compact set K, there exists T, $R > 0$ such that

$$P^{t,x}(\tau_R \leq t - (t_0 + 1)) > \frac{1}{2}, \quad t > t_0 + T > t_0 + 1, \quad x \in K.$$

Therefore we have

$$\begin{aligned} P^{t,x}(X_{t-t_0} \in G) &\geq P^{t,x}(\tau_R \leq t-(t_0+1), X_{t-t_0} \in G) \\ &= \int_{\overline{B}_R} \int_{t_0+1}^{t_0+T} P^{t,x}(\tau_R \in t-ds, X_{\tau_R} \in dy) P^{s,y}(X_{s-t_0} \in G) \\ &\geq \inf_{y \in \overline{B}_R, t_0+1 \leq s \leq t_0+T} P^{s,y}(X_{s-t_0} \in G) P^{t,x}(\tau_R \leq t-(t_0+1)) \\ &\geq \frac{1}{2} \inf_{t_0+1 \leq s \leq t_0+T, y \in \overline{B}_R} P^{s,y}(X_{s-t_0} \in G) \geq \frac{1}{2}\delta_G > 0 \end{aligned}$$

by using a strong Markov property. □

Proof of Theorem 3.4. Now we give the proof of Theorem 3.4. By Lemma 3.1 there exists a subsequence $\{t_k\}$ such that $\xi(t_k, x)$ converges uniformly on each compact set. Put $\bar{\xi}(x) = \lim_{k\to\infty} \xi(t_k, x) \geq 0$ and assume that $\bar{\xi}(x) \neq$ constant. Set $m_2 = \inf_{x \in R^N} \bar{\xi}(x)$. Then there exists $x_1 \in R^N$ such that $m_1 = \bar{\xi}(x_1) > m_2$. Setting $2\sigma = m_1 - m_2$, there exists an open neighborhood G of x_1 such that $\bar{\xi}(x) > m_1 - \sigma$, $x \in G$. Take a point x_2 such that $\bar{\xi}(x_2) < m_2 + \sigma\delta_G$, where δ_G is a constant defined in Lemma 3.7. For each $\epsilon > 0$ and $R > 0$ there exist $\bar{t}$ such that

$$|\bar{\xi}(\bar{t}, x) - \bar{\xi}(x)| < \epsilon, \quad |x| < R.$$

Take a compact set K such that $x_2 \in K$ and sufficiently large R such that $G \subset B_R$ and $B_{R_1} \subset B_R$, where R_1 is defined in Lemma 3.6 with $t_0 = \bar{t}$, ϵ, and K. Then for $x \in K$ and sufficiently large t we have

$$\begin{aligned} \xi(t,x) &= E^{t,x}[\xi(\bar{t}, X_{t-\bar{t}})] \\ &= E^{t,x}[\xi(\bar{t}, X_{t-\bar{t}}); X_{t-\bar{t}} \in G] + E^{t,x}[\xi(\bar{t}, X_{t-\bar{t}}); X_{t-\bar{t}} \in G^c \cap B_R] \\ &+ E^{t,x}[\xi(\bar{t}, X_{t-\bar{t}}); X_{t-\bar{t}} \in B_R^c] \\ &\geq (m_1 - \sigma - \epsilon) P^{t,x}(X_{t-\bar{t}} \in G) + (m_2 - \epsilon) P^{t,x}(X_{t-\bar{t}} \in G^c \cap B_R) \\ &\geq \sigma\delta_G + (m_2 - \epsilon)(1 - \epsilon) \\ &\geq \sigma\delta_G + m_2 - \epsilon(m_2 + 1). \end{aligned}$$

Letting $t = t_k$ tend to ∞ we obtain

$$\bar{\xi}(x) \geq \sigma\delta_G + m_2 - \epsilon(m_2 + 1), \quad x \in K, \quad \epsilon > 0.$$

Hence we have $\bar{\xi}(x_2) \geq \sigma\delta_G + m_2$, which is a contradiction. □

3.3. Principal eigenvalue and large deviation. We illustrate an example of Theorem 3.4. Let us consider equation (3.12) with $a^{ij} = \delta^{ij}$, $b^i = 0$, and $\kappa = 1 - \theta > 0$, namely

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{1}{2}\Delta u - \frac{1-\theta}{2}|\nabla u|^2 + V, \\ u(0,x) = 0. \end{cases} \tag{3.26}$$

It corresponds to the risk-sensitive control problem

$$\begin{aligned} \text{minimize} \quad & \frac{1}{\theta} \log E_x[e^{\theta \int_0^{T-t} \{V(X_s) + \frac{1}{2}|z_s|^2 ds\}}] \\ \text{subject to} \quad & dX_s = dB_s + Z_s ds, \quad X_0 = x, \end{aligned}$$

where the control region $Z = R^N$. Let u be a nonnegative solution to (3.26) and for each $T > 0$ we set $w(t,x) = u(T-t,x)$ and consider the SDE

$$\begin{cases} dY_s = dB_s - \nabla w(s, Y_s)ds, \\ Y_0 = x. \end{cases}$$

Then, owing to Proposition 2.4, we have

$$J_+(t,x) = e^{\theta w(t,x)} = E_x[e^{\theta \int_0^{T-t}\{V(Y_s)+\frac{1}{2}|\nabla w(s,Y_s)|^2\}ds}].$$

We furthermore have by Theorem 3.4

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{\theta T} \log J_+(0,x;T) &= \lim_{T\to\infty} \frac{u(T,x)}{T} \\ &= \lim_{T\to\infty} \frac{\partial u}{\partial t}(T,x) = \chi, \end{aligned}$$

where χ is a constant defined by the equation

$$\chi = \frac{1}{2}\Delta v - \frac{1-\theta}{2}|\nabla v|^2 + V. \tag{3.27}$$

On the other hand, $u(t,x)$ has the representation

$$u(t,x) = -\frac{1}{1-\theta} \log E_x[e^{-(1-\theta)\int_0^t V(B_s)ds}]$$

and we see that

$$\lim_{T\to\infty} \frac{u(T,x)}{T} = \frac{\lambda_1(\theta)}{1-\theta},$$

where λ_1 is the principal eigenvalue of $-\frac{1}{2}\Delta + (1-\theta)V$. Thus we see that $\chi = \frac{\lambda_1(\theta)}{1-\theta}$, hence

$$\lim_{T\to\infty} \frac{1}{T} \log J_+(0,x;T) = \frac{\theta\lambda_1(\theta)}{1-\theta}.$$

On the other hand, by Theorem 3.4 $u(T,x) - u(T,0)$ converges to the solution v to (3.27). We consider the temporally homogeneous diffusion process (P_x, X_t) associated with the Dirichlet form on $L^2(\rho^2 dx)$ defined by

$$\mathcal{E}(f,g) = \frac{1}{2}\int \nabla f \cdot \nabla g \rho^2 dx,$$

where $\rho = e^{-v}$ (cf. Fukushima [10]). Then the Donsker–Varadhan large deviation principle asserts that

$$\begin{aligned} &\lim_{T\to\infty} \frac{1}{T} \log E_x[e^{\theta \int_0^T \{V(X_s)+\frac{1}{2}|\nabla v|^2(X_s)\}ds}] \\ &\quad = \sup_f \left[\theta \int \left\{V(x) + \frac{1}{2}|\nabla v|^2\right\} f^2\rho^2 dx - \mathcal{E}(f,f)\right] \\ &\quad = \frac{\theta\lambda_1(\theta)}{1-\theta} \end{aligned}$$

(cf. Deuschel and Stroock [6], Donsker and Varadhan [7], [8]).

REFERENCES

[1] A. BENSOUSSAN AND J. FREHSE, *On Bellman equations of ergodic type with quadratic growth Hamiltonian*, Pitman Res. Notes Math. Ser., 148 (1987), pp. 13–26.

[2] ———, *On Bellman equations of ergodic control in R^N*, J. Reine Angew. Math., 429 (1992), pp. 125–160.

[3] A. BENSOUSSAN AND J. H. VAN SCHUPPEN, *Optimal control of partially observable stochastic systems with an exponential of integral performance index*, SIAM J. Control Optim., 23 (1985), pp. 599–613.

[4] A. BENSOUSSAN AND H. NAGAI, *An ergodic control problem arising from the principal eigenfunction of an elliptic operator*, J. Math. Soc. Japan, 43 (1991), pp. 49–65.

[5] E. B. DAVIES, *Heat Kernels and Spectral Theory*, Cambridge University Press, London, 1989.

[6] J. M. DEUSCHEL AND D. W. STROOCK, *Large Deviations*, Academic Press, New York, 1989.

[7] M. D. DONSKER AND S. R. S. VARADHAN, *Asymptotic evaluation of certain Markov process expectation for large time,* III, Comm. Pure Appl. Math., 29 (1976), pp. 389–461.

[8] ———, *On the principal eigenvalue of second-order elliptic differential operators*, Comm. Pure Appl. Math., 29 (1976), pp. 595–621.

[9] W. H. FLEMING AND W. M. MCENEANEY, *Risk sensitive control with ergodic cost criteria*, Proc. 31th CDC Conference, Tucson, AZ, 1992, pp. 2048–2052.

[10] M. FUKUSHIMA, *Dirichlet Forms and Markov Processes*, North Holland–Kodansha, Amsterdam, 1980.

[11] F. GIMBERT, *Problemes de Neumann quasilineaires*, J. Funct. Anal., 68 (1985), pp. 65–72.

[12] K. GLOVER AND J. C. DOYLE, *State-space formulae for all stabilizing controllers that satisfy an H^∞-norm bound and relation to risk sensitivity*, Systems Control Lett., 11 (1986), pp. 167–172.

[13] R. Z. HAS'MINSKII, *Stochastic Stability of Differential Equations*, Sijthoff Noordhoff, Alphen aan den Rijn, 1980.

[14] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North Holland–Kodansha, Amsterdam, 1989.

[15] D. H. JACOBSON, *Optimal stochastic linear systems with exponential performance criteria and relation to deterministic differential games*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 124–131.

[16] N. V. KRYLOV, *Controlled diffusion processes*, Springer-Verlag, New York, Heidelberg, Berlin, 1980.

[17] O. A. LADYZENSKAYA, V. A. SOLONIKOV, AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, AMS Transl. of Math. Monographs, Providence, RI, 1968.

[18] P. LI AND S. T. YAU, *On the parabolic kernel of the Schrödinger operator*, Acta Math., 156 (1986), pp. 153–201.

[19] P. L. LIONS, *Quelques remarques sur les problemes elliptiques quasi lineaires du second ordre*, J. d'Analyse Math., 45 (1985), pp. 234–254.

[20] R. S. LIPTSER AND A. N. SHIRYAEV, *Statistics of Random Processes* I, Springer-Verlag, New York, 1977.

[21] H. NAGAI, *Ergodic control problems on the whole Euclidean space and convergence of symmetric diffusions*, Forum Math., 4 (1992), pp. 159–173.

[22] ———, *A remark on parabolic Harnack inequalities*, Bull. London Math. Soc., 24 (1992), pp. 469–474.

[23] T. RUNOLFSSON, *Stationary risk-sensitive LQG control and its relation to LQG and H-infinity control*, Proc. 29th CDC Conference, Honolulu, HI, 1990, pp. 1018–1023.

[24] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer-Verlag, New York, 1979.

[25] P. WHITTLE, *Risk-sensitive linear/quadratic/Gaussian control*, Adv. Appl. Prob., 13 (1981), pp. 764–777.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 102–123, January 1996

005

OPTIMAL CONTROL OF THE BLOWUP TIME*

EMMANUEL N. BARRON† AND WENXIONG LIU‡

Abstract. The problem of optimal control of the blowup time of a system of nonlinear controlled ordinary differential equations is considered in this paper. The blowup time is defined to be the first time that the norm of the trajectory becomes infinite. When one seeks to maximize the blowup time the pair $(V(x), \Omega)$ comes under consideration, where $x \in R^n \mapsto V(x) \in [0, \infty]$ is the value function and $\Omega \subset R^n$ is the blowup set. This is the set of initial points from which finite time blowup will occur for any control. We prove that (V, Ω) is the unique viscosity solution of the equation $1 + \max_z D_x V(x) \cdot f(x, z) = 0, x \in \Omega$ and conditions $\lim_{|x| \to \infty} V(x) = 0, \lim_{x \to \partial\Omega} V(x) = +\infty$. Finally, we derive the Pontryagin maximum principle for an optimal control. Some generalizations are also discussed.

Key words. blowup time, optimal control, viscosity solutions, Pontryagin principle

AMS subject classifications. 49C20, 49C05, 35B99

1. Introduction.

In the modern theory of reaction diffusion equations the phenomenon of blowup in finite time is under study by many researchers too numerous to list here. The usual problem considered is modelled by the semilinear parabolic equation $u_t - \Delta u = u^p$, $p > 1$. It is known that for certain initial data the solution will blow up, i.e., become infinite, in finite time. It is the superlinear growth in the term u^p that leads to the explosion. Models with superlinear growth arise in many contexts including thermal and chemical explosions, population dynamics, biological processes, and some models in economics.

We pose the natural problem of controlling a system which may blow up in finite time. In certain circumstances it is obviously of interest to maximize the time at which the blowup will occur. For example, one may want to raise the temperature in a chemical reaction as much as possible prior to the actual explosion. In other circumstances one might want to minimize the blowup time. Fuel efficiency in a car engine is one case where minimizing the blowup time is desirable. These are natural problems to consider as optimal control problems. In addition, it may be of practical importance to maximize the blowup time under the worst possible environmental assumptions, or vice versa. In this case a differential game model of blowup would be appropriate.

To initiate the optimal blowup problem, in this paper we will consider optimal control of the blowup time when the dynamical system governing the underlying process is a system of ordinary differential equations. Assuming superlinear growth in the dynamics leads to the possibility that the trajectories may blow up in finite time. Most, if not all, prior work in optimal control with ordinary differential equations has assumed the trajectories will exist globally in time. This is not our interest here.

A notable study of optimal control of systems which may blow up is the book by J. L. Lions [12]. Lions considers the question of optimally controlling a distributed system which, for a fixed control, may blow up in finite time. The model problem he considers is $u_t - \Delta u - u^3 = z$ where z is the control. But Lions views the problem in a

* Received by the editors March 3, 1993; accepted for publication (in revised form) August 6, 1994. This research was supported in part by NSF grant DMS-9300805.

† Department of Mathematical Sciences, Loyola University of Chicago, Chicago, IL 60626 (enb@math.luc.edu). This research was supported in part by NSF grant DMS-9102967.

‡ Department of Mathematical Sciences, Loyola University of Chicago, Chicago, IL 60626 (wliu@math.luc.edu).

way completely different from our point of view. He considers the *pair* (u, z) together as *admissible* if and only if u exists globally in time. The requirement of global existence in time is imposed on the control functions. This is a different problem from what we consider in this paper since blowup is then precluded. Our interest is precisely the case when the solution may in fact blow up in finite time.

The focus in this paper is the system $d\xi/d\tau = f(\xi(\tau), \zeta(\tau))$, $\tau > 0$ with $\xi(0) = x \in R^n$. The control function is $\zeta(\cdot)$. The trajectory may blow up at time $T_x(\zeta)$, where we indicate the dependence on the initial point x as well as the control which is used. The blowup time is considered as a map $x \mapsto T_x(\zeta) \in [0, \infty]$. For any fixed control it is certainly possible that the blowup time is infinite, i.e., finite time blowup does not occur. Indeed this will happen at equilibrium points of the system. If we are trying to maximize the blowup time we might seek to steer the trajectory toward equilibrium points.

When the goal is to maximize the blowup time, one considers the value function

$$V(x) = \sup_{\zeta} T_x(\zeta), \qquad V : R^n \to [0, \infty].$$

The blowup set is defined to be the set of initial points of the trajectory at which finite time blowup will occur for any control used: $\Omega = \{x \in R^n : V(x) < +\infty\}$. Of course, Ω is not known a priori. We will characterize the pair (V, Ω) as the unique continuous viscosity solution of the Hamilton–Jacobi equation

$$1 + \max_{z \in Z} D_x V(x) \cdot f(x, z) = 0, \qquad x \in \Omega,$$

satisfying the conditions

$$\lim_{|x| \to \infty} V(x) = 0, \qquad \lim_{x \to \partial\Omega} V(x) = +\infty.$$

If no control is used, so $f = f(x)$, this equation will also hold without the max. It may be useful to know the problem satisfied by the blowup time even for the uncontrolled case.

The theory of necessary conditions for an optimal control for the blowup problem is developed in §5. The Pontryagin principle is derived on the basis of the Hamilton–Jacobi equation.

In a subsequent paper [8] we look at the problem of controlled diffusions which may explode. This problem has the distinctive feature that for nondegenerate diffusions, if finite time blowup occurs at any point $x \in R^n$, then, because of the properties of diffusions, finite time blowup will occur everywhere. Thus, $\Omega = R^n$ for the case of nondegenerate diffusions.

Our original motivation for studying optimal control of blowups was the problem governed by a distributed system, i.e., a partial differential equation (PDE). This problem turns out to be substantially more difficult than the case of ordinary differential equations because of the difficulties with obtaining the regularity of the blowup time with respect to the initial data. Since the initial data is now in a function space we need to determine the Frechet or Gateaux differentiability of the blowup time and the regularity of the associated value function. Differentiability of the blowup time is not known even for the uncontrolled case with the semilinear parabolic model problem $u_t - \Delta u = u^p$. On the other hand, it may be possible to bypass this difficulty using the Crandall and Lions theory of viscosity solutions in infinite dimensions extended by [15]. We hope to return to this problem in a future paper.

Finally, by transforming all of our dynamics from R^n to the stereographic sphere in R^{n+1}, in principle it looks like our problem can be formulated as the problem of maximizing the time at which we hit the north pole, which is mapped to points in R^n with infinite norm. Of course when we want to minimize the blowup time this would be the same as minimizing the time at which we hit the north pole. We see that there is an intimate connection between our problem of blowup and the minimum time problem studied so effectively by Bardi, Soravia, Falcone, Staicu, and others [3]–[6], [14]. Indeed, if we formally consider the target set $\mathcal{T} = \{\infty\}$, then our blowup problem is nothing more than maximizing the time to hit $\mathcal{T}$. Using this point of view, we adapt some arguments of [3]–[5], and [14] in order to prove some of our basic results.

For the convenience of the reader we record here the definition of viscosity solution.

DEFINITION 1.1. *A lower (upper) semicontinuous continuous function v is a viscosity subsolution (supersolution) of*

$$H(x, v, Dv) = 0, \quad x \in \Omega,$$

where Ω is an open set, if for any $\varphi \in C^1(\Omega)$

$$x_0 \in \arg\max(v - \varphi)(x) \Longrightarrow H(x_0, v(x_0), D\varphi(x_0)) \geq 0$$

(respectively, for any $\varphi \in C^1(\Omega)$ and

$$x_0 \in \arg\min(v - \varphi)(x) \Longrightarrow H(x_0, v(x_0), D\varphi(x_0)) \leq 0).$$

The reader is encouraged to look at [9], [10], and especially [11] for the primary results in viscosity solution theory.

2. Basic properties of the blowup time. Consider the autonomous controlled system of ordinary differential equations

$$\begin{aligned} &d\xi(t)/dt = f(\xi(t), \zeta(t)), \ t > 0, \\ &\xi(0) = x \in R^n. \end{aligned} \tag{2.1}$$

The control functions $\zeta(\cdot)$ are chosen from the class of functions $\mathcal{Z} = \{\zeta : [0, \infty) \to Z \mid \zeta \text{ is Lebesgue measureable}\}$, where Z is a fixed compact subset of some euclidean space R^q. Z is referred to as the control set. A solution of (2.1) which starts at $x \in R^n$ is denoted by $\xi_x(t)$ or by $\xi(t; x)$ to indicate the dependence on the initial condition.

We will assume that $f : R^n \times Z \to R^n$ is jointly continuous and is C^1 in x uniformly in $z \in Z$. This guarantees that f is locally Lipschitz in x. That is, for any $x_0 \in R^n$ and $\delta > 0$ there is a constant $K(x_0, \delta)$ such that

$$|f(x, z) - f(x', z)| \leq K(x_0, \delta)|x - x'|, \quad \forall x, x' \in B_\delta(x_0), \tag{2.2}$$

where $B_\delta(x_0)$ denotes, here and generally, the ball of radius δ centered at x_0.

It is well known that if f is assumed to be *uniformly* Lipschitz continuous then there is for each control $\zeta \in \mathcal{Z}$ a unique *global* solution of (2.1) for any initial position $x \in R^n$. This is also true if one assumes that f has linear growth in x. Since we are only assuming that f is locally Lipschitz in x, we can only conclude in general that, for each control $\zeta \in \mathcal{Z}$, there is a unique solution on a maximal interval of existence,

which we shall denote by $[0, T_x(\zeta))$. The time $T_x(\zeta) \in [0, +\infty]$ is called the *blowup time* for the initial position x when the control used is ζ. When the control is fixed we will also denote the blowup time as simply $T(x)$. We wish to study the blowup time as a function $T : R^n \to [0, \infty]$. If $T(x) = +\infty$ for some $x \in R^n$ then this says that the trajectory does not blow up in finite time. Conversely, if $T(x) < \infty$ for some $x \in R^n$, ξ leaves every compact subset of R^n in finite time.

When $T(x) < +\infty$ the only possible behavior of the trajectory is to blow up to $+\infty$ at time $T(x)$. Therefore, we may also describe T using

$$T(x) = \inf \left\{ s \in [0, \infty] : \lim_{t \to s-0} \|\xi(t)\| = +\infty \right\}.$$

Finally, the fact that finite time blowup means that ξ leaves every compact set in finite time motivates the useful characterization

$$T(x) = \lim_{R \to \infty} \tau_x^R(\zeta),$$

where τ_x^R is the exit time of the trajectory from $\overline{B_R(x)}$. The limit exists (since $R \mapsto \tau^R$ is not decreasing) in the extended sense and can be $+\infty$, if, for example, a trajectory stays in a compact set for all time. This characterization is particularly useful in §3.

The starting points of trajectories which are of interest to us are those for which blowup is inevitable no matter what control is used.

DEFINITION 2.1. *The blowup set $\Omega \subset R^n$ is defined by*

$$\Omega = \left\{ x \in R^n : \sup_{\zeta \in \mathcal{Z}} T_x(\zeta) < +\infty \right\}.$$

For a fixed control $\zeta \in \mathcal{Z}$ define the set

$$\Omega(\zeta) = \{x \in R^n : T_x(\zeta) < +\infty\}.$$

Obviously, $\Omega \subset \Omega(\zeta)$, $\forall \zeta \in \mathcal{Z}$.

We will need the following assumptions.

(Ai) For some $p > 1$,

$$\frac{x \cdot f(x, z)}{|x|^{p+1}} \to 1 \qquad \text{as } |x| \to \infty, \tag{2.3}$$

uniformly in $z \in Z$.

(Aii) For some $M > 0$,

$$|f(x, z)| \leq M(1 + |x|^p). \tag{2.4}$$

Conditions (2.2), (2.3), and (2.4) will be assumed to hold throughout this paper.

Condition (Ai) means that $f(\cdot, z)$ behaves like $|x|^p$ for any $z \in Z$ when $|x|$ is large. For example, $f(x, z) = |x|^{p-1}x + |x|^{q-1}x + g(z)$, where $1 < q < p$ and g is continuous, satisfies these conditions.

We may also consider dynamical systems which grow exponentially fast as $|x| \to \infty$. In that case we need an assumption like $x \cdot f(x, z)/(|x|e^{|x|}) \to 1$ as $|x| \to \infty$.

In what follows we will use the notation that $C^k(A, B)$ is the class of k times continuously differentiable functions from A to B. When $k = 0$ this is the class of

continuous functions and will be denoted simply as $C(A,B)$. If $B = R^1$ we will write $C^k(A)$.

Under condition (2.3) we will prove that the blowup set Ω, and therefore also $\Omega(\zeta)$, $\forall \zeta \in \mathcal{Z}$, contains points of sufficiently large norm for any ζ and is therefore always nonempty.

PROPOSITION 2.1. *If* (2.3) *holds, then* Ω *is nonempty and unbounded. In fact, there is a constant* $K > 0$ *such that* $\{|x| \geq K\} \subset \Omega$.

Proof. By (2.3), for any $\varepsilon > 0$, there is a $K > 0$ so that when $|x| \geq K$,

$$1 - \varepsilon \leq \frac{x \cdot f(x,z)}{|x|^{p+1}} \leq 1 + \varepsilon. \tag{2.5}$$

In particular this is true for any fixed ε, say $\varepsilon = \frac{1}{2}$.

Consider the set $M \equiv \{x \in R^n : |x| \geq K\}$. We will prove that M is an invariant set for any control $\zeta \in \mathcal{Z}$ for the corresponding trajectory given by (2.1). Indeed, define $\Phi(x) = K - |x|$. Then $M \equiv \Phi^{-1}(-\infty, 0]$, $\Phi \in C^1(M, R^1)$, and $D_x\Phi(x) = -x/|x| \neq 0$ for any $|x| = K$. For any $z \in Z$, if $|x| = K$ we have

$$D_x\Phi(x) \cdot f(x,z) = \frac{-x \cdot f(x,z)}{|x|} = \frac{-x \cdot f(x,z)}{|x|^{p+1}} K^p \leq -K^p(1-\varepsilon) < 0.$$

Therefore, $\max_{z \in Z}(D_x\Phi(x) \cdot f(x,z)) \leq 0$ if $|x| = K$. Consequently, by [1, p. 218], for example, M is (positively) invariant and this is true for any control.

Now, let $x \in M$ be the starting point of the trajectory ξ which corresponds to an arbitrary control $\zeta \in \mathcal{Z}$. We will prove that this trajectory blows up in finite time. Define $\gamma(t) = |\xi(t)|$. Then, since M is invariant, $\gamma(t) \geq K$ for all $0 \leq t \leq T_x(\zeta)$. Consequently, by (2.5),

$$\gamma'(t) = \frac{\xi(t) \cdot f(\xi(t), \zeta(t))}{|\xi(t)|} \geq (1-\varepsilon)|\xi(t)|^p = (1-\varepsilon)\gamma(t)^p. \tag{2.6}$$

Set $\eta(\cdot)$ as the solution of

$$\begin{aligned} \frac{d\eta(t)}{dt} &= (1-\varepsilon)\,|\eta(t)|^p, \quad t > 0, \\ \eta(0) &= |x|. \end{aligned}$$

The solution of this problem is given by

$$\eta(t) = [|x|^{1-p} + t\,(1-\varepsilon)\,(1-p)]^{1/1-p}, \tag{2.7}$$

which blows up at the finite time

$$T^\eta(x) = \frac{1}{|x|^{p-1}\,(p-1)\,(1-\varepsilon)} > 0. \tag{2.8}$$

Using (2.6) and the fact that $\gamma(0) = |x|$, we deduce that $\gamma(t) \geq \eta(t)$ for all $0 \leq t \leq T_x(\zeta) \wedge T^\eta(x)$. Then, it must be the case that $T_x(\zeta) \leq T^\eta(x) < +\infty$. Consequently, since ζ was arbitrary, $x \in \Omega$.

We have shown that $M \subset \Omega$. Therefore, Ω is an unbounded set. Finally, in the last statement of the proposition observe that K is found for any $0 < \varepsilon < 1$ such that the first inequality in (2.5) holds. □

COROLLARY 2.2. $\lim_{|x|\to\infty} T_x(\zeta) = 0$, *uniformly in* $\zeta \in \mathcal{Z}$.

Proof. This follows immediately from the proof of the proposition noting that $\lim_{|x|\to\infty} T^\eta(x) = 0$. □

We will need the following estimates.

LEMMA 2.3. *Fix a control* $\zeta \in \mathcal{Z}$ *and an initial point* $x \in \Omega(\zeta)$. *Set* $\beta = 1/(p-1)$. *For any* $\varepsilon > 0$, *there is a* $\delta > 0$, *independent of the control* ζ, *such that at any time* t *for which* $T_x(\zeta) > t > T_x(\zeta) - \delta$, *we have*

$$\beta^\beta((1+\varepsilon)(T_x - t))^{-1/(p-1)} \le |\xi(t)| \tag{2.9}$$

and, for all $0 < t < T_x$,

$$|\xi(t)| \le \max\{\beta^\beta((1-\varepsilon)(T_x - t))^{-1/(p-1)},\ K\}, \tag{2.10}$$

where K *is given so that* (2.5) *holds.*

Proof. From (2.1) it follows that

$$\frac{1}{2}(|\xi(\tau)|^2)' = \xi(\tau)\cdot f(\xi(\tau),\zeta(\tau)). \tag{2.11}$$

Dividing the above equation by $|\xi(\tau)|^{p+1}$ and integrating it over (t, T_x), we get after change of variable

$$\int_{|\xi(t)|}^{\infty} \frac{1}{u^p}du = \frac{1}{p-1}|\xi(t)|^{1-p} = \int_t^T \frac{\xi(s)\cdot f(\xi(s),\zeta(s))}{|\xi(s)|^{p+1}}ds. \tag{2.12}$$

According to (2.5), when $|\xi(t)| \ge K$

$$1 - \varepsilon \le \frac{\xi(t)\cdot f(\xi(t),\zeta(t))}{|\xi(t)|^{p+1}} \le 1 + \varepsilon. \tag{2.13}$$

Furthermore, by the proof of Proposition 2.1, if $|\xi(s)| \ge K$, then $|\xi(t)| \ge K$ for all $s \le t \le T_x$. Let $s \ge 0$ be the first time that $|\xi| \ge K$. On $[0, s]$, $|\xi| \le K$. On $[s, T_x]$ we integrate (2.13) from t to T_x, with $s \le t < T_x$, use (2.12), and rearrange the terms to get

$$|\xi(t)| \le \beta^\beta((1-\varepsilon)(T_x - t))^{-1/(p-1)},\ s \le t < T_x.$$

Since $|\xi(t)| \le K$ on $[0, s]$ we combine these two bounds to obtain (2.10).

To obtain (2.9), because of the fact that $|\xi(t)| \to \infty$ as $t \to T_x$, there exists $\gamma = \gamma(\zeta) > 0$ such that (2.13) holds whenever $T_x > t > T_x - \gamma$. Using (2.12), (2.9) follows for $T_x > t > T_x - \gamma$ again by integration of (2.13) and rearrangement.

We need this estimate with γ independent of ζ. To get it, choose $\delta > 0$ so that for any time t with $\delta > T_x(\zeta) - t$ we have that

$$K \le \beta^\beta((1+\varepsilon)(T_x - t))^{-1/(p-1)}.$$

Notice that δ is independent of ζ but t does depend on ζ. Now, if $\gamma = \gamma(\zeta) \ge \delta$ we are done. If $\gamma(\zeta) < \delta$, we set $\eta(\zeta) = \inf\{\gamma : |\xi(t)| \ge K, t \in (T_x(\zeta) - \gamma, T_x(\zeta))\}$. Clearly $\eta(\zeta) \ge \delta$ and (2.9) is true for all $t \in (T_x(\zeta) - \eta(\zeta), T_x(\zeta))$. This completes the proof. □

The next lemma gives us a condition under which we know that a trajectory is uniformly bounded on a given interval.

LEMMA 2.4. *Fix a control $\zeta \in \mathcal{Z}$ and fix $x \in R^n$.*

(a) *If $T(x) = +\infty$ then there is a constant $C > 0$ such that $||\xi_x|| \leq C$ on the time interval $[0, \infty)$. In fact $C = K$, and K is given so that* (2.5) *holds for some fixed $0 < \varepsilon < 1$.*

(b) *If $T(x) < +\infty$, then on $[0, T(x) - \delta]$, there is a constant C_δ such that $||\xi_x|| \leq C_\delta$, for any $\delta > 0$.*

Proof. (a) If $T(x) = +\infty$ then $\xi_x(t)$ exists for all $t \in [0, \infty)$. If the conclusion of (a) is not true, then there is a time t so that $|\xi_x(t)| > K$, where $K > 0$ is fixed so that (2.5) holds. By Proposition 2.1, this implies that $x \in \Omega(\zeta)$ which contradicts $T(x) = +\infty$.

(b) If $T(x) < +\infty$, set $S = T(x)$ and let $S - \delta > 0$. Then the proof is immediate from (2.10) of Lemma 2.3 □

For the next estimate, which is used later to prove the differentiability of the blowup time, we will need the following additional assumption.

(Aiii) For some $0 < \alpha < p + 2$,

$$| \, |x|^2 (f + x \, f_x) - (p+1)(x \cdot f)x \, | \leq M|x|^\alpha \qquad \text{as } |x| \to \infty, \tag{2.14}$$

and, for some $\varepsilon > 0$,

$$| \, f_x(x, z) \, | \leq p(1 + \varepsilon) \, |x|^{p-1}, \qquad \text{as } |x| \to \infty. \tag{2.15}$$

When the matrix f_x is symmetric (2.15) is not necessary, since then (2.14) implies (2.15). Notice that (Aiii) states that the derivative of f behaves like $p|x|^{p-1}$ for large $|x|$.

It is well known from the standard theory of ordinary differential equations (see for example [1]) that on the interval of existence $[0, T_x(\zeta))$, $y(t) \equiv D_x\xi(t; x)$ exists and satisfies the system

$$dy/dt = D_x f(\xi(t), \zeta(t)) \, y(t), \qquad 0 < t < T_x(\zeta), \tag{2.16}$$

$$y(0) = I. \tag{2.17}$$

Using (2.16), (2.17), and (2.15) we can obtain the asymptotic behavior of $D_x\xi(t; x)$ as t approaches the blowup time $T_x(\zeta)$.

LEMMA 2.5. *Assume that* (Aiii) *holds in addition to the basic assumptions. Let $x \in \Omega(\zeta)$. For any $\varepsilon > 0$, there exists an $M > 0$ and $\delta > 0$, independent of the control, such that when $T_x(\zeta) > t > T_x(\zeta) - \delta$ we have*

$$|D_x\xi(t; x)| \leq M \, (T_x(\zeta) - t)^{-p(1+\varepsilon)/(p-1)}.$$

Proof. Consider, say, $y_j(t) = D_{x_j}\xi(t; x)$. From (2.16),

$$\frac{1}{2}(|y|^2)' = y^t D_x f(x, z) y,$$

where y^t is the transpose of y. Using Lemma 2.3 and (2.15), we see that for any $\varepsilon > 0$,

$$|y|' \leq \frac{p(1+\varepsilon)}{(p-1)(T_x(\zeta) - t)} |y|$$

for $t > T_x(\zeta) - \delta$. The conclusion of the lemma follows by integration. □

LEMMA 2.6. *Let $\zeta \in \mathcal{Z}$ be fixed. The set $\Omega(\zeta)$ is open and $T_x(\zeta) = T(x)$ is locally bounded on $\Omega(\zeta)$.*

Proof. Let $x_0 \in \Omega(\zeta)$ and let ξ be the associated trajectory on $[0, T(x_0))$. Let K be such that (2.5) holds. If $\Omega(\zeta)$ is not open then there is a sequence of points $\{x_n\}$ such that $x_n \to x_0$ and $T(x_n) = +\infty$. Thus, the trajectory starting at x_n with the fixed control ζ, $\xi_n = \xi(\cdot; x_n)$ exists for all time. Obviously, we must have $|x_n| \leq K, \forall n$. From Lemma 2.4, $\{\xi_n\}$ is uniformly bounded (by K), for every n. Let $M > 0$ be arbitrary. On $[0, M]$ ξ_n and also ξ_n' are uniformly bounded. Since $x_n \to x_0$, there is a subsequence, with the same notation, such that $\xi_n \to \xi^*$ uniformly on $[0, M]$ and ξ^* is the trajectory starting from x_0 associated with ζ. By uniqueness, $\xi^* = \xi$. But then $T(x_0) > M$ and since M was arbitrary, this means that $T(x_0) = +\infty$, a contradiction.

Now we prove that $T(x_0)$ is locally bounded. Indeed, if this is not the case, then we can find a sequence of starting points $x_n \to x_0$ such that $\lim_{n\to\infty} T(x_n) = +\infty$. Let $M = T(x_0)$ and n sufficiently large that $T(x_n) \geq M + 1$. Then $\{\xi_n\}$ is uniformly bounded on $[0, M + \frac{1}{2}]$ and so a subsequence converges uniformly to $\xi(\cdot; x_0)$ on this interval. This contradicts the fact that ξ blows up at time $T(x_0) = M$. □

The next lemma proves that for any fixed control the blowup time is a continuous function of the starting position.

LEMMA 2.7. *Fix a control $\zeta \in \mathcal{Z}$. $T(x)$ is continuous in $x \in \Omega(\zeta)$.*

Proof. For $n = 1, 2, \ldots$, let $x_n \in \Omega(\zeta)$ with $x_n \to x \in \Omega(\zeta)$. Let ξ_n and ξ be the corresponding solutions of (2.1) for the same control ζ, and let T_n and T denote the corresponding blowup times. Lemma 2.6 allows us to assume that $T_n < +\infty$ for all n.

We first prove that $\liminf_n T_n \geq T$. To this end, we want to show that for any $\delta > 0$, $T_n \geq T - \delta$ holds for all sufficiently large n. Suppose it is not the case. Then there exists a subsequence, again labeled as T_n, such that $T_n \leq T - \delta$. By taking yet another subsequence, we may assume that $T_n \to T_0 \leq T - \delta$. Given $\varepsilon > 0$, by Lemma 2.4, the sequence $\{\xi_n\}$ is uniformly bounded in $[0, T_0 - \varepsilon]$ for all large n. Now the argument is completed just as in the proof of Lemma 2.6.

Next we show that $\limsup_n T_n \leq T$. We again argue by contradiction. So assume that there exists a subsequence T_n such that $\lim_n T_n \geq T + \delta$ for some $\delta > 0$. From Lemma 2.4, it follows that $\{\xi_n\}$ is uniformly bounded in $[0, T)$. The argument we used above implies that $\xi_n \to \xi$ uniformly in $[0, T)$; hence ξ is also bounded in $(0, T)$, which is a contradiction to the fact that T is the blowup time of ξ. The proof is complete. □

The next lemma tells us that if we also assume condition (Aiii), the blowup time is, in fact, continuously differentiable for each fixed control.

LEMMA 2.8. *Assume, in addition to the basic assumptions, condition* (Aiii). *Fix a control $\zeta \in \mathcal{Z}$. The blowup time $T_x(\zeta) = T(x)$ is a C^1 function of $x \in \Omega(\zeta)$.*

Proof. Divide (2.11) by $1 + |\xi(\tau)|^{p+1}$ and integrate the result over $[0, T(x))$ to get

$$\int_{|x|}^{\infty} \frac{u}{u^{p+1} + 1}\, du = \int_0^{T(x)} \frac{\xi(s) \cdot f(\xi(s), \zeta(s))}{1 + |\xi(s)|^{p+1}}\, ds. \tag{2.18}$$

The equation (2.18) defines T as a function of x implicitly. Indeed, if we set

$$g(t, x) = \begin{cases} \dfrac{\xi(t; x) \cdot f(\xi(t; x), \zeta(t))}{|\xi(t; x)|^{p+1} + 1}, & \text{if } t < T(x), \\ 1, & \text{if } t \geq T(x), \end{cases}$$

and

$$F(T,x) = \int_x^\infty \frac{u}{u^{p+1}+1}\,du - \int_0^T g(t,x)\,dt,$$

then $F(T(x),x) = 0$ and

$$\frac{\partial F}{\partial T}(T(x),x) = -\lim_{|x|\to\infty} \frac{x\cdot f(x,z)}{|x|^{p+1}+1} = -1.$$

If we can show further that $D_x F(T,x)$ exists and is continuous, then the conclusion of the lemma follows from the classical implicit function theorem in calculus.

Since for t near $T(x)$, $t > T(x) - \delta$, using (Aii) and (Aiii)

$$\begin{aligned}
&\left|\frac{\partial g(x,t)}{\partial x_i}\right| = \\
&\quad\left|\frac{[\xi_{x_i}\cdot f(\xi,\zeta) + \xi\cdot(f_x(\xi,\zeta)\xi_{x_i})](|\xi|^{p+1}+1) - (p+1)|\xi|^{p-1}(\xi\cdot\xi_{x_i})(\xi\cdot f(\xi,\zeta))}{(1+|\xi|^{p+1})^2}\right| \\
&\le \frac{||\xi|^2(f+\xi\, f_x) - (p+1)(\xi\cdot f)\,\xi||\xi_{x_i}|\,|\xi|^{p-1}}{(1+|\xi|^{p+1})^2} + \frac{|f+\xi\cdot f_x|\,|\xi_{x_i}|}{(1+|\xi|^{p+1})^2} \\
&\le \frac{M|\xi_{x_i}|}{|\xi|^{p+3-\alpha}+1},
\end{aligned}$$

we conclude, by using Lemma 2.5, that

$$\int_0^{T(x)} \frac{\partial g(x,t)}{\partial x}\,dt = \int_0^{T(x)} (T(x)-t)^{-\mu}\,dt < \infty,$$

where $0 < \mu < 1$. Combining this fact with Lemma 2.7, we see that $D_x F$ is continuous and therefore $T(x)$ is C^1. □

REMARK 2.1. *Differentiate* (2.18) *with respect to* x *to obtain*

$$\begin{aligned}
|D_x T(x)| &\le \frac{|x|}{|x|^{p+1}+1} + \int_0^{T(x)} |D_x g|\,dt \\
&\le \frac{|x|}{|x|^{p+1}+1} + M\int_0^{T(x)} (T(x)-t)^{-\mu}\,dt.
\end{aligned} \tag{2.19}$$

Observe that (2.19) *shows that* $|D_x T(x)| \to 0$ *as* $|x| \to \infty$.

3. The optimal control problem. In this section we will formulate the basic optimal control problem for the blowup time, namely the problem of maximizing the blowup time. This corresponds to the case in which one desires to delay the inevitable explosion as long as possible.

In this section we will assume only the basic assumptions (2.2), (Ai), and (Aii).

DEFINITION 3.1. *The value function* $V : R^n \to [0,\infty]$ *is defined by*

$$V(x) = \sup\{T_x(\zeta) : \zeta \in \mathcal{Z}\}.$$

The blowup set can then be expressed as $\Omega = \{x \in R^n : V(x) < +\infty\}$. Of course, the blowup set is not known a priori.

The main objective in this section is to prove that the value function is continuous. We begin by establishing that the blowup set is open and V is locally bounded on this set.

We introduce the relaxed control problem which will be needed for the proof. Let $M(Z)$ denote the space of bounded measures on Z. Viewing $M(Z)$ as the dual space of $C(Z)$ = continuous functions on Z, we endow $M(Z)$ with the weak star topology of $C(Z)^*$. Let the space of relaxed controls be given by

$$\widehat{\mathcal{Z}} = \{\mu \in L^\infty([0,\infty); M(Z)) \mid \mu(\tau) \text{ is a probability measure a.e. } \tau \in [0,\infty)\}.$$

Let $\mathcal{M}(Z)$ be the set of probability measures on Z. Then we may write that $\widehat{\mathcal{Z}} = L^\infty([0,\infty); \mathcal{M}(Z))$, the space of essentially bounded, Lebesgue measureable maps $\mu : [0,\infty) \to \mathcal{M}(Z)$.

For any relaxed control $\mu(\cdot) \in \widehat{\mathcal{Z}}$ there is a relaxed trajectory given by

$$\widehat{\xi}(\tau) = x + \int_0^\tau \int_Z f(\widehat{\xi}(s), z)\, \mu(s, dz) ds \tag{3.1}$$

on the maximal interval of existence $[0, \widehat{T}_x(\mu))$. We are using the notation that $\widehat{T}_x(\mu)$ is the blowup time for the relaxed trajectory starting from x when the relaxed control is μ. To simplify the presentation, given any $\mu \in \mathcal{M}(Z)$ define the function

$$\widehat{f}(x, \mu) = \int_Z f(x, z)\, \mu(dz). \tag{3.2}$$

It is clear that $\widehat{f}$ enjoys the same continuity and growth properties as does f, i.e., $\widehat{f}$ satisfies (2.2), (Ai), and (Aii). Consequently, the lemmas of the preceding section also hold for the relaxed problem.

Define the relaxed value function $\widehat{V} : R^n \to [0,\infty]$ as

$$\widehat{V}(x) = \sup_{\mu \in \widehat{\mathcal{Z}}} \widehat{T}_x(\mu).$$

Denote the relaxed blowup set $\widehat{\Omega} = \{x \in R^n : \widehat{V}(x) < +\infty\}$.

We are now ready to prove the following proposition.

PROPOSITION 3.1. *The blowup set Ω is open and V is locally bounded on Ω.*

Proof. Clearly, $V(x) \le \widehat{V}(x)$ and $\widehat{\Omega} \subset \Omega$. We will first establish that $\Omega = \widehat{\Omega}$.

Let $x_0 \in \Omega$. Then $T_{x_0}(\zeta) \le V(x_0) = M < +\infty$ for any $\zeta \in \mathcal{Z}$. If there is a $\mu \in \widehat{\mathcal{Z}}$ for which $T_{x_0}(\mu) \ge M + 1$, this says that the relaxed trajectory $\widehat{\xi}(\cdot)$, corresponding to μ, starting from x_0, exists for all $t \in [0, M + \frac{1}{2}]$. Then there is a constant K (see Lemma 2.4) such that $|\widehat{\xi}| \le K$ on this interval. Since $f(x,z)$ is uniformly Lipschitz on $B_K(0) \times Z$ we may apply the relaxation theorem [2] to obtain, for any given $\varepsilon > 0$, an ordinary trajectory $\xi(\cdot)$ starting from x_0 and a control $\zeta \in \mathcal{Z}$, such that $||\xi - \widehat{\xi}|| < \varepsilon$ in the sup norm on $[0, M + \frac{1}{2}]$. Therefore $M = V(x_0) \ge T_{x_0}(\zeta) > M + \frac{1}{2}$. This is a contradiction and we conclude that $\Omega = \widehat{\Omega}$.

Next, for any sequence of starting points $\{x_n\}$ such that $x_n \to x_0$, and relaxed controls $\{\mu_n\} \subset \widehat{\mathcal{Z}}$, we know that, at least on a subsequence, $\mu_n \to \mu \in \widehat{\mathcal{Z}}$ weak-*. Denote the associated relaxed trajectories by $\widehat{\xi}_n$ starting at x_n and $\widehat{\xi}$ starting at x_0. We claim that

$$\liminf_{n \to \infty} T_{x_n}(\mu_n) \le T_{x_0}(\mu). \tag{3.3}$$

To prove this claim, suppose it is not true. Then, there is a $\delta > 0$ for which

$$\liminf_{n\to\infty} T_{x_n}(\mu_n) \geq T_{x_0}(\mu) + \delta.$$

Hence, by Lemma 2.4 a subsequence of $\{\widehat{\xi}_n\}$, still denoted as $\{\widehat{\xi}_n\}$, is uniformly bounded on $[0, T_{x_0}(\mu) + \gamma]$ for any $0 < \gamma < \delta$. Therefore, a further subsequence converges uniformly to $\widehat{\xi}$ on $[0, T_{x_0}(\mu) + \gamma]$. But then $\widehat{\xi}$ is also bounded on this time interval, which contradicts the fact that $T_{x_0}(\mu)$ is the blowup time of $\widehat{\xi}$. We have proved that (3.3) must hold.

Now, to show that Ω, or equivalently, $\widehat{\Omega}$, is open, we must find an open neighborhood of x_0 contained in $\widehat{\Omega}$. If this cannot be done then we can find a sequence of points $\{x_n\}$ and a sequence of relaxed controls $\{\mu_n\} \subset \widehat{\mathcal{Z}}$, such that $T_{x_n}(\mu_n) \to +\infty$, and $x_n \to x_0$ as $n \to \infty$.

Set $T_0 \equiv \widehat{V}(x_0) < +\infty$. Select $\varepsilon > 0$, and fix N at least large enough so that $T_{x_n}(\mu_n) > T_0 + 1$ and $x_n \in B_\varepsilon(x_0)$, for all $n \geq N$. Let $\widehat{\xi}_n$ be the relaxed trajectory associated with μ_n starting from x_n. Since $T_{x_n}(\mu_n) > T_0 + 1$, we have that $\{\widehat{\xi}_n\}$ is uniformly bounded on $[0, T_0]$. Thus, there is a relaxed control $\mu \in \widehat{\mathcal{Z}}$ and corresponding relaxed trajectory $\widehat{\xi}$ starting from x_0 such that on a convergent subsequence $\mu_n \to \mu$ weak-* and $\widehat{\xi}_n \to \widehat{\xi}$ uniformly on $[0, T_0]$. Using (3.3), we have

$$T_0 + 1 < \liminf_{n\to\infty} T_{x_n}(\mu_n) \leq T_{x_0}(\mu) \leq \widehat{V}(x_0) = T_0,$$

which is a contradiction. Hence $\widehat{\Omega} = \Omega$ is open.

The proof that V is locally bounded follows the same line of proof. Since $V \leq \widehat{V}$ we may show that $\widehat{V}$ is locally bounded. If $\widehat{V}$ is not locally bounded at $x_0 \in \Omega$, this would again mean that there is a sequence of points $\{x_n\}$, $x_n \to x_0$, and a sequence of relaxed controls μ_n, such that $T_{x_0}(\mu_n) < +\infty$ but $\lim_{n\to\infty} T_{x_n}(\mu_n) = +\infty$. □

REMARK 3.1. *It is not difficult to verify, using the relaxation theorem that* $V = \widehat{V}$.

We are now ready to prove the following theorem.

THEOREM 3.2. *The function* $V(x) = \sup_{\zeta\in\mathcal{Z}} T_x(\zeta)$ *is continuous on* Ω. *Furthermore,* $V(x) \to 0$ *as* $|x| \to \infty$.

Proof. For any $R > 0$ let $\mathcal{T}_R = \{x : |x| < R\}$. Set $\Omega_R = \Omega \cap \mathcal{T}_R = \{x \in \Omega : |x| < R\}$. Let $\tau_x^R(\zeta)$ denote the first exit time of the trajectory from $\mathcal{T}_R$ when we use the control $\zeta \in \mathcal{Z}$. More precisely, $\tau_x^R(\zeta)$ is the first time that ξ_x is in $\partial\mathcal{T}_R$. If the trajectory never exits $\mathcal{T}_R$ we set $\tau_x^R(\zeta) = +\infty$.

It follows from the definition of blowup time that

$$T_x(\zeta) = \lim_{R\to\infty} \tau_x^R(\zeta). \tag{3.4}$$

Clearly, $\tau_x^R(\zeta)$ increases when R increases. We claim that

$$V(x) = \lim_{R\to\infty} V_R(x), \qquad x \in \Omega, \tag{3.5}$$

where V_R is the maximal (over controls) exit time from $\mathcal{T}_R$, that is, $V_R(x) = \sup_{\zeta\in\mathcal{Z}} \tau_x^R(\zeta)$.

Using assumption (Ai), one has that $f(x,z) \cdot x/|x| > (1-\varepsilon)R^p > 0$ for all sufficiently large $|x| = R$, say $R > R_0$. Indeed, R_0 can be taken as the constant K from Proposition 2.1. Since $\nu = x/|x|$ is the outward pointing normal to the boundary

of the ball of radius R, we have $\nu \cdot f > 0$. It follows from [5, Prop. 5.2] that this implies that $V_R(x)$ is continuous on Ω_R for all $R > R_0$.

For any $R > R_0$, from [5, Thm. 6.1] V_R is the unique continuous viscosity solution of the problem

$$\max_{z \in Z} (f(x,z) \cdot DV_R) + 1 = 0, \quad x \in \Omega_R, \tag{3.6}$$

$$\lim_{x \to \partial\Omega_R^\infty} V_R(x) = \infty, \tag{3.7}$$

and

$$V_R(x) = 0, \quad |x| = R, \tag{3.8}$$

where $\partial\Omega_R^\infty \equiv \partial\Omega_R - \{|x| = R\}$. From Proposition 2.1, $\{|x| > R_0\} \subset \Omega$ and so $\partial\Omega_R^\infty = \partial\Omega$ for $R > R_0$. Consequently, (3.7) implies that for all $R > R_0$,

$$\lim_{x \to \partial\Omega} V_R(x) = \infty. \tag{3.9}$$

It is clear that $R \leq R'$ implies that $V_R \leq V_{R'}$ and $\Omega_{R'} \subset \Omega_R$. Furthermore, $V(x) \geq V_R(x)$ for all $R > 0$.

Let $x \in \Omega$. Using (3.4) we have, for a given $\delta > 0$, the existence of a control $\zeta^* \in \mathcal{Z}$, and $R' > 0$ such that

$$V(x) \leq T_x(\zeta^*) + \delta \leq \tau_x^{R'}(\zeta^*) + 2\delta \leq V_{R'}(x) + 2\delta \leq \lim_{R \to \infty} V_R(x) + 2\delta,$$

which allows us to conclude that (3.5) is true on Ω.

Now we will show that the convergence in (3.5) is uniform on compact subsets of Ω, which allows us to conclude that $V(x)$ is continuous on Ω.

First, from Corollary 2.2, we immediately have that $V(x) \to 0$ as $|x| \to \infty$. In addition, simply from $V(x) \geq V_R(x) \geq 0$, we see that $V_R(x) \to 0$ as $|x| \to \infty$, uniformly in R. Consequently, for any $\delta > 0$, there is an R_δ such that $0 \leq V_R(x) \leq \delta$ if $|x| \geq R_\delta$. We may also assume that R_δ is sufficiently large such that (2.5) holds. For any $R_2 > R_1 > R_\delta$, we have that if $|x| = R_1$

$$0 \leq V_{R_2}(x) - V_{R_1}(x) = V_{R_2}(x) \leq \delta.$$

Therefore, $V_{R_1} \geq V_{R_2} - \delta$ on $\{|x| = R_1\}$. We have that in Ω_{R_1}, V_{R_1} is a supersolution and $V_{R_2} - \delta$ is a subsolution. By the comparison principle for viscosity solutions of (3.6)–(3.8), for example, [14, Thm. 2.1], we conclude that

$$0 \leq V_{R_2}(x) - V_{R_1}(x) \leq \delta, \quad x \in \Omega_{R_1}$$

and therefore this inequality holds also for all $x \in \Omega_{R_\delta} \subset \Omega_{R_1}$. We have shown that for any sequence $\{R_k\}$, with $R_k \to \infty$ as $k \to \infty$, the sequence $\{V_{R_k}(x)\}$ is a Cauchy sequence, uniformly in $x \in \Omega_{R_1}$, for any $R_1 > R_\delta$. Therefore, $V_{R_k}(x) \to V(x)$ uniformly on compact subsets of Ω. □

4. The Bellman equation for the value function. We have seen in the previous section that V is a continuous function on Ω. The main goal of this section is to characterize V by proving that it must satisfy, in the viscosity sense, a Hamilton–Jacobi equation with a free boundary. Then we will prove that V is the only viscosity solution of the problem.

In this section we assume only that the basic assumptions (2.2), (Ai), and (Aii) hold.

THEOREM 4.1. *The value function V is a continuous viscosity solution of the free boundary problem*

$$1 + \max_{z \in Z} D_x V(x) \cdot f(x,z) = 0, \quad x \in \Omega, \tag{4.1}$$

$$\lim_{x \to \partial\Omega} V(x) = +\infty, \tag{4.2}$$

$$\lim_{|x| \to \infty} V(x) = 0. \tag{4.3}$$

Proof. Referring to the proof of Theorem 3.2, we have that $V \geq V_R$ and $V_R \to \infty$ as $x \to \partial\Omega$, and (4.2) follows.

We need to verify that V solves (4.1). This follows immediately from the proof of Theorem 3.2 since we showed there that V is the uniform limit of V_R and V_R is the solution of (3.6). We will give a direct proof however. The proof is based on the dynamic programming principle:

$$V(x) = \sup_{\zeta \in \mathcal{Z}} (t \wedge T_x(\zeta) + \chi_{t \leq T_x(\zeta)} V(\xi_x(t))) \tag{4.4}$$

for any $t \geq 0$, where $\xi_x(\cdot)$ is the trajectory starting at x at time 0 associated with the control $\zeta \in \mathcal{Z}$. We use the notation that χ_A is the characteristic function of the set A and $a \wedge b = \min(a,b)$.

The proof of (4.4) is standard and will not be given here (see, for example, [3]).

Now suppose that $V - \varphi$ achieves a zero minimum at the point $x_0 \in \Omega$, with $\varphi \in C^1(\Omega)$. Let $z \in Z$ be arbitrary and set $\zeta(t) \equiv z$ for all time. The corresponding trajectory starting at x_0 will be denoted by $\xi(\cdot)$. Since the interval of existence of the trajectory is open [1, Thm. 8.3], for any $0 < t < T_x(z)$, sufficiently small, using (4.4)

$$\begin{aligned} \varphi(x_0) = V(x_0) &= \sup_{\zeta \in \mathcal{Z}} (t \wedge T_{x_0}(\zeta) + \chi_{t \leq T_{x_0}(\zeta)} V(\xi(t))) \\ &\geq t \wedge T_{x_0}(z) + \chi_{t \leq T_{x_0}(z)} V(\xi(t)) \\ &\geq t \wedge T_{x_0}(z) + \chi_{t \leq T_{x_0}(z)} \varphi(\xi(t)) \\ &= t + \varphi(\xi(t)). \end{aligned}$$

We have used the fact that, since f is locally Lipschitz, for any $R > 0$ there is a $\delta > 0$ independent of $z \in Z$ such that $\xi(t) \in B_R(x_0)$ if $0 < t < \delta$. Therefore, for t sufficiently small,

$$1 + \frac{\varphi(\xi(t)) - \varphi(x_0)}{t} \leq 0$$

and so, letting $t \to 0+$, using the fact that φ is smooth and ξ is differentiable (since we are using a constant control), we obtain from the arbitrariness of $z \in Z$ that

$$1 + \max_{z \in Z}(D_x\varphi(x_0) \cdot f(x_0, z)) \leq 0. \tag{4.5}$$

This proves that V is a supersolution of (4.1).

Suppose next that $V - \varphi$ achieves a zero maximum at the point $x_0 \in \Omega$, with $\varphi \in C^1(\Omega)$. Suppose that

$$1 + \max_{z \in Z}(D_x\varphi(x_0) \cdot f(x_0, z)) \leq -C < 0. \tag{4.6}$$

Let $\zeta \in \mathcal{Z}$ be arbitrary. For a given $\delta > 0$, set $M = \sup\{|f(x, z)| : x \in B_\delta(x_0), z \in Z\}$. Then the trajectory $\xi(\cdot)$ starting at x_0 will exist at least for $t \in [0, \delta/M)$ (see, for example, [1]). Notice that this is true independently of the control ζ which is chosen. Thus, we know that $\xi(t) \in B_{\delta\, M}(x_0)$ for all $0 \leq t < t_0 < T_{x_0}(\zeta)$ for some $t_0 > 0$ independent of ζ. Using (4.6) we have that

$$1 + D_x\varphi(x_0) \cdot f(x_0, \zeta(t)) \leq -C, \qquad 0 < t < t_0.$$

Then, for sufficiently small t, say $0 < t < t_1 < t_0$,

$$1 + D_x\varphi(\xi(t)) \cdot f(\xi(t), \zeta(t)) \leq -C/2, \qquad 0 < t < t_1 < T_{x_0}(\zeta),$$

since φ is smooth and $f(x, z)$ is locally Lipschitz. Now we integrate this from 0 to t_1 to get

$$t_1 + \varphi(\xi(t_1)) - \varphi(x_0) \leq -(C/2)t_1.$$

Rearranging this tells us that

$$t_1 + \varphi(\xi(t_1)) \leq \varphi(x_0) - (C/2)t_1 < \varphi(x_0).$$

But this contradicts the dynamic programming principle (4.4). Therefore, V is a subsolution of (4.1) as well. □

We will prove next that V is the only continuous viscosity solution of (4.1) which satisfies (4.2) and (4.3). The proof of uniqueness is based on the uniqueness theorems of Bardi and Soravia [5], [14] for pursuit-evasion differential games. Viewing the optimal control problem as a one-player differential game, we will state their theorem in the form suitable for our problem.

THEOREM 4.2 (see [5, Thm. 3.1]). *Assume that $f(x, z)$ is uniformly Lipschitz. Let Λ be an open set containing the closed target set $\mathcal{T}$. Let $U \in C(\Lambda - \mathcal{T})$ be a viscosity solution of*

$$\begin{aligned} &1 + \max_{z \in Z} DU \cdot f(x, z) = 0, \qquad x \in \Lambda - \mathcal{T}, \\ &U(x) = g(x), \quad x \in \partial\mathcal{T}, \qquad U(x) \to +\infty, \quad x \to \partial\Lambda. \end{aligned} \tag{4.7}$$

Define $W : R^n \to R^1 \cup \{\infty\}$ by

$$W(x) = \sup_{\zeta \in \mathcal{Z}} g(\xi(\sigma_x)) + \sigma_x(\zeta),$$

where g is a given continuous function and $\sigma_x(\zeta) = \inf\{t : \xi_x(t) \in \partial\mathcal{T}\}$ and $\sigma_x(\zeta) = +\infty$ if $\partial\mathcal{T}$ is never achieved. Then $U \equiv W$ and $\Lambda = \{x : W(x) < +\infty\}$.

Clearly this theorem gives us the uniqueness result for the blowup time problem with the target set taken to be $\mathcal{T} = \{\infty\}$. All we need to do is phrase it in terms which do not use the point at infinity.

We are ready to prove the following theorem.

THEOREM 4.3. *Assume* (2.2), (Ai), *and* (Aii). *Let* Λ *be any open set. Let* $U \in C(\Lambda)$ *be a viscosity solution of*

$$\begin{cases} 1 + \max_{z \in Z} (DU \cdot f(x,z)) = 0, & x \in \Lambda, \\ \lim_{|x| \to \infty} U(x) = 0, \qquad \lim_{x \to \partial\Lambda} U(x) = +\infty. \end{cases} \tag{4.8}$$

Then $U = V$ *and* $\Lambda = \Omega$. *That is,* U *must be the value function for the blowup time and* Λ *must be the blowup set* Ω.

Proof. Let $\varepsilon > 0$. Define the target set $\mathcal{T}_K = \{x : |x| \geq K\}$, where K is fixed large enough so that $|U(x)| < \varepsilon$ and so that $\mathcal{T}_K \subset \Omega$. This can be done according to Proposition 2.1. Then U is a viscosity solution of (4.7) if we take $\mathcal{T} = \mathcal{T}_K$ and $g(x) = U(x)$ restricted to $\{|x| = K\}$. Consequently, using Bardi and Soravia's theorem

$$U(x) = \sup_{\zeta \in \mathcal{Z}} \left(g(\xi_x(\tau_x^K)) + \tau_x^K(\zeta) \right) \tag{4.9}$$

and $\Lambda = \{x : \sup_{\zeta \in \mathcal{Z}} \left(g(\xi_x(\tau_x^K)) + \tau_x^K(\zeta) \right) < +\infty\}$. Here $\tau_x^K(\zeta)$ is the exit time from $\Lambda - \mathcal{T}_K$. Note that Bardi and Soravia's theorem requires uniform Lipschitz continuity of the vector field $f(x,\cdot)$. Since we are working in the set $\Lambda - \mathcal{T}_K$, which is bounded, under assumption (2.2), f is uniformly Lipschitz in that set.

Now, observe that $U(x) < +\infty$ if and only if $V(x) < +\infty$. Indeed, if $V(x) < +\infty$ any trajectory must exit the ball of radius K before it explodes, so $\tau_x^K(\zeta) < V(x) < +\infty$, which implies by (4.9) that $U(x) < +\infty$. If $U(x) < +\infty$ then an arbitrary trajectory hits $\mathcal{T}_K$ in finite time. Once it hits $\mathcal{T}_K$, since we chose K large enough so that $\mathcal{T}_K \subset \Omega$, it is in the blowup set Ω. Thus the starting position of the trajectory must have been in Ω as well. Therefore $\Lambda = \Omega$. Furthermore, we let $\varepsilon \to 0$, and therefore $K \to \infty$ in (4.9), and conclude that $U = V$ (refer to (3.5) in the proof of Theorem 3.2). □

The next result we want to prove gives us a way of calculating V and Ω without having to solve the free boundary problem (4.8). To this end, following Bardi and Soravia [5], we consider the Kruzhkov transform of V:

$$w(x) = \psi(V(x)) \equiv 1 - \exp(-V(x)), \qquad x \in \Omega. \tag{4.10}$$

The Kruzhkov transform was introduced for viscosity solutions in [9].

The proof of the following lemma is a simple consequence of the properties of $\psi(\cdot)$ and the definition of viscosity solution (cf. [3], [9]).

LEMMA 4.4. *The function* w *is a viscosity solution of*

$$1 - w + \max_{z \in Z} D_x w(x) \cdot f(x,z) = 0, \qquad x \in \Omega = \{x \in R^n : w(x) < 1\}, \tag{4.11}$$

$$\lim_{x \to \partial\Omega} w(x) = 1, \tag{4.12}$$

$$\lim_{|x| \to \infty} w(x) = 0. \tag{4.13}$$

The fact that w goes to 0 when $|x| \to \infty$ will be the key property that gives a unique solution of the problem (4.11)–(4.13).

The problem with dealing with (4.11) directly is the fact that it holds only on Ω which is unknown a priori since it too depends on w. But Bardi and Falcone in [4] showed how to avoid this problem when it also arose for the minimum time problem. We will do the same thing here. Namely, we consider the same problem for w but we solve it in R^n:

$$1 - w + \max_{z \in Z} D_x w(x) \cdot f(x,z) = 0, \qquad x \in R^n. \tag{4.14}$$

We drop the boundary condition (4.12) and only impose the condition (4.13).

LEMMA 4.5. *Assume* (2.2), (Ai), *and* (Aii). *A continuous viscosity solution of* (4.14), (4.13) *is given by*

$$w(x) = \begin{cases} 1 - \exp(-V(x)), & \text{if } x \in \Omega, \\ 1, & \text{if } x \in R^n - \Omega. \end{cases} \tag{4.15}$$

Proof. We know that w is continuous and bounded. Also, by Proposition 2.1, $w(x) \to 0$ as $|x| \to \infty$. Noting that ψ is an increasing function, we may write w as follows:

$$w(x) = \sup_{\zeta \in \mathcal{Z}} \int_0^{T_x(\zeta)} e^{-s}\, ds, \qquad x \in R^n,$$

which shows that w itself is the value function of a control problem up to the time of blowup with a discount factor. Using a standard dynamic programming argument, we can now prove in a straightforward way that w is a viscosity solution of (4.14). □

Now we will prove that (4.15) is the only solution of (4.14) satisfying (4.13). We phrase the theorem in the form of a comparison principle.

THEOREM 4.6. *Let u be a lower semicontinuous and v an upper semicontinuous function on R^n such that u is a bounded viscosity subsolution and v is a bounded viscosity supersolution of* (4.14) *both satisfying condition* (4.13). *Then $u \le v$ on R^n. In particular, if both u and v are continuous viscosity solutions, then $u \equiv v$.*

REMARK 4.1. *In general, when the function f has superlinear growth in x this theorem is false without the condition* (4.13).

REMARK 4.2. *The proof of the last assertion of the theorem follows from Theorem* 4.3. *Indeed, if u and v are both continuous viscosity solutions of* (4.14), *then the pairs $(U, \Omega(U))$ and $(V, \Omega(V))$, where $U = -\log(1-u), \Omega(U) = \{x : u(x) < 1\}$ and $V = -\log(1-v), \Omega(V) = \{x : v(x) < 1\}$ are both continuous viscosity solutions of* (4.8). *Then, by Theorem* 4.3, *$\Omega(U) = \Omega(V) = \Omega$, the blowup set, and $U = V$ must be the value function for the blowup time. The proof below gives us the comparison principle for discontinuous solutions which is used in* [8].

Proof. The proof is standard, with the only new point being that boundary condition (4.13) allows us to find points of maximum of the doubled function, Φ below, in a compact set. We sketch the proof for completeness.

Define the function $\vartheta(x,y) = u(x) - v(y)$. Let $\varphi \in C^1(R^{2n})$ and let (x_0, y_0) be a maximum point of $\vartheta - \varphi$. Then $x \mapsto \vartheta(x, y_0) - \varphi(x, y_0)$ achieves a maximum at x_0

and $y \mapsto -\vartheta(x_0, y) + \varphi(x_0, y)$ achieves a minimum at y_0. Since u is a subsolution and v is a supersolution of (4.14), we have that

$$1 - u(x_0) + \max_{z \in Z} D_x \varphi(x_0, y_0) \cdot f(x_0, z) \geq 0, \tag{4.16}$$

and

$$1 - v(y_0) + \max_{z \in Z} (-D_y \varphi(x_0, y_0)) \cdot f(y_0, z) \leq 0. \tag{4.17}$$

Subtract (4.17) from (4.16) to get

$$-\vartheta(x_0, y_0) + \max_{z \in Z} D_x \varphi(x_0, y_0) \cdot f(x_0, z) - \max_{z \in Z} (-D_y \varphi(x_0, y_0)) \cdot f(y_0, z) \geq 0.$$

This says that ϑ is a viscosity subsolution of

$$-\vartheta(x_0, y_0) + \max_{z \in Z} D_x \vartheta(x_0, y_0) \cdot f(x_0, z) - \max_{z \in Z} (-D_y \vartheta(x_0, y_0)) \cdot f(y_0, z) = 0. \tag{4.18}$$

Suppose now that there exists a point $x' \in R^n$ for which $u(x') > v(x')$. Then $\vartheta(x', x') > 0$. Consider the function

$$\Phi(x, y) = \vartheta(x, y) - \frac{1}{2\varepsilon} |x - y|^2, \qquad \varepsilon > 0.$$

For each $\varepsilon > 0$, Φ is a continuous function and $\Phi(x', x') > 0$. Since $u \to 0$ and $v \to 0$ as $|x|, |y| \to \infty$, we have that $\limsup_{|x|,|y| \to \infty} \Phi(x, y) \leq 0$. Therefore, for each $\varepsilon > 0$, Φ achieves a positive maximum at, say, $(x_\varepsilon, y_\varepsilon) \in R^{2n}$. Furthermore, we have that if we set

$$M_\varepsilon = \sup_{(x,y) \in R^{2n}} \Phi(x, y),$$

then $\lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} |x_\varepsilon - y_\varepsilon|^2 = 0$ and $\lim_{\varepsilon \to 0} M_\varepsilon = \sup_{x \in R^n} \vartheta(x, x) > 0$.

Using the test function $\frac{1}{2\varepsilon} |x - y|^2$ for the function ϑ, since ϑ is a viscosity subsolution of (4.18) we have that

$$-\vartheta(x_\varepsilon, y_\varepsilon) + \max_{z \in Z} \frac{1}{\varepsilon} (x_\varepsilon - y_\varepsilon) \cdot f(x_\varepsilon, z) - \max_{z \in Z} \frac{1}{\varepsilon} (x_\varepsilon - y_\varepsilon) \cdot f(y_\varepsilon, z) \geq 0. \tag{4.19}$$

Given any $\delta > 0$, we may assume that ε is chosen sufficiently small, for example, $\varepsilon < \varepsilon_0$, so that $|x_\varepsilon - y_\varepsilon| \leq \delta$. Define the point $z_\varepsilon \in Z$ by

$$\max_{z \in Z} \frac{1}{\varepsilon} (x_\varepsilon - y_\varepsilon) \cdot f(x_\varepsilon, z) = \frac{1}{\varepsilon} (x_\varepsilon - y_\varepsilon) \cdot f(x_\varepsilon, z_\varepsilon).$$

Then rearranging equation (4.19) and using the local Lipschitz continuity of f (uniform in z), we obtain

$$\begin{aligned}
\vartheta(x_\varepsilon, y_\varepsilon) &\leq \max_{z \in Z} \frac{1}{\varepsilon} (x_\varepsilon - y_\varepsilon) \cdot f(x_\varepsilon, z) - \max_{z \in Z} \frac{1}{\varepsilon} (x_\varepsilon - y_\varepsilon) \cdot f(y_\varepsilon, z) \\
&\leq \frac{1}{\varepsilon} (x_\varepsilon - y_\varepsilon) \cdot f(x_\varepsilon, z_\varepsilon) - \frac{1}{\varepsilon} (x_\varepsilon - y_\varepsilon) \cdot f(y_\varepsilon, z_\varepsilon) \\
&= \frac{1}{\varepsilon} (x_\varepsilon - y_\varepsilon) \cdot (f(x_\varepsilon, z_\varepsilon) - f(y_\varepsilon, z_\varepsilon)) \\
&\leq \frac{1}{\varepsilon} |x_\varepsilon - y_\varepsilon|^2 \, K_\delta.
\end{aligned}$$

The last expression goes to 0 as $\varepsilon \to 0$. But

$$\lim_{\varepsilon \to 0} \vartheta(x_\varepsilon, y_\varepsilon) \geq \lim_{\varepsilon \to 0} \Phi(x_\varepsilon, y_\varepsilon) > 0.$$

This is a contradiction and so we have established that $u \leq v$ everywhere. □

Consequently, $w = 1 - e^{-V}$ is the only viscosity solution of (4.14) satisfying (4.13). Therefore, in order to find V, the most practical way to do this is to solve (4.14), (4.13) for w and then

$$V(x) = -\log(1 - w(x)), \qquad \Omega = \{x \in R^n : w(x) < 1\}$$

is the unique viscosity solution of (4.1), (4.2), and (4.3), i.e., V is the value function for the blowup time.

5. Pontryagin maximum principle. In this section we will derive the Pontryagin maximum principle for our problem. Our approach will use the results of the previous sections. It is based on the use of the Bellman equation. This approach was first developed for use in the finite horizon problem of Lagrange in [7].

Fix a point $y \in \Omega$ and suppose that there exists an optimal control for this initial point, $\zeta^* \in \mathcal{Z}$. Let ξ^* denote the optimal trajectory with $\xi^*(0) = y$. Since the control ζ^* is fixed, we will denote the blowup time from the starting point y as simply $T(y) < +\infty$. Finally, define the function $w \in C^1(\Omega, R^1)$ by $w(x) = T_x(\zeta^*)$. In other words, w gives the blowup time from starting position x when we use the fixed control ζ^*. From the results of §2, we know that if we assume also condition (Aiii), w is continuously differentiable.

THEOREM 5.1. *Assume* (2.2), (Ai)–(Aiii). *Set* $p(t) \equiv D_x w(\xi^*(t))$ *for* $0 \leq t \leq T(y)$. *Then* ζ^* *must satisfy the maximum principle*

$$\max_{z \in Z}(p(t) \cdot f(\xi^*(t), z)) = p(t) \cdot f(\xi^*(t), \zeta^*(t)) \tag{5.1}$$

for almost all $0 < t < T(y)$. *Furthermore,* p *is given as the solution of*

$$\frac{dp}{dt} = -p(t) \cdot D_x f(\xi^*(t), \zeta^*(t)), \qquad 0 < t < T(y) \tag{5.2}$$

with terminal condition

$$p(T(y)) = 0 \tag{5.3}$$

and p *satisfies the condition*

$$1 + p(t) \cdot f(\xi^*(t), \zeta^*(t)) = 0, \qquad a.e. \quad 0 \leq t < T(y). \tag{5.4}$$

Proof. Begin by observing that if $y \in \Omega$, then $\xi^*(t) \in \Omega$ for all $0 \leq t < T(y)$. Therefore, $t \mapsto w(\xi^*(t))$ is finite. Since the optimal trajectory is absolutely continuous in t and w is continuously differentiable, $p(\cdot)$ is absolutely continuous.

By definition of V and w, we have immediately that $w(x) \leq V(x)$ for every $x \in \Omega$ and also, since ξ^* is optimal, $w(\xi^*(t)) = V(\xi^*(t))$ for all $t \in [0, T(y)]$. But this says that $V - w$ achieves a minimum of zero at each point of the optimal trajectory $\xi^*(t)$. Since w is continuously differentiable, it may be used as a test function in the

definition of supersolution for (4.1). Thus, at each point of differentiability of ξ^*, that is, for almost every $t \in [0, T(y)]$, we have that

$$1 + \max_{z \in Z} D_x w(\xi^*(t)) \cdot f(\xi^*(t), z) \leq 0. \tag{5.5}$$

On the other hand, since $w(\xi^*(t)) = T(\xi^*(t))$ is a.e. differentiable, we have that

$$1 + D_x w(\xi^*(t)) \cdot f(\xi^*(t), \zeta^*(t)) = 0, \qquad \text{a.e. } 0 \leq t < T(y). \tag{5.6}$$

This follows easily from the fact that for any $t < \tau < T(y)$, we have that $T(\xi^*(t)) = T(\xi^*(\tau; t, \xi^*(t))) + (\tau - t)$. By definition of p, (5.6) is the same as (5.4). It also follows from (5.4) that p cannot be identically zero.

From (5.6), we obtain

$$1 + \max_{z \in Z} D_x w(\xi^*(t)) \cdot f(\xi^*(t), z) \geq 0. \tag{5.7}$$

Then, combining (5.7) with (5.5) and using (5.6) allow us to conclude that (5.1) must be true.

Next, it follows from Remark 2.1 (see (2.19)) that $p(T(y)) = 0$.

Finally, we need to verify (5.2). We will use the fact that the function $\Phi(t, s) \equiv D_x \xi^*(t; s, x)$ satisfies the linear variational systems

$$\partial \Phi / \partial t = D_x f(\xi^*(t), \zeta^*(t)) \cdot \Phi, \qquad \partial \Phi / \partial s = -\Phi \cdot D_x f(\xi^*(s), \zeta^*(s)), \qquad 0 \leq s < t, \tag{5.8}$$

and $\Phi(s, s) = \mathbf{1}$. Again, using the fact that for any $t < \tau < T(y)$, $T(\xi^*(t)) = T(\xi^*(\tau; t, \xi^*(t))) + (\tau - t)$ and (5.8), we compute

$$\begin{aligned} p(t) = D_x w(\xi^*(t)) &= DT(\xi^*(\tau; t, \xi^*(t))) \frac{\partial \xi^*(\tau; t, \xi^*(t))}{\partial x} \\ &= DT(\xi^*(\tau; t, \xi^*(t)))(1 + \int_t^\tau \frac{\partial \xi^*(\tau; r, \xi^*(r))}{\partial x} \cdot D_x f(\xi^*(r), \zeta^*(r))\, dr) \\ &= DT(\xi^*(\tau; t, \xi^*(t))) + \int_t^\tau p(r) \cdot D_x f(\xi^*(r), \zeta^*(r))\, dr. \end{aligned}$$

The last equality follows from the fact that $T(\xi^*(\tau; t, \xi^*(t))) = T(\xi^*(\tau; r, \xi^*(r)))$ for all $t \leq r < \tau$. Letting $\tau \to T(y)$, since $DT(\xi^*(\tau; t, \xi^*(t))) \to 0$ (see (2.19) and Remark 2.1) we have that

$$p(t) = \int_t^{T(y)} p(r) \cdot D_x f(\xi^*(r), \zeta^*(r))\, dr. \tag{5.9}$$

Even though the integrand in (5.9) has a singularity at $r = T(y)$ (because $|\xi^*(T(y))| = +\infty$), the integral exists because the left side is always finite, according to (5.4) and the fact that $p(r) \to 0$ as $r \to T(y)$. □

6. Example. In this section we will present a simple example to illustrate the results of this paper. We look at the problem with dynamics

$$\frac{d\xi}{dt} = \xi^2 + \zeta, \qquad t > 0, \tag{6.1}$$

$$\xi(0) = x \in R^1. \tag{6.2}$$

The control set is taken to be $Z = [-1, 1]$. It is clear that our assumptions (2.2), (Ai)–(Aiii) are satisfied. Observe that for any $\zeta \leq 0$, the dynamics have equilibrium solutions which certainly do not blow up in finite time. At equilibrium starting points we have $T(x) = +\infty$.

The Bellman equation for the maximal blowup time $V(x)$ is

$$1 + \max_{-1 \leq z \leq 1} V'(x)(x^2 + z) = 1 + V'(x)x^2 + |V'(x)| = 0, \qquad x \in \Omega. \tag{6.3}$$

The maximum is achieved with $z = V'(x)/|V'(x)|$. Clearly, from (6.3), it is not possible for $V'(x) = 0$ on Ω.

Now we claim that $\Omega = (1, \infty)$. To see formally why this is true, we begin by noting that, from (6.3), it is not possible to have $V'(x) \geq 0$. Thus, $V'(x) < 0$ on Ω and so $z \equiv -1$, and (6.3) becomes.

$$1 + V'(x)(x^2 - 1) = 0, \qquad x \in \Omega. \tag{6.4}$$

If $|x| \leq 1$, (6.4) cannot be satisfied. Therefore, $\Omega \subset \{|x| > 1\}$. Next, if $x \to -1, x \in \Omega$, and -1 is a boundary point, we know that $V(x) \to +\infty$. But then $V'(x) \to +\infty$, which contradicts the fact that $V'(x) < 0$. Thus, $\Omega \subset \{x > 1\}$. We also know that $(\lambda, \infty) \subset \Omega$ for some $\lambda \geq 1$. Suppose $\lambda > 1$. From the fact that $\lambda^2 - 1 > 0$ and $V'(x) \to -\infty$ as $x \to \lambda$ we obtain a contradiction to (6.4). We conclude that $\Omega = (1, \infty)$.

The unique viscosity solution of (6.4) on $(1, \infty)$ converging to 0 at ∞ and growing to ∞ at $x = 1$ is given by

$$V(x) = \begin{cases} \frac{1}{2} \log \frac{x+1}{x-1}, & \text{if } x > 1, \\ +\infty, & \text{otherwise.} \end{cases} \tag{6.5}$$

Therefore, we have discovered that the value function is given by (6.5) and the blowup set is $\Omega = (1, \infty)$. Finally, the optimal control is $\zeta^*(t) \equiv -1$.

Now we illustrate the use of the maximum principle on this problem. Fix $x \in \Omega$. The optimal control, with assocoiated optimal trajectory ξ^*, satisfies

$$\max_{-1 \leq z \leq 1} p(t)(\xi^2(t) + z) = p(t)\xi^2(t) + |p(t)|,$$

with $\zeta^*(t) = p(t)/|p(t)|$ and $p(t)$ is given by the solution of the adjoint equation

$$\frac{dp}{dt} = -p(t) \cdot 2\xi^*(t), \qquad 0 < t < T(x), \qquad p(T(x)) = 0,$$

where $T(x) = T_x(\zeta^*)$ is the maximal blowup time for position x. The solution of this problem is

$$p(t) = p(0) \exp\left(- \int_0^t 2\xi^*(r)\, dr \right)$$

since $\int_0^{T(x)} \xi^*(r)\, dr = +\infty$. Finally, using condition (5.4) we obtain $1 + p(t)\xi^*(t)^2 + |p(t)| = 0$ and this implies that $p(t) < 0$. Thus, at $t = 0$ we get $p(0) = -1/(x^2 - 1)$ Therefore,

$$p(t) = \frac{-1}{x^2 - 1} \exp\left(- \int_0^t 2\xi^*(r)\, dr \right)$$

is the adjoint variable. The optimal control is again seen to be $\zeta^*(t) = -1$.

Finally we shall look at a two-dimensional problem:

$$\frac{d\xi}{d\tau} = -\eta + \xi(\xi^2 + \eta^2 - \zeta), \tag{6.6}$$

$$\frac{d\eta}{d\tau} = \xi + \eta(\xi^2 + \eta^2 - \zeta), \tag{6.7}$$

if $\tau > 0$ and $(\xi(0), \eta(0)) = (x, y) \in R^2$ We assume that the control set is $Z = [-1, 1]$. The Bellman equation for the value function $V(x, y)$ becomes

$$1 + x\, V_y - y\, V_x + (x^2 + y^2)\,(x\, V_x + y\, V_y) + |x\, V_x + y\, V_y| = 0. \tag{6.8}$$

The feedback control is

$$z = -\frac{x\, V_x + y\, V_y}{|x\, V_x + y\, V_y|}.$$

It is clearly a much more difficult problem to determine the blowup set Ω from equation (6.8) than in the one-dimensional case. Fortunately, we can calculate that the solution of (6.8) satisfying the conditions

$$\lim_{x^2+y^2\to\infty} V(x, y) = 0, \qquad \lim_{(x,y)\to\partial\Omega} V(x, y) = +\infty$$

is given by

$$V(x, y) = \begin{cases} \frac{1}{2}\log\frac{x^2+y^2}{x^2+y^2-1}, & \text{if } x^2 + y^2 > 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

The blowup set is therefore $\Omega = \{x^2 + y^2 > 1\}$ and the optimal control is $\zeta^*(t) \equiv +1$.

7. Some generalizations. 1. The results of this paper carry over with very little change to the more general problem with value function

$$V(x) = \sup_{\zeta\in\mathcal{Z}} \int_0^{T_x(\zeta)} h(\xi(r), \zeta(r))\, dr + g(T_x(\zeta)),$$

where $g(0) = 0$.

2. When we want to minimize the blowup time, $V(x) = \inf_{\zeta\in\mathcal{Z}} T_x(\zeta)$, then the blowup set $\Omega = \{x \in R^n : V(x) < \infty\}$ is now the set in which the blowup time is finite for some control ζ rather than for every control. The Bellman equation is the same as that in §3 with $\max_{z\in Z}$ replaced by $\min_{z\in Z}$.

3. The corresponding differential game also is of interest. In this case we have two opposing players, ζ and η, where ζ is trying to minimize the blowup time and η is attempting to maximize it. We have an upper value, $V^+(x)$, in which player ζ makes the first move. For the lower value, $V^-(x)$, the maximizer makes the first move. The game has value if $V^+ = V^-$. From a practical point of view the differential game is the appropriate model when the designer is very conservative. That is, if one wants to minimize the blowup time under the worst possible circumstances, one should assume that there is an opposing player who has the opposite goal. In many cases the opposing player can be taken as nature.

The upper value V^+ is the viscosity solution of

$$1 + \min_{z \in Z} \max_{y \in Y} D_x V^+(x) \cdot f(x, y, z) = 0$$

and the lower value satisfies the same equation with min and max interchanged. The proofs of these statements, as well as the various continuity results regarding the value functions, are substantially different than those in this paper.

4. It is also possible to consider a nonautonomous version of our problem. That is, suppose that the vector field depends on time, $f = f(t, x, z)$. In that case, we consider the value function as depending on the initial time and state of the problem, $V = V(t, x)$, as does the blowup time, $T_{t,x}(\zeta)$. This would be the first time in the finite time interval $[t, S]$ that the trajectory becomes infinite. If the trajectory does not blow up within that time interval, then the blowup time is defined as S. The Bellman equation becomes

$$V_t + \max_{z \in Z} D_x V \cdot f(t, x, z) = 0, \qquad (t, x) \in \Omega,$$

where the blowup set $\Omega = \{(t, x) \in [0, S) \times R^n : V(t, x) < +\infty\}$. We now have the additional terminal condition that $V(S, x) = S$.

Acknowledgments. The authors are indebted to the referees whose careful reading of the original version of the paper turned up several errors. In addition, this final version was significantly improved by the use of the results of Bardi and Soravia, which allowed us to drop the use of assumption (Aiii) in the proof of continuity of V and the uniqueness of viscosity solutions.

REFERENCES

[1] H. AMANN, *Ordinary Differential Equations*, Walter de Gruyter & Co., Berlin, Germany, 1990.
[2] J. P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, New York, 1984.
[3] M. BARDI, *A boundary value problem for the minimum time function*, SIAM J. Control Optim., 27 (1989), pp. 776–785.
[4] M. BARDI AND M. FALCONE, *An approximation scheme for the minimum time function*, SIAM J. Control Optim., 28 (1990), pp. 950–965.
[5] M. BARDI AND P. SORAVIA, *Hamilton Jacobi equations with singular boundary conditions on a free boundary and applications to differential games*, Trans. Amer. Math. Soc., 325 (1991), pp. 205–229.
[6] M. BARDI AND V. STAICU, *The Bellman equation for time optimal control of non controllable nonlinear systems*, Acta Appl. Math., 31 (1993), pp. 201–223.
[7] E. N. BARRON AND R. JENSEN, *The Pontryagin maximum principle from dynamic programming and viscosity solutions to first order partial differential equations*, Trans. Amer. Math. Soc., 298 (1986), pp. 635–641.
[8] E. N. BARRON, R. JENSEN, AND W. LIU, *Optimal control of the blowup time of a diffusion*, Math. Methods Mod. Appl. Sci., to appear.
[9] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
[10] M. G. CRANDALL, L. C. EVANS, AND P.-L. LIONS, *Some properties of viscosity solutions of Hamilton Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.
[11] M. G. CRANDALL, H. ISHII, AND P.-L. LIONS, *A user's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.
[12] J. L. LIONS, *Control of Distributed Singular Systems*, Gauthier–Villars, Paris, 1985.
[13] W. LIU, *The blowup rate of solutions of semilinear heat equations*, J. Differential Equations, 77 (1989), pp. 104–122.
[14] P. SORAVIA, *Comparison results for Hamilton Jacobi Isaacs equations and applications to pursuit evasion problems*, preprint.
[15] D. TATARU, *Viscosity solutions for the dynamic programming equations*, Appl. Math. Optim., 25 (1992), pp. 109–126.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 124–160, January 1996

006

A SMOOTH CONVERSE LYAPUNOV THEOREM FOR ROBUST STABILITY*

YUANDAN LIN†, EDUARDO D. SONTAG‡, AND YUAN WANG§

Abstract. This paper presents a converse Lyapunov function theorem motivated by robust control analysis and design. Our result is based upon, but generalizes, various aspects of well-known classical theorems. In a unified and natural manner, it (1) allows arbitrary bounded time-varying parameters in the system description, (2) deals with global asymptotic stability, (3) results in smooth (infinitely differentiable) Lyapunov functions, and (4) applies to stability with respect to not necessarily compact invariant sets.

Key words. nonlinear stability, stability with respect to sets, Lyapunov function techniques, robust stability

AMS subject classifications. 93D05, 93D09, 93D20, 34D20

1. Introduction.

This work is motivated by problems of robust nonlinear stabilization. One of our main contributions is to provide a statement and proof of a converse Lyapunov function theorem in a form particularly useful for the study of such feedback control analysis and design problems. We provide a single (and natural) unified result that

1. applies to stability with respect to not necessarily compact invariant sets;
2. deals with global (as opposed to merely local) asymptotic stability;
3. results in smooth (infinitely differentiable) Lyapunov functions;
4. most importantly, applies to stability in the presence of bounded time-varying parameters in the system.

(This last property is sometimes called "total stability" and it is equivalent to the stability of an associated differential inclusion.)

The interest in stability with respect to possibly noncompact sets is motivated by applications to areas such as output control (one needs to stabilize with respect to the zero set of the output variables) and Luenberger-type observer design ("detectability" corresponds to stability with respect to the diagonal set $\{(x, x)\}$, as a subset of the composite state/observer system). Such applications and others are explored in [16, Chap. 5].

Smooth Lyapunov functions, as opposed to merely continuous or once-differentiable ones, are required in order to apply "backstepping" techniques in which a feedback law is built by successively taking directional derivatives of feedback laws obtained for a simplified system. (See for instance [9] for more on backstepping design.)

Finally, the effect of parameter uncertainty and the study of associated Lyapunov functions are topics of interest in robust control theory. An application of the result

* Received by the editors December 9, 1993; accepted for publication (in revised form) August 6, 1994.

† Department of Mathematics, Florida Atlantic University, Boca Raton, FL 33431 (yuandan@polya.math.fau.edu). The research of this author was supported in part by US Air Force grant AFOSR-91-0346.

‡ Department of Mathematics, Rutgers University, New Brunswick, NJ 08903 (sontag@hilbert.rutgers.edu). The research of this author was supported in part by US Air Force grant AFOSR-91-0346.

§ Department of Mathematics, Florida Atlantic University, Boca Raton, FL 33431 (ywang@polya.math.fau.edu). The research of this author was supported in part by NSF grant DMS-9108250.

proved in this paper to the study of "input to state stability" is provided in [27].

1.1. Organization of paper. The paper is organized as follows. The next section provides the basic definitions and the statement of the main result. Actually, two versions are given, one that applies to global asymptotic stability with respect to arbitrary invariant sets, but assuming completeness of the system (that is, global existence of solutions for all inputs) and another version which does not assume completeness but only applies to the special case of compact invariant sets (in particular, to the usual case of global asymptotic stability with respect to equilibria).

Equivalent characterizations of stability by means of decay estimates have proved very useful in control theory (see e.g. [25]) and this is the subject of §3. Some technical facts about Lyapunov functions, including a result on the smoothing of such functions around an attracting set, are given in §4. After this, §5 establishes some basic facts about complete systems needed for the main result.

Section 6 contains the proof of the main result for the general case. Our proof is based upon, and follows to a great extent, the outline of the one given by Wilson in [31], who provided in the late 1960s a converse Lyapunov function theorem for local asymptotic stability with respect to closed sets. There are however some major differences from that work: we want a global rather than a local result, and several technical issues appear in that case; moreover, and most importantly, we have to deal with parameters, which makes the careful analysis of uniform bounds of paramount importance. (In addition, even for the case of no parameters and local stability, several critical steps in the proof are only sketched in [31], especially those concerning Lipschitz properties and smoothness around the attracting set. Later the author of [21] rederived the results, but only for the case when the invariant set is compact. Thus it seems useful to have an expository detailed and self-contained proof in the literature.) A needed technical result on smoothing functions, also based closely on [31], is placed in an appendix for convenience. Section 7 deals with the compact case, essentially by reparameterization of trajectories.

An example, motivated by related work of Tsinias and Kalouptsidis in [7] and [29], is given in §8 to show that the analogous theorems are false for unbounded parameters.

Obviously in a topic such as this one, there are many connections to previous work. While it is likely that we have missed many relevant references, we discuss in §9 some relationships between our work and other results in the literature. Relations to work using "prolongations" are particularly important, and are detailed further in §10.

2. Definitions and statements of main results. Consider the following system:

$$\dot{x}(t) = f(x(t),\, d(t))\,, \tag{1}$$

where for each $t \in \mathbb{R}$, $x(t) \in \mathbb{R}^n$ and $d(t) \in \mathcal{D}$, and where $\mathcal{D}$ is a compact subset of $\mathbb{R}^m$, for some positive integers n and m. The map $f : \mathbb{R}^n \times \mathcal{D} \to \mathbb{R}^n$ is assumed to satisfy the following two properties:

- f is continuous.
- f is locally Lipschitz on x uniformly on d, that is, for each compact subset K of $\mathbb{R}^n$ there is some constant c so that $|f(x,\mathbf{d}) - f(z,\mathbf{d})| \leq c\,|x - z|$ for all $x, z \in K$ and all $\mathbf{d} \in \mathcal{D}$, where $|\cdot|$ denotes the usual Euclidian norm.

Note that these properties are satisfied, for instance, if f extends to a continuously differentiable function on a neighborhood of $\mathbb{R}^n \times \mathcal{D}$.

Let $\mathcal{M}_\mathcal{D}$ be the set of all measurable functions from $\mathbb{R}$ to $\mathcal{D}$. We will call functions $d \in \mathcal{M}_\mathcal{D}$ *time-varying parameters*. For each $d \in \mathcal{M}_\mathcal{D}$, we denote by $x(t, x_0, d)$ (and sometimes simply by $x(t)$ if there is no ambiguity from the context) the solution at time t of (1) with $x(0) = x_0$. This is defined on some maximal interval $(T^-_{x_0,d}, T^+_{x_0,d})$ with $-\infty \leq T^-_{x_0,d} < 0 < T^+_{x_0,d} \leq +\infty$.

Sometimes we will need to consider time-varying parameters d that are defined only on some interval $I \subseteq \mathbb{R}$ with $0 \in I$. In those cases, by abuse of notation, $x(t, x_0, d)$ will still be used, but only times $t \in I$ will be considered.

The system is said to be *forward complete* if $T^+_{x_0,d} = +\infty$ for all x_0 and all $d \in \mathcal{M}_\mathcal{D}$. It is *backward complete* if $T^-_{x_0,d} = -\infty$ for all x_0 and all $d \in \mathcal{M}_\mathcal{D}$, and it is *complete* if it is both forward and backward complete.

We say that a closed set $\mathcal{A}$ is an *invariant set* for (1) if

$$\forall x_0 \in \mathcal{A},\ \forall d \in \mathcal{M}_\mathcal{D},\ T^+_{x_0,d} = +\infty \ \text{ and } \ x(t, x_0, d) \in \mathcal{A}, \quad \forall t \geq 0.$$

Remark 2.1. An equivalent formulation of invariance is in terms of the associated differential inclusion

$$\dot{x} \in F(x)\,, \tag{2}$$

where $F(x) = \{f(x, \mathbf{d}),\ \mathbf{d} \in \mathcal{D}\}$. The set $\mathcal{A}$ is invariant for (1) if and only if it is invariant with respect to (2) (see e.g. [1]). The notions of stability to be considered later can be rephrased in terms of (2) as well.

We will use the following notation: for each nonempty subset $\mathcal{A}$ of $\mathbb{R}^n$ and each $\xi \in \mathbb{R}^n$, we denote

$$|\xi|_\mathcal{A} \stackrel{\text{def}}{=} d(\xi, \mathcal{A}) = \inf_{\eta \in \mathcal{A}} d(\xi, \eta)\,,$$

the common point-to-set distance, and $|\xi|_{\{0\}} = |\xi|$ is the usual norm.

Let $\mathcal{A} \subseteq \mathbb{R}^n$ be a closed, invariant set for (1). We emphasize that we do not require $\mathcal{A}$ to be compact. We will assume throughout this work that the following mild property holds:

$$\sup_{\xi \in \mathbb{R}^n} \{|\xi|_\mathcal{A}\} = \infty\,. \tag{3}$$

This is a minor technical assumption, satisfied in all examples of interest, which will greatly simplify our statements and proofs. (Of course, this property holds automatically whenever $\mathcal{A}$ is compact, and in particular in the important special case in which $\mathcal{A}$ reduces to an equilibrium point.)

DEFINITION 2.2. *System* (1) *is* (absolutely) uniformly globally asymptotically stable (UGAS) *with respect to the closed invariant set* $\mathcal{A}$ *if it is forward complete and the following two properties hold:*

1. *Uniform Stability. There exists a* $\mathcal{K}_\infty$*-function* $\delta(\cdot)$ *such that for any* $\varepsilon \geq 0$,

$$|x(t, x_0, d)|_\mathcal{A} \leq \varepsilon \ \text{ for all } d \in \mathcal{M}_\mathcal{D}, \ \text{ whenever } |x_0|_\mathcal{A} \leq \delta(\varepsilon) \text{ and } t \geq 0\,. \tag{4}$$

2. *Uniform Attraction. For any* $r, \varepsilon > 0$, *there is a* $T > 0$, *such that for every* $d \in \mathcal{M}_\mathcal{D}$,

$$|x(t, x_0, d)|_\mathcal{A} < \varepsilon \tag{5}$$

whenever $|x_0|_\mathcal{A} < r$ *and* $t \geq T$. □

For the definitions of the standard comparison classes of $\mathcal{K}_\infty$- and $\mathcal{KL}$-functions, we refer the reader to the appendices.

Observe that when $\mathcal{A}$ is compact the forward completeness assumption is redundant, since in that case property (4) already implies that all solutions are bounded.

In the particular case in which the set $\mathcal{D}$ consists of just one point, the above definition reduces to the standard notion of set asymptotic stability of differential equations. (Note, however, that this definition differs from those in [3] and [31], which are not global.) If, in addition, $\mathcal{A}$ consists of just an equilibrium point x_0, this is the usual notion of global asymptotic stability for the solution $x(t) \equiv x_0$.

Remark 2.3. It is an easy exercise to verify that an equivalent definition results if one replaces $\mathcal{M}_\mathcal{D}$ by the subset of piecewise constant time-varying parameters.

Remark 2.4. Note that the uniform stability condition is equivalent to the statement that there is a $\mathcal{K}_\infty$-function φ so that

$$|x(t, x_0, d)|_\mathcal{A} \leq \varphi(|x_0|_\mathcal{A}), \quad \forall x_0, \ \forall t \geq 0, \quad \text{and } \forall d \in \mathcal{M}_\mathcal{D}\,.$$

(Just let $\varphi = \delta^{-1}$.)

The following characterization of the UGAS property will be extremely useful.

PROPOSITION 2.5. *The system* (1) *is UGAS with respect to a closed, invariant set* $\mathcal{A} \subseteq \mathbb{R}^n$ *if and only if it is forward complete and there exists a* $\mathcal{KL}$*-function* β *such that, given any initial state* x_0*, the solution* $x(t, x_0, d)$ *satisfies*

$$|x(t, x_0, d)|_\mathcal{A} \leq \beta(|x_0|_\mathcal{A}\,, t)\,, \quad \textit{for any } t \geq 0\,, \tag{6}$$

for any $d \in \mathcal{M}_\mathcal{D}$.

Observe that when $\mathcal{A}$ is compact the forward completeness assumption is again redundant, since in that case property (6) implies that solutions are bounded.

Next we introduce Lyapunov functions with respect to sets. For any differentiable function $V : \mathbb{R}^n \longrightarrow \mathbb{R}$, we use the standard Lie derivative notation

$$L_{f_\mathbf{d}} V(\xi) \stackrel{\text{def}}{=} \frac{\partial V(\xi)}{\partial x} \cdot f_\mathbf{d}(\xi)\,,$$

where for each $\mathbf{d} \in \mathcal{D}$, $f_\mathbf{d}(\cdot)$ is the vector field defined by $f(\cdot, \mathbf{d})$. By "smooth" we always mean infinitely differentiable.

DEFINITION 2.6. *A* Lyapunov function *for the system* (1) *with respect to a nonempty, closed, invariant set* $\mathcal{A} \subseteq \mathbb{R}^n$ *is a function* $V : \mathbb{R}^n \longrightarrow \mathbb{R}$ *such that* V *is smooth on* $\mathbb{R}^n \backslash \mathcal{A}$ *and satisfies*

1. *there exist two* $\mathcal{K}_\infty$*-functions* α_1 *and* α_2 *such that for any* $\xi \in \mathbb{R}^n$,

$$\alpha_1(|\xi|_\mathcal{A}) \leq V(\xi) \leq \alpha_2(|\xi|_\mathcal{A})\,; \tag{7}$$

2. *there exists a continuous, positive definite function* α_3 *such that for any* $\xi \in \mathbb{R}^n \backslash \mathcal{A}$, *and any* $\mathbf{d} \in \mathcal{D}$,

$$L_{f_\mathbf{d}} V(\xi) \leq -\alpha_3(|\xi|_\mathcal{A})\,. \tag{8}$$

A smooth *Lyapunov function is one which is smooth on all of* $\mathbb{R}^n$.

Remark 2.7. Continuity of V on $\mathbb{R}^n \backslash \mathcal{A}$ and property 1 in Definition 2.6 imply:

- V is continuous on all of $\mathbb{R}^n$;
- $V(x) = 0 \iff x \in \mathcal{A}$; and
- $V : \mathbb{R}^n \stackrel{\text{onto}}{\longrightarrow} \mathbb{R}_{\geq 0}$ (recall the assumption in equation (3)).

Our main results will be two converse Lyapunov theorems. The first one is for general closed, invariant sets and assumes completeness of the system.

THEOREM 2.8. *Assume that the system* (1) *is complete. Let* $\mathcal{A} \subseteq \mathbb{R}^n$ *be a nonempty, closed, invariant subset for this system. Then,* (1) *is UGAS with respect to* $\mathcal{A}$ *if and only if there exists a smooth Lyapunov function* V *with respect to* $\mathcal{A}$.

The following result does not assume completeness but instead applies only to compact $\mathcal{A}$.

THEOREM 2.9. *Let* $\mathcal{A} \subseteq \mathbb{R}^n$ *be a nonempty, compact, invariant subset for the system* (1). *Then,* (1) *is UGAS with respect to* $\mathcal{A}$ *if and only if there exists a smooth Lyapunov function* V *with respect to* $\mathcal{A}$.

3. Some preliminaries about UGAS. It will be useful to have a restatement of the second condition in the definition of UGAS stated in terms of uniform attraction times.

LEMMA 3.1. *The uniform attraction property defined in Definition* 2.2 *is equivalent to the following: there exists a family of mappings* $\{T_r\}_{r>0}$ *with*

- *for each fixed* $r > 0$, $T_r : \mathbb{R}_{>0} \xrightarrow{\text{onto}} \mathbb{R}_{>0}$ *is continuous and is strictly decreasing;*
- *for each fixed* $\varepsilon > 0$, $T_r(\varepsilon)$ *is (strictly) increasing as* r *increases and* $\lim_{r\to\infty} T_r(\varepsilon) = \infty$;

such that, for each $d \in \mathcal{M}_{\mathcal{D}}$,

$$|x(t, x_0, d)|_{\mathcal{A}} < \varepsilon \quad \text{whenever } |x_0|_{\mathcal{A}} < r \text{ and } t \geq T_r(\varepsilon)\,. \tag{9}$$

Proof. Sufficiency is clear. Now we show the necessity part. For any $r, \varepsilon > 0$, let

$$A_{r,\varepsilon} \stackrel{\text{def}}{=} \{T \geq 0 : \forall\, |x_0|_{\mathcal{A}} < r,\ \forall t \geq T, \forall d \in \mathcal{M}_{\mathcal{D}},\ |x(t, x_0, d)|_{\mathcal{A}} < \varepsilon\} \subseteq \mathbb{R}_{\geq 0}\,. \tag{10}$$

Then from the assumptions, $A_{r,\varepsilon} \neq \emptyset$ for any $r, \varepsilon > 0$. Moreover,

$$A_{r,\varepsilon_1} \subseteq A_{r,\varepsilon_2}, \text{ if } \varepsilon_1 \leq \varepsilon_2, \text{ and } A_{r_2,\varepsilon} \subseteq A_{r_1,\varepsilon}, \text{ if } r_1 \leq r_2\,.$$

Now define $\bar{T}_r(\varepsilon) \stackrel{\text{def}}{=} \inf A_{r,\varepsilon}$. Then $\bar{T}_r(\varepsilon) < \infty$, for any $r, \varepsilon > 0$, and it satisfies

$$\bar{T}_r(\varepsilon_1) \geq \bar{T}_r(\varepsilon_2), \text{ if } \varepsilon_1 \leq \varepsilon_2, \text{ and } \bar{T}_{r_1}(\varepsilon) \leq \bar{T}_{r_2}(\varepsilon), \text{ if } r_1 \leq r_2.$$

So we can define for any $r, \varepsilon > 0$,

$$\tilde{T}_r(\varepsilon) \stackrel{\text{def}}{=} \frac{2}{\varepsilon}\int_{\varepsilon/2}^{\varepsilon} \bar{T}_r(s)\, ds\,. \tag{11}$$

Since $\bar{T}_r(\cdot)$ is decreasing, $\tilde{T}_r(\cdot)$ is well defined and is locally absolutely continuous. Also

$$\tilde{T}_r(\varepsilon) \geq \frac{2}{\varepsilon}\bar{T}_r(\varepsilon)\int_{\varepsilon/2}^{\varepsilon} ds = \bar{T}_r(\varepsilon)\,. \tag{12}$$

Furthermore,

$$\begin{aligned}
\frac{d\tilde{T}_r(\varepsilon)}{d\varepsilon} &= -\frac{2}{\varepsilon^2}\int_{\varepsilon/2}^{\varepsilon} \bar{T}_r(s)\, ds + \frac{2}{\varepsilon}\left(\bar{T}_r(\varepsilon) - \frac{1}{2}\bar{T}_r\left(\frac{\varepsilon}{2}\right)\right) \\
&= \frac{1}{\varepsilon}\left[\bar{T}_r(\varepsilon) - \frac{2}{\varepsilon}\int_{\varepsilon/2}^{\varepsilon} \bar{T}_r(s)\, ds\right] + \frac{1}{\varepsilon}\left[\bar{T}_r(\varepsilon) - \bar{T}_r\left(\frac{\varepsilon}{2}\right)\right] \\
&= \frac{1}{\varepsilon}\left[\bar{T}_r(\varepsilon) - \tilde{T}_r(\varepsilon)\right] + \frac{1}{\varepsilon}\left[\bar{T}_r(\varepsilon) - \bar{T}_r\left(\frac{\varepsilon}{2}\right)\right] \leq 0, \quad \text{a.e.},
\end{aligned} \tag{13}$$

hence $\tilde{T}_r(\cdot)$ decreases (not necessarily strictly). Since $\bar{T}_{(\cdot)}(\varepsilon)$ increases, from the definition, $\tilde{T}_{(\cdot)}(\varepsilon)$ also increases. Finally, define

$$T_r(\varepsilon) \stackrel{\text{def}}{=} \tilde{T}_r(\varepsilon) + \frac{r}{\varepsilon}. \tag{14}$$

Then it follows that

- for any fixed r, $T_r(\cdot)$ is continuous, maps $\mathbb{R}_{>0} \stackrel{\text{onto}}{\longrightarrow} \mathbb{R}_{>0}$, and is strictly decreasing;
- for any fixed ε, $T_r(\varepsilon)$ is increasing as r increases, and $\lim_{r\to\infty} T_r(\varepsilon) = \infty$.

So the only thing left to be shown is that T_r defined by (14) satisfies (9). To do this, pick any x_0 and t with $|x_0|_{\mathcal{A}} < r$ and $t \geq T_r(\varepsilon)$. Then

$$t \geq T_r(\varepsilon) > \tilde{T}_r(\varepsilon) \geq \bar{T}_r(\varepsilon)\,.$$

Hence, by the definition of $\bar{T}_r(\varepsilon)$, $|x(t, x_0, d)|_{\mathcal{A}} < \varepsilon$, as claimed. □

3.1. Proof of characterization via decay estimate. We now provide a proof of Proposition 2.5.

[$\Longleftarrow$] Assume that there exists a $\mathcal{KL}$-function β such that (6) holds. Let

$$c_1 \stackrel{\text{def}}{=} \sup \beta(\cdot, 0) \leq \infty\,,$$

and choose $\delta(\cdot)$ to be any $\mathcal{K}_\infty$-function with

$$\delta(\varepsilon) \leq \bar{\beta}^{-1}(\varepsilon), \quad \text{for any } 0 \leq \varepsilon < c_1\,,$$

where $\bar{\beta}^{-1}$ denotes the inverse function of $\bar{\beta}(\cdot) \stackrel{\text{def}}{=} \beta(\cdot, 0)$. (If $c_1 = \infty$, we can simply choose $\delta(\varepsilon) \stackrel{\text{def}}{=} \bar{\beta}^{-1}(\varepsilon)$.) Clearly $\delta(\varepsilon)$ is the desired $\mathcal{K}_\infty$-function for the uniform stability property.

The uniform attraction property follows from the fact that for every fixed r, $\lim_{t\to\infty} \beta(r, t) = 0$.

[$\Longrightarrow$] Assume that (1) is UGAS with respect to the closed set $\mathcal{A}$, and let δ be as in the definition. Let $\varphi(\cdot)$ be the $\mathcal{K}$-function $\delta^{-1}(\cdot)$. As mentioned in Remark 2.4, it follows that $|x(t, x_0, d)|_{\mathcal{A}} \leq \varphi(|x_0|_{\mathcal{A}})$ for any $x_0 \in \mathbb{R}^n$, any $t \geq 0$, and any $d \in \mathcal{M}_{\mathcal{D}}$.

Let $\{T_r\}_{r\in(0,\infty)}$ be as in Lemma 3.1, and for each $r \in (0, \infty)$ denote $\psi_r \stackrel{\text{def}}{=} T_r^{-1}$. Then, for each $r \in (0, \infty)$, $\psi_r : \mathbb{R}_{>0} \longrightarrow \mathbb{R}_{>0}$ is again continuous, onto, and strictly decreasing. We also write $\psi_r(0) = +\infty$, which is consistent with the fact that

$$\lim_{t\to 0^+} \psi_r(t) = +\infty\,.$$

(Note: The property that $T_{(\cdot)}(t)$ increases to ∞ is not needed here.)

CLAIM. *For any $|x_0|_{\mathcal{A}} < r$, any $t \geq 0$, and any $d \in \mathcal{M}_{\mathcal{D}}$, $|x(t, x_0, d)|_{\mathcal{A}} \leq \psi_r(t)$.*

Proof. It follows from the definition of the maps T_r that, for any r, $\varepsilon > 0$, and for any $d \in \mathcal{M}_{\mathcal{D}}$,

$$|x_0|_{\mathcal{A}} < r, \quad t \geq T_r(\varepsilon) \implies |x(t, x_0, d)|_{\mathcal{A}} < \varepsilon\,.$$

As $t = T_r(\psi_r(t))$ if $t > 0$, we have, for any such x_0 and d,

$$|x(t, x_0, d)|_{\mathcal{A}} < \psi_r(t)\,, \ \forall t > 0\,. \tag{15}$$

The claim follows by combining (15) and the fact that $\psi_r(0) = +\infty$. □

Now for any $s \geq 0$ and $t \geq 0$, let

$$\bar{\psi}(s, t) \stackrel{\text{def}}{=} \min\left\{\varphi(s), \inf_{r \in (s, \infty)} \psi_r(t)\right\}. \tag{16}$$

Because of the definition of φ and the above claim, we have, for each x_0, $d \in \mathcal{M}_{\mathcal{D}}$, and $t \geq 0$,

$$|x(t, x_0, d)|_{\mathcal{A}} \leq \bar{\psi}(|x_0|_{\mathcal{A}}, t). \tag{17}$$

If $\bar{\psi}$ were of class $\mathcal{KL}$, we would be done. This may not be the case, so we next majorize $\bar{\psi}$ by such a function.

By its definition, for any fixed t, $\bar{\psi}(\cdot, t)$ is an increasing function (not necessarily strictly). Also, because for any fixed $r \in (0, \infty)$, $\psi_r(t)$ decreases to 0 (this follows from the fact that $\psi_r : \mathbb{R}_{>0} \stackrel{\text{onto}}{\longrightarrow} \mathbb{R}_{>0}$ is continuous and strictly decreasing), it follows that

$$\text{for any fixed } s, \quad \bar{\psi}(s, t) \text{ decreases to 0 as } t \to \infty.$$

Next we construct a function $\tilde{\psi} : \mathbb{R}_{[0, \infty)} \times \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ with the following properties:

- for any fixed $t \geq 0$, $\tilde{\psi}(\cdot, t)$ is continuous and strictly increasing;
- for any fixed $s \geq 0$, $\tilde{\psi}(s, t)$ decreases to 0 as $t \to \infty$;
- $\tilde{\psi}(s, t) \geq \bar{\psi}(s, t)$.

Such a function $\tilde{\psi}$ always exists; for instance, it can be obtained as follows. Define first

$$\hat{\psi}(s, t) \stackrel{\text{def}}{=} \int_s^{s+1} \bar{\psi}(\varepsilon, t)\, d\varepsilon. \tag{18}$$

Then $\hat{\psi}(\cdot, t)$ is an absolutely continuous function on every compact subset of $\mathbb{R}_{\geq 0}$, and it satisfies

$$\hat{\psi}(s, t) \geq \bar{\psi}(s, t) \int_s^{s+1} d\varepsilon = \bar{\psi}(s, t).$$

It follows that

$$\frac{\partial \hat{\psi}(s, t)}{\partial s} = \bar{\psi}(s+1, t) - \bar{\psi}(s, t) \geq 0, \text{ a.e.},$$

and hence $\hat{\psi}(\cdot, t)$ is increasing. Also since for any fixed s, $\bar{\psi}(s, \cdot)$ decreases, so does $\hat{\psi}(s, \cdot)$. Note that

$$\bar{\psi}(s, t) \leq \bar{\psi}(s, 0) = \min\left\{\inf_{r \in (s, \infty)} \psi_r(0), \varphi(s)\right\} = \varphi(s)$$

(recall that $\psi_r(0) = +\infty$), so by the Lebesgue-dominated convergence theorem, for any fixed $s \geq 0$,

$$\lim_{t \to \infty} \hat{\psi}(s, t) = \int_s^{s+1} \lim_{t \to \infty} \bar{\psi}(\varepsilon, t)\, d\varepsilon = 0.$$

Now we see that the function $\hat{\psi}(s, t)$ satisfies all of the requirements for $\tilde{\psi}(s, t)$ except possibly for the strictly increasing property. We define $\tilde{\psi}$ as follows:

$$\tilde{\psi}(s, t) \stackrel{\text{def}}{=} \hat{\psi}(s, t) + \frac{s}{(s+1)(t+1)}.$$

Clearly it satisfies all the desired properties.

Finally, define

$$\beta(s, t) \stackrel{\text{def}}{=} \sqrt{\varphi(s)}\sqrt{\tilde{\psi}(s, t)}.$$

Then it follows that $\beta(s, t)$ is a $\mathcal{KL}$-function, and for all x_0, t, d,

$$|x(t, x_0, d)|_{\mathcal{A}} \leq \sqrt{\varphi(|x_0|_{\mathcal{A}})}\sqrt{\tilde{\psi}(|x_0|_{\mathcal{A}}, t)} \leq \beta(|x_0|_{\mathcal{A}}, t),$$

which concludes the proof of Proposition 2.5.

4. Some preliminaries about Lyapunov functions. In this section we provide some technical results about set Lyapunov functions. A lemma on differential inequalities is also given, for later reference.

Remark 4.1. One may assume in Definition 2.6 that all of $\alpha_1, \alpha_2, \alpha_3$ are smooth in $(0, +\infty)$ and of class $\mathcal{K}_\infty$. For α_1 and α_2, this is proved simply by finding two functions $\tilde{\alpha}_1, \tilde{\alpha}_2$ in $\mathcal{K}_\infty$, smooth in $(0, +\infty)$ so that

$$\tilde{\alpha}_1(s) \leq \alpha_1(s) \leq \alpha_2(s) \leq \tilde{\alpha}_2(s), \text{ for all } s.$$

For α_3, a new Lyapunov function W and a function $\tilde{\alpha}_3$ which satisfies (8) with respect to W, but is smooth in $(0, +\infty)$ and of class $\mathcal{K}_\infty$, can be constructed as follows. First, pick $\tilde{\alpha}_3$ to be any $\mathcal{K}_\infty$-function, smooth in $(0, +\infty)$, such that

$$\tilde{\alpha}_3(s) \leq s\alpha_3(s), \ \forall s \in [0, \alpha_1^{-1}(1)].$$

This is possible since α_3 is positive definite. Then let

$$\gamma : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$$

be a $\mathcal{K}_\infty$-function, smooth in $(0, +\infty)$, such that

- $\gamma(r) \geq \alpha_1^{-1}(r)$ for all $r \in [0, 1]$;
- $\gamma(r) > \tilde{\alpha}_3(\alpha_1^{-1}(r))/\alpha_3(\alpha_1^{-1}(r))$ for all $r > 1$.

Now define $\beta(s) \stackrel{\text{def}}{=} \int_0^s \gamma(r)\, dr$. Note that β is a $\mathcal{K}_\infty$-function, smooth in $(0, +\infty)$. Let $W(\xi) \stackrel{\text{def}}{=} \beta(V(\xi))$. This is smooth on $\mathbb{R}^n \backslash \mathcal{A}$, and $\beta \circ \alpha_1$, $\beta \circ \alpha_2$ bound W as in equation (7). Moreover,

$$\beta'(V(\xi)) = \gamma(V(\xi)) \geq \gamma(\alpha_1(|\xi|_{\mathcal{A}})),$$

so

$$L_{f_d} W(\xi) = \beta'(V(\xi)) L_{f_d} V(\xi) \leq -\gamma(\alpha_1(|\xi|_{\mathcal{A}}))\alpha_3(|\xi|_{\mathcal{A}}). \tag{19}$$

We claim that this is bounded by $-\tilde{\alpha}_3(|\xi|_{\mathcal{A}})$. Indeed, if $s \stackrel{\text{def}}{=} |\xi|_{\mathcal{A}} \leq \alpha_1^{-1}(1)$, then from the first item above and the definition of $\tilde{\alpha}_3$,

$$\gamma(\alpha_1(s)) \geq s \geq \frac{\tilde{\alpha}_3(s)}{\alpha_3(s)};$$

if instead $s > \alpha_1^{-1}(1)$, then from the second item, also

$$\gamma(\alpha_1(s)) \geq \frac{\tilde{\alpha}_3(s)}{\alpha_3(s)}\,.$$

In either case, $\gamma(\alpha_1(s))\alpha_3(s) \geq \tilde{\alpha}_3(s)$, as desired. From now on, whenever necessary, we assume that α_1, α_2, α_3 are $\mathcal{K}_\infty$-functions, smooth in $(0, +\infty)$.

4.1. Smoothing of Lyapunov functions. When dealing with control system design, one often needs to know that V can be taken to be globally smooth, rather than just smooth outside of $\mathcal{A}$.

PROPOSITION 4.2. *If there is a Lyapunov function for* (1) *with respect to* $\mathcal{A}$, *then there is also a smooth such Lyapunov function.*

The proof relies on constructing a smooth function of the form $W = \beta \circ V$, where

$$\beta : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$$

is built using a partition of unity.

Again let $\mathcal{A} \subseteq \mathbb{R}^n$ be nonempty and closed. For a multi-index $\varrho = (\varrho_1, \varrho_2, \ldots, \varrho_n)$, we use $|\varrho|$ to denote $\sum_{i=1}^n \varrho_i$. The following regularization result will be needed; it generalizes to arbitrary $\mathcal{A}$ the analogous (but simpler, due to compactness) result for equilibria given in [13, Thm. 6].

LEMMA 4.3. *Assume that* $V : \mathbb{R}^n \longrightarrow \mathbb{R}_{\geq 0}$ *is* C^0, *the restriction* $V|_{\mathbb{R}^n \setminus \mathcal{A}}$ *is* C^∞, *and also* $V|_{\mathcal{A}} = 0$, $V|_{\mathbb{R}^n \setminus \mathcal{A}} > 0$. *Then there exists a* $\mathcal{K}_\infty$*-function* β, *smooth on* $(0, \infty)$ *and so that* $\beta^{(i)}(t) \to 0$ *as* $t \to 0^+$ *for each* $i = 0, 1, \ldots$, *and having* $\beta'(t) > 0$, $\forall t > 0$, *such that*

$$W \stackrel{\text{def}}{=} \beta \circ V$$

is a C^∞ *function on all of* $\mathbb{R}^n$.

Proof. Let $K_1, K_2, \ldots$, be compact subsets of $\mathbb{R}^n$ such that $\mathcal{A} \subseteq \bigcup_{i=1}^\infty \operatorname{int}(K_i)$. For any $k \geq 1$, let

$$I_k \stackrel{\text{def}}{=} \left(\frac{1}{k+2}, \frac{1}{k}\right) \subseteq \mathbb{R}$$

and $I_0 \stackrel{\text{def}}{=} I_1$. Pick for any $k \geq 1$ a smooth (C^∞) function $\gamma_k : \mathbb{R}_{>0} \to [0, 1]$ satisfying

- $\gamma_k(t) = 0$ if $t \notin I_k$; and
- $\gamma_k(t) > 0$ if $t \in I_k$.

Define for any $k \geq 1$,

$$\mathcal{G}_k \stackrel{\text{def}}{=} \left\{ x \in \mathbb{R}^n : \; x \in \bigcup_{i=1}^k K_i, \; V(x) \in \operatorname{clos} I_k \right\}.$$

Then $\mathcal{G}_k$ is compact (because of compactness of the sets K_i and continuity of V). Observe that each derivative $\gamma_k^{(i)}$ has a compact support included in $\operatorname{clos} I_k$, so it is bounded. For each $k = 1, 2, \ldots$, let $c_k \in \mathbb{R}$ satisfy

1. $c_k \geq 1$;
2. $c_k \geq |(D^\varrho V)(x)|$ for any multi-index $|\varrho| \leq k$ and any $x \in \mathcal{G}_k$; and
3. $c_k \geq |\gamma_k^{(i)}(t)|$, for any $i \leq k$ and any $t \in \mathbb{R}_{>0}$.

Choose the sequence d_k to satisfy

$$0 < d_k < \frac{1}{2^k(k+1)!c_k^k}\,,\ k = 1, 2, \ldots. \tag{20}$$

Let $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a C^∞ function such that $\alpha \equiv 0$ on $\left[0, \frac{1}{3}\right]$ and $\alpha \geq 1$ on $\left[\frac{1}{2}, \infty\right)$. Define $\gamma(0) \stackrel{\text{def}}{=} 0$ and

$$\gamma(t) \stackrel{\text{def}}{=} \sum_{k=1}^{\infty} d_k \gamma_k(t) + \alpha(t)\,,\ \forall t > 0\,. \tag{21}$$

Notice that for any $t \in (0, 1)$, if $k \stackrel{\text{def}}{=} \left\lfloor \frac{1}{t} \right\rfloor \geq 1$ denotes the largest integer $\leq \frac{1}{t}$, then $t \in I_{k-1}$ and

$$t \notin I_j \quad \text{if} \quad j \neq k, k-1\,.$$

Hence the sum in (21) consists of at most three terms (for $t \geq 1$ the sum is just $\gamma = \alpha$), and so γ is C^∞ at each $t \in (0, \infty)$.

CLAIM. *For any $i \geq 0$, $\lim_{t \to 0^+} \gamma^{(i)}(t) = 0$.*

Proof. Fix any $i \geq 0$. Given any $\varepsilon > 0$, let $k_0 \in \mathbb{Z}$ be such that $\varepsilon > \frac{1}{k_0} > 0$. Let

$$T \stackrel{\text{def}}{=} \min\left\{\frac{1}{k_0}, \frac{1}{i+1}, \frac{1}{3}\right\}.$$

We will show that $t \in (0, T) \implies \left|\gamma^{(i)}(t)\right| < \varepsilon$. Indeed, as $0 < t < \min\left\{\frac{1}{k_0}, \frac{1}{i+1}, \frac{1}{3}\right\}$, it follows that $k \stackrel{\text{def}}{=} \left\lfloor \frac{1}{t} \right\rfloor \geq \max\{i+1, k_0, 3\}$. So

$$\gamma^{(i)}(t) \leq d_{k-1}\gamma_{k-1}^{(i)}(t) + d_k \gamma_k^{(i)}(t)\,,$$

and noticing that

$$i \leq k-1 < k \implies c_k \geq \left|\gamma_k^{(i)}(t)\right|\,,\ c_{k-1} \geq \left|\gamma_{k-1}^{(i)}(t)\right|\,,$$

we have

$$\left|\gamma^{(i)}(t)\right| \leq d_{k-1}c_{k-1} + d_k c_k \leq \frac{1}{2k!} + \frac{1}{2(k+1)!} < \frac{1}{k!} < \frac{1}{k} \leq \frac{1}{k_0} < \varepsilon\,,$$

as wanted.

Note also that if $t \geq \frac{1}{2}$, then $\gamma(t) \geq \alpha(t) \geq 1 > 0$; and if $t \in \left(0, \frac{1}{2}\right)$, then $\gamma(t) \geq d_{k-1}\gamma_{k-1}(t) > 0$ with $k \stackrel{\text{def}}{=} \left\lfloor \frac{1}{t} \right\rfloor \geq 2$, so the function

$$\beta(t) \stackrel{\text{def}}{=} \int_0^t \gamma(s)\, ds \tag{22}$$

is also a $\mathcal{K}_\infty$-function, smooth on $(0, \infty)$. Furthermore, β satisfies $\beta^{(i)}(t) \to 0$ as $t \to 0^+$ for each $i = 0, 1, \ldots$.

Finally, we show that $W = \beta \circ V$ is C^∞. For this, it is enough to show that $D^{\varrho_0} W(x_n) \to 0$ as $x_n \to \bar{x} \in \partial\mathcal{A}$, for each multi-index ϱ_0 and each sequence $\{x_n\} \subseteq$

$\mathbb{R}^n \backslash \mathcal{A}$ converging to a point $\bar{x}$ in the boundary of $\mathcal{A}$. (In general—see, e.g., [4, p. 52]—if $\mathcal{A} \subseteq \mathbb{R}^n$ is closed and $\varphi : \mathbb{R}^n \longrightarrow \mathbb{R}$ satisfies that $\varphi|_{\mathcal{A}} = 0$, $\varphi|_{\mathbb{R}^n \backslash \mathcal{A}}$ is C^∞, and for each boundary point a of $\mathcal{A}$ and all multi-indices $\varrho = (\varrho_1, \varrho_2, \ldots, \varrho_n)$, it holds that $\lim_{\substack{x \to a \\ x \notin \mathcal{A}}} D^\varrho \varphi(x) = 0$, then φ is C^∞ on $\mathbb{R}^n$.)

Pick one such ϱ_0 and any sequence $\{x_n\}$ with $x_n \to \bar{x} \in \partial\mathcal{A}$. If $|\varrho_0| = 0$, one only needs to show that $W(x_n) \to 0$, which follows easily from the fact that $\beta \in \mathcal{K}_\infty$ and $V(x_n) \to 0$. So from now on, we can assume that $|\varrho_0| \stackrel{\text{def}}{=} i \geq 1$. As $\mathcal{A} \subseteq \cup_{j=0}^{\infty} \text{int}\, K_j$, $\bar{x} \in \text{int}\, K_l$ for some l, and without loss of generality we may assume that there is some fixed l so that

$$\cdot x_n \in K_l, \quad \text{for all } n.$$

Pick any $\varepsilon > 0$. We will show that there exists some N such that

$$n > N \implies |D^{\varrho_0} W(x_n)| < \varepsilon\,.$$

Let $k \in \mathbb{Z}$ be so that

$$k > \max\left\{i,\, \log_2\left(\frac{1}{\varepsilon}\right),\, l\right\}$$

and let $T \in (0, \frac{1}{3})$ be such that $T < \frac{1}{k+2}$. Observe that if $t < T$, then $t \notin I_1 \cup \cdots \cup I_k$.

As V is C^0 everywhere, $V = 0$ at $\mathcal{A}$, $V(x_n) \to V(\bar{x}) = 0$. So there exists N such that $V(x_n) < T$ whenever $n > N$. Fix an N like this. Then for any $n > N$,

$$\gamma_s^{(j)}\left(V(x_n)\right) = 0,\ \forall j,\ \forall s = 1, 2, \ldots, k$$

(since γ_s vanishes outside I_s). Pick any $j \in \mathbb{N}$ with $j \leq i$, any $h \in \mathbb{N}$ with $h \leq i$, and $\varrho_1, \ldots, \varrho_h$ multi-indices such that $|\varrho_\mu| \leq i,\ \forall \mu = 1, \ldots, h$. Then for any $q \in \mathbb{N}$ with $q > k$, by the way we chose c_k,

$$\left|\gamma_q^{(j)}\left(V(x_n)\right)\right| \leq c_q\,,$$

since $q > k > i \geq j$. Also, if $V(x_n) \in I_q$, then again by the properties of the sequence c_k,

$$|D^{\varrho_\mu} V(x_n)| \leq c_q$$

(since $q > k > l$ and $x_n \in K_l$ imply $x_n \in K_1 \cup \cdots \cup K_q$, and $|\varrho_\mu| \leq i < k < q$). Therefore, for such q, if $V(x_n) \in I_q$,

$$\left|\gamma_q^{(j)}\left(V(x_n)\right)\right| |D^{\varrho_1} V(x_n)| \cdots |D^{\varrho_h} V(x_n)| \leq c_q^{h+1} \leq c_q^{i+1} < c_q^q\,. \tag{23}$$

If instead it were the case that $V(x_n) \notin I_q$, then $\gamma_q^{(j)}\left(V(x_n)\right) = 0$, and hence the inequality (23) still holds. Since

$$\gamma^{(j)}\left(V(x_n)\right) = \sum_{q=k+1}^{\infty} d_q \gamma_q^{(j)}\left(V(x_n)\right)\,,$$

we also have

$$\left|\gamma^{(j)}\left(V(x_n)\right)\right| \left|D^{\varrho_1}V(x_n)\right| \cdots \left|D^{\varrho_h}V(x_n)\right| \leq \sum_{q=k+1}^{\infty} d_q c_q^q < \sum_{q=k+1}^{\infty} \frac{1}{2^q(q+1)!}$$

$$< \left(\sum_{q=k+1}^{\infty} \frac{1}{2^q}\right) \frac{1}{(k+1)!} = \frac{1}{2^k(k+1)!} < \frac{\varepsilon}{(k+1)!}\,. \tag{24}$$

Now observe that

$$\left(D^{\varrho_0}W\right)(x) = \left(D^{\varrho_0}(\beta \circ V)\right)(x)$$

is a sum of $\leq i!$ terms (recall $0 < i = |\varrho_0|$), each of which is of the form

$$\beta^{(p)}\left(V(x)\right)\left(D^{\varrho_1}V\right)(x) \cdots \left(D^{\varrho_h}V\right)(x)\,,$$

where $0 < p \leq i$, $h \leq i$, and each $|\varrho_\mu| \leq i$. Each

$$\beta^{(p)}\left(V(x)\right) = \gamma^{(j)}\left(V(x)\right)\,,\ j = p-1 \leq i-1\,,$$

so (24) applies, and we conclude

$$\left|\left(D^{\varrho_0}W\right)(x_n)\right| \leq i!\frac{\varepsilon}{(k+1)!} < \varepsilon\,,$$

(since $k > i$.) □

Now let us return to the proof of Proposition 4.2.

Proof of Proposition 4.2. Assume $\mathcal{A}$, V, and $\alpha_1, \alpha_2, \alpha_3$ are as defined in Definition 2.6. Let β, W be as in Lemma 4.3. We show that W is a smooth Lyapunov function as required.

Let $\hat{\alpha}_i \stackrel{\text{def}}{=} \beta \circ \alpha_i, i = 1, 2$. These are again $\mathcal{K}_\infty$-functions, and they satisfy

$$\hat{\alpha}_1(|\xi|_{\mathcal{A}}) \leq W(\xi) \leq \hat{\alpha}_2(|\xi|_{\mathcal{A}})\,.$$

We define, for $s > 0$,

$$\check{\beta}(s) \stackrel{\text{def}}{=} \min_{t \in [\alpha_1(s),\, \alpha_2(s)]} \beta'(t) > 0\,.$$

Also let $\check{\beta}(0) \stackrel{\text{def}}{=} 0$. Define $\hat{\alpha}_3(s) \stackrel{\text{def}}{=} \check{\beta}(s)\alpha_3(s)$. Then $\hat{\alpha}_3$ is a continuous, positive definite function. Also, for any $\xi \in \mathbb{R}^n \backslash \mathcal{A}$,

$$\begin{aligned} L_{f_{\mathrm{d}}}W(\xi) = \beta'(V(\xi))L_{f_{\mathrm{d}}}V(\xi) &\leq -\beta'(V(\xi))\alpha_3(|\xi|_{\mathcal{A}}) \\ &\leq -\check{\beta}(|\xi|_{\mathcal{A}})\alpha_3(|\xi|_{\mathcal{A}}) = -\hat{\alpha}_3(|\xi|_{\mathcal{A}})\,, \end{aligned}$$

which concludes the proof of Proposition 4.2. □

4.2. A useful estimate. The following lemma establishes a useful comparison principle.

LEMMA 4.4. *For each continuous and positive definite function α, there exists a $\mathcal{KL}$-function $\beta_\alpha(s, t)$ with the following property: if $y(\cdot)$ is any (locally) absolutely*

continuous function defined for $t \geq 0$ and with $y(t) \geq 0$ for all t, and $y(\cdot)$ satisfies the differential inequality

$$\dot{y}(t) \leq -\alpha(y(t)), \quad \textit{for almost all} \quad t \tag{25}$$

with $y(0) = y_0 \geq 0$, then it holds that

$$y(t) \leq \beta_\alpha(y_0, t)$$

for all $t \geq 0$.

Proof. Define for any $s > 0$, $\eta(s) \stackrel{\text{def}}{=} -\int_1^s \frac{dr}{\alpha(r)}$. This is a strictly decreasing differentiable function on $(0, \infty)$. Without loss of generality, we will assume that $\lim_{s\to 0^+} \eta(s) = +\infty$. If this were not the case, we could consider instead the following function:

$$\bar{\alpha}(s) \stackrel{\text{def}}{=} \min\{s, \alpha(s)\}.$$

This function is again continuous, positive definite, satisfies $\bar{\alpha}(s) \leq \alpha(s)$ for any $s \geq 0$, and

$$\lim_{s\to 0^+} \int_s^1 \frac{dr}{\bar{\alpha}(r)} \geq \lim_{s\to 0^+} \int_s^1 \frac{dr}{r} = +\infty .$$

Moreover, if $\dot{y}(t) \leq -\alpha(y(t))$ then also $\dot{y}(t) \leq -\bar{\alpha}(y(t))$, so $\beta_{\bar{\alpha}}$ could be used to bound solutions.

Let

$$0 < a \stackrel{\text{def}}{=} -\lim_{s\to+\infty} \eta(s) .$$

Then the range of η, and hence also the domain of η^{-1}, is the open interval $(-a, \infty)$. (We allow the possibility that $a = \infty$.) For $(s, t) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$, define

$$\beta_\alpha(s, t) \stackrel{\text{def}}{=} \begin{cases} 0, & \text{if } s = 0, \\ \eta^{-1}(\eta(s) + t), & \text{if } s > 0. \end{cases}$$

We claim that for any $y(\cdot)$ satisfying the conditions in the lemma,

$$y(t) \leq \beta_\alpha(y_0, t), \quad \text{for all } t \geq 0 . \tag{26}$$

As $\dot{y}(t) \leq -\alpha(y(t))$, it follows that $y(t)$ is nonincreasing, and if $y(t_0) = 0$ for some $t_0 \geq 0$, then $y(t) \equiv 0$, $\forall t \geq t_0$. Without loss of generality, assume that $y_0 > 0$. Let

$$t_0 \stackrel{\text{def}}{=} \inf\{t : \ y(t) = 0\} \leq +\infty .$$

It is enough to show (26) holds for $t \in [0, t_0)$.

As η is strictly decreasing, we only need to show that $\eta(y(t)) \geq \eta(y_0) + t$, that is,

$$-\int_1^{y(t)} \frac{dr}{\alpha(r)} \geq -\int_1^{y_0} \frac{dr}{\alpha(r)} + t ,$$

which is equivalent to

$$\int_{y(t)}^{y_0} \frac{dr}{\alpha(r)} \geq t . \tag{27}$$

From (25), one sees that

$$\int_0^t \frac{\dot{y}(\tau)}{\alpha(y(\tau))}\, d\tau \;\leq\; -\int_0^t d\tau \;=\; -t\,.$$

Changing variables in the integral, this gives (27).

It only remains to show that β_α is of class $\mathcal{KL}$. The function β_α is continuous since both η and η^{-1} are continuous in their domains, and $\lim_{r\to\infty} \eta^{-1}(r) = 0$. It is strictly increasing in s for each fixed t since both η and η^{-1} are strictly decreasing. Finally, $\beta_\alpha(s,\, t) \to 0$ as $t \to \infty$ by construction. So β_α is a $\mathcal{KL}$-function. □

5. Some properties of complete systems. We first need to establish some technical properties that hold for complete systems, and in particular a Lipschitz continuity fact.

For each $\xi \in \mathbb{R}^n$ and $T > 0$, let

$$\mathcal{R}^T(\xi) \stackrel{\text{def}}{=} \{\eta :\; \eta = x(T,\, \xi, d),\; d \in \mathcal{M}_{\mathcal{D}}\}\,.$$

This is the reachable set of (1) from ξ at time T. We use $\mathcal{R}^{\leq T}(\xi)$ to denote $\bigcup_{0\leq t\leq T} \mathcal{R}^t(\xi)$. If S is a subset of $\mathbb{R}^n$, we write

$$\mathcal{R}^T(S) \stackrel{\text{def}}{=} \bigcup_{\xi\in S} \mathcal{R}^T(\xi)\,,\quad \mathcal{R}^{\leq T}(S) \stackrel{\text{def}}{=} \bigcup_{\xi\in S} \mathcal{R}^{\leq T}(\xi)\,.$$

In what follows we use $\overline{S}$ to denote the closure of S for any subset S of $\mathbb{R}^n$.

PROPOSITION 5.1. *Assume that* (1) *is forward complete. Then for any compact subset* K *of* $\mathbb{R}^n$ *and any* $T > 0$, *the set* $\overline{\mathcal{R}^{\leq T}(K)}$ *is compact.*

To prove Proposition 5.1, we first need to make a couple of technical observations.

LEMMA 5.2. *Let* K *be a compact subset of* $\mathbb{R}^n$ *and let* $T > 0$. *Then the set* $\overline{\mathcal{R}^{\leq T}(K)}$ *is compact if and only if* $\overline{\mathcal{R}^{\leq T}(\xi)}$ *is compact for each* $\xi \in K$.

Proof. It is clear that the compactness of $\overline{\mathcal{R}^{\leq T}(K)}$ implies the compactness of $\overline{\mathcal{R}^{\leq T}(\xi)}$ for any $\xi \in K$.

Now assume, for $T > 0$ and a compact set K, that $\overline{\mathcal{R}^{\leq T}(\xi)}$ is compact for each $\xi \in K$. Pick any $\xi \in K$, and let $\mathcal{U} = \{\eta :\; d(\eta,\, \overline{\mathcal{R}^{\leq T}(\xi)}) < 1\}$. Then $\overline{\mathcal{U}}$ is compact. Let C be a Lipschitz constant for f with respect to x on $\overline{\mathcal{U}}$, and let $r = e^{-CT}$. For each $d \in \mathcal{M}_{\mathcal{D}}$ and each η with $|\eta - \xi| < r$, let $\tilde{t} = \inf\{t \geq 0 :\; |x(t,\, \eta,\, d) - x(t,\, \xi,\, d)| \geq 1\}$. Then, using Gronwall's lemma, one can show that $\tilde{t} \geq T$, from which it follows that

$$\mathcal{R}^{\leq T}(\eta) \subseteq \overline{\mathcal{U}}, \quad \forall |\eta - \xi| < r.$$

Thus, for each $\xi \in K$, there is a neighborhood $\mathcal{V}_\xi$ of ξ such that $\overline{\mathcal{R}^{\leq T}(\mathcal{V}_\xi)}$ is compact. By compactness of K, it follows that $\overline{\mathcal{R}^{\leq T}(K)}$ is compact. □

LEMMA 5.3. *For any subset* S *of* $\mathbb{R}^n$ *and any* $T > 0$,

$$\mathcal{R}^T(\overline{S}) \subseteq \overline{\mathcal{R}^T(S)},\quad \mathcal{R}^{\leq T}(\overline{S}) \subseteq \overline{\mathcal{R}^{\leq T}(S)}\,.$$

In particular, $\overline{\mathcal{R}^{\leq T}(\overline{S})} = \overline{\mathcal{R}^{\leq T}(S)}$.

Proof. The first conclusion follows from the continuity of solutions on initial states; see [26, Thm. 1]. The second is immediate from there. □

We now return to the proof of Proposition 5.1. By Lemma 5.2, it is enough to show that $\overline{\mathcal{R}^{\leq T}(\xi)}$ is compact for each $\xi \in \mathbb{R}^n$ and each $T > 0$. Pick any $\xi_0 \in \mathbb{R}^n$, and let

$$\tau = \sup\{T \geq 0 : \ \overline{\mathcal{R}^{\leq T}(\xi_0)} \text{ is compact } \}.$$

Note that $\tau > 0$. This is because $|x(t, \xi_0, d) - \xi_0| \leq 1$ for any $0 \leq t < 1/M$ and any $d \in \mathcal{M}_{\mathcal{D}}$, where

$$M = \max\{|f(\xi, \mathbf{d})| : \ |\xi - \xi_0| \leq 1, \mathbf{d} \in \mathcal{D}\}.$$

We must show that $\tau = \infty$.

Assume that $\tau < \infty$. Using the same argument as above, one can show that if $\overline{\mathcal{R}^{\leq t}(\xi_0)}$ is compact for some $t > 0$ then there is some $\delta > 0$ such that $\overline{\mathcal{R}^{\leq (t+\delta)}(\xi_0)}$ is compact. From here it follows that $\overline{\mathcal{R}^{\leq \tau}(\xi_0)}$ is not compact. By definition, $\overline{\mathcal{R}^{\leq t}(\xi_0)}$ is compact for any $t < \tau$.

Let $\tau_1 = \tau/2$. Then there is some $\eta_1 \in \overline{\mathcal{R}^{\tau_1}(\xi_0)}$ such that $\overline{\mathcal{R}^{\leq (\tau-\tau_1)}(\eta_1)}$ is not compact; otherwise, by Lemma 5.2, $\overline{\mathcal{R}^{\leq (\tau-\tau_1)}(\overline{\mathcal{R}^{\tau_1}(\xi_0)})}$ would be compact. This, in turn, would imply that $\overline{\mathcal{R}^{\leq \tau}(\xi_0)}$ is compact, since

$$\mathcal{R}^{\leq \tau}(\xi_0) \subseteq \mathcal{R}^{\leq \tau_1}(\xi_0) \bigcup \mathcal{R}^{\leq (\tau-\tau_1)}(\mathcal{R}^{\tau_1}(\xi_0)) \subseteq \overline{\mathcal{R}^{\leq \tau_1}(\xi_0)} \bigcup \overline{\mathcal{R}^{\leq (\tau-\tau_1)}\left(\overline{\mathcal{R}^{\tau_1}(\xi_0)}\right)}.$$

On the other hand, combining Lemma 5.3 with the fact that $\overline{\mathcal{R}^{\leq t}(\mathcal{R}^{\tau_1}(\xi_0))}$ is compact for any $0 \leq t < \tau - \tau_1$, one sees that $\overline{\mathcal{R}^{\leq t}(\eta_1)}$ is compact for any $0 \leq t < \tau - \tau_1$.

Since $\eta_1 \in \overline{\mathcal{R}^{\tau_1}(\xi_0)}$, there exists a sequence $\{z_n\} \to \eta_1$ with $z_n \in \mathcal{R}^{\tau_1}(\xi_0)$. Assume, for each n, that $z_n = x(\tau_1, \xi_0, d_n)$ for some $d_n \in \mathcal{M}_{\mathcal{D}}$. For each $d \in \mathcal{M}_{\mathcal{D}}$ and each $s \in \mathbb{R}$, we use d_s to denote the function defined by $d_s(t) = d(s + t)$. Then by uniqueness, one has that for each n, $x(s, z_n, (d_n)_{\tau_1}) \in K_1$ for any $-\tau_1 \leq s \leq 0$, where $K_1 = \overline{\mathcal{R}^{\leq \tau_1}(\xi_0)}$. We want to claim next that, by compactness of K_1 and Gronwall's lemma,

$$|x(-\tau_1, \eta_1, (d_n)_{\tau_1}) - \xi_0| = |x(-\tau_1, \eta_1, (d_n)_{\tau_1}) - x(-\tau_1, z_n, (d_n)_{\tau_1})| \longrightarrow 0, \quad \text{as } n \to \infty.$$

The only potential problem is that the solution $x(-\tau_1, \eta_1, (d_n)_{\tau_1})$ may fail to exist a priori. However, it is possible to modify $f(x, \mathbf{d})$ outside a neighborhood of $K_1 \times \mathcal{D}$ so that it now has compact support and is hence globally bounded. The modified dynamics is complete. Now the above limit holds for the modified system, and a fortiori it also holds for the original system.

Choose n_0 such that

$$|x(-\tau_1, \eta_1, (d_{n_0})_{\tau_1}) - \xi_0| < \frac{1}{2}. \tag{28}$$

Let $v_1 = d_{n_0}$, and let $\eta_0 = x(-\tau_1, \eta_1, (d_{n_0})_{\tau_1})$. Then, by continuity on initial conditions, there is a neighborhood $\mathcal{U}_1$ of η_1 contained in $B(\eta_1, 1)$ such that

$$|x(-\tau_1, \xi, (v_1)_{\tau_1}) - \eta_0| < \frac{1}{2}, \quad \forall \xi \in \mathcal{U}_1, \tag{29}$$

where $B(\eta, r)$ denotes the open ball centered at η with radius r. Combining (28) and (29), one has

$$x(-\tau_1, \xi, (v_1)_{\tau_1}) \in \mathcal{U}_0, \quad \forall \xi \in \mathcal{U}_1,$$

where $\mathcal{U}_0 = B(\xi_0, 1)$.

Let $\tau_2 = \tau_1/2 = (\tau - \tau_1)/2$. Applying the above argument with ξ_0 replaced by η_1, τ replaced by $(\tau - \tau_1)$, and τ_1 replaced by τ_2, one shows that there exists some $\eta_2 \in \overline{\mathcal{R}^{\tau_2}(\eta_1)}$ such that $\overline{\mathcal{R}^{\leq t}(\eta_2)}$ is compact for any $0 \leq t < \tau - \sigma_2$, and $\overline{\mathcal{R}^{\leq(\tau-\sigma_2)}(\eta_2)}$ is not compact, where $\sigma_2 = \tau_1 + \tau_2$, and there exist some v_2 defined on $[0, \tau_2)$ and some neighborhood $\mathcal{U}_2$ of η_2 contained in $B(\eta_2, 1)$, such that

$$x(-\tau_2, \xi, (v_2)_{\tau_2}) \in \mathcal{U}_1, \quad \forall \xi \in \mathcal{U}_2 .$$

By induction, one can get for each $k \geq 1$ a point η_k, a neighborhood $\mathcal{U}_k$ of η_k contained in $B(\eta_k, 1)$, and a function v_k defined on $[0, \tau_k)$ (where $\tau_k = 2^{-k}\tau$) such that

- $\overline{\mathcal{R}^{\leq(\tau-\sigma_k)}(\eta_k)}$ is not compact, where $\sigma_k = \tau_1 + \tau_2 + \cdots + \tau_k = \tau(1 - 2^{-k}) \to \tau$;
- $x(-\tau_k, \xi, (v_k)_{\tau_k}) \in \mathcal{U}_{k-1}$, for any $\xi \in \mathcal{U}_k$.

Now define v on $[0, \tau)$ by concatenating all the v_k's. That is, $v(t) = v_k(t)$ for $t \in [\sigma_{k-1}, \sigma_k)$ (with $\sigma_0 \stackrel{\text{def}}{=} 0$). Then $v \in \mathcal{M}_{\mathcal{D}}$. For each k, let

$$\zeta_k = x(-\sigma_k, \eta_k, (v^k)_{\sigma_k}) ,$$

where v^k is the restriction of v to $[0, \sigma_k)$. By induction,

$$x\Big(-(\sigma_k - \sigma_i), \eta_k, (v^k)_{\sigma_k}\Big) \in \mathcal{U}_{k-i} ,$$

for each $0 \leq i \leq k$, from which it follows that $\zeta_k \in \mathcal{U}_0$ for each k. By compactness of $\overline{\mathcal{U}_0}$, there exists some subsequence of $\{\zeta_k\}$ converging to some point $\zeta_0 \in \mathbb{R}^n$. For ease of notation, we still use $\{\zeta_k\}$ to denote this convergent subsequence. Our aim is next to prove that the solution starting at ζ_0 and applying the measurable function v does not exist for time τ, contradicting forward completeness.

First notice that for any compact set S, there exists some k such that $\eta_k \notin S$. Otherwise, assume that there exists some compact set S such that $\eta_k \in S$ for all k. Let $S_1 = \{\eta : d(\eta, S) \leq 1\}$. The compactness of S implies that there exists some $\delta > 0$ such that

$$\mathcal{R}^{\leq t}(\eta) \subseteq S_1$$

for any $\eta \in S$ and any $t \in [0, \delta]$. In particular, it implies that $\overline{\mathcal{R}^{\leq(\tau-\sigma_k)}(\eta_k)} \subseteq S_1$ for k large enough so that $\tau - \sigma_k < \delta$. This contradicts the fact that $\overline{\mathcal{R}^{\leq(\tau-\sigma_k)}(\eta_k)}$ is not compact for each k.

Assume that $x(\tau, \zeta_0, v)$ is defined. By continuity on initial conditions, this would imply that $x(t, \zeta_k, v)$ is defined for all $t \leq \tau$ and for all k large enough, and that it converges uniformly to $x(t, \zeta_0, v)$. Thus, $x(t, \zeta_k, v)$ remains in a compact set for all $t \in [0, \tau]$ and all k. But

$$x(\sigma_k, \zeta_k, v) = x(\sigma_k, \zeta_k, v^k) = \eta_k ,$$

contradicting what was just proved. So $x(\tau, \zeta_0, v)$ is not defined, which contradicts the forward completeness of the system. □

Remark 5.4. For $T > 0$ and $\xi \in \mathbb{R}^r$, let

$$\mathcal{R}^{-T}(\xi) = \{\eta : \ \eta = x(-T, \xi, d), \ d \in \mathcal{M}_{\mathcal{D}}\} \quad \text{and} \quad \mathcal{R}^{\geq -T}(\xi) = \bigcup_{t \in [-T, 0]} \mathcal{R}^t(\xi) .$$

These are the reachable sets from ξ for the time-reversed system

$$\dot{x}(t) = -f(x(t), d(t)) . \tag{30}$$

Similarly, one defines $\mathcal{R}^{-T}(S)$ and $\mathcal{R}^{\geq -T}(S)$ for subsets S of $\mathbb{R}^n$. If (1) is backward complete, that is, if (30) is forward complete, and applying Proposition 5.1 to (30), one concludes, for system (1), that $\overline{\mathcal{R}^{\geq -T}(K)}$ is compact for any $T > 0$ and any compact subset K of $\mathbb{R}^n$. In particular, for systems that are (forward and backward) complete,

$$\overline{\mathcal{R}^{\geq -T}(K) \bigcup \mathcal{R}^{\leq T}(K)}$$

is compact for any compact set K and any $T > 0$.

Combining the above conclusion and Gronwall's lemma, one has the following fact.

PROPOSITION 5.5. *Assume that* (1) *is complete. For any fixed* $T > 0$ *and any compact* $K \subseteq \mathbb{R}^n$, *there is a constant* $C > 0$ (*which only depends on the set* K *and* T), *such that for the trajectories* $x(t, x_0, d)$ *of the system* (1),

$$|x(t, \xi, d) - x(t, \eta, d)| \leq C|\xi - \eta|$$

for any $\xi, \eta \in K$, *any* $|t| \leq T$, *and any* $d \in \mathcal{M}_{\mathcal{D}}$.

6. Proof of the first converse Lyapunov theorem.

Proof. [$\Longleftarrow$] Pick any $x_0 \in \mathbb{R}^n$ and any $d \in \mathcal{M}_{\mathcal{D}}$, and let $x(\cdot)$ be the corresponding trajectory. Then we have

$$\frac{dV(x(t))}{dt} \leq -\alpha_3(|x(t)|_{\mathcal{A}}) \leq -\alpha(V(x(t))) , \text{ a.e. } t \geq 0 ,$$

where α is the $\mathcal{K}_\infty$-function defined by

$$\alpha(\cdot) \stackrel{\text{def}}{=} \alpha_3(\alpha_2^{-1}(\cdot)) .$$

Now let β_α be the $\mathcal{KL}$-function as in Lemma 4.4 with respect to α, and define

$$\beta(s, t) \stackrel{\text{def}}{=} \alpha_1^{-1}\big(\beta_\alpha(\alpha_2(s), t)\big) . \tag{31}$$

Then β is a $\mathcal{KL}$-function, since both α_1 and α_2 are $\mathcal{K}_\infty$-functions. By Lemma 4.4,

$$V(x(t)) \leq \beta_\alpha\big(V(x_0), t\big) , \quad \text{for any } t \geq 0 .$$

Hence

$$|x(t)|_{\mathcal{A}} \leq \beta(|x_0|_{\mathcal{A}} , t) , \text{ for any } t \geq 0 .$$

Therefore the system (1) is UGAS with respect to $\mathcal{A}$, by Proposition 2.5.

[$\Longrightarrow$] We will show the existence of a not necessarily smooth Lyapunov function; then the existence of a smooth function will follow from Proposition 4.2. Assume that the system is UGAS with respect to the set $\mathcal{A}$. Let δ and T_r be as in Definition 2.2 and Lemma 3.1.

Define $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ by

$$g(\xi) \stackrel{\text{def}}{=} \inf_{t \leq 0, d \in \mathcal{M}_{\mathcal{D}}} \{|x(t, \xi, d)|_{\mathcal{A}}\} . \tag{32}$$

Note that, by uniqueness of solutions, for each $t_0 > 0$ and each d, it holds that

$$x(t - t_0,\, x(t_0,\, \xi,\, d),\, d_{t_0}) \;=\; x(t,\, \xi, d)\,,$$

where d_{t_0} is defined by $d_{t_0}(t) = d(t + t_0)$. Pick any $d \in \mathcal{M}_{\mathcal{D}}$, $\xi \in \mathbb{R}^n$, and $t_1 > 0$. Let $\xi_1 = x(t_1,\, \xi,\, d)$. Then for any $t < 0$, and $v \in \mathcal{M}_{\mathcal{D}}$,

$$x(t,\, \xi, v) \;=\; x(t - t_1,\, \xi_1,\, v_{t_1} \# d_{t_1})\,,$$

where

$$v_{t_1} \# d_{t_1}(s) \;=\; \begin{cases} d(s + t_1), & \text{if } -t_1 \le s \le 0, \\ v(s + t_1), & \text{if } s < -t_1. \end{cases}$$

Thus,

$$\begin{aligned} g(\xi) &= \inf_{t \le 0, v \in \mathcal{M}_{\mathcal{D}}} |x(t,\, \xi,\, v)|_{\mathcal{A}} = \inf_{t \le 0, d \in \mathcal{M}_{\mathcal{D}}} \left|x(t - t_1,\, \xi_1,\, v_{t_1} \# d_{t_1})\right|_{\mathcal{A}} \\ &= \inf_{\tau \le -t_1, v \in \mathcal{M}_{\mathcal{D}}} \left|x(\tau,\, \xi_1,\, v_{t_1} \# d_{t_1})\right|_{\mathcal{A}} \;\ge\; \inf_{\tau \le 0, v \in \mathcal{M}_{\mathcal{D}}} \left|x(\tau,\, \xi_1,\, v)\right|_{\mathcal{A}} \\ &= g(\xi_1)\,. \end{aligned}$$

This implies that

$$g\big(x(t,\, \xi,\, d)\big) \le g(\xi), \quad \forall t > 0\,, \quad \forall d \in \mathcal{M}_{\mathcal{D}}\,. \tag{33}$$

Also one has

$$\delta(|\xi|_{\mathcal{A}}) \le g(\xi) \le |\xi|_{\mathcal{A}}\,. \tag{34}$$

The second half of (34) is obvious from $x(0,\, \xi,\, d) = \xi$. On the other hand, if the first half were not true, then there would be some $d \in \mathcal{M}_{\mathcal{D}}$ and some $t_0 \le 0$ such that

$$\delta(|\xi|_{\mathcal{A}}) > |x(t_0,\, \xi,\, d)|_{\mathcal{A}}\,.$$

Pick any $0 < \varepsilon < |\xi|_{\mathcal{A}}$ so that $|x(t_0,\, \xi,\, d)|_{\mathcal{A}} < \delta(\varepsilon)$. By the uniform stability property, applied with $t = -t_0$ and $x_0 = x(t_0,\, \xi,\, d)$,

$$|\xi|_{\mathcal{A}} = |x(-t_0,\, x(t_0,\, \xi,\, d),\, d_{t_0})|_{\mathcal{A}} < |\xi|_{\mathcal{A}}\,,$$

which is a contradiction.

For any $0 < \varepsilon < r$, define $K_{\varepsilon, r} \stackrel{\text{def}}{=} \{\xi \in \mathbb{R}^n :\; \varepsilon \le |\xi|_{\mathcal{A}} < r\}$.

FACT 1. *For all ε and r with $0 < \varepsilon < r$, there exists $q_{\varepsilon, r} \le 0$, such that*

$$\xi \in K_{\varepsilon, r},\; d \in \mathcal{M}_{\mathcal{D}}, \quad \textit{and} \quad t < q_{\varepsilon, r} \;\;\Longrightarrow\;\; |x(t,\, \xi,\, d)|_{\mathcal{A}} \ge r\,.$$

Proof. If the statement were not true, then there would exist ε, r with $0 < \varepsilon < r$ and three sequences $\{\xi_k\} \subseteq K_{\varepsilon, r}$, $\{t_k\} \subseteq \mathbb{R}$, and $d_k \in \mathcal{M}_{\mathcal{D}}$ with $\lim_{k \to \infty} t_k = -\infty$ such that for all k

$$|x(t_k,\, \xi_k, d_k)|_{\mathcal{A}} < r\,.$$

Pick k large enough so that $-t_k > T_r(\varepsilon)$. Then by the uniform attraction property,

$$|\xi_k|_{\mathcal{A}} = |x(-t_k,\, x(t_k,\, \xi_k,\, d_k),\, (d_k)_{t_k})|_{\mathcal{A}} < \varepsilon\,,$$

which is a contradiction. This proves the fact.

Therefore, for any $\xi \in K_{\varepsilon,r}$,

$$g(\xi) = \inf\{|x(t, \xi, d)|_{\mathcal{A}} : \ t \in [q_{\varepsilon,r}, 0], d \in \mathcal{M}_{\mathcal{D}}\}.$$

LEMMA 6.1. *The function $g(\xi)$ is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$ and continuous everywhere.*

Proof. Fix any $\xi_0 \in \mathbb{R}^n \backslash \mathcal{A}$, and let $s = |\xi_0|_{\mathcal{A}}/2$. Let $\bar{B}(\xi_0, s)$ denote the closed ball centered at ξ_0 and with radius s. Then $\bar{B}(\xi_0, s) \subseteq K_{\sigma,r}$ for some $0 < \sigma < r$. Pick a constant C as in Proposition 5.5 with respect to this closed ball and $T = |q_{\sigma,r}|$. Pick any $\zeta, \eta \in \bar{B}(\xi_0, s)$. For any $\varepsilon > 0$, there exist some $d_{\eta,\varepsilon}$ and $t_{\eta,\varepsilon} \in [q_{\sigma,r}, 0]$ such that $g(\eta) \geq |x(t_{\eta,\varepsilon}, \eta, d_{\eta,\varepsilon})|_{\mathcal{A}} - \varepsilon$. Thus

$$g(\zeta) - g(\eta) \leq |x(t_{\eta,\varepsilon}, \zeta, d_{\eta,\varepsilon})|_{\mathcal{A}} - |x(t_{\eta,\varepsilon}, \eta, d_{\eta,\varepsilon})|_{\mathcal{A}} + \varepsilon \leq C|\zeta - \eta| + \varepsilon. \tag{35}$$

Note that (35) holds for all $\varepsilon > 0$, so it follows that

$$g(\zeta) - g(\eta) \leq C\,|\zeta - \eta|\,.$$

Similarly, $g(\eta) - g(\zeta) \leq C|\zeta - \eta|$. This proves that g is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$.

Note that g is 0 on $\mathcal{A}$, and for $\xi \in \mathcal{A}$, $\eta \in \mathbb{R}^n$,

$$|g(\eta) - g(\xi)| = |g(\eta)| \leq |\eta|_{\mathcal{A}} \leq |\eta - \xi|\,,$$

thus g is globally continuous. (We are not claiming that g is locally Lipschitz on $\mathbb{R}^n$, though.) □

Now define $U : \mathbb{R}^n \longrightarrow \mathbb{R}_{\geq 0}$ by

$$U(\xi) \stackrel{\text{def}}{=} \sup_{t \geq 0, d \in \mathcal{M}_{\mathcal{D}}} \Big\{ g\big(x(t, \xi, d)\big)\, k(t) \Big\}, \tag{36}$$

where $k : R_{\geq 0} \longrightarrow \mathbb{R}_{>0}$ is any strictly increasing, smooth function that satisfies:

- there are two constants $0 < c_1 < c_2 < \infty$ such that $k(t) \in [c_1, c_2]$ for all $t \geq 0$;
- there is a bounded, positive decreasing, continuous function $\tau(\cdot)$, such that

$$k'(t) \geq \tau(t) \quad \text{for all} \quad t \geq 0\,.$$

(For instance, $(c_1 + c_2 t)/(1 + t)$ is one example of such a function.) Observe that

$$U(\xi) \leq \sup_{t \geq 0}(g(\xi)\, k(t)) \leq c_2 g(\xi) \leq c_2\, |\xi|_{\mathcal{A}}\,, \tag{37}$$

and

$$U(\xi) \geq \sup_{d \in \mathcal{M}_{\mathcal{D}}} g(x(t, \xi, d))\, k(t)\big|_{t=0} \geq c_1 g(\xi) \geq c_1 \delta(|\xi|_{\mathcal{A}})\,. \tag{38}$$

For any $\xi \in \mathbb{R}^n$, since

$$|x(t, \xi, d)|_{\mathcal{A}} \leq \beta(|\xi|_{\mathcal{A}}, t), \ \ \forall d, \ \ \forall t \geq 0,$$

for some $\mathcal{KL}$-function β, and $0 \leq g(x(t, \xi, d)) \leq |x(t, \xi, d)|_{\mathcal{A}}$ for all $t \geq 0$, it follows that

$$\lim_{t \to +\infty} \sup_{d} g(x(t, \xi, d)) = 0\,.$$

Thus there exists some $\tau_\xi \in [0, \infty)$ such that

$$U(\xi) = \sup_{0\le t\le \tau_\xi, d\in\mathcal{M}_\mathcal{D}} g(x(t, \xi, d))\, k(t)\,.$$

In fact, we can get the following explicit bound.

FACT 2. *For any* $0 < |\xi|_\mathcal{A} < r$,

$$U(\xi) = \sup_{0\le t\le t_\xi, d\in\mathcal{M}_\mathcal{D}} g\big(x(t, \xi, d)\big)k(t)\,,$$

where $t_\xi = T_r(\frac{c_1}{2c_2}\delta(|\xi|_\mathcal{A}))$.

Proof. If the statement is not true, then for any $\varepsilon > 0$, there exists some $t_\varepsilon > T_r(\frac{c_1}{2c_2}\delta(|\xi|_\mathcal{A}))$ and some d_ε such that

$$U(\xi) \le g(x(t_\varepsilon, \xi, d_\varepsilon))k(t_\varepsilon) + \varepsilon\,.$$

So we have

$$\begin{aligned}\delta(|\xi|_\mathcal{A}) &\le \frac{1}{c_1}U(\xi) \le \frac{1}{c_1}g(x(t_\varepsilon, \xi, d_\varepsilon))\,k(t_\varepsilon) + \frac{\varepsilon}{c_1}\\ &\le \frac{c_2}{c_1}g(x(t_\varepsilon, \xi, d_\varepsilon)) + \frac{\varepsilon}{c_1} \le \frac{c_2}{c_1}|x(t_\varepsilon, \xi, d_\varepsilon)|_\mathcal{A} + \frac{\varepsilon}{c_1} < \frac{\delta(|\xi|_\mathcal{A})}{2} + \frac{\varepsilon}{c_1}\,.\end{aligned}$$

Taking the limit as ε tends to 0 results in a contradiction.

For any compact set $K \subseteq \mathbb{R}^n\backslash\mathcal{A}$, let

$$t_K \stackrel{\text{def}}{=} \max_{\xi\in K} t_\xi < \infty\,.$$

(Finiteness follows from Fact 2, as $K \subseteq \{\xi :\ 0 < |\xi|_\mathcal{A} < r\}$ for some $r > 0$.)

LEMMA 6.2. *The function* $U(\cdot)$ *defined by* (36) *is locally Lipschitz on* $\mathbb{R}^n\backslash\mathcal{A}$ *and continuous everywhere.*

Proof. For $\xi_0 \notin \mathcal{A}$, pick up a compact neighborhood K_0 of ξ_0 so that $K_0 \cap \mathcal{A} = \emptyset$. By (38), one knows that

$$U(\xi) > r_0\,,\quad \forall \xi \in K_0\,,$$

for some constant $r_0 > 0$. Let $r_1 = r_0/(2c_2)$ and let

$$K_1 = K_0 \bigcap \Big\{\eta :\ |\eta - \xi_0| \le \frac{r_1}{4C}\Big\}\,,$$

where C is a constant such that

$$|x(t, \xi, d) - x(t, \eta, d)| \le C\,|\xi - \eta|\,,\quad \forall \xi, \eta \in K_0,\ 0 \le t \le t_{K_0},\ d \in \mathcal{M}_\mathcal{D}. \tag{39}$$

In what follows we will show that there exists some $L > 0$ such that for any $\xi, \eta \in K_1$, it holds that

$$|U(\xi) - U(\eta)| \le L\,|\xi - \eta|\,. \tag{40}$$

First of all, for any $\xi \in K_1$ and any $\varepsilon \in (0, r_0/2)$, there exists $t_{\xi,\varepsilon} \in [0, t_{K_0}]$ and $d_{\xi,\varepsilon} \in \mathcal{M}_\mathcal{D}$ such that

$$U(\xi) \le g(x(t_{\xi,\varepsilon}, \xi, d_{\xi,\varepsilon}))k(t_{\xi,\varepsilon}) + \varepsilon \le c_2\,|x(t_{\xi,\varepsilon}, \xi, d_{\xi,\varepsilon})|_\mathcal{A} + \varepsilon\,,$$

from which it follows that

$$|x(t_{\xi,\varepsilon},\, \xi,\, d_{\xi,\varepsilon})|_{\mathcal{A}} \geq r_1 \,.$$

It follows from (39) that for any $\eta \in K_1$,

$$|x(t_{\xi,\varepsilon},\, \eta,\, d_{\xi,\varepsilon})|_{\mathcal{A}} \geq |x(t_{\xi,\varepsilon},\, \xi,\, d_{\xi,\varepsilon})|_{\mathcal{A}} - |x(t_{\xi,\varepsilon},\, \xi,\, d_{\xi,\varepsilon}) - x(t_{\xi,\varepsilon},\, \eta,\, d_{\xi,\varepsilon})| \geq \frac{r_1}{2} \,.$$

By Proposition 5.1 one knows that there exists some compact set K_2 such that

$$x(t,\, \xi,\, d) \in K_2, \quad \forall \xi \in K_1,\ \forall t \in [0,\, t_{K_1}], \text{ and } \forall d \in \mathcal{M}_{\mathcal{D}} \,.$$

Again, applying Lemma 6.1 to the compact set $K_2 \bigcap \{\zeta :\ |\zeta|_{\mathcal{A}} \geq r_1/2\}$, one sees that

$$|g(x(t_{\xi,\varepsilon},\, \xi,\, d_{\xi,\varepsilon})) - g(x(t_{\xi,\varepsilon},\, \eta,\, d_{\xi,\varepsilon}))| \leq C_1\, |x(t_{\xi,\varepsilon},\, \xi,\, d_{\xi,\varepsilon}) - x(t_{\xi,\varepsilon},\, \eta,\, d_{\xi,\varepsilon})| \,,$$

for some $C_1 > 0$. Therefore, we have the following:

$$\begin{aligned} U(\xi) - U(\eta) &\leq g(x(t_{\xi,\varepsilon},\, \xi,\, d_{\xi,\varepsilon}))k(t_{\xi,\varepsilon}) + \varepsilon - g(x(t_{\xi,\varepsilon},\, \eta,\, d_{\xi,\varepsilon}))k(t_{\xi,\varepsilon}) \\ &\leq c_2\, |g(x(t_{\xi,\varepsilon},\, \xi,\, d_{\xi,\varepsilon})) - g(x(t_{\xi,\varepsilon},\, \eta,\, d_{\xi,\varepsilon}))| + \varepsilon \\ &\leq C_1 c_2\, |x(t_{\xi,\varepsilon},\, \xi,\, d_{\xi,\varepsilon}) - x(t_{\xi,\varepsilon},\, \eta,\, d_{\xi,\varepsilon})| + \varepsilon \\ &\leq L\, |\xi - \eta| + \varepsilon \,, \end{aligned}$$

for some constant L that depends only on the compact set K_1. Note that the above holds for any $\varepsilon \in (0,\, r_0/2)$, thus,

$$U(\xi) - U(\eta) \leq L\, |\xi - \eta| \,, \quad \forall \xi, \eta \in K_1 \,.$$

By symmetry, one proves (40).

To prove the continuity of U on $\mathbb{R}^n$, note that for any $\xi \in \mathcal{A}$, it holds that $U(\xi) = 0$, and so for all $\eta \in \mathbb{R}^n$

$$|U(\xi) - U(\eta)| = U(\eta) \leq c_2\, |\eta|_{\mathcal{A}} \leq c_2\, |\xi - \eta| \,.$$

The proof of Lemma 6.2 is thus concluded. □

We next start proving that U decreases along trajectories. Now pick any $\xi \notin \mathcal{A}$. Let $h_0 > 0$ be such that

$$|x(t,\, \xi,\, \mathbf{d})|_{\mathcal{A}} \geq \frac{|\xi|_{\mathcal{A}}}{2}, \quad \forall \mathbf{d} \in \mathcal{D},\ \forall t \in [0,\, h_0] \,,$$

where $\mathbf{d}$ denotes the constant function $d(t) \equiv \mathbf{d}$. Such an h_0 exists by continuity. Pick any $h \in [0,\, h_0]$. For each $\mathbf{d} \in \mathcal{D}$, let $\eta_{\mathbf{d}} = x(h,\, \xi,\, \mathbf{d})$. For any $\varepsilon > 0$, there exist some $t_{\mathbf{d},\varepsilon}$ and $d_{\mathbf{d},\varepsilon} \in \mathcal{M}_{\mathcal{D}}$ such that

$$\begin{aligned} U(\eta_{\mathbf{d}}) &\leq g(x(t,\, \eta_{\mathbf{d}},\, d_{\mathbf{d},\varepsilon}))k(t_{\mathbf{d},\varepsilon}) + \varepsilon \\ &= g(x(t_{\mathbf{d},\varepsilon} + h,\, \xi,\, \tilde{d}_{\mathbf{d},\varepsilon}))k(t_{\mathbf{d},\varepsilon} + h) \left(1 - \frac{k(t_{\mathbf{d},\varepsilon} + h) - k(t_{\mathbf{d},\varepsilon})}{k(t_{\mathbf{d},\varepsilon} + h)}\right) + \varepsilon \\ &\leq U(\xi) \left(1 - \frac{k(t_{\mathbf{d},\varepsilon} + h) - k(t_{\mathbf{d},\varepsilon})}{c_2}\right) + \varepsilon \,, \end{aligned} \tag{41}$$

where $\tilde{d}_{\mathbf{d},\varepsilon}$ is the concatenation of $\mathbf{d}$ and $d_{\mathbf{d},\varepsilon}$. Still for these ξ and h, and for any $r > |\xi|_{\mathcal{A}}$, define

$$T^r_{\xi,h} \stackrel{\text{def}}{=} \max_{0\le \bar{t}\le h, \mathbf{d}\in\mathcal{D}} T_r\left(\frac{c_1}{2c_2}\delta(|x(\bar{t},\,\xi,\,\mathbf{d})|_{\mathcal{A}})\right). \tag{42}$$

CLAIM. *$t_{\mathbf{d},\varepsilon} + h \le T^r_{\xi,h}$, for all $\mathbf{d}\in\mathcal{D}$ and for all $\varepsilon \in (0,\, \frac{c_1}{2}\delta(\frac{|\xi|_{\mathcal{A}}}{2}))$.*

Proof. If this were not true, then there would exist some $\tilde{\mathbf{d}}$ and some $\tilde{\varepsilon}\in(0,\,\frac{c_1}{2}\delta(\frac{|\xi|_{\mathcal{A}}}{2}))$ such that $t_{\tilde{\mathbf{d}},\tilde{\varepsilon}} + h > T^r_{\xi,h}$, and hence in particular for $\bar{t} = h$ and $\mathbf{d} = \tilde{\mathbf{d}}$ it holds that

$$t_{\tilde{\mathbf{d}},\tilde{\varepsilon}} + h > T_r\left(\frac{c_1}{2c_2}\delta(|\eta_{\tilde{\mathbf{d}}}|_{\mathcal{A}})\right),$$

which implies that

$$\left|x(t_{\tilde{\mathbf{d}},\tilde{\varepsilon}},\,\eta_{\tilde{\mathbf{d}}},\,d_{\tilde{\mathbf{d}},\tilde{\varepsilon}})\right|_{\mathcal{A}} = \left|x(t_{\tilde{\mathbf{d}},\tilde{\varepsilon}}+h,\,\xi,\,v)\right|_{\mathcal{A}} < \frac{c_1}{2c_2}\delta(|\eta_{\tilde{\mathbf{d}}}|_{\mathcal{A}}),$$

where v is the concatenated function defined by

$$v(t) = \begin{cases} \tilde{\mathbf{d}}, & \text{if } 0\le t\le h,\\ d_{\tilde{\mathbf{d}},\tilde{\varepsilon}}(t-h), & \text{if } t > h.\end{cases}$$

Using (38), one has

$$\begin{aligned}\delta\left(|\eta_{\tilde{\mathbf{d}}}|_{\mathcal{A}}\right) &\le \frac{1}{c_1}U(\eta_{\tilde{\mathbf{d}}}) \le \frac{1}{c_1}g(x(t_{\tilde{\mathbf{d}},\tilde{\varepsilon}},\,\eta_{\tilde{\mathbf{d}}},\,d_{\tilde{\mathbf{d}},\tilde{\varepsilon}}))k(t_{\tilde{\mathbf{d}},\tilde{\varepsilon}}) + \frac{\tilde{\varepsilon}}{c_1}\\ &\le \frac{c_2}{c_1}\left|x(t_{\tilde{\mathbf{d}},\tilde{\varepsilon}},\,\eta_{\tilde{\mathbf{d}}},\,d_{\tilde{\mathbf{d}},\tilde{\varepsilon}})\right|_{\mathcal{A}} + \frac{\tilde{\varepsilon}}{c_1} < \frac{1}{2}\delta\left(|\eta_{\tilde{\mathbf{d}}}|_{\mathcal{A}}\right) + \frac{\tilde{\varepsilon}}{c_1},\end{aligned}$$

which is a contradiction, since $\tilde{\varepsilon} < \frac{c_1}{2}\delta(\frac{|\xi|_{\mathcal{A}}}{2}) \le (c_1\delta(|\eta_{\tilde{\mathbf{d}}}|_{\mathcal{A}}))/2$. This proves the claim.

From (41), we have for any $\mathbf{d}\in\mathcal{D}$ and for any $\varepsilon > 0$ small enough,

$$U(x(h,\,\xi,\,\mathbf{d})) - U(\xi) \le -U(\xi)\frac{(k(t_{\mathbf{d},\varepsilon}+h) - k(t_{\mathbf{d},\varepsilon}))}{c_2} + \varepsilon = -\frac{U(\xi)}{c_2}\,k'(t_{\mathbf{d},\varepsilon}+\theta h)h + \varepsilon,$$

where θ is some number in $(0,\,1)$. Hence, by the assumptions made on the function k, we have

$$U(x(h,\,\xi,\,\mathbf{d})) - U(\xi) \le -\frac{U(\xi)}{c_2}\tau(t_{\mathbf{d},\varepsilon}+\theta h)h + \varepsilon \le -\frac{U(\xi)}{c_2}\tau(T^r_{\xi,h})\,h + \varepsilon.$$

Again, since ε can be chosen arbitrarily small, we have

$$U(x(h,\,\xi,\,\mathbf{d})) - U(\xi) \le -\frac{U(\xi)}{c_2}\tau(T^r_{\xi,h})h, \quad \forall \mathbf{d}\in\mathcal{D}.$$

Thus we showed that for any $\mathbf{d}$ and any $h > 0$ small enough,

$$\frac{U(x(h,\,\xi,\,\mathbf{d})) - U(\xi)}{h} \le -\frac{U(\xi)}{c_2}\tau(T^r_{\xi,h}).$$

Since U is locally Lipschitz on $\mathbb{R}^n\backslash\mathcal{A}$, it is differentiable almost everywhere in $\mathbb{R}^n\backslash\mathcal{A}$, and hence for any $\mathbf{d}\in\mathcal{D}$ and for any $r>|\xi|_{\mathcal{A}}$,

$$\begin{aligned} L_{f_{\mathbf{d}}}U(\xi) &= \lim_{h\to 0^+}\frac{U(x(h,\xi,\mathbf{d}))-U(\xi)}{h} \le -\lim_{h\to 0^+}\frac{U(\xi)}{c_2}\tau(T^r_{\xi,h}) \\ &= -\frac{U(\xi)}{c_2}\tau\left(\lim_{h\to 0^+}T^r_{\xi,h}\right) = -\frac{U(\xi)}{c_2}\tau\left(T_r\left(\frac{c_1}{2c_2}\delta(|\xi|_{\mathcal{A}})\right)\right) \\ &\le -\frac{c_1\delta(|\xi|_{\mathcal{A}})}{c_2}\tau\left(T_r\left(\frac{c_1}{2c_2}\delta(|\xi|_{\mathcal{A}})\right)\right) \qquad (43) \\ &= -\bar{\alpha}_r(|\xi|_{\mathcal{A}}),\ \text{a.e.}\ , \qquad (44) \end{aligned}$$

where

$$\bar{\alpha}_r(s) = \frac{c_1\delta(s)}{c_2}\tau\left(T_r\left(\frac{c_1}{2c_2}\delta(s)\right)\right).$$

Now define the function $\bar{\alpha}$ by

$$\bar{\alpha}(s) = \sup_{r>s}\bar{\alpha}_r(s).$$

Note that $\bar{\alpha}_r(0)=0$ for any $r>0$, so $\bar{\alpha}(0)=0$. Also, applying to $r=2s$, we have

$$\bar{\alpha}(s) \ge \frac{c_1\delta(s)}{c_2}\tau\left(T_{2s}\left(\frac{c_1}{2c_2}\delta(s)\right)\right) > 0$$

for all $s>0$. Notice that (44) holds for any $r>|\xi|_{\mathcal{A}}$, so it follows that for every $\mathbf{d}\in\mathcal{D}$, $L_{f_{\mathbf{d}}}U(\xi)\le -\bar{\alpha}(|\xi|_{\mathcal{A}})$ for almost all $\xi\in\mathbb{R}^n\backslash\mathcal{A}$. Now let

$$\check{\alpha}(s) = \frac{c_1\delta(s)}{c_2}\int_{2s}^{2s+1}\tau\left(T_r\left(\frac{c_1}{2c_2}\delta(s)\right)\right)dr$$

for $s>0$, and let $\check{\alpha}(0)=0$. Then $\check{\alpha}$ is continuous on $[0,\infty)$ (the continuity at $s=0$ is because τ is bounded and $\delta(0)=0$), and for $s>0$, it holds that

$$0<\check{\alpha}(s)\le\frac{c_1\delta(s)}{c_2}\tau\left(T_{2s}\left(\frac{c_1}{2c_2}\delta(s)\right)\right)$$

because of the monotonicity properties of T and τ. Furthermore,

$$L_{f_{\mathbf{d}}}U(\xi)\le-\bar{\alpha}(|\xi|_{\mathcal{A}})\le-\check{\alpha}(|\xi|_{\mathcal{A}}),$$

for almost all $\xi\in\mathbb{R}^n\setminus\mathcal{A}$.

By Theorem B.1 provided in the appendix, there exists a C^∞ function $V:\mathbb{R}^n\backslash\mathcal{A}\longrightarrow\mathbb{R}_{\ge 0}$ such that for almost all $\xi\in\mathbb{R}^n\backslash\mathcal{A}$,

$$|V(\xi)-U(\xi)|<\frac{U(\xi)}{2}\ \text{ and }\ L_{f_{\mathbf{d}}}V(\xi)\le-\frac{1}{2}\check{\alpha}(|\xi|_{\mathcal{A}}),\ \forall\mathbf{d}\in\mathcal{D}.$$

Extend V to $\mathbb{R}^n$ by letting $V|_{\mathcal{A}}=0$ and again denote the extension by V. Note that V is continuous on $\mathbb{R}^n$. So V is a Lyapunov function, as desired, with $\alpha_1(s)=\frac{c_1}{2}\delta(s)$, $\alpha_2(s)=\frac{3c_2}{2}s$ and $\alpha_3(s)=\frac{1}{2}\check{\alpha}(s)$. □

7. Proof of the second converse Lyapunov theorem. We need a couple of lemmas. The first one is trivial, so we omit its proof.

LEMMA 7.1. *Let $f : \mathbb{R}^n \times \mathcal{D} \longrightarrow \mathbb{R}^n$ be continuous, where $\mathcal{D}$ is a compact subset of $\mathbb{R}^l$. Then there exists a smooth function $a_f : \mathbb{R}^n \longrightarrow \mathbb{R}$, with $a_f(x) \geq 1$ everywhere, such that $|f(x, \mathbf{d})| \leq a_f(x)$ for all x and all $\mathbf{d}$.*

Now for any given system

$$\Sigma : \quad \dot{x} = f(x, \mathbf{d}),$$

not necessarily complete, consider the following system:

$$\Sigma_b : \quad \dot{x} = \frac{1}{a_f(x)} f(x, \mathbf{d}).$$

Note that the system Σ_b is complete since $\frac{|f(x,\mathbf{d})|}{a_f(x)} \leq 1$ for all $x, \mathbf{d}$. We let $x_b(\cdot, x_0, d)$ denote the trajectory of Σ_b corresponding to the initial state x_0 and the time-varying parameter d. The following result is a simple consequence of the fact that the trajectories of Σ are the same as those of Σ_b up to a rescaling of time. We provide the details to show clearly that the uniformity conditions are not violated.

LEMMA 7.2. *Assume that $\mathcal{A}$ is a compact set. Suppose that system Σ is UGAS with respect to $\mathcal{A}$. Then, system Σ_b is UGAS with respect to $\mathcal{A}$.*

Proof. Pick a time-varying parameter $d \in \mathcal{M}_{\mathcal{D}}$ and an initial state $x_0 \in \mathbb{R}^n$. Let $\gamma_b(t)$ denote $x_b(t, x_0, d)$. Let $\tau_{\gamma_b}(t)$ denote the solution for $t \geq 0$ of the following initial value problem:

$$\dot{\tau} = a_f(\gamma_b(\tau)), \quad \tau(0) = 0. \tag{45}$$

Since a_f is smooth, and γ_b is Lipschitz, $a_f \circ \gamma_b$ is locally Lipschitz as well. It follows that a unique $\tau_{\gamma_b}(t)$ is at least defined in some interval $[0, \bar{t})$. Note that τ_{γ_b} is strictly increasing, so $\bar{t} < +\infty$ would imply $\lim_{t \to \bar{t}^-} \tau_{\gamma_b}(t) = +\infty$.

CLAIM. *For every trajectory γ_b of Σ_b, $\tau_{\gamma_b}(t)$ is defined for all $t \geq 0$.*

Proof. If the claim is not true, then there exist some trajectory γ_b of Σ_b and some $t_1 > 0$ such that $\lim_{t \to t_1^-} \tau_{\gamma_b}(t) = \infty$. Now for $t \in [0, t_1)$, one has

$$\begin{aligned} \frac{d}{dt}\gamma_b(\tau_{\gamma_b}(t)) &= \frac{1}{a_f(\gamma_b(\tau_{\gamma_b}(t)))} f\Big(\gamma_b\big(\tau_{\gamma_b}(t)\big), d\big(\tau_{\gamma_b}(t)\big)\Big) \frac{d}{dt}\tau_{\gamma_b}(t) \\ &= f\Big(\gamma_b\big(\tau_{\gamma_b}(t)\big), d\big(\tau_{\gamma_b}(t)\big)\Big). \end{aligned} \tag{46}$$

Thus $\gamma_b(\tau_{\gamma_b}(t))$ is a solution of Σ on $[0, t_1)$. By the stability of Σ, it follows that

$$|\gamma_b(\tau_{\gamma_b}(t))|_{\mathcal{A}} < \delta^{-1}(|x_0|_{\mathcal{A}}), \quad t \in [0, t_1),$$

where $x_0 = \gamma_b(0)$, and δ is the function for Σ as defined in Definition 2.2. (Cf. Remark 2.4.) Let $c = \delta^{-1}(|x_0|_{\mathcal{A}})$, and let $M = \sup_{|\xi|_{\mathcal{A}} \leq c} a_f(\xi)$. ($M$ is finite because the set $\{\xi : |\xi|_{\mathcal{A}} \leq c\}$ is a compact set.) From here one sees that $|\tau_{\gamma_b}(t)| \leq Mt_1$ for any $t \in [0, t_1)$. This is a contradiction. Thus $\tau_{\gamma_b}(t)$ is defined for all $t \geq 0$. This proves the claim.

Since $a_f(s) \geq 1$ and, for every trajectory γ_b of Σ_b, $\tau_{\gamma_b}(0) = 0$, it follows that $\tau_{\gamma_b}(\cdot) \in \mathcal{K}_\infty$ for each trajectory γ_b of Σ_b. From (46), one also sees that if $\gamma_b(t)$ is a trajectory of Σ_b, then $\gamma_b(\tau_{\gamma_b}(t))$ is a trajectory of Σ, and furthermore,

$$|\gamma_b(\tau_{\gamma_b}(s))|_{\mathcal{A}} < \varepsilon \quad \forall s \geq 0, \quad \text{if } |\gamma_b(0)|_{\mathcal{A}} \leq \delta(\varepsilon).$$

It follows that

$$|\gamma_b(t)|_{\mathcal{A}} = \left|\gamma_b(\tau_{\gamma_b}(\tau_{\gamma_b}^{-1}(t)))\right|_{\mathcal{A}} < \varepsilon, \quad \forall t \geq 0, \quad \text{whenever } |\gamma_b(0)|_{\mathcal{A}} \leq \delta(\varepsilon)\,.$$

This shows that condition (1) of Definition 2.2 holds for Σ_b, with the same function δ.

Fix any $r, \varepsilon > 0$. Pick any x_0 with $|x_0|_{\mathcal{A}} < r$ and any $d \in \mathcal{M}_{\mathcal{D}}$. Again let $\gamma_b(t)$ denote the corresponding trajectory of Σ_b. Then

$$|\gamma_b(t)|_{\mathcal{A}} = \left|\gamma_b(\tau_{\gamma_b}(\tau_{\gamma_b}^{-1}(t)))\right|_{\mathcal{A}} < \delta^{-1}(r), \quad \forall t \geq 0\,.$$

Let

$$L = \sup\{a_f(\xi) : \ |\xi|_{\mathcal{A}} \leq \delta^{-1}(r)\}.$$

Then one sees that $|\dot{\tau}(t)| \leq L$, which implies that $\tau_{\gamma_b}(t) \leq Lt$ for all $t \geq 0$. Note that for the given $r, \varepsilon > 0$, by the UGAS property for Σ, there exists $T > 0$ such that for every $d \in \mathcal{M}_{\mathcal{D}}$,

$$|\gamma_b(\tau_{\gamma_b}(s))|_{\mathcal{A}} < \varepsilon$$

whenever $|\gamma_b(0)|_{\mathcal{A}} < r$ and $s \geq T$. This implies that

$$|\gamma_b(t)|_{\mathcal{A}} < \varepsilon$$

whenever $|\gamma_b(0)|_{\mathcal{A}} < r$ and $t \geq \tau_{\gamma_b}(T)$. Combining this with the fact that $\tau_{\gamma_b}(t) \leq Lt$, one proves that for any $d \in \mathcal{M}_{\mathcal{D}}$, it holds that

$$|\gamma_b(t)|_{\mathcal{A}} < \varepsilon$$

whenever $|\gamma_b(0)|_{\mathcal{A}} < r$ and $t \geq LT$. Hence we conclude that Σ_b is UGAS. □

In Lemma 7.2, the assumption that $\mathcal{A}$ is compact is crucial. Without this assumption, the conclusion may fail as the following example shows.

Example 7.3. Consider the following system Σ:

$$\dot{x} = -(1+y^2)\tanh x, \quad \dot{y} = y^4\,. \tag{47}$$

(Here f is independent of d.) Let $\mathcal{A} = \{(x, y) : \ x = 0\}$. Clearly the system is UGAS with respect to $\mathcal{A}$. For this system, a natural choice of a_f is $2 + y^4$. Thus, the corresponding Σ_b is as follows:

$$\dot{x} = -(\tanh x)\frac{1+y^2}{2+y^4}, \quad \dot{y} = \frac{y^4}{2+y^4}.$$

However, the system Σ_b is not UGAS with respect to $\mathcal{A}$. This can be seen as follows. Assume that Σ_b is UGAS. Then for $\varepsilon = \frac{1}{2}$, there exists some $T > 0$ such that for any solution $(x(t), y(t))$ of Σ_b with $x(0) = 1$, it holds that

$$|x(t)| < \frac{1}{2}, \quad \forall t \geq T\,. \tag{48}$$

Since $(1+y^2)/(2+y^4) \to 0$ as $y \to \infty$, it follows that there exists some $y_0 > 0$ such that

$$\left|\frac{1+y^2}{2+y^4}\right| < \frac{1}{3T}, \quad \forall y \geq y_0\,.$$

Now consider the trajectory $(x(t), y(t))$ of Σ_b with $x(0) = 1, y(0) = y_0$, where y_0 is as above. Clearly $y(t) \geq y_0$ for all $t \geq 0$, and thus,

$$\dot{x} = -(\tanh x)\frac{1+y^2}{2+y^4} \geq -(\tanh x)\,\frac{1}{3T} \geq -\frac{1}{3T}\,,$$

which implies that

$$|x(T)| \;\geq\; 1 - \frac{1}{3T}\,T = \frac{2}{3}\,.$$

This contradicts (48). From here one sees that Σ_b is not UGAS with respect to $\mathcal{A}$.

We now prove Theorem 2.9.

The proof of the sufficiency part is the same as in the proof of Theorem 2.8. Observe that the fact that $V(\xi)$ is nonincreasing along trajectories implies, by compactness of $\mathcal{A}$, that trajectories are bounded, so $x(t)$ is defined for all $t \geq 0$. We now prove necessity.

Let a_f be a function for f as in Lemma 7.1, and let Σ_b be the corresponding system. Then by Lemma 7.2, one knows that the system Σ_b is UGAS. Applying Theorem 2.8 to the complete system Σ_b, one knows that there exists a smooth Lyapunov function V for Σ_b such that

$$\alpha_1(|\xi|_{\mathcal{A}}) \leq V(\xi) \leq \alpha_2(|\xi|_{\mathcal{A}})\,, \quad \forall \xi \in \mathbb{R}^n\,,$$

and

$$L_{\tilde{f}_{\mathbf{d}}} V(\xi) \leq -\alpha_3(|\xi|_{\mathcal{A}})\,, \quad \forall \xi \notin \mathcal{A}\,,\ \forall \mathbf{d} \in \mathcal{D}\,,$$

for some $\mathcal{K}_\infty$ functions α_1, α_2 and some positive definite function α_3, where

$$\tilde{f}_{\mathbf{d}}(\xi) \;=\; \frac{f(\xi, \mathbf{d})}{a_f(\xi)}\,.$$

Since $a_f(\xi) \geq 1$ everywhere, it follows that

$$L_{f_{\mathbf{d}}} V(\xi) \leq -\alpha_3(|\xi|_{\mathcal{A}}), \quad \forall \xi \notin \mathcal{A},\ \forall \mathbf{d} \in \mathcal{D}\,.$$

Thus, one concludes that V is also a Lyapunov function of Σ.

8. An example. In general, for a noncompact parameter value set $\mathcal{D}$, the converse Lyapunov theorem will fail, even if the vector fields $f(\xi, \mathbf{d})$ are locally Lipschitz uniformly on $\mathbf{d}$ on any compact subset of $\mathcal{D}$ (for instance, if f is smooth everywhere). To illustrate this fact, consider the common case of systems affine in controls:

$$\dot{x} = f(x) + g(x)\mathbf{d},$$

where for simplicity we consider only the unconstrained single-input case, that is, $\mathcal{D} = \mathbb{R}$. Assume that there would exist a Lyapunov function V for this system in the sense of Definition 2.6. Then, calculating Lie derivatives, we have that, in particular,

$$L_f V(\xi) + \mathbf{d} L_g V(\xi) < 0, \quad \forall \xi \neq 0, \quad \forall \mathbf{d} \in \mathbb{R},$$

which implies that

$$L_g V(\xi) \;=\; 0, \quad \forall \xi \neq 0.$$

Thus V must be constant along all the trajectories of the differential equation

$$\dot{x} = g(x).$$

In general, such a property will contradict the properness or the positive definiteness of V, unless the vector field g is very special. As a way to construct counterexamples, consider the following property of a vector field g, which is motivated by the prolongation ideas in [28].

Consider the closure $W(\xi_0)$ of the trajectory through ξ_0 with respect to the vector field g. Note that if $\xi_1 \in W(\xi_0)$, then the fact that V is constant on trajectories, coupled with continuity of V, implies that $V(\xi_1) = V(\xi_0)$. Now assume that there is a chain $\xi_0, \xi_1, \xi_2, \ldots$ so that for each $i = 1, 2, \ldots$, $\xi_i \in W(\xi_{i-1})$. Then we conclude that $V(\xi_i) = V(\xi_0)$ for all i. If the sequence $\{\xi_i\}$ converges to zero (and $\xi_0 \neq 0$) or diverges to infinity, we contradict positive definiteness or properness of V, respectively. For an example, take the following two-dimensional system, which was used in [7] to show essentially the same fact.

Let $\mathfrak{S}$ be the spiral that describes the solution of the differential equation

$$\dot{x} = -x - y, \quad \dot{y} = x - y,$$

passing through the point $(1, 0)$. Explicitly, $\mathfrak{S}$ can be parameterized as $x = e^{-t}\cos t$, $y = e^{-t}\sin t$, $-\infty < t < \infty$. In polar coordinates, the spiral is given by $r = e^{-\theta}$, $-\infty < \theta < \infty$. Let $a(x, y)$ be any nonnegative smooth function which is zero exactly on the closure of the spiral $\mathfrak{S}$ (that is, $\mathfrak{S}$ plus the origin). (Such a function always exists since any closed subset of Euclidean space can be described as the zero set of a smooth function; see for instance [6].) Now consider the system

$$\begin{aligned} \dot{x} &= -x - y + xa(x, y)\mathbf{d}, \\ \dot{y} &= \ \ x - y + ya(x, y)\mathbf{d}. \end{aligned} \tag{49}$$

Note that the system is smooth everywhere. Let $\mathcal{D} = \mathbb{R}$, and let $\mathcal{A}$ be the origin. In polar coordinates, the system (49) on $\mathbb{R}^2\backslash\{0\}$ satisfies the equations

$$\dot{r} = -r + ra(r\cos\theta, r\sin\theta)\mathbf{d}, \quad \dot{\theta} = 1. \tag{50}$$

(This can be seen as a system on $\mathbb{R}_{>0} \times S^1$.) In polar coordinates, then, the trajectory passing through $(r, \theta) = (1, 0)$ is precisely the spiral $r = e^{-\theta}$, for any $d \in \mathcal{M}_{\mathcal{D}}$. Pick any trajectory $(r(t), \theta(t))$ with $(r(0), \theta(0)) = (r_0, \theta_0)$, where $\theta_0 \in [0, 2\pi)$. Then there exists some integer $k \geq 0$ such that $r_0 < e^{-\theta_0 + 2k\pi}$.

CLAIM. *It holds that*

$$r(t) < e^{-\theta_0 + 2k\pi - t} \leq e^{2k\pi - t}, \quad \forall t \geq 0. \tag{51}$$

Assume that (51) is not true. Then there exists some $t_1 > 0$ such that

$$r(t_1) = e^{-\theta_0 + 2k\pi - t_1}.$$

Note that we also have $\theta(t_1) = \theta_0 + t_1$. Now let $(\bar{r}(t), \bar{\theta}(t)) = (e^{-\theta_0 + 2k\pi - t}, \theta_0 - 2k\pi + t)$. Then $(\bar{r}(t), \bar{\theta}(t))$ is a trajectory of the system, and furthermore, $(\bar{r}(0), \bar{\theta}(0))$ and $(r(0), \theta_0)$ are different points since $\bar{r}(0) \neq r(0)$. However, the points $(r(t_1), \theta(t_1))$ and $(\bar{r}(t_1), \bar{\theta}(t_1))$ are the same point on the xy plane. This violates the uniqueness of solutions. Therefore, (51) holds for $t \geq 0$.

Note that in the above discussion, one can always choose $k \leq r_0 + 1$. It then follows from (51) that for any trajectory of the system with $r(0) = r_0$, it holds that

$$r(t) \leq e^{2(r_0+1)\pi - t}, \quad \forall t \geq 0, \quad \forall d. \tag{52}$$

Thus we conclude that the system is UGAS.

However, this system fails to admit a Lyapunov function. In this example, the vector field g is $(xa(x,y), ya(x,y))$. Consider the sequence of points in the xy plane $\{\xi_k\}$ with $\xi_k = (e^{2k\pi}, 0)$ for $k \geq 0$. Note that for each $k \geq 1$,

$$\xi_k \in W(\xi_{k-1}^j),$$

where $\xi_k^j = (e^{2k\pi} + \frac{1}{j}, 0)$. Therefore, $V(\xi_k) = V(\xi_{k-1}^j)$ for any j and any k. This implies that

$$V(\xi_k) = V(\xi_0), \quad \forall k \geq 1,$$

contradicting the properness of V. This shows that it is impossible for the system to have a Lyapunov function.

It is worthwhile to note that by the same argument, one sees that not only is there no smooth Lyapunov function for the system, but also there is not even a Lyapunov function which is merely continuous (in the sense that V is not even smooth away from $\mathcal{A}$, and the Lie derivative condition is replaced by a condition asking that V should decrease along trajectories).

In [17], a simple example is given illustrating that uniform global asymptotic stability with respect merely to *constant* parameters is also not sufficient to guarantee the existence of Lyapunov functions.

9. Relation to other work. The study of smooth converse Lyapunov theorems has a long history. In the special case of stability with respect to equilibria, and for systems without parameters, the first complete work was that done in the early 1950s by Massera and Kurzweil; see for instance the papers [18] and [13]. (Although we are more general because we deal with set stability and time-varying parameters, there is one important aspect in which our results are weaker than some of this classical work, especially that of Kurzweil: we assume enough regularity on the original system so that there are unique solutions and there is continuous dependence. We do so because lack of regularity is not an issue in the main applications in which we are interested. Of course, the proofs become much simpler under regularity assumptions.) In the late 1960s, Wilson, in [31], extended the Massera and Kurzweil results to a converse Lyapunov function theorem for local asymptotic stability with respect to closed sets. But some details of critical steps were omitted in [31]. In 1990, Nadzieja [21] rederived the results given in [31] for the special case when the invariant set is compact. As explained earlier, our proof is modeled along the lines of [31]. See also the textbooks [32] and [12] for many of these classical results.

Nondifferentiable Lyapunov functions have been studied in many papers and textbooks. Among these we may mention the classic book [3] by Bhatia and Szegö, as well as Zubov's work (see for instance [33]), which study in detail continuous Lyapunov function characterizations for global asymptotic stability with respect to arbitrary closed invariant sets. Also, in [29] and [28] and related work, the authors obtained the existence of continuous Lyapunov functions for systems which are stable, uniformly on parameters (or inputs) and with respect to compact sets, assuming various

additional conditions involving prolongations of dynamical systems. (The next section provides some more details on the prolongation approach.) Many results on converse Lyapunov functions with respect to sets can also be found in the many books and articles by Lakshmikantham and several coauthors. For instance, in [14, Thm. 3.4.1], a Massera-type proof is provided of a general converse theorem on local asymptotic stability with respect to two $\mathcal{K}$ functions that provides a Lipschitz Lyapunov function. As the authors point out, their theorem immediately provides a set-stability result (when using distance to the set as one of the comparison functions). In a very recent work [22], the author considered asymptotic stability for systems with merely measurable right-hand sides, and proved the existence of locally Lipschitz Lyapunov functions for such systems. Note that in our case, we obtained the existence of locally Lipschitz Lyapunov functions as an intermediate result, but our regularity assumption on the vector fields made it possible to obtain the existence of smooth Lyapunov functions.

The questions addressed in this paper are related to studies of "total stability," which typically ask about the preservation of stability when considering a new system $\dot{x} = f(x) + R(x,t)$, where $R(x,t)$ is a perturbation. (Sometimes the original system may be allowed to be time varying, that is, it has equations $\dot{x} = f(x,t)$; in that case, its stability can in turn be interpreted in terms of stability of the set $\{x = 0\}$ for the extended system $\dot{x} = f(x,z)$, $\dot{z} = 1$.) In [15], Lefschetz discussed stability with respect to equilibria under perturbations (referred to by the author as quasi-stability). In [12] and [32] one can find such studies and relationships to the special case of $\dot{x} = f(x) + d(t)$, with results proved regarding stability under integrable perturbations (not arbitrary bounded ones).

Under suitable technical conditions, systems with time-varying parameters can also be treated as general dynamical systems, or general control systems, as in [24], [33], [23], [10], [11]. In these works, systems were defined in terms of set-valued maps associated with reachable sets (or attainable sets). A similar treatment was also adopted in [29] and related work, where the prolongation sets of reachable sets were used to study stability. In [23], the author established the existence of different types of Lyapunov functions (not necessarily continuous) for both stability and weak stability with respect to closed invariant sets, where "weak stability" means the existence of a stable trajectory from every point outside the invariant set. In [10], the author provided Lyapunov characterizations for both local asymptotic stability and weak asymptotic stability. See [11] for an excellent survey of work along these lines.

It is also possible to reformulate stability for systems with time-varying parameters in terms of differential inclusions, as explained earlier; see for example [1] and [2]. The first of these books employs Lyapunov functions in sufficiency characterizations of viability properties (not the same as stability with respect to all solutions), while the second one (see Chapter 6, and especially §4) shows various converse theorems that result in nondifferentiable Lyapunov functions, connecting their existence with the solution of optimal control problems. In a recent work [20], one can find conclusions analogous to those in this paper but only for the very special case of linear differential inclusions, resulting in homogeneous "quasiquadratic" Lyapunov functions. Finally, let us mention the work [19] on systems with time-varying parameters, in which the author established, under the assumption of *exponential* stability, the existence of differentiable Lyapunov functions on compact sets, for the special case of equilibria.

10. Relations to stability of prolongations. In [7], [8], [28]–[30], the authors considered various notions of stability for systems of the type (1) (with $\mathcal{D}$ not nec-

essarily compact). These properties are defined in terms of the "prolongations" of the original system. The above papers investigated the relationships between such stability notions and the existence of continuous, not necessarily smooth, Lyapunov functions. In this section, we briefly discuss relations between UGAS stability and the notions considered in those papers, with the purpose of clarifying relations to this related previous work. For more details on the definitions and elementary properties of prolongation maps and the corresponding stability concepts, we refer the reader to the papers mentioned above.

We start with some abstract definitions. Let $F : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to 2^{\mathbb{R}^n}$, $(\xi, t) \mapsto F(\xi, t) \subseteq \mathbb{R}^n$ be any map from $\mathbb{R}^n \times \mathbb{R}_{\geq 0}$ to the set of subsets of $\mathbb{R}^n$. Associated to F, one defines $\mathfrak{D}F$ and $\mathfrak{J}F$ by

$$\mathfrak{D}F(\xi, t) = \big\{\eta \in \mathbb{R}^n : \text{ there exist sequences } \xi_n,\ \eta_n \in \mathbb{R}^n, \text{ and } t_n \geq 0 \text{ with } \xi_n \to \xi,\ \eta_n \to \eta,\ t_n \to t,\ \eta \in F(\xi_n, t_n)\big\},$$

$$\mathfrak{J}F(\xi, t) = \Big\{\eta \in \mathbb{R}^n : \text{ there exist } t_1, t_2, \ldots, t_k \geq 0 \text{ with } \sum_{i=1}^{k} t_i = t, \text{ such that } \eta \in F\big(F\big(\ldots F(F(\xi_n, t_1), t_2)\ldots, t_{k-1}\big), t_k\big)\Big\},$$

where $F(S, t) \stackrel{\text{def}}{=} \bigcup_{\xi \in S} F(\xi, t)$ for any subset S of $\mathbb{R}^n$.

The map F is called *cluster* if $\mathfrak{D}F = F$, and F is called *transitive* if $\mathfrak{J}F = F$.

For any system (1), consider the reachable set $\mathcal{R}^t(\xi)$ defined in §5, seen now as a set-valued map. The prolongation map Γ associated with (1) is then defined by letting $\Gamma(\xi, t)$ be the smallest set containing $\mathcal{R}^t(\xi)$ such that Γ is both transitive and cluster. For further discussion regarding the definition of the map Γ, we refer the reader to [28] and to the other papers mentioned above.

For subsets A and B of $\mathbb{R}^n$, we denote the usual distance between the two sets by $d(A, B) = \inf\{d(\xi, \eta) : \xi \in A,\ \eta \in B\}$. We say that a system (1) is T-stable (we use here the "T" for the name of the author of [28] who, in turn, was inspired by previous work [8]) with respect to a closed, invariant set $\mathcal{A}$ if the following two properties hold:

- There exists a $\mathcal{K}_\infty$-function $\delta(\cdot)$ such that for any $\varepsilon > 0$,

$$d(\Gamma(\xi, t), \mathcal{A}) < \varepsilon, \quad \text{whenever } |\xi|_{\mathcal{A}} \leq \delta(\varepsilon), \text{ and } t \geq 0;$$

- For any $r, \varepsilon > 0$, there is a $T > 0$ such that

$$d(\Gamma(\xi, t), \mathcal{A}) < \varepsilon, \quad \text{whenever } |\xi|_{\mathcal{A}} < r, \text{ and } t \geq T.$$

Note that this is the same as what is called "global absolute asymptotic stability" (global AAS) in [28] for the special case when $\mathcal{A}$ is compact. Clearly, if a system is T-stable, then it is UGAS. It was shown in [28], under some extra technical assumptions but without the compactness of $\mathcal{D}$, that global AAS implies the existence of a continuous, not necessarily smooth, Lyapunov function (meaning that V is globally merely continuous; the condition $L_{f_d} V(\xi) \leq -\alpha_3(|\xi|_{\mathcal{A}})$ is replaced by a condition that V should decrease along trajectories).

We will show next that, at least when $\mathcal{D}$ is compact, UGAS implies (and is therefore equivalent to) T-stability. So in what follows in this section, we assume that $\mathcal{D}$ is compact, and also that all systems involved are forward complete. We first need the following fact.

LEMMA 10.1. *For system* (1), $\Gamma(\xi, t) = \overline{\mathcal{R}^t(\xi)}$ *for any* $\xi \in \mathbb{R}^n$ *and any* $t \geq 0$.

Proof. First note that the cluster property of Γ implies that $\Gamma(\xi, t)$ is closed for each $\xi \in \mathbb{R}^n$ and each $t \geq 0$. Thus it is enough to show that the map $\mathfrak{R} : (\xi, t) \mapsto \overline{\mathcal{R}^t(\xi)}$ is cluster and transitive.

Take $\xi_0 \in \mathbb{R}^n$ and $\tau > 0$. (The case when $t = 0$ is trivial.) Pick $\eta_0 \in \mathfrak{D}\mathfrak{R}(\xi_0, \tau)$. Then, by definition, there exist sequences $\{\xi_n\}$, $\{\eta_n\}$, and $\{t_n\}$ with $t_n \geq 0$ such that $\xi_n \to \xi_0$, $\eta_n \to \eta_0$, $t_n \to \tau$, and $\eta_n \in \overline{\mathcal{R}^{t_n}(\xi_n)}$.

Note then that for each n, there exists d_n such that

$$|\eta_n - x(t_n, \xi_n, d_n)| < \frac{1}{n}.$$

Let $\zeta_n = x(t_n, \xi_n, d_n)$. Then $\zeta_n \in \mathcal{R}^{t_n}(\xi_n)$ and $\zeta_n \to \eta_0$. Let K_0 be a compact set such that $\xi_n \in K_0$ for each n, and let $T > 0$ be such that $t_n \leq T$ for any n. Then by Proposition 5.1, there exists a compact set K_1 such that $\mathcal{R}(K_0, T) \subseteq K_1$. Let L be a Lipschitz constant for f with respect to states in K_1. Then it follows from Gronwall's Lemma that, for n large enough so that $|\xi_n - \xi_0| < e^{-LT}$, it holds that

$$|x(t, \xi_0, d_n) - x(t, \xi_n, d_n)| \leq |\xi_0 - \xi_n| e^{LT}$$

for any $0 \leq t \leq T$. Let $\kappa_n = x(\tau, \xi_0, d_n)$. Then

$$\begin{aligned} |\kappa_n - \zeta_n| &= |x(\tau, \xi_0, d_n) - x(t_n, \xi_n, d_n)| \\ &\leq |x(\tau, \xi_0, d_n) - x(\tau, \xi_n, d_n)| + |x(\tau, \xi_n, d_n) - x(t_n, \xi_n, d_n)| \\ &\leq |\xi_0 - \xi_n| e^{\tau L} + M|\tau - t_n|, \end{aligned}$$

where $M = \max\{|f(\xi, \mathbf{d})|,\ d(\xi, K_1) \leq 1,\ \mathbf{d} \in \mathcal{D}\}$. It then follows that $\kappa_n \in \mathcal{R}^\tau(\xi_0)$ for each n and $\kappa_n \to \eta_0$. Thus, we conclude that $\eta_0 \in \overline{\mathcal{R}^\tau(\xi_0)}$. Hence we showed that $\mathfrak{D}\overline{\mathcal{R}^\tau(\xi_0)} = \overline{\mathcal{R}^\tau(\xi_0)}$ for any $\tau > 0$ and any $\xi_0 \in \mathbb{R}^n$, that is, the map $\mathfrak{R}$ is cluster.

To show the transitivity of $\mathfrak{R}$, first note that, by induction, it is enough to show that

$$\mathfrak{R}(\mathfrak{R}(\xi, t_1),\ t_2) \subseteq \mathfrak{R}(\xi, t_1 + t_2) \tag{53}$$

for any $\xi \in \mathbb{R}^n$ and any $t_1, t_2 \geq 0$.

Applying Lemma 5.3 to $S = R^{t_1}(\xi)$, together with the fact that

$$\mathcal{R}^{t_2}\left(\mathcal{R}^{t_1}(\xi)\right) = \mathcal{R}^{t_1+t_2}(\xi),$$

one immediately gets (53). □

Rewriting the definition of UGAS in terms of reachable sets, one has that a system (1) is UGAS if and only if the following properties hold:

- There exists a $\mathcal{K}_\infty$-function $\delta(\cdot)$ such that for any $\varepsilon > 0$,

$$d\left(\mathcal{R}^t(\xi), \mathcal{A}\right) < \varepsilon, \quad \text{whenever } |\xi|_{\mathcal{A}} \leq \delta(\varepsilon), \text{ and } t \geq 0;$$

- For any $r, \varepsilon > 0$, there is a $T > 0$ such that

$$d\left(\mathcal{R}^t(\xi), \mathcal{A}\right) < \varepsilon, \quad \text{whenever } |\xi|_{\mathcal{A}} < r, \text{ and } t \geq T.$$

The following conclusion then follows immediately from the continuity of the function $\xi \mapsto d(\xi, \mathcal{A})$ and Lemma 10.1:

PROPOSITION 10.2. *For compact $\mathcal{D}$, a system* (1) *is* UGAS *with respect to $\mathcal{A}$ if and only if it is T-stable.*

Remark 10.3. In the special case when $\mathcal{A}$ is compact, a UGAS system is always forward complete. Thus in that case Proposition 10.2 is still true without completeness.

Remark 10.4. The compactness condition on $\mathcal{D}$ is essential. Without the compactness of $\mathcal{D}$, Proposition 10.2 is in general not true. For instance, the system defined by (50) in §8 is UGAS with respect to the origin $(0, 0)$. However the system is not T-stable, since $\Gamma(0, t) = \mathbb{R}^2$ for any $t > 0$. Note that for this example, $\overline{R^t(0, t)} = \{0\}$ for any $t > 0$ which is different from $\Gamma(0, t)$. The inconsistency with the conclusion of Lemma 10.1 is caused by the noncompactness of $\mathcal{D}$.

Appendix A. Some basic definitions. In this section we recall some standard concepts from stability theory.

A function $\gamma : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ is:

- a *$\mathcal{K}$-function* if it is continuous, strictly increasing and $\gamma(0) = 0$;
- a *$\mathcal{K}_\infty$-function* if it is a $\mathcal{K}$-function and also $\gamma(s) \to \infty$ as $s \to \infty$;
- a *positive definite* function if $\gamma(s) > 0$ for all $s > 0$, and $\gamma(0) = 0$.

A function $\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ is a *$\mathcal{KL}$-function* if:

- for each fixed $t \geq 0$ the function $\beta(\cdot, t)$ is a $\mathcal{K}$-function, and
- for each fixed $s \geq 0$ it is decreasing to zero as $t \to \infty$.

Note that we are not requiring β to be continuous in both variables simultaneously; however it turns out in our results that this stronger property will usually hold.

Appendix B. Smooth approximations of locally Lipschitz functions. In the proof of the converse Lyapunov theorem, we used a parameterized version of an approximation theorem given in [31]. For convenience of reference, and to make this work self-contained and expository, we next provide the needed variation of the theorem and its proof. (Several details, missing in the proof in [31], have been included as well.)

THEOREM B.1. *Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathcal{D}$ be a compact subset of $\mathbb{R}^l$, and assume given:*

- *a locally Lipschitz function $\Phi : \mathcal{O} \longrightarrow \mathbb{R}$;*
- *a continuous map $f : \mathbb{R}^n \times \mathcal{D} \longrightarrow \mathbb{R}^n$, $(x, \mathbf{d}) \mapsto f(x, \mathbf{d})$ which is locally Lipschitz on x uniformly on $\mathbf{d}$;*
- *a continuous function $\alpha : \mathcal{O} \longrightarrow \mathbb{R}$ and continuous functions $\mu, \nu : \mathcal{O} \longrightarrow \mathbb{R}_{>0}$*

such that for each $\mathbf{d} \in \mathcal{D}$,

$$L_{f_\mathbf{d}}\Phi(\xi) \leq \alpha(\xi), \quad a.e.\ \xi \in \mathcal{O}, \tag{B.54}$$

where $f_\mathbf{d}$ is the vector field defined by $f_\mathbf{d}(\cdot) = f(\cdot, \mathbf{d})$. (Recall that $\nabla\Phi$ is defined a.e., since Φ is locally Lipschitz, by Rademacher's theorem, see e.g. [5, p. 216].) *Then there exists a smooth function $\Psi : \mathcal{O} \longrightarrow \mathbb{R}$ such that*

$$|\Phi(\xi) - \Psi(\xi)| < \mu(\xi), \quad \forall \xi \in \mathcal{O}$$

and for each $\mathbf{d} \in \mathcal{D}$,

$$L_{f_\mathbf{d}}\Psi(\xi) \leq \alpha(\xi) + \nu(\xi), \quad \forall \xi \in \mathcal{O}.$$

To prove the theorem, we first need some easy facts about regularization. Let $\psi : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a smooth nonnegative function which vanishes outside of the unit

disk and satisfies

$$\int_{\mathbb{R}^n} \psi(s)\, ds \;=\; 1\,.$$

For any measurable, locally essentially bounded function $\Phi : \mathcal{O} \longrightarrow \mathbb{R}$ and $0 < \sigma \leq 1$, define the function Φ_σ by convolution with $\frac{1}{\sigma^n}\,\psi(\frac{s}{\sigma})$, that is:

$$\Phi_\sigma(\xi) \;\stackrel{\text{def}}{=}\; \int_{\mathbb{R}^n} \Phi(\xi + \sigma s)\psi(s)\, ds\,. \tag{B.55}$$

We think of this function as defined only for those ξ so that $\xi + \sigma s \in \mathcal{O}$ for all $|s| \leq 1$. Note that the integral is finite, as the integrand is essentially bounded and of compact support. The following observation is a standard approximation exercise, so we omit its proof.

LEMMA B.2. *For each compact subset K of $\mathcal{O}$, there exists some $\sigma_0 > 0$ such that Φ_σ is defined on K, and smooth there, for all $\sigma < \sigma_0$. Moreover, if Φ is continuous, then Φ_σ approaches Φ uniformly on K, as σ tends to 0.*

Now assume that Φ is a locally Lipschitz function. Then, for each $\mathbf{d} \in \mathcal{D}$, $L_{f_\mathbf{d}}\Phi$ is defined almost everywhere, and furthermore, on any compact subset $K \subseteq \mathcal{O}$,

$$|L_{f_\mathbf{d}}\Phi(\xi)| \;\leq\; k\,|f(\xi, \mathbf{d})|\,, \quad \text{a.e. } \xi \in K\,,\ \forall \mathbf{d} \in \mathcal{D}\,,$$

where k is a Lipschitz constant for Φ on K. Therefore, for each $\mathbf{d}$ (omitting from now on the $\mathbb{R}^n$ in integrals)

$$\left(L_{f_\mathbf{d}}\Phi\right)_\sigma(\xi) = \int \left(L_{f_\mathbf{d}}\Phi\right)(\xi + \sigma s)\psi(s)\, ds$$

is well defined as long as $\xi + \sigma s \in \mathcal{O}$ for all $|s| \leq 1$. Applying Lemma B.2 to $(L_{f_\mathbf{d}}\Phi)_\sigma$, this is smooth for any $\sigma > 0$ small.

Suppose that for all $\mathbf{d} \in \mathcal{D}$,

$$L_{f_\mathbf{d}}\Phi(\xi) \leq \alpha(\xi)\,, \quad \text{a.e. } \xi \in \mathcal{O}\,, \tag{B.56}$$

for some continuous function α. Pick any compact subset $K \subseteq \mathcal{O}$. On this set K, we have

$$\begin{aligned}\left(L_{f_\mathbf{d}}\Phi\right)_\sigma(\xi) &= \int \left(L_{f_\mathbf{d}}\Phi\right)(\xi + \sigma s)\psi(s)\, ds \;\leq\; \int \alpha(\xi + \sigma s)\psi(s)\, ds \\ &\leq \alpha(\xi) + \max_{|s|\leq 1,\, \xi \in K} |\alpha(\xi + \sigma s) - \alpha(\xi)|\,.\end{aligned}$$

From here we get the following conclusion.

LEMMA B.3. *For any compact subset K of $\mathcal{O}$, $(L_{f_\mathbf{d}}\Phi)_\sigma$ is a C^∞ function defined on K for all σ small enough, and, if* (B.56) *holds for all $\mathbf{d} \in \mathcal{D}$ and all $\xi \in \mathcal{O}$, then for any $\varepsilon > 0$ given, there exists some $\sigma_0 > 0$ such that*

$$(L_{f_\mathbf{d}}\Phi)_\sigma(\xi) \;\leq\; \alpha(\xi) + \varepsilon$$

for all $\sigma \leq \sigma_0$, all $\mathbf{d} \in \mathcal{D}$, and all $\xi \in K$.

The following lemma illustrates the relationship between $L_{f_\mathbf{d}}(\Phi_\sigma)$ and $(L_{f_\mathbf{d}}\Phi)_\sigma$.

LEMMA B.4. *On any compact subset K of $\mathcal{O}$,*

$$\sup_{\mathbf{d}\in\mathcal{D}, \xi \in K} \left|L_{f_\mathbf{d}}(\Phi_\sigma)(\xi) - \left(L_{f_\mathbf{d}}\Phi\right)_\sigma(\xi)\right| \;\to\; 0$$

as σ tends to 0.

Proof. For each $\xi \in \mathcal{O}$, we use $\varphi(t, \xi, \mathbf{d})$ to denote the solution of the differential equation

$$\dot{x} = f(x, \mathbf{d})$$

with the initial condition $\varphi(0, \xi, \mathbf{d}) = \xi$. It follows from the assumptions on f and compactness of K and $\mathcal{D}$ that there exist some compact neighborhood V of K and some $\tau_1 > 0$ and $\sigma_0 > 0$ such that $\varphi(t, \xi + \sigma s, \mathbf{d}) \in V$ for all $\xi \in K$, $|s| \leq 1$, $\sigma \leq \sigma_0$, $\mathbf{d} \in \mathcal{D}$, and $|t| \leq \tau_1$.

For the Lipschitz function Φ, we have, for all $\xi, \mathbf{d}$ and $\sigma \leq \sigma_0$,

$$\begin{aligned} L_{f_{\mathbf{d}}}(\Phi_\sigma)(\xi) &= \frac{d}{dt}\Big|_{t=0} \Phi_\sigma(\varphi(t, \xi, \mathbf{d})) = \frac{d}{dt}\Big|_{t=0} \int \Phi(\varphi(t, \xi, \mathbf{d}) + \sigma s)\psi(s)\, ds \\ &= \lim_{t \to 0} \frac{1}{t} \int (\Phi(\varphi(t, \xi, \mathbf{d}) + \sigma s) - \Phi(\xi + \sigma s))\psi(s)\, ds\,, \end{aligned}$$

and

$$(L_{f_{\mathbf{d}}}\Phi)_\sigma(\xi) = \int L_{f_{\mathbf{d}}}\Phi(\xi + \sigma s)\psi(s)\, ds \tag{B.57}$$

$$= \int \frac{d}{dt}\Big|_{t=0} \Phi(\varphi(t, \xi + \sigma s, \mathbf{d}))\psi(s)\, ds \tag{B.58}$$

$$= \lim_{t \to 0} \frac{1}{t} \int [\Phi(\varphi(t, \xi + \sigma s, \mathbf{d})) - \Phi(\xi + \sigma s)]\, \psi(s)\, ds\,. \tag{B.59}$$

Notice that the integrand in (B.57) equals that in (B.58) almost everywhere on s (for each fixed ξ and σ) and that (B.59) follows from (B.58) because of the Lebesgue dominated convergence theorem and the following fact:

$$\begin{aligned} &\frac{1}{|t|} |\Phi(\varphi(t, \xi + \sigma s, \mathbf{d})) - \Phi(\xi + \sigma s)|\, \psi(s) \\ &\leq \frac{k}{|t|} |\varphi(t, \xi + \sigma s, \mathbf{d}) - (\xi + \sigma s)|\, \psi(s) \leq kC\psi(s)\,, \quad \forall t \in [-\tau_1, \tau_1]\,, \end{aligned}$$

where $C \stackrel{\text{def}}{=} \max_{\xi \in V, \mathbf{d} \in \mathcal{D}} |f(\xi, \mathbf{d})|$ and k is a Lipschitz constant for Φ on V.

Now one sees that

$$L_{f_{\mathbf{d}}}(\Phi_\sigma)(\xi) - (L_{f_{\mathbf{d}}}\Phi)_\sigma(\xi) = \lim_{t \to 0} \frac{1}{t} \int [\Phi(\varphi(t, \xi, \mathbf{d}) + \sigma s) - \Phi(\varphi(t, \xi + \sigma s, \mathbf{d}))]\psi(s)\, ds.$$

Thus it is enough to show that for any $\varepsilon > 0$, there exist some $\delta > 0$ and $\tau^* > 0$ such that the above integral is bounded by ε for all $\mathbf{d} \in \mathcal{D}$, $\xi \in K$, $|t| < \tau^*$, and $\sigma < \delta$. This is basically a standard argument on continuous dependence on initial conditions, but we provide the details. For $0 \leq \tau \leq \tau_1$, let

$$\gamma(\tau) \stackrel{\text{def}}{=} \sup \{|f(\varphi(t, \zeta, \mathbf{d}), \mathbf{d}) - f(\zeta, \mathbf{d})| : \ |t| \leq \tau, \ \zeta \in V, \ \mathbf{d} \in \mathcal{D}\}\,.$$

Then $\gamma(0) = 0$, and γ is nondecreasing and continuous at $t = 0$, because

$$|f(\varphi(t, \zeta, \mathbf{d}), \mathbf{d}) - f(\zeta, \mathbf{d})| \leq C_3 |\varphi(t, \zeta, \mathbf{d}) - \zeta| \leq C_3 C_4 |t|\,,$$

where C_3 is a (uniform) Lipschitz constant for f on V_1, C_4 is an upper bound for $|f(\xi, \mathbf{d})|$ on V_1, and V_1 is some compact neighborhood of V such that $\varphi(t, \zeta, \mathbf{d}) \in V_1$ for any $\zeta \in V$, $\mathbf{d} \in \mathcal{D}$, and $|t| \leq \tau_1$. For any $\zeta \in V$, $\mathbf{d} \in \mathcal{D}$, and $|t| \leq \tau_1$,

$$|\varphi(t, \zeta, \mathbf{d}) - (\zeta + tf(\zeta, \mathbf{d}))| \leq \int_0^{|t|} \gamma(\tau)\, d\tau \leq |t|\, \gamma(|t|)\,.$$

Now for $\xi \in K$, we have

$$\begin{aligned}
&|\Phi(\varphi(t, \xi, \mathbf{d}) + \sigma s) - \Phi(\varphi(t, \xi + \sigma s), \mathbf{d})| \\
&\quad \leq k\, |\varphi(t, \xi, \mathbf{d}) + \sigma s - \varphi(t, \xi + \sigma s, \mathbf{d})| \\
&\quad \leq k\, |\xi + \sigma s + tf(\xi, \mathbf{d}) - (\xi + \sigma s + tf(\xi + \sigma s, \mathbf{d}))| \\
&\qquad + k\, |\varphi(t, \xi, \mathbf{d}) - (\xi + tf(\xi, \mathbf{d}))| + k|\varphi(t, \xi + \sigma s, \mathbf{d}) \\
&\qquad - (\xi + \sigma s + tf(\xi + \sigma s, \mathbf{d}))| \\
&\quad \leq k\, |t|\, |f(\xi, \mathbf{d}) - f(\xi + \sigma s, \mathbf{d})| + 2k\, |t|\, \gamma(|t|)\,. \qquad \text{(B.60)}
\end{aligned}$$

Finally, for $\varepsilon > 0$, let δ and τ^* be such that

$$\gamma(\tau) < \frac{\varepsilon}{3k} \quad \text{and} \quad |f(\xi, \mathbf{d}) - f(\xi + \sigma s, \mathbf{d})| < \frac{\varepsilon}{3k}$$

for any $\xi \in K$, $\mathbf{d} \in \mathcal{D}$, $|s| \leq 1$, $\sigma < \delta$, and $|t| < \tau^*$. It then follows from (B.60) that

$$\frac{1}{|t|} \int [\Phi(\varphi(t, \xi, \mathbf{d}) + \sigma s) - \Phi(\varphi(t, \xi + \sigma s, \mathbf{d}))]\psi(s)\, ds < \int \varepsilon \psi(s)\, ds = \varepsilon$$

for any $\xi \in K$, $\mathbf{d} \in \mathcal{D}$, $|t| < \tau^*$, and $\sigma < \delta$, which implies

$$|L_{f_\mathbf{d}}(\Phi_\sigma)(\xi) - (L_{f_\mathbf{d}}\Phi)_\sigma(\xi)| < \varepsilon$$

for any $\sigma < \sigma_0$, $\mathbf{d} \in \mathcal{D}$, and $\xi \in K$. □

Combining the previous three lemmas, we obtain the following conclusion.

LEMMA B.5. *Let K be a compact subset of $\mathcal{O}$. Then for any given $\varepsilon > 0$, there exists some smooth function Ψ defined on K such that*

$$|\Psi(\xi) - \Phi(\xi)| < \varepsilon \quad \textit{and} \quad L_{f_\mathbf{d}}\Psi(\xi) \leq \alpha(\xi) + \varepsilon$$

for all $\xi \in K$, $\mathbf{d} \in \mathcal{D}$.

Now we are ready to complete the proof of Theorem B.1. For the open subset $\mathcal{O}$ of $\mathbb{R}^n$, let $\{\mathcal{U}_i\}$ be a locally finite, countable cover of $\mathcal{O}$ with $\bar{\mathcal{U}}_i$ compact and $\bar{\mathcal{U}}_i \subseteq \mathcal{O}$. Let $\{\beta_i\}$ be a partition of unity on $\mathcal{O}$ subordinate to $\{\mathcal{U}_i\}$. For any given positive functions $\mu(\cdot)$ and $\nu(\cdot)$, let

$$\varepsilon_i \stackrel{\text{def}}{=} \min\left\{\inf_{\xi \in \mathcal{U}_i} \mu(\xi),\ \inf_{\xi \in \mathcal{U}_i} \nu(\xi)\right\}.$$

For each i, it follows from Lemma B.5 that there exists some smooth function Ψ_i defined on $\bar{\mathcal{U}}_i$ such that

$$|\Phi(\xi) - \Psi_i(\xi)| < \frac{\varepsilon_i}{2^{i+1}(1 + \tau_i)} \quad \text{and} \quad L_{f_\mathbf{d}}\Psi_i(\xi) \leq \alpha(\xi) + \frac{\varepsilon_i}{2}$$

on $\bar{\mathcal{U}}_i$, where $\tau_i \stackrel{\text{def}}{=} \max\{|L_{f_{\mathbf{d}}}\beta_i(\xi)| : \xi \in \bar{\mathcal{U}}_i, \mathbf{d} \in \mathcal{D}\}$. We define $\Psi = \sum_i \beta_i \Psi_i$. Clearly Ψ is a smooth function defined on $\mathcal{O}$, and

$$\begin{aligned}|\Psi(\xi) - \Phi(\xi)| &\leq \sum_{j \in \mathcal{J}_\xi} \beta_j(\xi)\, |\Psi_j(\xi) - \Phi(\xi)| \\ &< \max_{j \in \mathcal{J}_\xi} \varepsilon_j \;\leq\; \mu(\xi)\,,\end{aligned}$$

where $\mathcal{J}_\xi \stackrel{\text{def}}{=} \{j : \xi \in \mathcal{U}_j\}$.

For $L_{f_{\mathbf{d}}}\Psi$, one has

$$\begin{aligned}L_{f_{\mathbf{d}}}\Psi(\xi) &= L_{f_{\mathbf{d}}}\Phi(\xi) + L_{f_{\mathbf{d}}}\left(\sum_i \beta_i(\Psi_i - \Phi)\right)(\xi) \\ &= L_{f_{\mathbf{d}}}\Phi(\xi) + \sum \left(L_{f_{\mathbf{d}}}\beta_i\right)(\Psi_i - \Phi)(\xi) + \sum \beta_i \left(L_{f_{\mathbf{d}}}\Psi_i(\xi) - L_{f_{\mathbf{d}}}\Phi(\xi)\right) \\ &= \sum_{j \in \mathcal{J}_\xi} \left(L_{f_{\mathbf{d}}}\beta_j\right)(\Psi_j - \Phi)(\xi) + \sum_{j \in \mathcal{J}_\xi} \beta_j L_{f_{\mathbf{d}}}\Psi_j(\xi) \\ &< \sum_{j \in \mathcal{J}_\xi} \frac{\varepsilon_j}{2^{j+1}} + \sum_{j \in \mathcal{J}_\xi} \beta_j(\xi)\left(\alpha(\xi) + \frac{\varepsilon_i}{2}\right) \\ &\leq \frac{1}{2}\max_{j \in \mathcal{J}_\xi}\{\varepsilon_j\} + \alpha(\xi) + \frac{1}{2}\max_{j \in \mathcal{J}_\xi}\{\varepsilon_j\} \\ &\leq \alpha(\xi) + \nu(\xi)\,.\end{aligned}$$

We conclude that Ψ is the desired function. □

Acknowledgments. We wish to thank the Institute for Mathematics and Its Applications for providing an excellent research environment during the Special Year in Control Theory (1992–1993); part of this work was completed while the authors visited the IMA. We also wish to thank John Tsinias and Randy Freeman for useful comments, and most especially Héctor Sussmann for help with the proof of Proposition 5.1.

REFERENCES

[1] J.-P. Aubin, *Viability Theory*, Birkhäuser Boston, Cambridge, MA, 1991.
[2] J.-P. Aubin and A. Cellina, *Differential Inclusions: Set-Valued Maps and Viability Theory*, Springer-Verlag, New York, 1984.
[3] N. P. Bhatia and G. P. Szegö, *Stability Theory of Dynamical Systems*, Springer-Verlag, New York, 1970.
[4] T. Bröcker, *Differentiable Germs and Catastrophes*, Cambridge University Press, Cambridge, 1975.
[5] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.
[6] M. W. Hirsch, *Differential Topology*, Springer-Verlag, New York, 1976.
[7] N. Kalouptsidis, *Accessibility and Stability Theory of Nonlinear Control Systems*, Ph.D. thesis, Washington University, Saint Louis, Missouri, 1977.
[8] N. Kalouptsidis and D. L. Elliott, *Stability analysis of the orbits of control systems*, Math. Systems Theory, 15 (1982), pp. 323–342.
[9] I. Kanellakopoulos, P. V. Kokotovic, and A. S. Morse, *A toolkit for nonlinear feedback design*, Systems Control Lett., 18 (1992), pp. 83–92.
[10] P. E. Kloeden, *Eventual stability in general control systems*, J. Differential Equations, 19 (1975), pp. 106–124.

[11] ———, *General control systems*, in Mathematical Control Theory (Proceedings, Canberra, Australia, 1977), A. Bold and B. Eckmann, eds., Springer-Verlag, New York, 1978, pp. 119–137.
[12] N. N. KRASOVSKIĬ, *Stability of Motion*, Stanford University Press, Stanford, CA, 1963.
[13] J. KURZWEIL, *On the inversion of Lyapunov's second theorem on stability of motion*, Amer. Math. Soc. Transl. Ser. 2, 24 (1956), pp. 19–77.
[14] V. LAKSHMIKANTHAM, S. LEELA, AND A. A. MARTYNYUK, *Stability Analysis of Nonlinear Systems*, Marcel Dekker, New York, 1989.
[15] S. LEFSCHETZ, *Differential Equations: Geometric Theory*, 2nd ed., Interscience, New York, 1963.
[16] Y. LIN, *Lyapunov Function Techniques for Stabilization*, Ph.D. thesis, Rutgers, The State University of New Jersey, New Brunswick, NJ, 1992.
[17] Y. LIN, E. D. SONTAG, AND Y. WANG, *Input to state stabilizability for parameterized families of systems*, International Journal of Robust and Nonlinear Control, 5 (1995), pp. 187–205.
[18] J. L. MASSERA, *Contributions to stability theory*, Ann. of Math., 64 (1956), pp. 182–206.
[19] A. M. MEILAKHS, *Design of stable control systems subject to parametric perturbation*, Avtomat. i Telemekh., 10 (1978), pp. 5–15.
[20] A. P. MOLCHANOV AND Y. S. PYANITSKIY, *Criteria of asymptotic stability of differential and difference inclusions encountered in control theory*, Systems Control Lett., 13 (1989), pp. 59–64.
[21] T. NADZIEJA, *Construction of a smooth Lyapunov function for an asymptotically stable set*, Czechoslovak Math. J., 115 (1990), pp. 195–199.
[22] L. ROSIER, *Inverse of Lyapunov's second theorem for measurable functions*, in Proceedings of Nonlinear Control Systems Design Symposium, Bordeaux, France, 1992, M. Fliess, ed., IFAC Publications, pp. 655–660.
[23] E. O. ROXIN, *Stability in general control systems*, J. Differential Equations, 1 (1965), pp. 115–150.
[24] P. SEIBERT, *Stability under perturbations in generalized dynamical systems*, in International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, J. P. LaSalle and S. Lefschetz, eds., Academic Press, New York, 1963, pp. 463–473.
[25] E. D. SONTAG, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 435–443.
[26] ———, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Springer-Verlag, New York, 1990.
[27] E. D. SONTAG AND Y. WANG, *On characterizations of the input-to-state stability property*, Systems Control Lett., 24 (1995), pp. 351–359.
[28] J. TSINIAS, *A Lyapunov description of stability in control systems*, Nonlinear Anal., 13 (1989), pp. 63–74.
[29] J. TSINIAS AND N. KALOUPTSIDIS, *Prolongations and stability analysis via Lyapunov functions of dynamical polysystems*, Math. Systems Theory, 20 (1987), pp. 215–233.
[30] J. TSINIAS, N. KALOUPTSIDIS, AND A. BACCIOTTI, *Lyapunov functions and stability of dynamical polysystems*, Math. Systems Theory, 19 (1987), pp. 333–354.
[31] F. W. WILSON, JR., *Smoothing derivatives of functions and applications*, Trans. Amer. Math. Soc., 139 (1969), pp. 413–428.
[32] T. YOSHIZAWA, *Stability Theory and the Existence of Periodic Solutions and Almost Periodic Solutions*, Springer-Verlag, New York, 1975.
[33] V. I. ZUBOV, *Methods of A. M. Lyapunov and Their Application*, English ed., Noordhoff, Groningen, The Netherlands, 1964.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 161–178, January 1996

007

DETERMINISTIC APPROXIMATION FOR STOCHASTIC CONTROL PROBLEMS*

R. SH. LIPTSER†, W. J. RUNGGALDIER‡, AND M. TAKSAR§

Abstract. We consider a class of stochastic control problems where uncertainty is due to driving noises of general nature as well as to rapidly fluctuating processes affecting the drift. We show that, when the noise "intensity" is small and the fluctuations become fast, the stochastic problems can be approximated by a deterministic one. We also show that the optimal control of the deterministic problem is asymptotically optimal for the stochastic problems.

Key words. stochastic and deterministic control, stochastic differential equations, weak convergence, asymptotic optimality

AMS subject classifications. 93E20, 93C15, 60B10, 60F17, 60G44, 49J15, 49K40, 49M45

1. Introduction.

There are only few stochastic control problems that can be solved in closed form. A lot of effort has therefore been put into developing approximation techniques for such problems. One approach in this direction is to consider, instead of the original model, a model where the underlying processes are replaced by simpler ones. This approach makes it possible to construct nearly optimal controls for the original model, based on the solution to the simpler model. This simpler model may involve underlying processes that are diffusions ("diffusion approximation"), but it may also simply be a deterministic model ("fluid approximation"). A general tool, especially for diffusion approximations, is techniques of weak convergence of random processes [1], [3], [6], [15] combined with an averaging principle [5]. This methodology is actively used in various practical problems of engineering, manufacturing, queuing, inventory, and others and is studied, e.g., in [7]–[13].

The underlying idea of this methodology is actually rather simple, but the mathematics required for its implementation are in general quite sophisticated. Although there exist some general approaches (see, e.g., [9]), in each particular case the rigorous verification of the convergence of the controlled systems requires specific technical tools and ideas.

In the present paper we apply "fluid approximation" techniques to a rather general stochastic control model with convex control cost function. In this model the controlled process X is described by a stochastic differential equation with respect to a general (not necessarily continuous) martingale M. The control affects the drift of X; this drift is furthermore affected by a rapidly fluctuating exogenous process ξ. To implement the

*Received by the editors August 2, 1993; accepted for publication (in revised form) August 18, 1994. This research was supported in part by Gruppo Nazionale per l'Analisi Funzionale e sue Applicazione of the Italian National Research Council..

†Department of Electrical Engineering Systems, Tel Aviv University, 69978 Ramat Aviv, Tel Aviv, Israel. Part of this research was performed while the author was at the University of Padova, Padova, Italy.

‡Dipartimento di Matematica Pura ed Applicata, Università di Padova, Via Belzoni 7, 35131 Padova, Italy.

§Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794-3600. The research of this author was supported by NSF grant DMS 9301200 and NATO scientific exchange grant CRG 900147. Part of this research was performed while the author was at the University of Padova, Padova, Italy.

approximation approach, we embed the given model into a family of similar models, parametrized by a small parameter $\varepsilon > 0$. We consider the case when the "intensity" of the random noise disturbance M becomes small with ε, while the "contaminating" process ξ fluctuates with increasing speed. For such a case the limiting model becomes deterministic, and it is possible to obtain asymptotically (as $\varepsilon \downarrow 0$) optimal controls for the prelimit models by using the optimal control of the limiting deterministic system.

Although we consider explicitly only the case when the controlled state process X can be completely observed, our results nevertheless hold in the same form when the state is only partially observed.

In a more formal way, we have a family of controlled stochastic systems, parametrized by a small (positive) parameter ε ($\varepsilon \downarrow 0$), with dynamics

$$dX_t^\varepsilon = \left[a(X_t^\varepsilon, \xi_{t/\varepsilon}) + b(X_t^\varepsilon)u^\varepsilon(t)\right] dt + dM_t^\varepsilon \tag{1.1}$$

and initial condition X_0^ε. Here $X^\varepsilon = (X_t^\varepsilon)$ is the controlled state (or signal) process, $\xi = (\xi_t)$ is the "contamination" process affecting the drift of X^ε, and $M^\varepsilon = (M_t^\varepsilon)$ is a process representing the noise in the system. The random function $u^\varepsilon = (u^\varepsilon(t))$ is the control that affects the drift of X^ε in a linear way and satisfies the usual requirements for admissibility (see Definition 2.1 below).

Given a finite horizon $T > 0$, with each control u^ε we associate the cost

$$J^\varepsilon(u^\varepsilon) = E\left\{\int_0^T [p(X_t^\varepsilon) + q(u^\varepsilon(t))]dt + r(X_T^\varepsilon)\right\}, \tag{1.2}$$

where $p(x), q(u)$, and $r(x)$ are nonnegative functions on the real line referred to as holding cost, control cost, and terminal cost functions, respectively. The objective is to find

$$V^\varepsilon = \inf_{u^\varepsilon} J^\varepsilon(u^\varepsilon) \tag{1.3}$$

and an optimal (minimizing) control. For practical purposes one may just as well be interested in finding a nearly optimal control or, as will be the case here, an asymptotically (as $\varepsilon \downarrow 0$) optimal control.

To describe the limiting control model, we assume that the following ergodic properties hold:

$$P - \lim_{\varepsilon \to 0} X_0^\varepsilon = x_0, \quad x_0 \in \mathbf{R}, \tag{1.4}$$

$$\overline{a}(x) = P - \lim_{t \to \infty} \frac{1}{t} \int_0^t a(x, \xi_s)ds; \quad x \in \mathbf{R}, \tag{1.5}$$

$$P - \lim_{\varepsilon \to 0} \sup_{t \le T} |M_t^\varepsilon| = 0. \tag{1.6}$$

In the next section we formulate conditions under which (1.4)–(1.6) are valid.

The dynamics of the limiting system is given by the following ordinary differential equation:

$$dx(t) = [\overline{a}(x(t)) + b(x(t))u(t)]\, dt; \quad x(0) = x_0. \tag{1.7}$$

Here $x(t)$ is a (deterministic) controlled process and $u(t)$ is a (deterministic) control. Define

$$j(u) := \int_0^T [p(x(t)) + q(u(t))]\, dt + r(x(T)) \tag{1.8}$$

and

$$v := \inf_u j(u), \tag{1.9}$$

where the infimum is taken over all (deterministic) measurable functions on $[0, T]$.

Our main results are the following two theorems.

THEOREM 1.1. *The following relation holds:*

$$\lim_{\varepsilon \to 0} V^\varepsilon = v.$$

THEOREM 1.2. *Let $u^*(t)$, $0 \le t \le T$, be an optimal deterministic control for* (1.7)–(1.9). *Then $u^*(t)$ is asymptotically optimal for* (1.1)–(1.3) *in the sense that*

$$\lim_{\varepsilon \to 0} |J^\varepsilon(u^*) - V^\varepsilon| = 0.$$

Remark 1. If for the limit model there exists a feedback control

$$u^*(t) = u^\circ(t, x^*(t)),$$

where $x^*(t)$ is the controlled process defined by the differential equation (1.7) with $u(t) = u^*(t)$, and the function $u^\circ(t, x)$ is Lipschitz continuous in x uniformly in $t \in [0, T]$, then the statement of Theorem 1.2 remains true with $u^\circ(t, X_t^\varepsilon)$ replacing $u^*(t)$; i.e., the feedback control $u^\circ(t, X_t^\varepsilon)$ is asymptotically optimal.

Remark 2. The results obtained here for the one-dimensional control problem can be extended to an n-dimensional problem. The motivation to consider just the scalar case is to present the main ideas in the simplest form.

The main contribution of this paper is twofold: from a more theoretical point of view we obtain a stability result for the optimal control of a deterministic system in the sense that this control is asymptotically optimal for a large class of stochastic control problems of a rather complicated nature. From a practical point of view our results allow one to compute an asymptotically optimal control for a variety of problems under quite general conditions, where a direct approach would be impossible.

The proof consists of two parts carried out in §§3 and 4: first we show that v is an asymptotically lower bound for the optimal cost functions V^ε. Then we show that the deterministic optimal control of the limiting problem can be applied to the pre-limit models, yielding asymptotically optimal cost. Results of more technical nature, interesting in their own right, are moved to appendices (§§5, 6, and 7).

2. Main assumptions and notations. For simplicity we assume $\varepsilon \in (0, 1]$. For each ε let $SB := (\Omega, \mathcal{F}, \mathbf{F}^\varepsilon = (\mathcal{F}_t^\varepsilon)_{t \ge 0}, P)$ be a fixed stochastic basis, where $(\Omega, \mathcal{F}, P)$ is a complete probability space and $\mathbf{F}^\varepsilon$ is a filtration satisfying the "usual assumptions" (see [2]). The initial value X_0^ε of the state process is $\mathcal{F}_0^\varepsilon$-measurable, while $(\xi_{t/\varepsilon}), (M_t^\varepsilon)$ are $\mathbf{F}^\varepsilon$-adapted.

DEFINITION 2.1. *The control process* $u^\varepsilon = (u^\varepsilon(t))_{t\geq 0}$ *is said to be admissible if it is* $\mathbf{F}^\varepsilon$*-adapted and*

$$\int_0^T |u^\varepsilon(t)|\,dt < \infty, \quad P-a.s. \tag{2.1}$$

Throughout the paper we make the following assumptions:

(A.1) The control cost function $q(u)$ is nonnegative *convex* satisfying
$q(u) \geq c|u|^{1+\gamma}, \quad c, \gamma > 0.$

(A.2) The cost functions $p(x)$ and $r(x)$ are continuous nonnegative satisfying
$p(x),\ r(x) \leq c_1(1+|x|^{\gamma_1}), \quad c_1, \gamma_1 > 0.$

(A.3) There exist $x_0 \in \mathbf{R}$ and positive constants c_2, γ_2 such that
(i) $P - \lim_{\varepsilon\to 0} X_0^\varepsilon = x_0$,
(ii) $E|X_0^\varepsilon|^{2n^*} < c_2$,
where n^* is the smallest integer such that $\gamma_1 < n^*$.

(A.4) The function $a(x,y)$ is measurable in (x,y) and satisfies the linear growth and Lipschitz conditions in x (uniformly in y); i.e., there exists $\ell > 0$ such that
(i) $|a(x,y)| \leq \ell(1+|x|), \quad x, y \in \mathbf{R}$,
(ii) $|a(x',y) - a(x'',y)| \leq \ell|x' - x''|, \quad x', x'', y \in \mathbf{R}$.

(A.5) The function $b(x)$ is bounded and Lipschitz; i.e.,
(i) $|b(x)| \leq \ell$,
(ii) $|b(x') - b(x'')| \leq \ell|x' - x''|, \quad x', x'' \in \mathbf{R}$.

(A.6) The random process $\xi = (\xi_t)_{t\geq 0}$ is ergodic; namely, there exists a probability measure $\lambda(dy)$ on $\mathbf{R}$ such that for any bounded and measurable function $g(y)$

$$P - \lim_{t\to\infty} \frac{1}{t}\int_0^t g(\xi_s)\,ds = \int_{\mathbf{R}} g(y)\,\lambda(dy).$$

(A.7) The process $M^\varepsilon = (M_t^\varepsilon)_{t\geq 0}$ is a square integrable martingale with paths in the Skorokhod space $D[0,\infty)$ whose predictable quadratic variation $\langle M^\varepsilon\rangle_t$ satisfies
(i) $\langle M^\varepsilon\rangle_t = \varepsilon\int_0^t m_s^\varepsilon\,ds$
with bounded density m_s^ε. The latter means that there exists a constant c_3 such that
(ii) $m_t^\varepsilon \leq c_3; \quad t \leq T \quad P-$a.s.
The jumps $\Delta M_s^\varepsilon := M_s^\varepsilon - \lim_{v\uparrow s} M_v^\varepsilon$ are bounded, i.e., there exists a constant $L > 0$ such that
(iii) $|\Delta M_t^\varepsilon| \leq L; \quad t \leq T, \quad \varepsilon \in (0,1]$.

Note that by assumptions (A.4) and (A.5) equation (1.1) has a unique strong solution X^ε for every admissible control u^ε. We shall refer to X^ε as the state process associated with u^ε. The only requirement for the "contamination" process ξ is its ergodicity; no stationarity of ξ or independence from other processes is required. We furthermore remark that our results remain valid if M^ε is any process with paths in D satisfying
i) $\sup_{t\leq T}|M_t^\varepsilon| \xrightarrow{P} 0 \quad \forall T > 0$ (see derivations (4.5) and (6.3) below),
ii) $\sup_\varepsilon E\sup_{t\leq T}|M_t^\varepsilon|^{2n} < \infty$, $n \geq 1$ (see §7).
In this more general case, a rigorous representation of the dynamics of the system should be made in the integral form below rather than in the differential form (1.1):

$$X_t^\varepsilon = X_0^\varepsilon + \int_0^t \left[a(X_s^\varepsilon, \xi_{s/\varepsilon}) + b(X_s^\varepsilon)u^\varepsilon(s)\right]ds + M_t^\varepsilon.$$

Finally, note that our assumptions on the cost functions are quite natural and represent a minimal set of assumptions for the problem to be meaningful: (A.1) guarantees that we stay within the classical control problems rather than having also to deal with singular controls (e.g., see [14]), while (A.2) is the usual polynomial growth condition assumption.

3. Asymptotic lower bound for the optimal cost functions. Let v and V^ε be the optimal cost functions, corresponding to the deterministic and the original control problems respectively (see (1.7)–(1.9) and (1.1)–(1.3)). The aim of this section is to prove the following theorem.

THEOREM 3.1. *Let the assumptions of* §2 *be satisfied. Then*

$$\liminf_{\varepsilon\to 0} V^\varepsilon \geq v.$$

Proof. We may limit ourselves to the case when $\liminf_{\varepsilon\to 0} J^\varepsilon(u^\varepsilon) = \beta < \infty$. Take a subsequence $\varepsilon_k \to 0$ $(k \to \infty)$ such that $\lim_k J^{\varepsilon_k}(u^{\varepsilon_k}) = \liminf_{\varepsilon\to 0} J^\varepsilon(u^\varepsilon)$. Then for k large enough

$$J^{\varepsilon_k}(u^{\varepsilon_k}) \leq 2\beta. \tag{3.1}$$

(For notational convenience we shall assume that (3.1) holds for all k.) From (3.1) and (1.2) it follows that

$$E\int_0^T q(u_t^{\varepsilon_k})\,dt < 2\beta. \tag{3.2}$$

Let X^{ε_k} be the state process associated with u^{ε_k}.

Given (3.2), we may apply Theorem 6.1 to conclude that the sequence $(X^{\varepsilon_k}, U^{\varepsilon_k}, ||U^{\varepsilon_k}||)$, $k \geq 1$, is relatively compact, where $U_t^{\varepsilon_k} = \int_0^t u^{\varepsilon_k}(s)ds$ and $||U^{\varepsilon_k}||_t = \int_0^t |u^{\varepsilon_k}(s)|ds$. Let $(X^{\varepsilon_{\tilde{k}}}, U^{\varepsilon_{\tilde{k}}}, ||U^{\varepsilon_{\tilde{k}}}||)$ be a weakly converging subsequence with limit $(X, U, ||U||)$. Then, by Theorem 6.1, we have

$$\begin{aligned} X_t &= x_0 + \int_0^t [\overline{a}(X_s) + b(X_s)u(s)]\,ds, \\ U(t) &= \int_0^t u(s)ds, \end{aligned} \tag{3.3}$$

where x_0 is the "limit of X_0^ε" (see assumption (A.3)), $\overline{a}(x)$ is defined in (1.5), and $b(x)$ is the same as in (1.1). Since

$$\liminf_{\varepsilon\to 0} J^\varepsilon(u^\varepsilon) = \lim_{\tilde{k}\to 0} J^{\varepsilon_{\tilde{k}}}(u^{\varepsilon_{\tilde{k}}}), \tag{3.4}$$

where $(\varepsilon_{\tilde{k}})$ is any subsequence of (ε_k), we use (3.4) with $(\varepsilon_{\tilde{k}})$ corresponding to the weakly converging sequence $(X^{\varepsilon_{\tilde{k}}}, U^{\varepsilon_{\tilde{k}}}, ||U^{\varepsilon_{\tilde{k}}}||)$. Then by Theorems 5.1 and 6.1 we get

$$\lim_{\tilde{k}} J^{\varepsilon_{\tilde{k}}}(u^{\varepsilon_{\tilde{k}}}) \geq E\left\{\int_0^T [p(X_t) + q(u(t))]\,dt + r(X_T)\right\}. \tag{3.5}$$

From (3.4) and (3.5) we derive

$$\liminf_{\varepsilon\to 0} J^\varepsilon \geq v. \tag{3.6}$$

If an optimal control exists, then the statement of the theorem is a consequence of (3.6). Otherwise we approximate the optimal value function by the cost associated with δ-optimal controls.

4. Proofs of Theorems 1.1 and 1.2. It follows from Theorem 3.1 that the lower limit of the optimal costs is bounded from below by the optimal cost corresponding to the deterministic model (1.7)–(1.9). The existence of an optimal control u^* for problem (1.7)–(1.9) can be shown by standard arguments (see the remark at the end of §6 or the proof of Theorem III.4.1 in [4]). Notice also that assumption (A.1) implies

$$\int_0^T |u^*(t)|^{1+\gamma} dt < \infty. \tag{4.1}$$

Next let $x^*(t)$ be the (deterministic) solution of (1.7) corresponding to the control $u^*(t)$ and $X^{*,\varepsilon} = (X_t^{*,\varepsilon})_{0\le t\le T}$ be the (stochastic) state process associated with the control $u_t^\varepsilon \equiv u^*(t)$ via (1.1).

We first show that

$$P - \lim_{\varepsilon\to 0} \sup_{t\le T} |X_t^{*,\varepsilon} - x^*(t)| = 0. \tag{4.2}$$

Let

$$\Delta_t^\varepsilon := |X_t^{*,\varepsilon} - x^*(t)|. \tag{4.3}$$

Using (1.1) and (1.7), we get the inequality

$$\Delta_t^\varepsilon \le |X_0^\varepsilon - x_0| + \int_0^t \left[\, |\overline{a}(X_s^{*,\varepsilon}) - \overline{a}(x^*(s))| + |b(X_s^{*,\varepsilon}) - b(x^*(s))||u^*(s)| \,\right] ds$$
$$+ \sup_{t\le T} \left| \int_0^t [a(X_s^{*,\varepsilon}, \xi_{s/\varepsilon}) - \overline{a}(X_s^{*,\varepsilon})]\, ds \right| + \sup_{t\le T} |M_t^\varepsilon|.$$

By the Lipschitzianity of $\overline{a}(x)$ and $b(x)$ (see assumptions (A.4) and (A.5)) it follows that

$$\Delta_t^\varepsilon \le \left\{ |X_0^\varepsilon - x_0| + \sup_{t\le T} \left| \int_0^t [a(X_s^{*,\varepsilon}, \xi_{s/\varepsilon}) - \overline{a}(X_s^{*,\varepsilon})]\, ds \right| + \sup_{t\le T} |M_t^\varepsilon| \right\}$$
$$+ \ell \int_0^t (1 + |u^*(s)|)\, |\Delta_s^\varepsilon|\, ds.$$

Therefore, by the Gronwall–Bellman inequality

$$\sup_{t\le T} |\Delta_t^\varepsilon| \le \left\{ |X_0^\varepsilon - x_0| + \sup_{t\le T} \left| \int_0^t [a(X_s^{*,\varepsilon}, \xi_{s/\varepsilon}) - \overline{a}(X_s^{*,\varepsilon})]\, ds \right| \right.$$
$$\left. + \sup_{t\le T} |M_t^\varepsilon| \right\} \exp\left(\ell \int_0^T [1 + |u^*(s)|]\, ds \right). \tag{4.4}$$

Now, by assumption (A.2) we have $P - \lim_{\varepsilon\to 0} |X_0^\varepsilon - x_0| = 0$; furthermore, using a similar argument as in the proof of (6.8) below, we get

$$P - \lim_{\varepsilon\to 0} \sup_{t\le T} \left| \int_0^t [a(X_s^{*,\varepsilon}, \xi_{s/\varepsilon}) - \overline{a}(X_s^{*,\varepsilon})]\, ds \right| = 0. \tag{4.5}$$

Finally, by assumption (A.7) and by Problem 1.9.2 in [15], $P-\lim_{\varepsilon\to 0}\sup_{t\le T}|M_t^\varepsilon| = 0$.

Thus, (4.2) holds. As a consequence of (4.2) we have

$$\begin{aligned} &P-\lim_{\varepsilon\to 0} p(X_t^{*,\varepsilon}) = p(x^*(t)), \quad t\in[0,T], \\ &P-\lim_{\varepsilon\to 0} r(X_T^{*,\varepsilon}) = r(x^*(T)). \end{aligned} \tag{4.6}$$

Next we need to prove that the families $p(X_t^{*,\varepsilon})$ of functions on $[0,T]\times\Omega$ and of random variables $r(X_T^{*,\varepsilon})$ are uniformly integrable with respect to the measures $dt\times dP$ and dP on $[0,T]\times\Omega$ and Ω, respectively. To this end it is sufficient to show that there exists a constant $c>0$ such that

$$E\,[p(X_t^{*,\varepsilon})]^2 \le c, \qquad E\,[r(X_T^{*,\varepsilon})]^2 \le c. \tag{4.7}$$

By assumption (A.2) we have $p(x), r(x) \le c_1(1+|x|^{\gamma_1})$. Let n^* be the smallest integer such that $\gamma_1 < n^*$. Evidently, (4.7) holds if there exists a constant c' such that

$$E\sup_{t\le T}|X_t^{*,\varepsilon}|^{2n^*} \le c'. \tag{4.8}$$

Using (1.1) as well as assumptions (A.4) and (A.5), we get

$$\sup_{s\le t}|X_s^{*,\varepsilon}| \le |X_0^\varepsilon| + \ell\int_0^t\left(1+\sup_{\tau\le s}|X_\tau^{*,\varepsilon}|\right)ds + \ell\int_0^T|u^*(t)|\,dt + \sup_{s\le T}|M_t^\varepsilon|.$$

The Gronwall–Bellman inequality implies

$$\sup_{s\le T}|X_s^{*,\varepsilon}| \le e^{\ell T}\left\{|X_0^\varepsilon| + \ell T + \ell\int_0^T|u^*(t)|\,dt + \sup_{s\le T}|M_t^\varepsilon|\right\}. \tag{4.9}$$

From (7.1) we have

$$E\sup_{t\le T}|M_t^\varepsilon|^{2n^*} \le \text{const}. \tag{4.10}$$

Inequality (4.8) is therefore a consequence of (4.9), (4.10), and assumption (A.3).

By virtue of (4.8) and Theorem 5.4 in [1]

$$\begin{aligned} \lim_{\varepsilon\to 0} J^\varepsilon(u^*) &= \lim_{\varepsilon\to 0} E\left\{\int_0^T[p(X_t^{*,\varepsilon}) + q(u^*(t))]\,dt + r(X_T^{*,\varepsilon})\right\} \\ &= \int_0^T[p(x^*(t)) + q(u^*(t))]\,dt + r(x^*(T)) = v. \end{aligned} \tag{4.11}$$

Since $V^\varepsilon \le J^\varepsilon(u^*)$, we have $\limsup_{\varepsilon\to 0} V^\varepsilon \le v$. This inequality and Theorem 3.1 imply Theorem 1.1, which together with (4.11), in turn implies Theorem 1.2.

5. Relative compactness of $(U^\varepsilon, ||U^\varepsilon||)$.

Let $q(u)$ be the control cost function from (1.2). Assume

$$\sup_{\varepsilon\le 1} E\int_0^T q(u^\varepsilon(t))dt < \infty. \tag{5.1}$$

Recall that $U^\varepsilon(t) = \int_0^t u^\varepsilon(s)ds$ and denote its total variation in the time interval $[0,t]$ by

$$||U^\varepsilon||_t = \int_0^t |u^\varepsilon(s)|ds. \tag{5.2}$$

The process $||U^\varepsilon||_t, 0 \le t \le T$ has paths in a subset of $C_{[0,T]}$ of continuous increasing functions $C^+_{[0,T]}$. Also, ρ will be used for designating of the uniform metric in $C_{[0,T]}$.

THEOREM 5.1. *Let assumption* (A.1) *and* (5.1) *be satisfied. Then the family of random processes* $(U^\varepsilon, ||U^\varepsilon||) = (U^\varepsilon(t), ||U^\varepsilon||_t)_{0\le t\le T}, \varepsilon \le 1$ *is relatively compact in the metric space* $\big(C_{[0,T]} \times C^+_{[0,T]}, \rho \times \rho\big)$.

If $(U^{\varepsilon_k}, ||U^{\varepsilon_k}||)$ *is any weakly converging sequence with limit* $(U, ||U||)$, *then there exists a measurable process* $(u(t))_{0\le t\le T}$ *such that*

1. $E\int_0^T |u(t)|^{1+\gamma}dt < \infty$;
2. *for any* $t \le T$ *and* P*-a.s.*

$$U(t) = \int_0^t u(s)ds, \quad ||U||_t = \int_0^t |u(s)|ds;$$

3.

$$\liminf_{k\to\infty} E\int_0^T q(u^{\varepsilon_k}(t)) \ge E\int_0^T q(u(t))dt. \tag{5.3}$$

Proof. Since $C^+_{[0,T]}$ is closed in $C_{[0,T]}$ in the metric ρ, by virtue of Prokhorov's theorem (see, e.g., [1]) only tightness of the family in $C_{[0,T]} \times C^+_{[0,T]}$ has to be checked. Due to Theorems 8.2 and 15.2 in [1], we verify two conditions:

$$\begin{aligned} &\lim_{c\to\infty} \limsup_{\varepsilon\to 0} P\big(\sup_{t\le T} ||U^\varepsilon||_t > c\big) = 0, \\ &\lim_{\delta\to 0} \limsup_{\varepsilon\to 0} P\Big(\sup_{t,s\le T:|t-s|\le\delta} \Big|||U^\varepsilon||_t - ||U^\varepsilon||_s\Big| > \nu\Big) = 0 \quad \forall \nu > 0 \end{aligned} \tag{5.4}$$

and the same conditions for U^ε. Conditions (A.1) and (5.1) imply

$$\sup_{\varepsilon\le 1} E\int_0^T |u^\varepsilon(t)|^{1+\gamma}dt < \infty. \tag{5.5}$$

Thereby, conditions (5.4) are verified by Hölder's inequality. Namely,

$$\sup_{t\le T} ||U^\varepsilon||_t = ||U^\varepsilon||_T \le T^{\gamma/(1+\gamma)}\Big(\int_0^T |u^\varepsilon(t)|^{(1+\gamma)}dt\Big)^{1/(1+\gamma)}, \tag{5.6}$$

and for any random $t, s \le T : |t-s| \le \delta$

$$\Big|||U^\varepsilon||_t - ||U^\varepsilon||_s\Big| \le \delta^{\gamma/(1+\gamma)}\Big(\int_0^T |u^\varepsilon(t)|^{(1+\gamma)}dt\Big)^{1/(1+\gamma)}. \tag{5.7}$$

We conclude by using Chebyshev's inequality. The validity of the conditions of the type (5.4) for U^ε is proved analogously.

Let $W(t)$ be any random process with paths from $C_{[0,T]}$ and let $I^n = \{s_i = \frac{iT}{2^n}, i = 0, 1, \ldots, 2^n\}, n \geq 1$ be subdivisions of the time interval $[0, T]$. Put

$$w_n(t) = \frac{W_{s_i} - W_{s_{i-1}}}{s_i - s_{i-1}}, \quad s_{i-1} \leq t < s_i. \tag{5.8}$$

It is known (see [16]) that under the assumption

$$\sup_n E \int_0^T |w_n(t)|^2 dt < \infty \tag{5.9}$$

the process $W(t)$ is absolutely continuous (with respect to Lebesgue measure $\Lambda(dt) = dt$); i.e., there exists a measurable process $w(t)$ such that for any $t \leq T$ and P-a.s.

$$W(t) = \int_0^t w(s) ds, \quad E \int_0^T |w(t)|^2 dt < \infty, \tag{5.10}$$

and additionally

$$w(t, \omega) = \lim_n w_n(t, \omega), \quad \Lambda \times P \text{ - a.s.} \tag{5.11}$$

The same proof shows that under the assumption that for some $\gamma > 0$

$$\sup_n E \int_0^T |w_n(t)|^{1+\gamma} dt < \infty \tag{5.12}$$

we have that (5.10) with $E \int_0^T |w(t)|^{1+\gamma} dt < \infty$ and (5.11) hold.

Let $W(t) \equiv U(t)$ and, correspondingly, $u_n(t) \equiv w_n(t)$. Therefore, statements 1. and 2. of Theorem 5.1 take place if, for γ the same as in (A.1),

$$\sup_n E \int_0^T |u_n(t)|^{1+\gamma} dt < \infty. \tag{5.13}$$

To this end, defining $u_n^{\varepsilon_k}(t)$ in the same way as $w_n(t)$ but with $W(t) \equiv U^{\varepsilon_k}(t)$, we find

$$\mathcal{E}_n(U^{\varepsilon_k}) = \int_0^T |u_n^{\varepsilon_k}(t)|^{1+\gamma} dt = \sum_{i=1}^{2^n} \left| \frac{\int_{\frac{i-1}{2^n}}^{\frac{i}{2^n}} u^{\varepsilon_k}(t) dt}{2^{-n}} \right|^{1+\gamma} 2^{-n}.$$

On the other hand, due to Jensen's inequality and assumption (A.1),

$$\begin{aligned} \sum_{i=1}^{2^n} \left| \frac{\int_{\frac{i-1}{2^n}}^{\frac{i}{2^n}} u^{\varepsilon_k}(t) dt}{2^{-n}} \right|^{1+\gamma} 2^{-n} &\leq \sum_{i=1}^{2^n} \int_{\frac{i-1}{2^n}}^{\frac{i}{2^n}} |u^{\varepsilon_k}(t)|^{1+\gamma} dt = \int_0^T |u^{\varepsilon_k}(t)|^{1+\gamma} dt \\ &\leq \frac{1}{c} \int_0^T q(u^{\varepsilon_k}(t)) dt. \end{aligned} \tag{5.14}$$

By virtue of the weak convergence of U^{ε_k} and assumption (5.1), for any $N \geq 1$ we get

$$E \min \left[N, \mathcal{E}_n(U)\right] = \lim_k E \min \left[N, \mathcal{E}_n(U^{\varepsilon_k})\right] \leq \sup_{\varepsilon \leq 1} E \int_0^T |u^{\varepsilon}(t)|^{1+\gamma} dt < \infty.$$

By the monotone convergence theorem, $\sup_n E\,\mathcal{E}_n(U) < \infty$, and thus, noting that $\mathcal{E}_n(U) = \int_0^T |u_n(t)|^{1+\gamma} dt$, we conclude that (5.13) holds.

To prove statement 3. of Theorem 5.1, introduce

$$\mathcal{E}_{n,q}(U^{\varepsilon_k}) = \int_0^T q(u_n^{\varepsilon_k}(t))dt = \sum_{i=1}^{2^n} q\left(\frac{\int_{\frac{i-1}{2^n}}^{\frac{i}{2^n}} u^{\varepsilon_k}(t)dt}{2^{-n}}\right)2^{-n}.$$

Since by Jensen's inequality

$$\int_0^T q(u_n^{\varepsilon_k}(t))dt = \sum_{i=1}^{2^n} q\left(\frac{\int_{\frac{i-1}{2^n}}^{\frac{i}{2^n}} u^{\varepsilon_k}(t)dt}{2^{-n}}\right)2^{-n} \le \int_0^T q(u^{\varepsilon_k}(t))dt,$$

we derive statement 3. by Fatou's lemma and by (5.11), reformulated for $u_n(t)$:

$$\begin{aligned}
\liminf_k E\int_0^T q(u^{\varepsilon_k}(t))dt &\ge \liminf_n \lim_{N\to\infty} \lim_k E\,\min\left[N, \mathcal{E}_{n,q}(U^{\varepsilon_k})\right] \\
&= \liminf_n \lim_{N\to\infty} E\,\min\left[N, \mathcal{E}_{n,q}(U)\right] \\
&= \liminf_n E\,\mathcal{E}_{n,q}(U) = \liminf_n E\int_0^T q(u_n(t))dt \\
&\ge E\int_0^T \liminf_n q(u_n(t))dt = E\int_0^T q(u(t))dt.
\end{aligned}$$

6. Relative compactness of $(X^\varepsilon, U^\varepsilon, ||U^\varepsilon||)$. Let $X^\varepsilon = (X_t^\varepsilon)_{t\ge 0}$ be defined as in (1.1) and $||U^\varepsilon||_t$ in (5.2). We consider the triple $(X^\varepsilon, U^\varepsilon, ||U^\varepsilon||) = (X_t^\varepsilon, U_t^\varepsilon, ||U^\varepsilon||_t)_{0\le t\le T}$ with values in $D_{[0,T]} \times C_{[0,T]} \times C^+_{[0,T]}$, where $D_{[0,T]}$ is Skorokhod's space.

THEOREM 6.1. *Let the assumptions of* §2 *and* (5.1) *be satisfied. Then the family* $(X^\varepsilon, U^\varepsilon, ||U^\varepsilon||), \varepsilon \le 1$, *is relatively compact in the metric space* $\left(D_{[0,T]} \times C_{[0,T]} \times C^+_{[0,T]}, \rho\times\rho\times\rho\right)$. *If* $(X^{\varepsilon_k}, U^{\varepsilon_k}, ||U^{\varepsilon_k}||)$ *is any weakly converging sequence with limit* $(X, U, ||U||)$, *then the statements of Theorem* 5.1 *hold and*

$$X_t = x_0 + \int_0^t [\,\bar{a}(X_s) + b(X_s)u(s)\,]\,ds, \quad t \le T, \tag{6.1}$$

where $\bar{a}(x)$ *is defined as in* (1.5) *and* $u(s)$ *is the process from Theorem* 5.1. *For any continuous nonnegative functions* $p(x)$ *and* $r(x)$,

$$\liminf_k E\left\{\int_0^T p(X_t^{\varepsilon_k})dt + r(X_T^{\varepsilon_k})\right\} \ge E\left\{\int_0^T p(X_t)dt + r(X_T)\right\}. \tag{6.2}$$

Proof. Parallel to X_t^ε, introduce a process $X_t^{\varepsilon,\circ}$ defined by (compare to (1.1))

$$X_t^{\varepsilon,\circ} = X_0^\varepsilon + \int_0^t \left[a(X_s^{\varepsilon,\circ}, \xi_{s/\varepsilon}) + b(X_s^{\varepsilon,\circ})u^\varepsilon(s)\right]ds. \tag{6.3}$$

Due to (1.1), (6.3), and assumptions (A.4) and (A.5), the process $Y_t^\varepsilon = \sup_{s\le t}\left|X_s^\varepsilon - X_s^{\varepsilon,\circ}\right|$ satisfies the inequality

$$Y_t^\varepsilon \le \ell\int_0^t Y_s^\varepsilon d[s + ||U^\varepsilon||_s] + \sup_{s\le T}|M_s^\varepsilon|, \quad t \le T,$$

and thus by the Gronwall–Bellman inequality we get

$$Y_T^\varepsilon \le \ell \sup_{s\le T} |M_s^\varepsilon| \exp\{\ell[T + ||U^\varepsilon||_T]\}.$$

By virtue of assumption (A.7) and Problem 1.9.2 in [15], $\sup_{t\le T} |M_t^\varepsilon| \to 0, \varepsilon \to 0$, in probability and $||U^\varepsilon||_T$ satisfies (5.4). Consequently $Y_T^\varepsilon \to 0, \varepsilon \to 0$, in probability, and by Theorem 4.1, Chapter 1 in [1] the result of the theorem remains true if its statements are proved only for the triple $(X^{\varepsilon,\circ}, U^\varepsilon, ||U^\varepsilon||)$.

By virtue of (5.4), it is sufficient to verify only the following two conditions (see Theorems 8.2 and 15.2 in [1]):

$$\lim_{c\to\infty} \limsup_{\varepsilon\to 0} P\left(\sup_{t\le T} |X_t^{\varepsilon,\circ}| > c\right) = 0$$

$$\lim_{\delta\to 0} \limsup_{\varepsilon\to 0} P\left(\sup_{t,s\le T:|t-s|\le\delta} |X_t^{\varepsilon,\circ} - X_s^{\varepsilon,\circ}| > \nu\right) = 0 \quad \forall \nu > 0. \tag{6.4}$$

It follows from (6.3) and assumptions (A.4) and (A.5) that for any $t \le T$

$$\sup_{s\le t} |X_s^{\varepsilon,\circ}| \le |X_0^\varepsilon| + \ell \int_0^t \left[1 + \sup_{\tau\le s} |X_\tau^{\varepsilon,\circ}|\right] ds + ||U^\varepsilon||_T,$$

and thus, using Gronwall–Bellman's inequality, we get

$$\sup_{s\le T} |X_s^{\varepsilon,\circ}| \le e^{\ell T}\big(|X_0^\varepsilon| + \ell||U^\varepsilon||_T\big).$$

Evidently, the first condition in (6.4) holds by the proof of Theorem 5.1 and by assumption (A.3.i).

For any $t - s < \delta$ we can apply assumptions (A.4) and (A.5) to write

$$\begin{aligned} |X_t^{\varepsilon,\circ} - X_s^{\varepsilon,\circ}| &\le \ell \int_{t\wedge s}^{t\vee s} \left[1 + \sup_{\tau\le T} |X_\tau^{\varepsilon,\circ}|\right] d\tau + \ell\big(||U^\varepsilon||_{t\vee s} - ||U^\varepsilon||_{t\wedge s}\big) \\ &\le \ell\delta \left[1 + \sup_{\tau\le T} |X_\tau^{\varepsilon,\circ}|\right] + \ell\big(||U^\varepsilon||_{t\vee s} - ||U^\varepsilon||_{t\wedge s}\big). \end{aligned}$$

Therefore, the validity of the second condition in (6.4) follows from the proof of Theorem 5.1 and from the first condition in (6.4), which has already been proved.

Let $(X^{\varepsilon_k,\circ}, U^{\varepsilon_k}, ||U^{\varepsilon_k}||)$, $k \ge 1$, be a weakly converging sequence with limit $(X, U, ||U||)$. Denote by Q the distribution of the limit $(X, U, ||U||)$; i.e., Q is a probability measure on $C_{[0,T]} \times C_{[0,T]} \times C^+_{[0,T]}$. For any element $(X, U, ||U||)$ from $C_{[0,T]} \times C_{[0,T]} \times C^+_{[0,T]}$ put

$$\Phi_t(X, U, ||U||) := X_t - x_0 - \int_0^t \overline{a}(X_s)\, ds - \int_0^t b(X_s)\, dU(s), \tag{6.5}$$

where the function $\overline{a}(x)$ is defined by (1.5) and x_0 is the same as in assumption (A.3.i). The second statement of Theorem 6.1. holds if

$$\sup_{t\le T} |\,\Phi_t(X, U, ||U||)\,| = 0 \qquad Q - \text{a.s.} \tag{6.6}$$

To prove the validity of (6.6), we show that the functional $\sup_{t\le T}|\,\Phi_t(X,U,||U||)\,|$ is continuous in the product-metric $\rho^3=\rho\times\rho\times\rho$. Let $(X^0,U^0,||U^0||)$ and $(X^n,U^n,||U^n||)$, $n\ge 1$, be elements of $C_{[0,T]}\times C_{[0,T]}\times C^+_{[0,T]}$ such that

$$\lim_n \rho^3((X^0,U^0,||U^0||),(X^n,U^n,||U^n||)=0.$$

We show that $\lim_n \sup_{t\le T}|\,\Phi_t(X^n,U^n,||U^n||)=\sup_{t\le T}|\,\Phi_t(X^0,U^0,||U^0||)$. Taking (6.5) into account, we get

$$\begin{aligned}
L^n &:= \Big|\sup_{t\le T}|\Phi_t(X^n,U^n,||U^n||)|-\sup_{t\le T}|\Phi_t(X^0,U^0,||U^0||)|\,\Big|\\
&\le \sup_{t\le T}\Big|\,\Phi_t(X^n,U^n,||U^n||)-\Phi_t(X^0,U^0,||U^0||)\,\Big|\\
&\le 2\sup_{t\le T}|X^n_t-X^0_t|+\int_0^T|\,\overline{a}(X^n_s)-\overline{a}(X^0_s)\,|\,ds\\
&\quad+\int_0^T|\,b(X^n_s)-b(X^0_s)\,|\,d||U^n||_s+\sup_{t\le T}\Big|\int_0^t b(X^0_s)\,d[U^n(s)-U^0(s)]\,\Big|.
\end{aligned}$$

Using the Lipschitzianity of the functions $\overline{a}(x)$ (it is inherited from $a(x,y)$; see (A.4.ii)) and $b(x)$, we obtain the following upper bound for L^n:

$$L^n\le\rho(X^n,X^0)\left\{2+\ell T+\ell||U^0||_T+\ell\rho(||U^n||,||U^0||)\right\}+L^n_b,$$

where

$$L^n_b:=\sup_{t\le T}\Big|\int_0^t b(X^0_s)\,d[U^n(s)-U^0(s)]\,\Big|.$$

The quantity L^n_b can be evaluated from above in the following way ($[\alpha]$ below stands for the integer part of α):

$$\begin{aligned}
L^n_b\le&\sup_{t\le T}\Big|\int_0^t b(X^0_{\frac{[Ns]}{N}})\,d[U^n(s)-U^0(s)]\,\Big|\\
&+\ell\sup_{|s'-s''|\le\frac{1}{N}}|X^0_{s'}-X^0_{s''}|\,[\,(||U^n||_T+||U^0||)_T\,].
\end{aligned}$$

Therefore, $\limsup_n L^n_b\le 2\ell||U^0||_T\sup_{|s'-s''|\le\frac{1}{N}}|X^0_{s'}-X^0_{s''}|\to 0$ for $N\to\infty$; i.e., $\sup_{t\le T}|\,\Phi_t(X,U,||U||)\,|$ is continuous functional.

Using this fact, the equality

$$Q\Big(\sup_{t\le T}|\,\Phi_t(X,U,||U||)\,|\ge\nu\Big)=\lim_k P\left(\sup_{t\le T}|\,\Phi_t(X^{\varepsilon_k,\circ},U^{\varepsilon_k},||U^{\varepsilon_k}||)\,|\ge\nu\right),\quad \nu>0,$$

is implied by the weak convergence mentioned above, and by the estimate

$$\sup_{t\le T}|\,\Phi_t(X^{\varepsilon_k,\circ},U^{\varepsilon_k},||U^{\varepsilon_k}||)\,|\le|X^{\varepsilon_k}_0-x_0|+\sup_{t\le T}\Big|\int_0^t[a(X^{\varepsilon_k}_s,\xi_{s/\varepsilon_k})-\overline{a}(X^{\varepsilon_k}_s)]ds\Big|,\tag{6.7}$$

we can conclude that (6.6) holds if the right-hand side of (6.7) goes to zero in probability as $k\to\infty$. Taking assumption (A.3.i) into account, for the validity of (6.6) only

$$P-\lim_k\sup_{t\le T}\Big|\int_0^t[a(X^{\varepsilon_k}_s,\xi_{s/\varepsilon_k})-\overline{a}(X^{\varepsilon_k}_s)]ds\Big|=0\tag{6.8}$$

has to be checked.

Evidently, for a piecewise constant function such that $\phi(t) = \phi\left(\frac{i}{n}\right)$ for $\frac{i}{n} \le t < \frac{i+1}{n}$,

$$P - \lim_{k\to\infty} \left| \int_0^t [\, a(\phi(s), \xi_{s/\varepsilon_k}) - \overline{a}(\phi(s)) \,]\, ds \right| = 0 \quad \forall t \le T \tag{6.9}$$

holds. Notice also that (6.9) remains true when $a(x,z)$ and $\overline{a}(x)$ are replaced with $a^{\pm}(x,z)$ and $\overline{a}^{\pm}(x)$, where $e^+ = \max[0,e]$ and $e^- = -\min[0,e]$. Then, by Problem 5.5.2 in [15] we get

$$P - \lim_{k\to\infty} \sup_{t\le T} \left| \int_0^t [\, a(\phi(s), \xi_{s/\varepsilon_k}) - \overline{a}(\phi(s)) \,]\, ds \right| = 0, \tag{6.10}$$

Then approximate the process $(X_t^{\varepsilon_k})_{0\le t\le T}$ by a sequence $X^{k,m,n} = (X_t^{k,m,n})_{0\le t\le T}$, $n \ge 1$, $m \ge 1$, where

$$X_t^{k,m,n} = \sum_{j=-\infty}^{\infty} \frac{j-1}{m}\, I\left(\frac{j-1}{m} \le X_{i/n}^{\varepsilon_k} < \frac{j}{m} \right), \quad \frac{i}{n} \le t < \frac{i+1}{n}.$$

The process $X^{k,m,n}$ has piecewise constant paths, and on the set $\{\sup_{t\le T} |X_t^{\varepsilon_k}| \le c\}$ the number of its paths is finite and does not depend on k. Therefore, using (6.10), we see that for any $c > 0$, $m \ge 1$, $n \ge 1$ and putting $\xi_s^k = \xi_{s/\varepsilon_k}$

$$P - \lim_k I\left(\sup_{t\le T} |X_t^{\varepsilon_k}| \le c \right) \sup_{t\le T} \left| \int_0^t [\, a(X_s^{k,m,n}, \xi_s^k) - \overline{a}(X_s^{k,m,n}) \,]\, ds \right| = 0. \tag{6.11}$$

On the other hand, taking into account the weak convergence of $(X_t^{\varepsilon_k})_{0\le t\le T}$, which implies $\lim_k \limsup_{c\to\infty} P\left(\sup_{t\le T} |X_t^{\varepsilon_k}| > c\right) = 0$, for the validity of (6.8) it remains to show that

$$P - \lim_{m,n\to\infty} \lim_k \int_0^T |\, a(X_s^{\varepsilon_k}, \xi_s^k) - a(X_s^{k,m,n}, \xi_s^k) \,|\, ds = 0,$$

$$P - \lim_{m,n\to\infty} \lim_k \int_0^T |\, \overline{a}(X_s^{\varepsilon_k}) - \overline{a}(X_s^{k,m,n}) \,|\, ds = 0.$$

Taking into account the Lipschitzianity of the function $a(x,y)$ (see assumption (A.4)), which is also inherited by the function $\overline{a}(x)$, it is sufficient to show

$$P - \lim_{m,n\to\infty} \lim_k \int_0^T |\, X_s^{\varepsilon_k} - X_s^{k,m,n} \,|\, ds = 0. \tag{6.12}$$

To this end, put $X_t^{k,n} = X_{\frac{[nt]}{n}}^{\varepsilon_k}$, where $[\alpha]$ is the integer part of α. Then

$$|\, X_s^{\varepsilon_k} - X_s^{k,m,n} \,| \le |\, X_s^{\varepsilon_k} - X_s^{k,n} \,| + |\, X_s^{k,n} - X_s^{k,m,n} \,|.$$

Obviously $|\, X_s^{k,n} - X_s^{k,m,n} \,| \le \frac{1}{m}$. Consequently

$$\int_0^T |\, X_s^{\varepsilon_k} - X_s^{k,m,n} \,|\, ds \le \frac{T}{m} + \int_0^T |\, X_s^{\varepsilon_k} - X_{\frac{[ns]}{n}}^{k} \,|\, ds \le \frac{T}{m} + T \sup_{s,t\le T:|s-t|\le 1/n} |X_t^{\varepsilon_k} - X_s^k|.$$

Therefore, for any $\nu > 0$

(6.13)
$$P\left(\int_0^T |X_s^{\varepsilon_k} - X_s^{k,m,n}|\,ds > \nu\right) \le P\left(\sup_{s,t\le T:|s-t|\le 1/n} |X_t^{\varepsilon_k} - X_s^k| > \frac{\nu}{T} - \frac{1}{m}\right).$$

As a result, (6.12) follows from weak convergence of $(X_t^{\varepsilon_k})_{0\le t\le T}$, which implies the convergence to zero of the right-hand side of (6.13).

It remains to prove (6.2). Due to the weak convergence of $(X_t^{\varepsilon_k})_{0\le t\le T}, k \ge 1$, we find

$$\liminf_k E\left\{\int_0^T p(X_t^{\varepsilon_k})dt + r(X_T^{\varepsilon_k})\right\} \ge \lim_k E\left\{\int_0^T \big(N \wedge p(X_t^{\varepsilon_k})\big)dt + N \wedge r(X_T^{\varepsilon_k})\right\}$$
$$= E\left\{\int_0^T \big(N \wedge p(X_t)\big)dt + N \wedge r(X_T)\right\} \quad \forall N \ge 1$$

and conclude by using the monotone convergence theorem.

Remark. The method of proof of Theorem 6.1 can be adapted to the following deterministic problem. Let $u^n(t)$, $n \ge 1$, be a sequence of measurable functions satisfying

$$\sup_n \int_0^T |u^n(t)|^{1+\gamma}dt < \infty, \qquad \gamma > 0.$$

For each n consider the differential equation

$$\frac{dx^n(t)}{dt} = \bar{a}(x^n(t)) + b(x^n)u^n(t)$$

with the initial condition $x^n(0) = x_0$. Put

$$U^n(t) = \int_0^t u^n(s)\,ds, \qquad ||U||_t^n = \int_0^t |u^n(s)|\,ds,$$

By the same technique as in the proof of Theorem 6.1, one can show that the family $(x^n(t), U^n(t), ||U||_t^n)_{0\le t\le T}$, $n \ge 1$ is uniformly bounded and equicontinuous. Then by the Arzelà–Ascoli theorem this family is relatively compact and there exists a subsequence $(x^{n_k}(t), U^{n_k}(t), ||U||_t^{n_k})_{0\le t\le T}$ converging uniformly to a limit

$$(x^0(t), U^0(t), ||U||_t^0)_{0\le t\le T}$$

with absolutely continuous $U^0(t)$; i.e., there exists a measurable function $u^0(t)$ such that $U^0(t) = \int_0^t u^0(s)ds$. Furthermore, $x^0(t)$ is the unique solution of the differential equation

$$\frac{dx^0(t)}{dt} = \bar{a}(x^0(t)) + b(x^0)u^0(t)$$

with the initial condition $x^0(0) = x_0$.

7. Upper bound for $E\sup_{t\le T}|M_t^\varepsilon|^{2n}$. In this section we prove, under assumption (A.7), that for any $n > 1$ and $T > 0$

(7.1)
$$\sup_{\varepsilon\le 1} E \sup_{t\le T} |M_t^\varepsilon|^{2n} < \infty.$$

In the case of $E\,|M_T^\varepsilon|^{2n} < \infty$, we can apply Doob's inequality (see, e.g., [15]) to obtain

$$E \sup_{t\le T} |M_t^\varepsilon|^{2n} \le \left(\frac{2n}{2n-1}\right)^{2n} E\,|M_T^\varepsilon|^{2n}.$$

Thus, it suffices to show that

$$\sup_{\varepsilon\le 1} E\,|M_T^\varepsilon|^{2n} < \infty. \tag{7.2}$$

We shall use the notations k, N_t, and V_t to denote a generic positive constant depending on (c_3, L, n), a local martingale, and a nondecreasing process (with paths in $D_{[0,\infty)}$), respectively, where N_t and V_t are adapted to the filtration $\mathbf{F}^\varepsilon$. (All these objects might be different in different formulas.)

To check the validity of (7.2), we shall show that $(M_t^\varepsilon)^{2n}$ admits the representation

$$(M_t^\varepsilon)^{2n} = k\int_0^t [1 + (M_s^\varepsilon)^{2n}]ds + N_t - V_t. \tag{7.3}$$

From (7.3) the desired result follows immediately. In fact, by Ito's formula we find

$$\begin{aligned} e^{-kt}(M_t^\varepsilon)^{2n} &= 1 - e^{-kt} + \int_0^t e^{-ks}dN_s - \int_0^t e^{-ks}dV_s \\ &\le 1 + \int_0^t e^{-ks}dN_s. \end{aligned} \tag{7.4}$$

The Ito integral $\int_0^t e^{-ks}dN_s$ is a local martingale. Denote its localizing sequence of stopping times by $(\tau_j)_{j\ge 1}$; i.e., for any $t > 0$, $E\int_0^{t\wedge\tau_j} e^{-ks}dN_s = 0$, $\quad j \ge 1$.

Therefore, from (7.4) it follows that

$$E\,e^{-k(T\wedge\tau_j)}(M_{T\wedge\tau_j}^\varepsilon)^{2n} \le 1, \quad j \ge 1,$$

and so we conclude by using Fatou's lemma.

Thus, only (7.3) has to be proved.

By Ito's formula

$$\begin{aligned} (M_t^\varepsilon)^{2n} &= 2n\int_0^t (M_{s-}^\varepsilon)^{2n-1}dM_s^\varepsilon + n(2n-1)\int_0^t (M_{s-}^\varepsilon)^{2n-2}d\langle M^{\varepsilon,c}\rangle_s \\ &\quad + \sum_{s\le t}\left[(M_s^\varepsilon)^{2n} - (M_{s-}^\varepsilon)^{2n} - 2n(M_{s-}^\varepsilon)^{2n-1}\Delta M_s^\varepsilon\right], \end{aligned} \tag{7.5}$$

where $\langle M^{\varepsilon,c}\rangle_t$ is the predictable quadratic variation of the continuous part of the martingale M_t^ε.

The representation (7.5) is nothing but

$$(M_t^\varepsilon)^{2n} = N_t + B_t \tag{7.6}$$

with the local martingale

$$N_t = 2n\int_0^t (M_{s-}^\varepsilon)^{2n-1}dM_s^\varepsilon \tag{7.7}$$

and the nondecreasing process

(7.8)
$$B_t = n(2n-1)\int_0^t (M^\varepsilon_{s-})^{2n-2} d\langle M^{\varepsilon,c}\rangle_s + \sum_{s\le t}\left[(M^\varepsilon_s)^{2n} - (M^\varepsilon_{s-})^{2n} - 2n(M^\varepsilon_{s-})^{2n-1}\Delta M^\varepsilon_s\right].$$

Denote by $\mu^\varepsilon(dt,dz)$ the measure of jumps of the martingale M^ε_t and by $\nu^\varepsilon(dt,dz)$ its compensator. Since $(R^\circ = R\setminus\{0\})$

$$\sum_{s\le t}\left[(M^\varepsilon_s)^{2n} - (M^\varepsilon_{s-})^{2n} - 2n(M^\varepsilon_{s-})^{2n-1}\Delta M^\varepsilon_s\right] = \int_0^t\int_{R^\circ}\left[(M^\varepsilon_{s-}+z)^{2n} - (M^\varepsilon_{s-})^{2n} - 2n(M^\varepsilon_{s-})^{2n-1}z\right]\mu^\varepsilon(ds,dz)$$

and the process

$$N_t = \int_0^t\int_{R^\circ}\left[(M^\varepsilon_{s-}+z)^{2n} - (M^\varepsilon_{s-})^{2n} - 2n(M^\varepsilon_{s-})^{2n-1}z\right]\mu^\varepsilon(ds,dz) - \int_0^t\int_{R^\circ}\left[(M^\varepsilon_{s-}+z)^{2n} - (M^\varepsilon_{s-})^{2n} - 2n(M^\varepsilon_{s-})^{2n-1}z\right]\nu^\varepsilon(ds,dz)$$

is a local martingale too, we arrive to a new decomposition of the type (7.6) with local martingale

$$N_t = 2n\int_0^t (M^\varepsilon_{s-})^{2n-1} dM^\varepsilon_s + \int_0^t\int_{R^\circ}\left[(M^\varepsilon_{s-}+z)^{2n} - (M^\varepsilon_{s-})^{2n} - 2n(M^\varepsilon_{s-})^{2n-1}\right][\mu^\varepsilon - \nu^\varepsilon](ds,dz) \tag{7.9}$$

and nondecreasing process

$$B_t = n(2n-1)\int_0^t (M^\varepsilon_{s-})^{2n-2} d\langle M^{\varepsilon,c}\rangle_s + \int_0^t\int_{R^\circ}\left[(M^\varepsilon_{s-}+z)^{2n} - (M^\varepsilon_{s-})^{2n} - 2n(M^\varepsilon_{s-})^{2n-1}z\right]\nu^\varepsilon(ds,dz). \tag{7.10}$$

Using the fact that $|\Delta M^\varepsilon_s| \le L$, we get $\nu^\varepsilon(ds,dz) = I(|z|\le L)\nu^\varepsilon(ds,dz)$. Therefore, by virtue of Taylor's expansion for the function $f(x) = x^{2n}$ and Hölder's inequality one can find a constant k such that

$$dB_t \le n(2n-1)(M^\varepsilon_{t-})^{2n-2} d\langle M^{\varepsilon,c}\rangle_t + k\int_{|z|\le L}(M^\varepsilon_{s-})^{2n-2}(1+z^2)\nu^\varepsilon(dt,dz).$$

Recall that the quadratic variation $[M^\varepsilon, M^\varepsilon]_t$ of M^ε_t is defined as

$$[M^\varepsilon, M^\varepsilon]_t = \langle M^{\varepsilon,c}\rangle_t + \sum_{s\le t}(\Delta M^\varepsilon_s)^2 = \langle M^{\varepsilon,c}\rangle_t + \int_0^t\int_{R^\circ} z^2\nu^\varepsilon(dt,dz).$$

Consequently, taking into account that $x^{2n-2} \le 1 + x^{2n}$ we obtain

$$dB_t \le 2[n(2n-1)+k](M^\varepsilon_{t-})^{2n-2} d[M^\varepsilon, M^\varepsilon]_t \le 2[n(2n-1)+k](1+M^\varepsilon_{t-})^{2n} d[M^\varepsilon, M^\varepsilon]_t.$$

Define a nondecreasing process

$$V_t = 2[n(2n-1)+k] \int_0^t (1 + M^\varepsilon_{s-})^{2n} d[M^\varepsilon, M^\varepsilon]_s - B_t.$$

Then for $(M^\varepsilon_t)^{2n}$ we have the following decomposition:

$$(M^\varepsilon_t)^{2n} = N_t + 2[n(2n-1)+k] \int_0^t (1 + M^\varepsilon_{s-})^{2n} d[M^\varepsilon, M^\varepsilon]_s - V_t, \tag{7.11}$$

where the local martingale N_t is defined in (7.9). Since $[M^\varepsilon, M^\varepsilon]_t - \langle M^\varepsilon \rangle_t$ is a local martingale, we arrive at a new representation for $(M^\varepsilon_t)^{2n}$:

$$(M^\varepsilon_t)^{2n} = N_t + 2[n(2n-1)+k] \int_0^t (1 + M^\varepsilon_{s-})^{2n} d\langle M^\varepsilon \rangle_s - V_t \tag{7.12}$$

with the same nondecreasing process V_t and a new local martingale N_t.

Due to assumption (A.7) we have (for $\varepsilon \le 1$) $d\langle M^\varepsilon \rangle_t \le c_3 dt$; i.e.,

$$V'_t = 2[n(2n-1)+k] \Big[\int_0^t (1 + M^\varepsilon_s)^{2n} c_3 ds - \int_0^t (1 + M^\varepsilon_{s-})^{2n} d\langle M^\varepsilon \rangle_s \Big] \tag{7.13}$$

is a nondecreasing process.

Thus, (7.3) is implied by (7.12) and (7.13).

REFERENCES

[1] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[2] C. DELLACHERIE, *Capacités et Processus Stochastiques*, Springer-Verlag, Berlin, 1972.

[3] S. N. ETHIER AND T. G. KURTZ, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.

[4] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, 1975.

[5] M. I. FREIDLIN AND A. D. WENTZELL, *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York, 1984.

[6] J. JACOD AND A. N. SHIRYAYEV, *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin, New York, 1987.

[7] E. V. KRICHAGINA AND M. I. TAKSAR, *Diffusion approximation for GI/G/1 controlled queues*, Queuing Systems Theory Appl., 12 (1992), pp. 333–368.

[8] E. V. KRICHAGINA, S. X. C. LOU, S. P. SETHI, AND M. I. TAKSAR, *Production control in a failure-prone manufacturing system: Diffusion approximation and asymptotic optimality*, Ann. Appl. Probab., 3 (1993), pp. 421–453.

[9] H. J. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes*, MIT Press, Cambridge, MA, 1984.

[10] H. J. KUSHNER AND W. J. RUNGGALDIER, *Nearly optimal state feedback controls for stochastic systems with wideband noise disturbances*, SIAM J. Control Optim., 25 (1987), pp. 298–315.

[11] H. J. KUSHNER AND K. M. RAMACHANDRAN, *Optimal and approximately optimal control policies for queues in heavy traffic*, SIAM J. Control Optim., 27 (1989), pp. 1293–1318.

[12] H. J. KUSHNER AND L. F. MARTINS, *Routing and singular control for queueing networks in heavy traffic*, SIAM J. Control Optim., 28 (1990), pp. 1209–1233.

[13] J. LEHOCZKY, S. SETHI, M. SONER, AND M. TAKSAR, *An asymptotic analysis of hierarchical control of manufacturing systems*, Math. Oper. Res., 16 (1991), pp. 596–608.

[14] J. LEHOCZKY AND S. SHREVE, *Absolutely continuous and singular stochastic control*, Stochastics, 17 (1986), pp. 91–109.

[15] R. SH. LIPTSER AND A. N. SHIRYAYEV, *Theory of Martingales*, Kluwer Academic Publishers, Dordrecht, 1989.

[16] A. D. WENTZELL, *Additive Functionals of a Multidimensional Wiener Process*, Soviet Mathematics 2 (1961), pp. 848–851.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 179–198, January 1996

008

FINITE-DIMENSIONAL FILTERS WITH NONLINEAR DRIFT IV: CLASSIFICATION OF FINITE-DIMENSIONAL ESTIMATION ALGEBRAS OF MAXIMAL RANK WITH STATE–SPACE DIMENSION 3*

JIE CHEN†, STEPHEN S.-T. YAU†, AND CHI-WAH LEUNG‡

Abstract. The idea of using estimation algebras to construct finite-dimensional nonlinear filters was first proposed by Brockett and Mitter independently. It turns out that the concept of estimation algebra plays a crucial role in the investigation of finite-dimensional nonlinear filters. In his talk at the International Congress of Mathematics in 1983, Brockett proposed a classification of all finite-dimensional estimation algebras. Chiou and Yau classify all finite-dimensional estimation algebras of maximal rank with dimension of the state space less than or equal to two. In this paper we succeed in classifying all finite-dimensional estimation algebras of maximal rank with state–space dimension equal to three. Thus from the Lie algebraic point of view, we have now understood generically all finite dimensional filters with state–space dimension less than four.

Key words. finite-dimensional filter, estimation algebra of maximal rank, nonlinear drift

AMS subject classifications. 17B30, 35J15, 60G35, 93E11

1. Introduction.

In the sixties and early seventies, the basic approach to nonlinear filtering theory was via the "innovation methods" originally proposed by Kailath and subsequently rigorously developed by Fujisaki, Kallianpur, and Kunita [FKK] in 1972. As pointed out by Mitter [Mi], the difficulty with this approach is that the innovations process is not, in general, explicitly computable (except in the well-known Kalman–Bucy case). In the late seventies, Brockett and Clark [BrCl], Brockett [Br], and Mitter [Mi] proposed the idea of using estimation algebras to construct finite-dimensional nonlinear filters. In a previous paper [Ya], Yau has studied the general class of nonlinear filtering systems which included both Kalman–Bucy and Benes filtering systems as special cases. He gives necessary and sufficient conditions for an estimation algebra of such filtering systems to be finite dimensional. Using the Wei–Norman approach, he constructed explicitly finite-dimensional recursive filters for such nonlinear filtering systems.

In his talk at the International Congress of Mathematics in 1983, Brockett proposed classification of all finite-dimensional estimation algebras. Since then, the concept of estimation algebras has proved to be an invaluable tool in the study of nonlinear filtering problems. In [ChYa], Chiou and Yau introduced the concept of an estimation algebra of maximal rank. They were able to classify all finite-dimensional estimation algebras of maximal rank with state–space dimension less than or equal to two. The novelty of their theorem is that there is no assumption on the drift term of the nonlinear filtering system. On the other hand, if the drift term has a potential function (i.e., drift term is a gradient vector field), then the corresponding estimation algebra is called exact. In [TWY], Tam, Wong, and Yau classified all finite-dimensional ex-

*Received by the editors June 30, 1993; accepted for publication (in revised form) August 19, 1994. This research was supported by U.S. Army grant DAAH0493G0006.

†Control and Information Laboratory, University of Illinois at Chicago, Box 4348, M/C 249, 851 South Morgan Street, Chicago, IL 60607 (u53651@uicvm.bitnet and u32790@uicvm.bitnet).

‡Department of Mathematics, National Central University, Chung-Li, Taiwan 32054, Republic of China (leung@math.ncu.edu.tw).

act estimation algebras of maximal rank with arbitrary state–space dimension. This paper is a natural continuation of [ChYa]. We shall classify all finite-dimensional estimation algebras of maximal rank with state–space dimension equal to 3 (without any assumption on the drift term). The following is our main theorem.

THEOREM 1 (main theorem). *Suppose that the state space of the filtering system* (2.0) *is of dimension three. If E is the finite-dimensional estimation algebra of maximal rank, then the drift term f must be a linear vector field (i.e., each component is a polynomial of degree one) plus a gradient vector field, and E is a real vector space of dimension eight with bases given by* 1, x_1, x_2, x_3, D_1, D_2, D_3, *and* L_0.

This kind of nonlinear filtering system was studied by Yau [Ya]. Therefore, from the Lie algebraic point of view, we have shown that the finite-dimensional filters considered in [Ya] are the most general.

Let $\omega_{ij} = \frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j}$, which was first introduced by Wong [Wo2]. Our strategy is to prove ω_{ij} constant for all i, j. Then we can apply the result of [Ya] to finish the proof. This involves two steps. The first step is to prove that ω_{ij} is a degree-one polynomial. The second step is to prove that ω_{ij} is a constant. Let n be the dimension of the state space. Unlike the case $n = 2$, where there is only one unknown, ω_{12}, the case $n = 3$ for the treatment of the first step is more difficult because there are three unknowns: ω_{12}, ω_{13}, and ω_{23}, and they cannot be separated and thus they cannot be treated individually. For the second step, which is the hard part of the paper, we have to introduce a new concept and technique in addition to the method used in [ChYa] to overcome the difficulties.

The paper is in essence a continuation of [Ya], [ChYa], and we strongly recommend that readers familiarize themselves with the results in [Ya], [ChYa]. However, every effort will be made to make this paper as self-contained as possible with minimal duplication of the previous papers.

2. Basic concepts. In this section, we shall recall some basic concepts and results from [Ya]. Consider a filtering problem based on the following signal observation model:

$$\begin{cases} dx(t) = f(x(t))dt + g(x(t))dv(t), & x(0) = x_0, \\ dy(t) = h(x(t))dt + dw(t), & y(0) = 0, \end{cases} \tag{2.0}$$

in which x, v, y, and w are, respectively, $\mathbf{R}^n$-, $\mathbf{R}^p$-, $\mathbf{R}^m$-, and $\mathbf{R}^m$-valued processes, and v and w have components which are independent, standard Brownian processes. We further assume that $n = p, f, h$ are C^∞ smooth, and that g is an orthogonal matrix. We shall refer to $x(t)$ as the state of the system at time t and to $y(t)$ as the observation at time t.

Let $\rho(t, x)$ denote the conditional density of the state given the observation $\{y(s) : 0 \leq s \leq t\}$. It is well known (see [DaMa], for example) that $\rho(t, x)$ is given by normalizing a function, $\sigma(t, x)$, which satisfies the Duncan–Mortensen–Zakai equation.

$$d\sigma(t, x) = L_0\sigma(t, x)dt + \sum_{i=1}^{m} L_i\sigma(t, x)dy_i(t), \qquad \sigma(0, x) = \sigma_0, \tag{2.1}$$

where

$$L_0 = \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2} - \sum_{i=1}^{n} f_i \frac{\partial}{\partial x_i} - \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} - \frac{1}{2}\sum_{i=1}^{m} h_i^2$$

and for $i = 1, \ldots, m$, L_i is the zero-degree differential operator of multiplication by h_i and σ_0 is the probability density of the initial point x_0. In this paper, we will assume σ_0 is a C^∞ function.

Equation (2.1) is a stochastic partial differential equation. The stochastic differential is a Stratonovich one, not an Ito one. In real applications, we are interested in constructing state estimators from observed sample paths with some property of robustness. Davis [Da] studied this problem and proposed some robust algorithms. In our case, his basic idea reduces to defining a new, unnormalized density

$$\xi(t,x) = \exp\Big(-\sum_{i=1}^{m} h_i(x)y_i(t)\Big)\sigma(t,x).$$

It is easy to show that $\xi(t,x)$ satisfies the following time-varying partial differential equation:

$$\frac{\partial \xi}{\partial t}(t,x) = L_0\xi(t,x) + \sum_{i=1}^{m} y_i(t)[L_0, L_i]\xi(t,x) \cdot \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i(t)y_j(t)[[L_0, L_i], L_j]\xi(t,x) \tag{2.2}$$

where $[\cdot,\cdot]$ is the Lie bracket defined as follows.

DEFINITION. *If X and Y are differential operators, the Lie bracket of X and Y, $[X,Y]$, is defined by $[X,Y]\phi = X(Y\phi) - Y(X\phi)$ for any C^∞ function ϕ.*

DEFINITION. *The estimation algebra E of a filtering problem* (2.0) *is defined as the Lie algebra generated by $\{L_0, L_1, \ldots, L_m\}$. E is said to be an estimation algebra of maximal rank if, for any $1 \le i \le n$, there exists a constant c_i such that $x_i + c_i$ is in E.*

Most of the known finite-dimensional estimation algebras are maximal. For example, if the equation (2.0) is linear, i.e., $f(x) = Ax$, $g(x) = B$, and $h(x) = Cx$, and if also (A, B, C) is minimal, then the corresponding estimation algebra is maximal [Ha].

In [Ya], the following proposition is proven.

PROPOSITION 1 (Yau). *$\omega_{ij} = \frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j}$ are constant functions for all i and j if and only if $(f_1, \ldots, f_n) = (l_1, \ldots, l_n) + (\frac{\partial \phi}{\partial x_1}, \ldots, \frac{\partial \phi}{\partial x_n})$, where $l_1, \ldots, l_n$ are polynomials of degree one and ϕ is a C^∞ function.*

We need the following basic result for later discussion.

THEOREM 2 (Ocone). *Let E be a finite-dimensional estimation algebra. If a function ξ is in E, then ξ is a polynomial of degree ≤ 2.*

Define

$$D_i = \frac{\partial}{\partial x_i} - f_i,$$

$$\eta = \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 + \sum_{i=1}^{m} h_i^2.$$

Then

$$L_0 = \frac{1}{2}\Big(\sum_{i=1}^{n} D_i^2 - \eta\Big).$$

The following theorem proved in [Ya] plays a fundamental role in the classification of finite-dimensional estimation algebras.

THEOREM 3 (Yau). *Let E be a finite-dimensional estimation algebra of* (2.0) *such that $\omega_{ij} = \frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j}$ are constant functions. If E is of maximal rank, then E is a real vector space of dimension $2n+2$ with bases given by $1, x_1, x_2, \ldots, x_n, D_1, D_2, \ldots, D_n$, and L_0.*

For the convenience of readers, we also list the following elementary lemmas without proof. The lemmas were proven in [Ya] and [ChYa].

LEMMA 4. (i) $[XY, Z] = X[Y, Z] + [X, Z]Y$ *where X, Y and Z are differential operators.*

(ii) $[gD_i, h] = g\frac{\partial h}{\partial x_i}$, *where $D_i = \frac{\partial}{\partial x_i} - f_i, g$ and h are functions defined on $\mathbf{R}^n$.*

(iii) $[gD_i, hD_j] = -gh\omega_{ij} + g\frac{\partial h}{\partial x_i}D_j - h\frac{\partial g}{\partial x_j}D_i$, *where $\omega_{ji} = [D_i, D_j] = \frac{\partial f_i}{\partial x_j} - \frac{\partial f_j}{\partial x_i}$.*

(iv) $[gD_i^2, h] = 2g\frac{\partial h}{\partial x_i}D_i + g\frac{\partial^2 h}{\partial x_i^2}$.

(v) $[D_i^2, hD_j] = 2\frac{\partial h}{\partial x_i}D_iD_j - 2h\omega_{ij}D_i + \frac{\partial^2 h}{\partial x_i^2}D_j - h\frac{\partial \omega_{ij}}{\partial x_i}$.

(vi) $[D_i^2, D_j^2] = 4\omega_{ji}D_jD_i + 2\frac{\partial \omega_{ji}}{\partial x_j}D_i + 2\frac{\partial \omega_{ji}}{\partial x_i}D_j + \frac{\partial^2 \omega_{ji}}{\partial x_i \partial x_j} + 2\omega_{ji}^2$.

(vii) $[D_k^2, hD_iD_j] = 2\frac{\partial h}{\partial x_k}D_kD_iD_j + 2h\omega_{jk}D_iD_k + 2h\omega_{ik}D_kD_j + \frac{\partial^2 h}{\partial x_k^2}D_iD_j + 2h\frac{\partial \omega_{jk}}{\partial x_i}D_k + h\frac{\partial \omega_{jk}}{\partial x_k}D_i + h\frac{\partial \omega_{ik}}{\partial x_k}D_j + h\frac{\partial^2 \omega_{jk}}{\partial x_i \partial x_k}$.

(viii) $[gD_iD_j, hD_k] = g\frac{\partial h}{\partial x_j}D_iD_k + g\frac{\partial h}{\partial x_i}D_jD_k + gh\omega_{kj}D_i + gh\omega_{ki}D_j + g\frac{\partial^2 h}{\partial x_i \partial x_j}D_k + gh\frac{\partial \omega_{kj}}{\partial x_i} - h\frac{\partial g}{\partial x_k}D_iD_j$.

LEMMA 5. (i) $[L_0, x_j + c_j] = D_j$, *where $L_0 = \frac{1}{2}(\sum_{i=1}^n D_i^2 - \eta)$.*

(ii) $[D_i, x_j + c_j] = \delta_{ij}$.

(iii) $[D_i, D_j] = \omega_{ji}$.

(iv) $Y_j := [L_0, D_j] = \sum_{i=1}^n (\omega_{ji}D_i + \frac{1}{2}\frac{\partial \omega_{ji}}{\partial x_i}) + \frac{1}{2}\frac{\partial \eta}{\partial x_j}$.

(v) $[Y_j, \omega_{kl}] = \sum_{i=1}^n \omega_{ji}\frac{\partial \omega_{kl}}{\partial x_i}$.

(vi) $[Y_j, D_k] = \sum_{i=1}^n (\omega_{ji}\omega_{ki} - \frac{\partial \omega_{ji}}{\partial x_k}D_i) - \frac{1}{2}\sum_{i=1}^n \frac{\partial^2 \omega_{ji}}{\partial x_k \partial x_i} - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_k \partial x_j}$.

Consider $\tilde{x} = Rx$, where R is an orthogonal matrix. Then (2.0) becomes

$$\text{(2.3)} \qquad \begin{cases} d\tilde{x}(t) = \tilde{f}(\tilde{x}(t))dt + \tilde{g}(\tilde{x}(t))d\tilde{v}(t), & \tilde{x}(0) = \tilde{x}_0 := Rx_0, \\ d\tilde{y}(t) = \tilde{h}(\tilde{x}(t))dt + d\tilde{w}(t), & \tilde{y}(0) = 0, \end{cases}$$

where

$$\tilde{f}(\tilde{x}) = Rf(x),\ \tilde{g}(\tilde{x}) = Rg(x),$$
$$\tilde{v} = v, \qquad \tilde{w} = w,$$
$$\tilde{y} = y, \qquad \tilde{h}(\tilde{x}) = h(x).$$

It was observed for instance in [TWY] and [ChYa] that $\tilde{L}_0 = \frac{1}{2}\sum_{i=1}^n \frac{\partial^2}{\partial \tilde{x}_i^2} - \sum_{i=1}^n \tilde{f}_i \frac{\partial}{\partial \tilde{x}_i} - \sum_{i=1}^n \frac{\partial \tilde{f}_i}{\partial \tilde{x}_i} - \frac{1}{2}\sum_{i=1}^m \tilde{h}_m^2$ is equal to L_0. Hence the Lie algebra $\tilde{E} = \langle \tilde{L}_0, \tilde{L}_1, \ldots, \tilde{L}_m \rangle_{L.A.}$ is isomorphic to $E = \langle L_0, L_1, \ldots, L_m \rangle_{L.A.}$.

Let $\Omega = (\omega_{ij})$ and $\tilde{\Omega} = (\tilde{\omega}_{ij})$ where $\omega_{ij} = \frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j}$ and $\tilde{\omega}_{ij} = \frac{\partial \tilde{f}_j}{\partial \tilde{x}_i} - \frac{\partial \tilde{f}_i}{\partial \tilde{x}_j}$. It was shown in [ChYa] that the following lemma is true.

LEMMA 6. $\tilde{\Omega} = R\Omega R^{-1}$.

3. Proof of the main theorem. In this section, we shall classify all finite-dimensional estimation algebras with maximal rank for dimension of state space equal to three. By Lemma 5, we know that ω_{ij} is in E and in view of Ocone's result, ω_{ij} is a polynomial of degree at most two for all i, j. The first step is to prove ω_{ij} is a

degree-one polynomial for all i, j. This step was carried out in detail in our Conference on Decision and Control paper [YaLe]. So we have

$$\begin{pmatrix} \omega_{12} \\ \omega_{13} \\ \omega_{23} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} c_{12} \\ c_{13} \\ c_{23} \end{pmatrix}.$$

Now we have to deal with the hard part of the proof. We are going to prove that ω_{ij}'s are constants. For this, we introduce an invariant $r_{\max}$ of the estimation algebra E as follows.

DEFINITION. *Let $p(x)$ be a quadratic polynomial. The rank of $p(x), r(p)$ is defined as the rank of the Hessian matrix $(\frac{\partial^2 p}{\partial x_i \partial x_j})$.*

Denote

$$\begin{aligned} Q &= \text{space of homogeneous polynomials of degree 2,} \\ P_i &= \text{space of polynomials of degree at most } i, \\ U_k &= \text{space of differential operators with order at most } k. \end{aligned}$$

LEMMA 7. *Let E be a finite-dimensional estimation algebra of maximal rank. Then $P_1 \subseteq E$. If $p(x)$ is a polynomial of degree two in E, then the homogeneous degree-two part of $p(x)$ is also in E.*

Proof. This follows immediately from Lemma 5 and the definition of maximal rank. □

DEFINITION. *Let $E_Q = E \cap Q$. Define $r_{\max} = \text{Max}\{\text{rank } p(x) : p(x) \in E_Q\}$.*

Remark. Observe that $r_{\max}$ is invariant under orthogonal change of coordinates and $0 \leq r_{\max} \leq 3$ in this paper.

3.1. Case $r_{\max} = 3$. There exists homogeneous $p(x) \in E$ with rank $(p(x)) = 3$. By applying an orthogonal change of coordinates, if necessary, we may assume without loss of generality that

$$p(x) = k_1 x_1^2 + k_2 x_2^2 + k_3 x_3^2,$$

where $k_i \neq 0$ for $i = 1, 2, 3$. There are three possibilities.

Case I: all k_i's are distinct. By Lemmas 4 and 5,

$$\begin{aligned} [D_i^2, x_j^2] &= \delta_{ij}(4x_j D_j + 2), \\ [L_0, p(x)] &= \frac{1}{2}\sum_{i=1}^{3}[D_i^2, p(x)] = \sum_{j=1}^{3}(2k_j x_j D_j + 1). \end{aligned}$$

So $\sum_{j=1}^{3} k_j x_j D_j \in E$ and

$$\left[\sum_{i=1}^{3} k_i x_i D_i, \frac{1}{2}p(x)\right] = \sum_{i=1}^{3} k_i^2 x_i^2 \in E.$$

Replacing $p(x)$ by $\sum_{i=1}^{3} k_i^2 x_i^2$, we deduce that $\sum_{i=1}^{3} k_i^3 x_i^2 \in E$. Since the matrix

$$\begin{pmatrix} k_1 & k_2 & k_3 \\ k_1^2 & k_2^2 & k_3^2 \\ k_1^3 & k_2^3 & k_3^3 \end{pmatrix}$$

is nonsingular, we conclude that $x_1^2, x_2^2, x_3^2 \in E$. Now for $i \neq j$,

$$\begin{aligned}[[L_0, x_i^2], [L_0, x_j^2] &= [2x_iD_i + 1, 2x_jD_j + 1]] \\ &= -4x_ix_j\omega_{ij} \in E.\end{aligned}$$

Since $x_ix_j\omega_{ij}$ is a polynomial of degree at most 2 by Ocone's result, we deduce that ω_{ij} is constant.

Case II: two of the k_i's are equal. In this case we may take $p(x) = k_1x_1^2 + k_2(x_2^2 + x_3^2)$. By evaluating $[[L_0, p(x)], p(x)]$, we can obtain $k_1^2x_1^2 + k_2^2(x_2^2 + x_3^2) \in E$. It follows that $x_1^2 \in E$ and $x_2^2 + x_3^2 \in E$. Since $[L_0, x_i^2] = 2x_iD_i + 1$, we have $x_1D_1, x_2D_2 + x_3D_3 \in E$. So we have

$$\begin{aligned}&- [x_1D_1, x_2D_2 + x_3D_3] = x_1x_2\omega_{12} + x_1x_3\omega_{13} \\ &= a_{11}x_1^2x_2 + a_{21}x_1^2x_3 + a_{12}x_1x_2^2 + a_{23}x_1x_3^2 + (a_{13} + a_{22})x_1x_2x_3 \qquad \mod P_2 \\ &\in E.\end{aligned}$$

By Ocone's result, $[x_1D_1, x_2D_2 + x_3D_3] \in P_2$. We deduce immediately that

$$a_{11} = a_{21} = a_{12} = a_{23} = 0, \quad a_{13} + a_{22} = 0.$$

Furthermore, from the cyclic relation $\frac{\partial\omega_{12}}{\partial x_3} + \frac{\partial\omega_{23}}{\partial x_1} + \frac{\partial\omega_{31}}{\partial x_2} = 0$, we have $a_{13} + a_{31} - a_{22} = 0$ and

$$A = \begin{pmatrix} 0 & 0 & -a_{22} \\ 0 & a_{22} & 0 \\ 2a_{22} & a_{32} & a_{33} \end{pmatrix}.$$

Recall that

$$\begin{aligned}Y_1 &= [L_0, D_1] = \omega_{12}D_2 + \omega_{13}D_3 \qquad \mod U_0, \\ Y_2 &= [L_0, D_2] = \omega_{21}D_1 + \omega_{23}D_3 \qquad \mod U_0, \\ Y_3 &= [L_0, D_3] = \omega_{31}D_1 + \omega_{32}D_2 \qquad \mod U_0.\end{aligned}$$

Then

$$\begin{aligned}&\frac{1}{2}[Y_2, x_2^2 + x_3^2] = \omega_{23}x_3 = a_{31}x_1x_3 + a_{32}x_2x_3 + a_{33}x_3^2 \qquad \mod P_1 \\ &- \frac{1}{2}[Y_3, x_2^2 + x_3^2] = \omega_{23}x_2 = a_{31}x_1x_2 + a_{32}x_2^2 + a_{33}x_2x_3 \qquad \mod P_1, \\ &\left[x_1D_1, \frac{1}{2}[Y_2, x_2^2 + x_3^2]\right] = a_{31}x_1x_3 \qquad \mod U_0, \\ &a_{31}(x_1D_3 + x_3D_1), a_{31}x_1x_3] = a_{31}^2(x_1^2 + x_3^2) \qquad \mod P_0.\end{aligned}$$

Choose k such that $k \neq \pm a_{31}^2, 0$. Then $a_{31}^2(x_1^2 + x_3^2) + k(x_2^2 + x_3^2) = a_{31}^2x_1^2 + kx_2^2 + (a_{31}^2 + k)x_3^2$ is in E. If $a_{31} \neq 0$, then we are back in Case I and we are done. So we have $a_{31} = 0 = a_{13} = a_{22}$ and

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & a_{32} & a_{33} \end{pmatrix}.$$

$a_{31} = 0$ implies $a_{32}x_2x_3 + a_{33}x_3^2$ and $a_{32}x_2^2 + a_{33}x_2x_3$ are in E.

$$
\begin{aligned}
&[L_0, a_{32}x_2x_3 + a_{33}x_3^2] = a_{32}(x_2D_3 + x_3D_2) + a_{33}(2x_3D_3 + 1)\\
&\qquad \Rightarrow a_{32}x_3D_2 + (a_{32}x_2 + 2a_{33}x_3)D_3 \in E,\\
&[L_0, a_{32}x_2^2 + a_{33}x_2x_3] = a_{32}(2x_2D_2 + 1) + a_{33}(x_2D_3 + x_3D_2)\\
&\qquad \Rightarrow (2a_{32}x_2 + a_{33}x_3)D_2 + a_{33}x_2D_3 \in E,\\
&[a_{32}x_3D_2 + (a_{32}x_2 + 2a_{33}x_3)D_3, a_{32}x_2x_3 + a_{33}x_3^2]\\
&\qquad = a_{32}^2x_3^2 + a_{32}x_2(a_{32}x_2 + 2a_{33}x_3) + 2a_{33}x_3(a_{32}x_2 + 2a_{33}x_3)\\
&\qquad = a_{32}^2x_2^2 + 4a_{32}a_{33}x_2x_3 + (a_{32}^2 + 4a_{33}^2)x_3^2 \in E, \qquad (3.1)\\
&[(2a_{32}x_2 + a_{33}x_3)D_2 + a_{33}x_2D_3, a_{32}x_2^2 + a_{33}x_2x_3]\\
&\qquad = (2a_{32}x_2 + a_{33}x_3)^2 + a_{33}^2x_2^2\\
&\qquad = (a_{33}^2 + 4a_{32}^2)x_2^2 + 4a_{32}a_{33}x_2x_3 + a_{33}^2x_3^2 \in E. \qquad (3.2)
\end{aligned}
$$

From (3.1) and (3.2), we have

$$(-a_{33}^2 - 3a_{32}^2)x_2^2 + (a_{32}^2 + 3a_{33}^2)x_3^2 \in E.$$

Recall that $x_2^2 + x_3^2 \in E$. If

$$\det \begin{pmatrix} 1 & 1 \\ -a_{33}^2 - 3a_{32}^2 & a_{32}^2 + 3a_{33}^2 \end{pmatrix} = 4(a_{32}^2 + a_{33}^2)$$

is nonzero, then x_2^2 and x_3^2 are in E. So ω_{ij} = constant for all i, j in view of the argument in Case I. On the other hand if the determinant above is zero, then $a_{32}^2 + a_{33}^2 = 0$, which implies $a_{32} = a_{33} = 0$. So $A = 0$, which means that ω_{ij}'s are constants.

Case III: all k_i's are the same. In this case, we may take $p(x) = x_1^2 + x_2^2 + x_3^2 \in E$. If there exists quadratic form $q(x)$ with $0 < \operatorname{rank}(q(x)) < 3$, we can find an orthogonal transformation R such that

$$
\begin{aligned}
q(x) &\longmapsto \tilde{q}(\tilde{x}) = \tilde{x}_1^2 \quad \text{or} \quad \tilde{x}_1^2 + \tilde{x}_2^2,\\
p(x) &\longmapsto \tilde{p}(\tilde{x}) = \tilde{x}_1^2 + \tilde{x}_2^2 + \tilde{x}_3^2,
\end{aligned}
$$

so that $\tilde{E}$ contains either $\tilde{x}_1^2, \tilde{x}_2^2 + \tilde{x}_3^2$ or $\tilde{x}_1^2 + \tilde{x}_2^2, \tilde{x}_3^2$, for which the proof in Case II works. Therefore we shall assume without loss of generality that $E_Q = \langle x_1^2 + x_2^2 + x_3^2 \rangle$.

Recall from Lemma 5, $Y_j = \sum_{i=1}^3 \omega_{ji}D_i \bmod U_0$ is in E.

$$
\begin{aligned}
&[Y_1, p(x)] = [\omega_{12}D_2 + \omega_{13}D_3, p(x)] = 2(x_2\omega_{12} + x_3\omega_{13})\\
&\qquad = 2(a_{11}x_1x_2 + a_{12}x_2^2 + a_{13}x_2x_3 + a_{21}x_1x_3 + a_{22}x_2x_3 + a_{23}x_3^2) \mod P_1.
\end{aligned}
$$

So $a_{11}x_1x_2 + a_{12}x_2^2 + (a_{13} + a_{22})x_2x_3 + a_{21}x_1x_3 + a_{23}x_3^2$ is in E_Q and hence equal to $c_1(x_1^2 + x_2^2 + x_3^2)$. Comparing coefficients of x_1^2 allows us to conclude that $c_1 = 0$. Thus $a_{11} = a_{21} = a_{12} = a_{23} = 0, a_{13} + a_{22} = 0$, and

$$A = \begin{pmatrix} 0 & 0 & a_{13} \\ 0 & -a_{13} & 0 \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

Similarly,

$$\begin{aligned}[Y_2, p(x)] &= [\omega_{21}D_1 + \omega_{23}D_3, p(x)] = 2(-x_1\omega_{12} + x_3\omega_{23}) \\ &= 2(-a_{11}x_1^2 - a_{12}x_1x_2 - a_{13}x_1x_3 + a_{31}x_1x_3 + a_{32}x_2x_3 + a_{33}x_3^2) \mod P_1 \\ &= 2((a_{31} - a_{13})x_1x_3 + a_{32}x_2x_3 + a_{33}x_3^2) \mod P_1.\end{aligned}$$

So $(a_{31} - a_{13})x_1x_3 + a_{32}x_2x_3 + a_{33}x_3^2$ is in E_Q and hence equal to $c_2(x_1^2 + x_2^2 + x_3^2)$. Thus $c_2 = 0$ and $a_{32} = a_{33} = 0, a_{31} = a_{13}$.

$$A = a_{13}\begin{pmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Finally, the cyclic relation $\frac{\partial\omega_{12}}{\partial x_3} + \frac{\partial\omega_{23}}{\partial x_1} + \frac{\partial\omega_{31}}{\partial x_2} = 0$ allows us to conclude that $a_{13} = 0$. Therefore A is a zero matrix and we are done.

3.2. Case $r_{\max} = 2$. There exists homogeneous polynomial $p(x) \in E$ with $\operatorname{rank}(p(x)) = 2$. Without loss of generality, we shall assume that

$$p(x) = k_1x_1^2 + k_2x_2^2,$$

where $k_1k_2 \neq 0$. We remark that E cannot contain x_3^2 since $r_{\max} = 2$.

Case I: $k_1 \neq k_2$. By evaluating $[[L_0, p(x)], p(x)]$, we can obtain $k_1^2x_1^2 + k_2^2x_2^2$ in E. It follows that x_1^2, x_2^2 are in E.

$$\begin{aligned}&[L_0, x_1^2] = 2x_1D_1 + 1 \qquad \text{and} \qquad [L_0, x_2^2] = 2x_2D_2 + 1 \\ &\Rightarrow x_1D_1 \in E \qquad \text{and} \qquad x_2D_2 \in E \\ &\Rightarrow x_1x_2\omega_{12} = -[x_1D_1, x_2D_2] \in E \\ &\Rightarrow \omega_{12} = c_{12} = \text{constant} \qquad \text{by Ocone's result.}\end{aligned}$$

LEMMA 8. *Suppose that x_1D_1, x_2D_2 are in E. If $q(x) = q_{11}x_1^2 + q_{12}x_1x_2 + q_{13}x_1x_3 + q_{22}x_2^2 + q_{23}x_2x_3 + q_{33}x_3^2$ is in E, then each individual $q_{ij}x_ix_j$ is in E.*

Proof.

$$\begin{aligned}[x_1D_1, q(x)] &= x_1\frac{\partial q}{\partial x_1} = 2q_{11}x_1^2 + q_{12}x_1x_2 + q_{13}x_1x_3 \\ [x_1D_1, [x_1D_1, q(x)]] &= 4q_{11}x_1^2 + q_{12}x_1x_2 + q_{13}x_1x_3.\end{aligned}$$

These imply $q_{11}x_1^2 \in E$ and $q_{12}x_1x_2 + q_{13}x_1x_3 \in E$.

$$[x_2D_2, q_{12}x_1x_2 + q_{13}x_1x_3] = q_{12}x_1x_2 \in E.$$

This implies $q_{13}x_1x_3 \in E$.

$$\begin{aligned}&[x_2D_2, q_{22}x_2^2 + q_{23}x_2x_3 + q_{33}x_3^2 = 2q_{22}x_2^2 + q_{23}x_2x_3 \in E \\ &[x_2D_2, 2q_{22}x_2^2 + q_{23}x_2x_3] = 4q_{22}x_2^2 + q_{23}x_2x_3 \in E.\end{aligned}$$

These imply $q_{22}x_2^2 \in E, q_{23}x_2x_3 \in E$ and $q_{23}x_3^2 \in E$. □

We now claim that $x_1x_3 \notin E$. If $x_1x_3 \in E$, then

$$\begin{aligned}
&[L_0, x_1x_3] = \frac{1}{2}[D_1^2 + D_2^2 + D_3^2, x_1x_3] = x_3D_1 + x_1D_3 \in E \\
&\Rightarrow\ [x_1D_3 + x_3D_1, x_1x_3] = x_1^2 + x_3^2 \in E \\
&\Rightarrow\ x_1^2 + x_2^2 + x_3^2 \in E \qquad \text{and} \qquad \operatorname{rank}(x_1^2 + x_2^2 + x_3^2) = 3.
\end{aligned}$$

This gives a contradiction. So we conclude that $x_1x_3 \notin E$. Similarly we conclude that $x_2x_3 \notin E$. Clearly $x_3^2 \notin E$. In view of Lemma 8, we have

$$\langle x_1^2, x_2^2 \rangle \subseteq E_Q \subseteq \langle x_1^2, x_2^2, x_1x_2 \rangle.$$

By Lemma 5,

$$-[x_1D_1, D_3] = x_1\omega_{13} = a_{21}x_1^2 + a_{22}x_1x_2 + a_{23}x_1x_3 \in E.$$

In view of Lemma 8, we have $a_{23}x_1x_3 \in E$, which implies $a_{23} = 0$. Similarly,

$$[x_2D_2, D_3] = x_2\omega_{23} = a_{31}x_1x_2 + a_{32}x_2^2 + a_{33}x_2x_3 \in E$$

implies $a_{33} = 0$. Then

$$A = \begin{pmatrix} 0 & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & 0 \end{pmatrix}.$$

Let $Z_1 = x_1D_1$,

$$\begin{aligned}
Z_2 :=& [L_0, Z_1] = \frac{1}{2}\sum_{i=1}^{3}\left(2\frac{\partial x_1}{\partial x_i}D_iD_1 - 2x_1\omega_{i1}D_i\right) \mod U_0 \\
=& D_1^2 + c_{12}x_1D_2 + x_1\omega_{13}D_3 \mod U_0, \\
Z_3 :=& [L_0, Z_2] = \frac{1}{2}\sum_{i=1}^{3}[D_i^2, D_1^2 + c_{12}x_1D_2 + x_1\omega_{13}D_3] \mod U_1 \\
=& \sum_{i=1}^{3}\left(2\omega_{1i}D_iD_1 + \frac{\partial(c_{12}x_1)}{\partial x_i}D_iD_2 + \frac{\partial(x_1\omega_{13})}{\partial x_i}D_iD_3\right) \mod U_1 \\
=& 2\omega_{12}D_2D_1 + 2\omega_{13}D_3D_1 + c_{12}D_1D_2 + \left(\omega_{13} + x_1\frac{\partial\omega_{13}}{\partial x_1}\right)D_1D_3 \\
&+ x_1\frac{\partial\omega_{23}}{\partial x_2}D_2D_3 \mod U_1 \\
=& 3c_{12}D_1D_2 + (4a_{21}x_1 + 3a_{22}x_2 + 3c_{13})D_1D_3 + a_{32}x_1D_2D_3 \mod U_1, \\
[Z_3, Z_1] =& [3c_{12}D_1D_2 + (4a_{21}x_1 + 3a_{22}x_2 + 3c_{13})D_1D_3 + a_{32}x_1D_2D_3, x_1D_1] \\
&\mod U_1 \\
=& 3c_{12}D_1D_2 + (4a_{21}x_1 + 3a_{22}x_2 + 3c_{13})D_1D_3 - 4a_{21}x_1D_1D_3 - a_{32}x_1D_2D_3 \\
&\mod U_1 \\
=& 3c_{12}D_1D_2 + 3a_{22}x_2 + 3c_{13})D_1D_3 - a_{32}x_1D_2D_3 \mod U_1 \\
Z_4 =& \frac{1}{2}[Z_3, Z_1] + \frac{1}{2}Z_3 = 3c_{12}D_1D_2 + (2a_{21}x_1 + 3a_{22}x_2 + 3c_{13})D_1D_3 \mod U_1, \\
[Z_4, Z_1] =& [3c_{12}D_1D_2 + (2a_{21}x_1 + 3a_{22}x_2 + 3c_{13})D_1D_3, x_1D_1] \mod U_1
\end{aligned}$$

$$
\begin{aligned}
&= 3c_{12}D_1D_2 + (2a_{21}x_1 + 3a_{22}x_2 + 3c_{13})D_1D_3 - 2a_{21}x_1D_1D_3 \mod U_1\\
&= 3c_{12}D_1D_2 + (3a_{22}x_2 + 3c_{13})D_1D_3 \mod U_1,\\
Z_5 &= \frac{1}{2}(Z_4 - [Z_4, Z_1]) = a_{21}x_1D_1D_3 \mod U_1,\\
[L_0, Z_5] &= \frac{1}{2}[D_1^2 + D_2^2 + D_3^2, a_{21}x_1D_1D_3] \mod U_2\\
&= a_{21}D_1^2D_3 \mod U_2,\\
[[L_0, Z_5], Z_5] &= [a_{21}D_1^2D_3, a_{21}x_1D_1D_3] \mod U_3\\
&= a_{21}^2D_1^2D_3^2 \mod U_3.
\end{aligned}
$$

By induction, we get infinite elements in E of the form

$$(-1)^n Ad_{Z_5}^n(L_0) = a_{21}^n D_1^2 D_3^n \mod U_{n+1}.$$

Since E is finite dimensional, we conclude that

$$a_{21} = 0. \tag{3.3}$$

$$
\begin{aligned}
W_1 &:= x_2D_2 \in E\\
W_2 &:= [L_0, W_1] = \frac{1}{2}\sum_{i=1}^{3}[D_i^2, x_2D_2] \mod U_0\\
&= \frac{1}{2}\sum_{i=1}^{3}\left(2\frac{\partial x_2}{\partial x_i}D_iD_2 - 2x_1\omega_{i2}D_i\right) \mod U_0\\
&= D_2^2 - x_1c_{12}D_1 + x_1\omega_{23}D_3 \mod U_0,\\
W_3 &:= [L_0, W_2]\\
&= \frac{1}{2}\sum_{i=1}^{3}[D_i^2, D_2^2 - c_{12}x_1D_1 + x_1\omega_{23}D_3] \mod U_1\\
&= \sum_{i=1}^{3}\left(2\omega_{2i}D_iD_2 - \frac{\partial(c_{12}x_1)}{\partial x_i}D_iD_1 + \frac{\partial(x_1\omega_{23})}{\partial x_i}D_iD_3\right) \mod U_1\\
&= 2\omega_{21}D_1D_2 + 2\omega_{23}D_3D_2 - c_{12}D_1^2 + \left(\omega_{23} + x_1\frac{\partial\omega_{23}}{\partial x_1}\right)D_1D_3\\
&\quad + x_1\frac{\partial\omega_{23}}{\partial x_2}D_2D_3 \mod U_1\\
&= (-2a_{22}x_2)D_1D_2 + (2a_{31}x_1 + 2a_{32}x_2)D_2D_3 - c_{12}D_1^2\\
&\quad + (2a_{31}x_1 + a_{32}x_2)D_1D_3 + a_{32}x_1D_2D_3 \mod U_1\\
&= (-2a_{22}x_2)D_1D_2 + [(2a_{31} + a_{32})x_1 + 2a_{32}x_2]D_2D_3 - c_{12}D_1^2 \qquad (3.4)\\
&\quad + (2a_{31}x_1 + a_{32}x_2)D_1D_3 \mod U_1,\\
[W_3, W_1] &= [(-2a_{22}x_2)D_1D_2 + ((2a_{31} + a_{32})x_1 + 2a_{32}x_2)D_2D_3 - c_{12}D_1^2\\
&\quad + (2a_{31}x_1 + a_{32}x_2)D_1D_3, x_2D_2] \mod U_1\\
&= -2a_{22}x_2D_1D_2 + ((2a_{31} + a_{32})x_1 + 2a_{32}x_2)D_2D_3 + 2a_{22}x_2D_1D_2\\
&\quad - 2a_{32}x_2D_2D_3 - a_{32}x_2D_1D_3 \mod U_1\\
&= (2a_{31} + a_{32})x_1D_2D_3 - a_{32}x_2D_1D_3 \mod U_1,
\end{aligned}
$$

$$
\begin{aligned}
[[W_3, W_1], Z_1] &= [(2a_{31} + a_{32}x_1D_2D_3 - a_{32}x_2D_1D_3, x_1D_1] \mod U_1 \\
&= -a_{32}x_2D_1D_3 - (2a_{31} + a_{32})x_1D_2D_3 \mod U_1 \\
-\frac{1}{2}([W_3, W_1] + [[W_3, W_1], Z_1]) &= a_{32}x_2D_1D_3 \mod U_1. \qquad (3.5) \\
\frac{1}{2}([W_3, W_1] - [[W_3, W_1], Z_1]) &= (2a_{31} + a_{32})x_1D_2D_3 \mod U_1. \qquad (3.6)
\end{aligned}
$$

It follows from (3.4), (3.5), and (3.6) that

$$
\begin{aligned}
W_4 &:= -2a_{22}x_2D_1D_2 + 2a_{32}x_2D_2D_3 - c_{12}D_1^2 + 2a_{31}x_1D_1D_3 \mod U_1, \\
[W_4, Z_1] &= [-2a_{22}x_2D_1D_2 + 2a_{32}x_2D_2D_3 - c_{12}D_1^2 + 2a_{31}x_1D_1D_3, x_1D_1] \\
&\quad \mod U_1 \\
&= -2a_{22}x_2D_1D_2 - 2c_{12}D_1^2 \mod U_1, \\
W_5 &:= -\frac{1}{2}[W_4, Z_1] \mod U_1 \\
&= a_{22}x_2D_1D_2 + c_{12}D_1^2 \mod U_1, \\
[L_0, W_5] &= \frac{1}{2}[D_1^2 + D_2^2 + D_3^2, a_{22}x_2D_1D_2 + c_{12}D_1^2] \mod U_2 \\
&= a_{22}D_1D_2^2 \mod U_2, \\
[[L_0, W_5], W_5] &= [a_{22}D_1D_2^2, a_{22}x_2D_1D_2 + c_{12}D_1^2] \mod U_3 \\
&= 2a_{22}^2D_1^2D_2^2 \mod U_3.
\end{aligned}
$$

By induction, we have

$$(-1)^n Ad_{W_5}^n(L_0) = 2^{n-1}a_{22}^n D_1^n D_2^2 \mod U_{n+1}.$$

Since E is finite dimensional, we conclude that

$$a_{22} = 0. \qquad (3.7)$$

By the cyclic relation $\frac{\partial \omega_{12}}{\partial x_3} + \frac{\partial \omega_{23}}{\partial x_1} + \frac{\partial \omega_{31}}{\partial x_2} = 0$, we get

$$a_{13} + a_{31} - a_{22} = 0.$$

From (3.3) and (3.7), we get $a_{31} = 0$. It follows that

$$
\begin{aligned}
W_4 &= 2a_{32}x_2D_2D_3 - c_{12}D_1^2 \mod U_1, \\
\left[L_0, \frac{1}{2}W_4\right] &= \frac{1}{2}\left[D_1^2 + D_2^2 + D_3^2, a_{32}x_2D_2D_3 - \frac{1}{2}c_{12}D_1^2\right] \mod U_2 \\
&= a_{32}D_2^2D_3 \mod U_2, \\
\left[\left[L_0, \frac{1}{2}W_4\right], \frac{1}{2}W_4\right] &= \left[a_{32}D_2^2D_3, a_{32}x_2D_2D_3 - \frac{1}{2}c_{12}D_1^2\right] \mod U_3 \\
&= 2a_{32}^2D_2^2D_3^2 \mod U_3, \\
(-1)^n Ad_{\frac{1}{2}W_4}^n(L_0) &= 2^{n-1}a_{32}^n D_2^2 D_3^n \mod U_{n+1}.
\end{aligned}
$$

Since E is finite dimensional, we have $a_{32} = 0$. Therefore, the ω_{ij}'s are constants for all i, j.

Case II: $k_1 = k_2$. Without loss of generality, we may take $p(x) = x_1^2 + x_2^2$. In view of Case I, we shall assume that E does not contain x_1^2, x_2^2.

LEMMA 9. *Under the Case* II *assumption,* $\langle x_1^2 + x_2^2 \rangle \subseteq E_Q \subseteq \langle x_1^2, x_2^2, x_1x_2 \rangle$.

Proof. Let $q(x) \in E_Q$. Then

$$q(x) = q_{11}x_1^2 + q_{22}x_2^2 + q_{33}x_3^2 + q_{12}x_1x_2 + q_{13}x_1x_3 + q_{23}x_2x_3.$$

Recall that $x_1D_1 + x_2D_2$ is in E. By applying $x_1D_1 + x_2D_2$ repeatedly to $q(x)$, we see immediately that $q_{11}x_1^2 + q_{22}x_2^2 + q_{12}x_1x_2$, $q_{13}x_1x_3 + q_{23}x_2x_3, q_{33}x_3^2 \in E$. These imply $q_{33} = 0$ (since $r_{\max} = 2$) and $\frac{1}{2}(x_1^2 + x_2^2) + (q_{13}x_1x_3 + q_{23}x_2x_3) \in E$.

$$Hess\left[\frac{1}{2}(x_1^2 + x_2^2) + (q_{13}x_1x_3 + q_{23}x_2x_3)\right] = \begin{pmatrix} 1 & 0 & q_{13} \\ 0 & 1 & q_{23} \\ q_{13} & q_{23} & 0 \end{pmatrix}.$$

The determinant of the above matrix is $-(q_{13}^2 + q_{23}^2)$. Since $r_{\max} = 2 < 3$, we have $q_{13}^2 + q_{23}^2 = 0$ which implies $q_{13} = 0 = q_{23}$. □

We deduce from Lemma 9 that $1 \leq \dim E_Q \leq 3$.

If $\dim E_Q = 3$, then $E_Q = \langle x_1^2, x_2^2, x_1x_2 \rangle$ and we are in Case I.

If $\dim E_Q = 2$, then we may take $E_Q = \langle x_1^2 + x_2^2, q_{11}x_1^2 + q_{12}x_1x_2 \rangle$. If $q_{12} = 0$, then E_Q contains both x_1^2 and x_2^2 and we are back in Case I. Therefore we can assume that $q_{12} \neq 0$. Furthermore if $q_{11} = 0$, then E_Q is actually $\langle x_1^2 + x_2^2, x_1x_2 \rangle$. We consider the following particular orthogonal transformation:

$$\tilde{x} = Rx \qquad R = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

such that it gives rise to

$$\begin{aligned} x &= R^T\tilde{x} \\ x_1^2 + x_2^2 &\longmapsto \tilde{x}_1^2 + \tilde{x}_2^2 \\ x_1x_2 &\longmapsto \frac{\tilde{x}_1^2 + \tilde{x}_2^2}{\sqrt{2}} \frac{-\tilde{x}_1^2 + \tilde{x}_2^2}{\sqrt{2}} = \frac{-\tilde{x}_1^2 + \tilde{x}_2^2}{2} \\ E_Q &\longmapsto \tilde{E}_Q = \left\langle \tilde{x}_1^2 + \tilde{x}_2^2, \frac{-\tilde{x}_1^2 + \tilde{x}_2^2}{2} \right\rangle = \langle \tilde{x}_1^2, \tilde{x}_2^2 \rangle. \end{aligned}$$

Thus, $\tilde{E}$ contains $\tilde{x}_1^2$ and $\tilde{x}_2^2$. By Case I, the $\tilde{\omega}_{ij}$'s are constants and so are the ω_{ij}'s as $\Omega = R^T\tilde{\Omega}R$. Hence we may also assume that $q_{11} \neq 0$. So $E_Q = \langle x_1^2 + x_2^2, x_1^2 + 2kx_1x_2 \rangle$ for $k \neq 0$. Observe that if we can find a quadratic form $p_0 \in E_Q$ with $r(p_0) = 1$, then there exists an orthogonal transformation such that E_Q is mapped into $\tilde{E}_Q$, which contains both $\tilde{x}_1^2$ and $\tilde{x}_2^2$, and we are done. So we try to find such a p_0 below. Consider

$$p_0 = \lambda(x_1^2 + x_2^2) + \sigma(x_1^2 + 2kx_1x_2).$$

Its underlying symmetric matrix is

$$A_{p_0} = \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \sigma \begin{pmatrix} 1 & k \\ k & 0 \end{pmatrix} = \begin{pmatrix} \lambda + \sigma & \sigma k \\ \sigma k & \lambda \end{pmatrix}$$

and $\det A_{p_0} = \lambda^2 + \sigma\lambda - \sigma^2 k^2$. Fix $\sigma \neq 0$ (say $\sigma = k$) and choose

$$\lambda = \sigma \frac{-1 + \sqrt{1+4k^2}}{2}.$$

Then $r(p_0) = 1$. We are done for $\dim E_Q = 2$.

If $\dim E_Q = 1$, then $E_Q = \langle x_1^2 + x_2^2 \rangle$. Recall from Lemma 5 that Y_j's are in E where $Y_1 = \omega_{12}D_2 + \omega_{13}D_3 \bmod U_0$, $Y_2 = \omega_{21}D_1 + \omega_{23}D_3 \bmod U_0$, and $Y_3 = \omega_{31}D_1 + \omega_{32}D_2 \bmod U_0$.

$$\begin{aligned}
\frac{1}{2}[Y_1, x_1^2 + x_2^2] &= x_2\omega_{12} = a_{11}x_1x_2 + a_{12}x_2^2 + a_{13}x_1x_3 \mod P_1 \\
&\Rightarrow \quad a_{11}x_1x_2 + a_{12}x_2^2 + a_{13}x_1x_3 \in E_Q = \langle x_1^2 + x_2^2 \rangle \\
&\Rightarrow \quad a_{11} = a_{12} = a_{13} = 0, \\
-\frac{1}{2}[Y_3, x_1^2 + x_2^2] &= x_1\omega_{13} + x_2\omega_{23} \\
&= a_{21}x_1^2 + (a_{22} + a_{31})x_1x_2 + a_{32}x_2^2 + a_{23}x_1x_3 + a_{33}x_2x_3 \mod P_1 \\
&\Rightarrow \quad a_{21}x_1^2 + (a_{22} + a_{31})x_1x_2 + a_{32}x_2^2 + a_{23}x_1x_3 + a_{33}x_2x_3 \in \langle x_1^2 + x_2^2 \rangle \\
&\Rightarrow \quad a_{21} = a_{32}, a_{22} + a_{31} = 0, a_{23} = a_{33} = 0.
\end{aligned}$$

By the cyclic relation $\frac{\partial \omega_{12}}{\partial x_3} + \frac{\partial \omega_{23}}{\partial x_1} + \frac{\partial \omega_{31}}{\partial x_2} = 0$, we have $a_{13} + a_{31} - a_{22} = 0$. It follows that $a_{22} = a_{31} = 0$ and

$$A = a_{21} \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

In order to prove that $a_{21} = 0$, we consider the following sequence of elements in E.

$$\begin{aligned}
K_1 &:= x_1D_1 + x_2D_2, \\
K_2 &:= [L_0, K_1] = \frac{1}{2}\sum_{i=1}^{3}[D_i^2, x_1D_1] + \frac{1}{2}\sum_{i=1}^{3}[D_i^2, x_2D_2] \mod U_0 \\
&= D_1^2 + x_1\omega_{12}D_2 + x_1\omega_{13}D_3 + D_2^2 + x_2\omega_{21}D_1 + x_2\omega_{23}D_3 \mod U_0 \\
&= D_1^2 + D_2^2 + x_2\omega_{21}D_1 + x_1\omega_{12}D_2 + (x_1\omega_{13} + x_2\omega_{23})D_3 \mod U_0, \\
K_3 &:= [L_0, K_2] \\
&= \frac{1}{2}\sum_{i=1}^{3}[D_i^2, D_1^2 + D_2^2 + x_2\omega_{21}D_1 + x_1\omega_{12}D_2 + (x_1\omega_{13} + x_2\omega_{23})D_3] \\
&\quad \bmod U_1 \\
&= 2(\omega_{12}D_1D_2 + \omega_{13}D_1D_3 + \omega_{21}D_2D_1 + \omega_{23}D_2D_3) \\
&\quad - c_{12}D_2D_1 + c_{12}D_1D_2 \\
&\quad + (\omega_{13}D_1D_3 + a_{21}x_1D_1D_3 + \omega_{23}D_2D_3 + a_{32}x_2D_2D_3) \mod U_1 \\
&= (3\omega_{13} + a_{21}x_1)D_1D_3 + (3\omega_{23} + a_{21}x_2)D_2D_3 \mod U_1 \\
&= 4a_{21}(x_1D_1 + x_2D_2)D_3 \mod U_1, \\
(-1)Ad_{K_3}(K_2) &= [K_2, K_3] \\
&= [D_1^2 + D_2^2, 4a_{21}(x_1D_1 + x_2D_2)D_3] \mod U_2 \\
&= 4a_{21}([D_1^2, x_1D_1] + [D_2^2, x_2D_2])D_3 \mod U_2 \\
&= 8a_{21}(D_1^2 + D_2^2)D_3 \mod U_2.
\end{aligned}$$

Inductively, we have

$$(-1)Ad^n_{K_3}(K_2) = (8a_{21})^n(D_1^2 + D_2^2)D_3^n \mod U_{n+1}.$$

Since $\dim E < \infty$, $a_{21} = 0$ and we have $A = 0$. So ω_{ij}'s are constant for all i, j.

3.3. Case $r_{\max} = 1$. In this case, we may assume that $p(x) = x_1^2 \in E$ and $E_Q = \langle x_1^2 \rangle$.

$$[Y_2, p(x)] = [\omega_{21}D_1 + \omega_{23}D_3, x_1^2] = 2\omega_{21}x_1 \in E_Q$$
$$[Y_3, p(x)] = [\omega_{31}D_1 + \omega_{32}D_2, x_1^2] = 2\omega_{31}x_1 \in E_Q.$$

Thus, ω_{12} and ω_{13} depend only on x_1 because $E_Q = \langle x_1^2 \rangle$. So

$$A = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & 0 & 0 \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

The cyclic relation $\frac{\partial \omega_{12}}{\partial x_3} + \frac{\partial \omega_{23}}{\partial x_1} + \frac{\partial \omega_{31}}{\partial x_2} = 0$ implies $a_{31} = 0$ and implies $a_{31} = 0$ and

$$A = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Now $x_1^2 \in E$ implies $x_1 D_1 \in E$. Let

$$\begin{aligned}
X_1 &:= x_1 D_1, \\
X_2 &:= [L_0, X_1] = \frac{1}{2}\sum_{i=1}^{3}[D_i^2, x_1 D_1] \mod U_0 \\
&= \frac{1}{2}\sum_{i=1}^{3}\left(2\frac{\partial x_1}{\partial x_i}D_iD_1 - 2x_1\omega_{i1}D_i + \frac{\partial^2 x_1}{\partial x_i^2}D_1\right) \mod U_0 \\
&= D_1^2 + x_1\omega_{12}D_2 + x_1\omega_{13}D_3 \mod U_0 \\
&= D_1^2 + (a_{11}x_1^2 + c_{12}x_1)D_2 + (a_{21}x_1^2 + c_{13}x_1)D_3 \mod U_0 \\
&= D_1^2 \mod U_1, \\
X_3 &:= [L_0, X_2] \\
&= \frac{1}{2}\sum_{i=1}^{3}[D_i^2, D_1^2 + (a_{11}x_1^2 + c_{12}x_1)D_2 + (a_{21}x_1^2 + c_{13}x_1)D_3] \mod U_1 \\
&= \sum_{i=1}^{3}\left(2\omega_{1i}D_1D_i + \frac{\partial(a_{11}x_1^2 + c_{12}x_1)}{\partial x_i}D_iD_2 + \frac{\partial(a_{21}x_1^2 + c_{13}x_1)}{\partial x_i}D_iD_3\right) \mod U_1 \\
&= 2(a_{11}x_1 + c_{12})D_1D_2 + 2(a_{21}x_1 + c_{13})D_1D_3 + (2a_{11}x_1 + c_{12})D_1D_2 \\
&\quad + (2a_{21}x_1 + c_{13})D_1D_3 \mod U_1 \\
&= (4a_{11}x_1 + 3c_{12})D_1D_2 + (4a_{21}x_1 + 3c_{13})D_1D_3 \mod U_1.
\end{aligned}$$

We are going to show that $a_{11} = 0$. Suppose $a_{11} \neq 0$. Denote $a = \frac{a_{21}}{a_{11}}$ and define

$$\alpha := \frac{1}{4a_{11}}X_3 = \left(x_1 + \frac{3c_{12}}{a_{11}}\right)D_1D_2 + \left(ax_1 + \frac{3c_{13}}{a_{11}}\right)D_1D_3 \mod U_1.$$

LEMMA 10. *If* $a_{11} \neq 0$, *let* $\alpha := \frac{1}{4a_{11}} X_3, a = \frac{a_{21}}{a_{11}}$. *Then for* $j \geq 1$

$$\frac{(-1)^j Ad_\alpha^j(X_2)}{2^j} = D_1^2(D_2 + aD_3)^j \mod U_{j+1}.$$

Proof. We shall prove this by induction.

$$\begin{aligned}
\frac{(-1)Ad_\alpha(X_2)}{2} &= \frac{1}{2}[X_2, \alpha] \\
&= \frac{1}{2}\left[D_1^2, \left(x_1 + \frac{3c_{12}}{a_{11}}\right)D_1D_2 + \left(ax_1 + \frac{3c_{13}}{a_{11}}\right)D_1D_3\right] \mod U_2 \\
&= \frac{\partial}{\partial x_1}\left(x_1 + \frac{3c_{12}}{a_{11}}\right)D_1^2D_2 + \frac{\partial}{\partial x_1}\left(ax_1 + \frac{3c_{13}}{a_{11}}\right)D_1^2D_3 \mod U_2 \\
&= D_1^2(D_2 + aD_3) \mod U_2, \\
\frac{(-1)^{j+1}Ad_\alpha^{j+1}(X_2)}{2^{j+1}} &= \left(-\frac{1}{2}\right)Ad_\alpha \frac{(-1)^j Ad_\alpha^j(X_2)}{2^j} \\
&= \frac{1}{2}\left[\frac{(-1)^j Ad_\alpha^j(X_2)}{2^j}, \alpha\right] \\
&= \frac{1}{2}\left[D_1^2(D_2 + aD_3)^j, \left(x_1 + \frac{3c_{13}}{a_{11}}\right)D_1D_2 + \left(ax_1 + \frac{3c_{13}}{a_{11}}\right)D_1D_3\right] \\
&\quad \mod U_{j+2} \\
&= D_1^2(D_2 + aD_3)^j D_2 + aD_1^2(D_2 + aD_3)^j D_3 \mod U_{j+2} \\
&= D_1^2(D_2 + aD_3)^j(D_2 + aD_3) \mod U_{j+2} \\
&= D_1^2(D_2 + aD_3)^{j+1} \mod U_{j+2}. \qquad \square
\end{aligned}$$

The above lemma implies that E is infinite dimensional, contradicting the finite-dimensionality of E. Hence $a_{11} = 0$. Then

$$\begin{aligned}
A &= \begin{pmatrix} 0 & 0 & 0 \\ a_{21} & 0 & 0 \\ 0 & a_{32} & a_{33} \end{pmatrix}, \\
X_2 &= D_1^2 + c_{12}x_1D_2 + (a_{21}x_1^2 + c_{13}x_1)D_3 \mod U_0, \\
X_3 &= 3c_{12}D_1D_2 + (4a_{21}x_1 + 3c_{13})D_1D_3 \mod U_1.
\end{aligned}$$

Next we shall see that $a_{21} = 0$. Suppose $a_{21} \neq 0$. Consider

$$\begin{aligned}
\beta := \frac{1}{4a_{21}}X_3 &= \frac{3c_{12}}{4a_{21}}D_1D_2 + \left(x_1 + \frac{3c_{13}}{4a_{21}}\right)D_1D_3 \mod U_1, \\
(-1)Ad_\beta(X_2) = [X_2, \beta] &= \left[D_1^2, \frac{3c_{12}}{4a_{21}}D_1D_2 + \left(x_1 + \frac{3c_{13}}{4a_{21}}\right)D_1D_3\right] \mod U_2 \\
&= 2D_1^2D_3 \mod U_2.
\end{aligned}$$

We claim that $(-1)^j Ad_\beta^j(X_2) = 2^j D_1^2 D_3^j \mod U_{j+1}$. This can be seen by induction.

$$\begin{aligned}
(-1)^{j+1}Ad_\beta^{j+1}(X_2) &= (-1)Ad_\beta((-1)Ad_\beta^j(X_2)) \\
&= [(-1)^j Ad_\beta^j(X_2), \beta] \\
&= \left[2^j D_1^2 D_3^j, \frac{3c_{12}}{4a_{21}}D_1D_2 + \left(x_1 + \frac{3c_{13}}{4a_{21}}\right)D_1D_3\right] \mod U_{j+2} \\
&= 2^{j+1}D_1^2D_3^{j+1} \mod U_{j+2}.
\end{aligned}$$

Then E contains an infinite-dimensional subspace, which is impossible. Hence $a_{21} = 0$ and

$$A = \begin{pmatrix} o & o & o \\ 0 & 0 & 0 \\ 0 & a_{32} & a_{33} \end{pmatrix}.$$

Consider the expression of $[Y_j, D_k]$ in Lemma 5(vi). Noting that ω_{ij}'s are linear, the following elements belong to E:

$$K_{jk} = \sum_{i=1}^{3} \omega_{ji}\omega_{ki} - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_k \partial x_j}.$$

K_{jk} is symmetric about j, k (Table 1) and is a polynomial of degree at most two, which in turn forces η to be a polynomial of degree at most four.

TABLE 1.

(j,k)	K_{jk}
(1,1)	$\omega_{12}^2 + \omega_{13}^2 - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1^2}$
(1,2)	$\omega_{13}\omega_{23} - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1 \partial x_2}$
(1,3)	$\omega_{12}\omega_{32} - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1 \partial x_3}$
(2,2)	$\omega_{21}^2 + \omega_{23}^2 - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_2^2}$
(2,3)	$\omega_{21}\omega_{31} - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_2 \partial x_3}$
(3,3)	$\omega_{31}^2 + \omega_{32}^2 - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_3^2}$

Recall our notation:

$$\begin{pmatrix} \omega_{12} \\ \omega_{13} \\ \omega_{23} \end{pmatrix} = A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} c_{12} \\ c_{13} \\ c_{23} \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & a_{32} & a_{33} \end{pmatrix}.$$

Since $K_{jk} \in P_2$ and $E_Q = \langle x_1^2 \rangle$ in this case $r_{\max} = 1$, we have

$$K_{jk} = k x_1^2 \mod P_1.$$

So we can form the following relationships:

$$\begin{aligned} \frac{1}{2}\frac{\partial^2 \eta}{\partial x_2^2} &= a_{32}^2 x_2^2 + 2a_{32}a_{33}x_2x_3 + a_{33}x_3^2 + a x_1^2 \mod P_1, \\ \frac{1}{2}\frac{\partial^2 \eta}{\partial x_3^2} &= a_{32}^2 x_2^2 + 2a_{32}a_{33}x_2x_3 + a_{33}x_3^2 + b x_1^2 \mod P_1, \\ \frac{1}{2}\frac{\partial^2 \eta}{\partial x_2 \partial x_3} &= c x_1^2 \mod P_1. \end{aligned}$$

Observe that the term $a_{33}x_3^2$ in $\frac{1}{2}\frac{\partial^2 \eta}{\partial x_2^2}$ must come from the $x_2^2 x_3^2$ term in η. Let η contain the term $\alpha x_2^2 x_3^2$. Since

$$\frac{1}{2}\frac{\partial^2 \alpha x_2^2 x_3^2}{\partial x_2^2} = \alpha x_3^2, \quad \frac{1}{2}\frac{\partial^2 \alpha x_2^2 x_3^2}{\partial x_3^2} = \alpha x_2^2, \quad \frac{1}{2}\frac{\partial^2 \alpha x_2^2 x_3^2}{\partial x_2 \partial x_3} = 2\alpha x_2 x_3,$$

by comparing coefficients we obtain

$$\alpha = a_{33}^2, \quad \alpha = a_{32}^2, \quad 2\alpha = 0.$$

So, $a_{32} = a_{33} = 0$ and accordingly

$$\Omega = O_{3\times 3} \mod P_0.$$

Hence Case $r_{\max} = 1$ is done.

3.4. Case $r_{\max} = 0$. In this case $E_Q = \phi$. All functions in E are automatically linear.

Recall that

$$m_{jk} = -\sum_{i=1}^{3} \omega_{ji}\omega_{ki} + \frac{1}{2}\frac{\partial^2 \eta}{\partial x_k \partial x_j} \in E_Q.$$

This expression is written in element form. It's more insightful to view it in matrix form.

Let $M = (m_{jk})_{3\times 3}$ and note that the Ω matrix is antisymmetric. Then we have

$$\begin{aligned} M &= -\Omega\Omega^T + \frac{1}{2}Hess(\eta) \\ &= \Omega^2 + \frac{1}{2}Hess(\eta), \end{aligned}$$

where $Hess(\eta) = (\frac{\partial^2 \eta}{\partial x_k \partial x_j})_{3\times 3}$ is the Hessian matrix of η.

Let $\Omega = Dx_1 + Bx_2 + Cx_3 \pmod{P_0}$, where $D = (\alpha_{ij})_{3\times 3}, B = (\beta_{ij})_{3\times 3}, C = (\gamma_{ij})_{3\times 3}$ are skew-symmetric matrices. We make use of $\Omega^2 + \frac{1}{2}Hess(\eta) = 0 \mod P_1$ to infer that $D = B = C = (0)_{3\times 3}$ as follows. Writing

$$\begin{aligned} H =& \Omega^2 \\ =& H_{11}x_1^2 + H_{22}x_2^2 + H_{33}x_3^2 + H_{12}x_1x_2 + H_{13}x_1x_3 + H_{23}x_2x_3 \\ =& D^2x_1^2 + B^2x_2^2 + C^2x_3^2 + (DB + BD)x_1x_2 + (DC + CD)x_1x_3 \\ &+ (BC + CB)x_2x_3, \end{aligned}$$

we have

$$H_{11} = D^2 = -DD^T, \qquad \text{where} \quad D = \begin{pmatrix} 0 & \alpha_{12} & \alpha_{13} \\ -\alpha_{12} & 0 & \alpha_{23} \\ -\alpha_{13} & -\alpha_{23} & 0 \end{pmatrix}.$$

So

$$H_{11} = -\begin{pmatrix} \alpha_{12}^2 + \alpha_{13}^2 & \alpha_{13}\alpha_{23} & -\alpha_{12}\alpha_{23} \\ \alpha_{13}\alpha_{23} & \alpha_{12}^2 + \alpha_{23}^2 & \alpha_{12}\alpha_{13} \\ -a_{12}a_{23} & a_{12}a_{13} & a_{13}^2 + a_{23}^2 \end{pmatrix}.$$

The other H_{ij} matrices can be obtained similarly and they are listed explicitly at the end of this section.

We consider terms in η and relationships derived from $\Omega^2 + \frac{1}{2}Hess(\eta) = 0 \mod P_1$ in terms of entries in H_{ij} matrices. The coefficient of $x_1^2x_2^2$ in $-\eta = H_{11}[2,2] =$

$H_{22}[1,1] = \frac{1}{2}H_{12}[1,2]$. Similarly, $H_{11}[3,3] = H_{33}[1,1] = \frac{1}{2}H_{13}[1,3]$ and $H_{22}[3,3] = H_{33}[2,2] = H_{23}[2,3]$ ($H_{ij}[p,q]$ means the (p,q)-entry of matrix H_{ij}). We have

$$\alpha_{12}^2 + \alpha_{23}^2 = \beta_{12}^2 + \beta_{13}^2 = \frac{1}{2}(\alpha_{13}\beta_{23} + \alpha_{23}\beta_{13}), \tag{3.8}$$

$$\alpha_{13}^2 + \alpha_{23}^2 = \gamma_{12}^2 + \gamma_{13}^2 = -\frac{1}{2}(\alpha_{12}\gamma_{23} + \alpha_{23}\gamma_{12}), \tag{3.9}$$

$$\beta_{13}^2 + \beta_{23}^2 = \gamma_{12}^2 + \gamma_{23}^2 = \frac{1}{2}(\beta_{12}\gamma_{13} + \beta_{13}\gamma_{12}). \tag{3.10}$$

Together with the simple majorization relationship between any two real numbers, $2ab \le a^2 + b^2$, we can rewrite (3.8), (3.9) and (3.10) to obtain

$$\begin{aligned}
2(\alpha_{12}^2 + \alpha_{23}^2 + \beta_{12}^2 + \beta_{13}^2) &= 2(\alpha_{13}\beta_{23} + \alpha_{23}\beta_{13}) \le \alpha_{13}^2 + \beta_{23}^2 + \alpha_{23}^2 + \beta_{13}^2, \\
2(\alpha_{13}^2 + \alpha_{23}^2 + \gamma_{12}^2 + \gamma_{13}^2) &= -2(\alpha_{12}\gamma_{23} + \alpha_{23}\gamma_{12}) \le \alpha_{12}^2 + \gamma_{23}^2 + \alpha_{23}^2 + \gamma_{12}^2, \\
2(\beta_{13}^2 + \beta_{23}^2 + \gamma_{12}^2 + \gamma_{23}^2) &= 2(\beta_{12}\gamma_{13} + \beta_{13}\gamma_{12}) \le \beta_{12}^2 + \gamma_{13}^2 + \beta_{13}^2 + \gamma_{12}^2.
\end{aligned}$$

Summing these three inequalities and simplifying, we have

$$\alpha_{12}^2 + \alpha_{13}^2 + 2\alpha_{23}^2 + \beta_{12}^2 + 2\beta_{13}^2 + \beta_{23}^2 + 2\gamma_{12}^2 + \gamma_{13}^2 + \gamma_{23}^2 \le 0,$$

which implies that

$$\alpha_{12} = \alpha_{13} = \alpha_{23} = \beta_{12} = \beta_{13} = \beta_{23} = \gamma_{12} = \gamma_{13} = \gamma_{23} = 0,$$

i.e.,

$$D = B = C = O_{3\times 3}.$$

Hence

$$\Omega = O_{3\times 3} \mod P_0.$$

Case $r_{\max} = 0$ is done.

For reference we list the H_{ij} matrices below:

$$H_{11} = -\begin{pmatrix} \alpha_{12}^2 + \alpha_{13}^2 & \alpha_{13}\alpha_{23} & -\alpha_{12}\alpha_{23} \\ \alpha_{13}\alpha_{23} & \alpha_{12}^2 + \alpha_{23}^2 & \alpha_{12}\alpha_{13} \\ -\alpha_{12}\alpha_{23} & \alpha_{12}\alpha_{13} & \alpha_{13}^2 + \alpha_{23}^2 \end{pmatrix};$$

$$H_{22} = -\begin{pmatrix} \beta_{12}^2 + \beta_{13}^2 & \beta_{13}\beta_{23} & -\beta_{12}\beta_{23} \\ \beta_{13}\beta_{23} & \beta_{12}^2 + \beta_{23}^2 & \beta_{12}\beta_{13} \\ -\beta_{12}\beta_{23} & \beta_{12}\beta_{13} & \beta_{13}^2 + \beta_{23}^2 \end{pmatrix};$$

$$H_{33} = -\begin{pmatrix} \gamma_{12}^2 + \gamma_{13}^2 & \gamma_{13}\gamma_{23} & -\gamma_{12}\gamma_{23} \\ \gamma_{13}\gamma_{23} & \gamma_{12}^2 + \gamma_{23}^2 & \gamma_{12}\gamma_{13} \\ -\gamma_{12}\gamma_{23} & \gamma_{12}\gamma_{13} & \gamma_{13}^2 + \gamma_{23}^2 \end{pmatrix};$$

$$H_{12} = -\begin{pmatrix} 2\alpha_{12}\beta_{12} + 2\alpha_{13}\beta_{13} & \alpha_{13}\beta_{23} + a_{23}b_{13} & -\alpha_{12}\beta_{23} - \beta_{23}\beta_{12} \\ \alpha_{13}\beta_{23} + \alpha_{23}\beta_{13} & 2\alpha_{12}\beta_{12} + 2\alpha_{23}\beta_{23} & \alpha_{12}\beta_{13} + \alpha_{13}\beta_{12} \\ -\alpha_{12}\beta_{23} - \alpha_{23}\beta_{12} & \alpha_{12}\beta_{13} + \alpha_{13}\beta_{12} & 2\alpha_{13}\beta_{13} + 2\alpha_{23}\beta_{23} \end{pmatrix};$$

$$H_{13} = -\begin{pmatrix} 2\alpha_{12}\gamma_{12} + 2\alpha_{13}\gamma_{13} & \alpha_{13}\gamma_{23} + \alpha_{23}\gamma_{13} & -\alpha_{12}\gamma_{23} - \alpha_{23}\gamma_{12} \\ \alpha_{13}\gamma_{23} + \alpha_{23}\gamma_{13} & 2\alpha_{12}\gamma_{12} + 2\alpha_{23}\gamma_{23} & \alpha_{12}\gamma_{13} + \alpha_{13}\gamma_{12} \\ -\alpha_{12}\gamma_{23} - \alpha_{23}\gamma_{12} & \alpha_{12}\gamma_{13} + \alpha_{13}\gamma_{12} & 2\alpha_{13}\gamma_{13} + 2\alpha_{23}\gamma_{23} \end{pmatrix};$$

$$H_{23} = -\begin{pmatrix} 2\beta_{12}\gamma_{12} + 2\beta_{13}\gamma_{13} & \beta_{13}\gamma_{23} + \beta_{23}\gamma_{13} & -\beta_{12}\gamma_{23} - \beta_{23}\gamma_{12} \\ \beta_{13}\gamma_{23} + \beta_{23}\gamma_{13} & 2\beta_{12}\gamma_{12} + 2\beta_{23}\gamma_{23} & \beta_{12}\gamma_{13} + \beta_{13}\gamma_{12} \\ -\beta_{12}\gamma_{23} - \beta_{23}\gamma_{12} & \beta_{12}\gamma_{13} + \beta_{13}\gamma_{12} & 2\beta_{13}\gamma_{13} + 2\beta_{23}\gamma_{23} \end{pmatrix}.$$

Acknowledgments. We would like to thank the referee for many helpful comments, especially for pointing out the very interesting Ph.D. thesis by M. Cohen de Lara [La], in which the links between finite-dimensional estimation algebras and finite-dimensional filters were discussed.

REFERENCES

[Be] V. Benes, *Exact finite dimensional filters for certain diffusions with nonlinear drift*, Stochastics, 5 (1981), pp. 65–92.

[BrCl] R. W. Brockett and J. M. C. Clark, *The geometry of the conditional density functions*, in Analysis and Optimization of Stochastic Systems, O. L. R. Jacobs, et al., eds., Academic Press, New York, 1980, pp. 299–309.

[Br] R. W. Brockett, *Nonlinear systems and nonlinear estimation theory*, in The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. S. Willems, eds., Reidel, Dordrecht, The Netherlands, 1981.

[ChMi] M. Chaleyat-Maurel and D. Michel, *Des resultats de non-existence de filtre de dimension finie*, Stochastics, 13 (1984), pp. 83–102.

[ChYa] W. L. Chiou and S. S.-T. Yau, *Finite dimensional filters with nonlinear drift* II: *Brockett's problem on classification of finite dimensional estimation algebras*, SIAM J. Control Optim., 32 (1994), pp. 297–310.

[Co] P. C. Collingwood, *Some remarks on estimation algebras*, Systems Control Lett., 7 (1986), pp. 217–224.

[Da] M. H. A. Davis, *On a multiplicative functional transformation arising in nonlinear filtering theory*, Z. Wahrsch Verw. Gebiete, 54 (1980), pp. 125–139.

[DaMa] M. H. A. Davis and S. I. Marcus, *An introduction to nonlinear filtering*, in The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. S. Willems, eds., D. Reidel, Dordrecht, The Netherlands, 1981.

[DTWY] R. T. Dong, L. F. Tam, W. S. Wong, and S. S.-T. Yau, *Structure and classification theorems of finite dimensional exact estimation algebras*, SIAM J. Control Optim., 29 (1991), pp. 866–877.

[Fr] A. Friedman, *Stochastic Differential Equations and Applications*, Vol 1, Academic Press, New York, 1975.

[FKK] M. Fujisaki, G. Kallianpur, and H. Kunita, *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math., 1 (1972), pp. 19–40

[Ha] M. Hazewinkel, *Lecture on linear and nonlinear filtering*, in Analysis and Estimation of Stochastic Mechanical Systems, CISM Courses and Lectures, 303, W. Shiehlen and W. Wedig, eds., Springer, Vienna, 1988.

[La] M. Cohen De Lara, *Contribution des Methodes geometriques au filtrage de dimension finie*, Ph.D. thesis, Ecole des mines de Paris, 1991.

[Mi] S. K. Mitter, *On the analogy between mathematical problems of nonlinear filtering and quantum physics*, Ricerche di Automatica, 10 (1979), pp. 163–216.

[Oc] D. L. Ocone, *Finite dimensional estimation algebras in nonlinear filtering*, in The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. S. Willems, eds., D. Reidel, Dordrecht, The Netherlands, 1981.

[St] S. Steinberg, *Applications of the Lie algebraic formulas of Baker, Campbell, Hausdorff and Zassenhaus to the calculation of explicit solutions of partial differential equations*, J. Differential Equations, 26 (1979), pp. 404–434.

[TWY] L. F. Tam, W. S. Wong, and S. S.-T. Yau, *On a necessary and sufficient condition for finite dimensionality of estimation algebras*, SIAM J. Control Optim., 28 (1990), pp. 173–185.

[W] E. Wong, *Stochastic Processes in Information and Dynamical Systems*, McGraw-Hill, New York, 1971.

[WeNo] J. Wei and E. Norman, *On global representations of the solutions of linear differential equations as a product of exponentials*, Proc. Amer. Math. Soc., 15 (1964), pp. 327–334.

[Wi] D. V. Widder, *The heat equation*, Mathematics 67, Academic Press, New York, 1975.

[Wo1] W. S. Wong, *New classes of finite dimensional nonlinear filters*, Systems Control Lett., 3 (1983), pp. 155–164.

[Wo2] W. S. WONG, *On a new class of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 79–83.

[Wo3] ———, *Theorems on the structure of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 117–124.

[YaCh] S. S.-T. YAU AND W. L. CHIOU, *Recent results on classification of finite dimensional estimation algebras: Dimension of State Space* ≤ 2, Proc. 30th IEEE Conference on Decision and Control, Brighton, England, 1991.

[Ya] S. S.-T. YAU, *Finite dimensional filters with nonlinear drift* I: *A class of filters including both Kalman-Bucy filters and Benes filters*, J. Math. Systems, Estim. Control, 4 (1994), pp. 181–203.

[YaLe] S. S.-T. YAU AND C.-W. LEUNG, *Recent result on classification of finite dimensional maximal rank estimation algebras with state space dimension* 3, in Proceedings of the 31st Conference on Decision and Control, Tucson, AZ, Dec. 1992, pp. 2247–2250.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 199–225, January 1996

009

DYNAMIC PROGRAMMING FOR NONLINEAR SYSTEMS DRIVEN BY ORDINARY AND IMPULSIVE CONTROLS*

MONICA MOTTA† AND FRANCO RAMPAZZO†

Abstract. A dynamic programming approach is considered for a class of minimum problems with impulses. The minimization domain consists of trajectories satisfying an ordinary differential equation whose right-hand side depends not only on a measurable control v but also on a second control u and on its time derivative $\dot{u}$. For this reason, the control u and the differential equation are called *impulsive*.

The value function of the considered minimum problem turns out to depend on the time, the state, the u variable, and the variation allowed to the impulsive control. It is shown that the value function satisfies, in a generalized sense, a dynamic programming equation (DPE), which is obtained from a dynamic programming principle involving space–time trajectories. Moreover the value function is the unique map-solving equation (DPE) satisfying either an inequality condition or a supersolution condition at each point of the boundary. Incidentally this extends a result by Barron, Jensen, and Menaldi [Nonlinear Anal., 21 (1993), pp. 241–268], where the impulsive control is scalar monotone and the corresponding vector field is independent of the state variable. Next, a maximum principle is proved, and the well-known relationship between adjoint variables and value function is suitably extended to impulsive control systems. A fully elaborated example concludes the paper.

Key words. impulsive control, minimum problem, dynamic programming

AMS subject classifications. 34A37, 49N25, 49L20, 49L25

1. Introduction.

The optimal control problem. This paper concerns the dynamic programming approach to minimum problems involving impulsive control systems of the form

$$\text{(E)} \qquad \begin{aligned} \dot{x} &= g_0(t,x,u,v) + \sum_{i=1}^{m} g_i(t,x,u,v)\dot{u}_i, \\ x(\bar{t}) &= \bar{x}, \end{aligned}$$

where the state x belongs to $\mathbb{R}^n$ and the controls u and v map a time interval $[\bar{t}, T]$ into a closed subset $U \subset \mathbb{R}^m$ and a compact subset $V \subset \mathbb{R}^q$, respectively. Moreover u is subject to the directional constraint $\dot{u} \in C$, where $C \subset \mathbb{R}^m$ is a closed cone. Optimum problems involving a dynamics of the form (E) arise in applications to rational mechanics [13]–[15], [35], economics [17], space navigation [25], [29], [33], and advertising strategy [20], [39].

Because of the presence of the derivative $\dot{u}$ on the right-hand side of (E) the state can jump in consequence of a discontinuity of the control u. However, the notion of solution to (E) is provided by the Carathéodory theory of ordinary differential equations only if the control u is absolutely continuous. Moreover, it is known—see, e.g., [9]–[12], [19], [21], [23], [28], [30], [32], [37], [41]—that whenever the fields $g_1, \ldots, g_m$ depend on x, u, and v, a mere measure-theoretic extension of this notion to the case of a discontinuous u does not agree with elementary requirements of continuity

* Received by the editors September 2, 1993; accepted for publication (in revised form) August 22, 1994.

† Dipartimento di Matematica Pura e Applicata, Università di Padova, via Belzoni 7, I-35131 Padova, Italy (motta@pdmath1.math.unipd.it), (rampazzo@pdmath1.math.unipd.it).

of the input–output map. In order to overcome this difficulty, in [10], [32], [34] one extends system (E) to the space–time system

$$x' = g_0(t,x,u,v)t' + \sum_{i=1}^{m} g_i(t,x,u,v)u_i', \tag{STE}$$

where the *controls* $t(s), u(s)$ are Lipschitz continuous and the superscript denotes differentiation with respect to the pseudo–time parameter $s \in [0,1]$. In this space–time setting a discontinuous control $u(t)$ is regarded as the space projection of a space–time control $t(s), u(s)$ whose first component $t(s)$ is allowed to be nondecreasing. We just recall—see, e.g., [10], [30], [32], [37]—that, because of the noncommutativity of the vector fields $g_1, \dots, g_m$, the evolution of x depends on the particular space–time control $t(s), u(s)$ which *completes* the graph of $u(t)$. Incidentally we remark that in the standard impulse control theory there is no need of considering space–time controls. Indeed, in that case the fields $g_1, \dots, g_m$ $(m = n)$ coincide with the canonical basis; in particular they commute. As a consequence each completion of a control u produces the same trajectory which in turn coincides with the unique trajectory resulting from the measure–theoretic approach; see, e.g., [3].

As a prototype of a minimum problem initially formulated for the original system (E) we consider an unconstrained Mayer problem with finite horizon and a bound on the total variation of u.

More precisely, let $\Phi : \mathbb{R}^n \times U \to \mathbb{R}$ be a continuous map, C be a closed cone of $\mathbb{R}^m$, and $K > 0$ be an upper bound for the total variation of the control u. For every $(\bar{t}, \bar{x}, \bar{u}, \bar{k}) \in [0,T) \times \mathbb{R}^n \times U \times [0,K]$, we consider the following problem:

$$(\mathcal{P}_{(\bar{t},\bar{x},\bar{u},\bar{k})}) \qquad \text{minimize } \{\Phi(x(T), u(T))\}$$

over all end points $(x(T), u(T))$ of (E) corresponding to control policies $(u(\cdot), v(\cdot))$, where $v : [\bar{t}, T] \to V$ is a Borel-measurable map and $u : [\bar{t}, T] \to U$ is an absolutely continuous map which satisfies

$$u(\bar{t}) = \bar{u}, \quad V_{\bar{t}}^{T}(u) \le K - \bar{k}, \quad \text{and} \quad \dot{u}(t) \in C \text{ for a.e. } t \in [\bar{t}, T].$$

($V_{\bar{t}}^{T}(u)$ denotes the total variation of $u(\cdot)$ on the interval $[\bar{t}, T]$.) Since the unbounded control $\dot{u}$ appears linearly on the right-hand side of (E), problem $\mathcal{P}_{(\bar{t},\bar{x},\bar{u},\bar{k})}$ does not display anyone of the standard coercivity assumptions which guarantee the existence of an optimal control. This justifies the introduction of the extended system (STE) and of the corresponding space–time reformulation of problem $\mathcal{P}_{(\bar{t},\bar{x},\bar{u},\bar{k})}$. Actually this extension is *proper*, i.e., the infimum of the original problem turns out to coincide with the infimum of the extended problem. Hence the value functions determined by the two problems coincide. Moreover the set of original controls is dense in the set of space–time controls, and under some further assumptions, there exists an optimal control $(t(s), u(s), v(s))$ for the extended problem; see [32].

The dynamic programming approach. We call *value function* the map $\mathcal{V} : [0,T) \times \mathbb{R}^n \times U \times [0,K] \to \mathbb{R}$ which associates the infimum of problem $\mathcal{P}_{(\bar{t},\bar{x},\bar{u},\bar{k})}$ to every $(\bar{t}, \bar{x}, \bar{u}, \bar{k})$. Actually $\mathcal{V}$ can be identified with the value function corresponding to the extended problem, for the two maps turn out to coincide on $[0,T) \times \mathbb{R}^n \times U \times [0,K]$. Moreover, in the extended setting $\mathcal{V}$ can be defined also at $\bar{t} = T$.

In the particular case when the control u is a scalar nondecreasing map (i.e., $m = 1$, $\dot{u} \in C \doteq \mathbb{R}^+$, and $u \equiv k \in [0,K]$) and the vector field g_1 *does not depend on* x *and* u, the dynamic programming approach has been already pursued by E. N. Barron,

R. Jensen, and J. L. Menaldi [8]. Their main result consisted in proving that the value function $\mathcal{V}$ is the unique continuous map which satisfies (in the viscosity sense) a certain Hamilton–Jacobi–Bellman equation together with the following, quite natural, Dirichlet conditions:

$(BC)_1$ $\mathcal{V}$ coincides with the value function of the corresponding nonimpulsive problem ($\dot{u} = 0$) on the strip $[0,T] \times \mathbb{R}^n \times \{K\}$ (where all the available variation of u has run out);

$(BC)_2$ $\mathcal{V}$ coincides with the value function of the corresponding purely impulsive problem ($g_0 = 0$) on the region $\{T\} \times \mathbb{R}^n \times [0,K]$(where no more time is available).

Moreover, Barron, Jensen, and Menaldi left the following questions as open problems:

a) *Can the well-known relationship between the adjoint variables of the maximum principle and the value function be extended in some way to impulsive problems?*

b) *Can we state a rigorous result (i.e., a verification theorem) which relates the dynamic programming equation with the problem of testing the optimality of a given control?*

c) *What can be said when g_1 depends also on x and u?*

This paper is also a trial to give an answer to the above questions, not only in the scalar control case but also in the general situation where u is vector valued. More precisely, we begin by proving that, under suitable assumptions on the set U and the cone C, the value function $\mathcal{V}$ is continuous on $[0,T] \times \mathbb{R}^n \times U \times [0,K]$. Next, via a dynamic programming principle involving space–time trajectories, we prove that the value function $\mathcal{V}$ is a viscosity solution on $[0,T[\times\mathbb{R}^n \times \overset{\circ}{U} \times [0,K[$ of the dynamic programming equation

$$\text{(DPE)} \qquad -H\left(t,x,u,\frac{\partial \mathcal{V}}{\partial t},\frac{\partial \mathcal{V}}{\partial x},\frac{\partial \mathcal{V}}{\partial u},\frac{\partial \mathcal{V}}{\partial k}\right) = 0,$$

where, for every $(p_t,p_x,p_u,p_k) \in \mathbb{R}^{1+n+m+1}$, the Hamiltonian function H is defined by

$$H(t,x,u,p_t,p_x,p_u,p_k) \doteq \min\left\{ \begin{array}{c} (p_t + p_x \cdot g_0(t,x,u,v))w_0 + \sum_{i=1}^{m}(p_x \cdot g_i(t,x,u,k) + p_{u_i})w_i + p_k|w|, \\ |(w_0,\dots,w_m)| = 1,\ w_0 \geq 0,\ w = (w_1,\dots,w_m) \in C,\ v \in V \end{array} \right\}.$$

Furthermore, $\mathcal{V}$ turns out to be the unique solution of (DPE) satisfying the following boundary conditions:

$(BC)'_1$ $\mathcal{V}$ is a (viscosity) *supersolution* of (DPE) at all points of $[0,T[\times\mathbb{R}^n \times \partial U \times [0,K[\cup [0,T[\times\mathbb{R}^n \times U \times \{K\}$;

$(BC)'_2$ at each boundary point $(T,x,u,k)\mathcal{V} \leq \Phi$ either $\mathcal{V}$ is a supersolution of (DPE) or it satisfies the relation $\mathcal{V}(T,x,u,k) = \Phi(x,u)$.

We remark that, unlike conditions $(BC)_1$ and $(BC)_2$ above, boundary conditions $(BC)'_1$ and $(BC)'_2$ do not involve any auxiliary minimum problem and refer only to the cost function Φ and to equation (DPE).

We also prove a verification theorem (Theorem 5.1), which incidentally provides a possible answer to the open question b) mentioned above.

Finally, by applying standard results to the space–time embedding, we are able to clarify the relationship occurring between the adjoint variables of the maximum principle and the value function $\mathcal{V}$. This provides a possible answer to the open

question a) above, while the answer to question c) is inherent to the general setting of the problem, for the vector fields $g_1, \dots, g_m$ *do depend* on x and u.

The paper ends with a simple, elaborated example where the theoretical results proved throughout the paper are explicitly applied to test the optimality of a feedback control previously computed by means of the maximum principle.

2. The minimum problem. Let us consider the control system

$$\dot{x} = g_0(t,x,u,v) + \sum_{i=1}^{m} g_i(t,x,u,v)\dot{u}_i(t), \tag{2.1}$$

$$x(\bar{t}) = \bar{x}, \qquad u(\bar{t}) = \bar{u} \tag{2.2}$$

defined on a time interval $[\bar{t}, T]$, where the state x ranges in $\mathbb{R}^n$ while the controls u and v take values on a closed arcwise connected subset $U \subset \mathbb{R}^m$ and a compact subset $V \subset \mathbb{R}^q$, respectively. Moreover the control u is subject to the directional constraint $\dot{u} \in C$, where $C \subset \mathbb{R}^m$ denotes a given closed cone.

Let K be a positive constant, and for every $\bar{k} \in [0, K]$ let us define the set

$$W_{K-\bar{k}}(\bar{t},\bar{u}) \doteq \left\{ \begin{array}{l} (u,v) \in AC([\bar{t},T],U) \times \mathcal{B}([\bar{t},T],V) \ : \ u(\bar{t}) = \bar{u}, \\ \dot{u}(t) \in C \text{ for a.e. } t \in [\bar{t},T] \text{ and } V_{\bar{t}}^{T}(u) \le K - \bar{k} \end{array} \right\}, \tag{2.3}$$

where $AC\big([\bar{t},T],U\big)$ denotes the set of absolutely continuous functions from $[\bar{t},T]$ into U, $\mathcal{B}\big([\bar{t},T],V\big)$ is the set of Borel-measurable functions from $[\bar{t},T]$ into V, and $V_{\bar{t}}^{T}(u)$ denotes the total variation of u on the interval $[\bar{t},T]$. We call $W_{K-\bar{k}}(\bar{t},\bar{u})$ the set of *admissible regular controls from* $(\bar{t},\bar{u})$ *such that the variation of* u *is less than or equal to* $K - \bar{k}$.

Let Φ be a continuous function defined on $\mathbb{R}^n \times U$. For any $(\bar{t},\bar{x},\bar{u},\bar{k}) \in [0,T[\times\mathbb{R}^n \times U \times [0,K]$ we consider the following minimum problem of Mayer type:

$$(\mathcal{P}_{(\bar{t},\bar{x},\bar{u},\bar{k})}) \qquad \underset{(u,v)\in W_{K-\bar{k}}(\bar{t},\bar{u})}{\text{minimize}} \Phi\big(x[\bar{t},\bar{x},\bar{u};u,v](T),u(T)\big),$$

where $x[\bar{t},\bar{x},\bar{u};u,v](\cdot)$ denotes the solution of (2.1), (2.2) corresponding to the control (u,v).

Throughout this paper we assume the following hypothesis **(H1)** on the vector fields $g_0, \dots, g_m$ and the function Φ:

(H1) $g_0, \dots, g_m$ and Φ are continuous in all of its variables, and there is a positive constant M such that

$$|g_i(t,x,u,v)| \le M\big(1 + |(x,u)|\big), \quad |\Phi(x,u)| \le M$$
$$\forall (t,x,u,v) \in [0,T] \times \mathbb{R}^n \times U \times V \quad (i = 0, \dots, m).$$

Moreover, for any compact subset $Q \subset \mathbb{R}^n \times U$ there is a constant L such that

$$|g_i(t,x,u,v) - g_i(t,\bar{x},u,v)| \le L|x - \bar{x}|$$
$$\forall (t,x,u,v),(t,\bar{x},u,v) \in [0,T] \times Q \times V \quad (i = 0, \dots, m).$$

In the following discussion, whenever the compact set Q is specified, we will denote by $\omega_{g_0}, \dots, \omega_{g_m}$ and ω_Φ the modulus of uniform continuity of the restrictions of the functions $g_0, \dots, g_m$ and Φ to $[0,T] \times Q \times V$ and Q, respectively.

Remark 2.1. The condition $|\Phi(x,u)| \le M$ implies that the value function is globally bounded, which turns out to be very convenient for applying the theory of viscosity

solutions. On the other hand one can skip such a limitation by replacing the cost function Φ with the bounded cost function $\arctan \Phi$. It is obvious that this transformation will not affect the essential character of the problem.

Since the right-hand side of (2.1) depends linearly on the derivative $\dot{u}$, in general no optimal controls can be found within the class $W_{K-\bar{k}}(\bar{t}, \bar{u})$. Hence, denoting the triple (t, x, u) by y, on the basis of the results in [10], [32], we embed (2.1) into the space–time system

$$y' = \hat{g}_0(y, v)t'(s) + \sum_{i=1}^{m} \hat{g}_i(y, v)u_i'(s) \tag{2.4}$$

together with the initial condition·

$$y(0) = (\bar{t}, \bar{x}, \bar{u}). \tag{2.5}$$

In (2.4) the superscript denotes differentiation with respect to the new parameter $s \in [0, 1]$ and for every $i = 0, \ldots, m$ the vector field $\hat{g}_i$ coincides with the ith column of the $(1 + n + m) \times (1 + m)$ matrix

$$\hat{G}(y, v) \doteq \begin{pmatrix} 1 & 0 & \ldots & 0 \\ g_{0_1} & g_{1_1} & \cdots & g_{m_1} \\ \vdots & \vdots & \ddots & \vdots \\ g_{0_n} & g_{1_n} & \cdots & g_{m_n} \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}.$$

DEFINITION 2.1. *The control system* (2.4) *is called the* space–time control system *relative to* (2.1), *and a map*

$$(t, u, v) : [0, 1] \to [\bar{t}, T] \times U \times V$$

is called a space–time control *for* (2.4), (2.5) *whenever the following hold:*

(i) $(t, u)(0) = (\bar{t}, \bar{u})$;

(ii) $(t, u) : [0, 1] \to [\bar{t}, T] \times U$ *is Lipschitz continuous and* $u'(s) \in C$ *for almost every* $s \in [0, 1]$;

(iii) $t : [0, 1] \to [\bar{t}, T]$ *is surjective and nondecreasing*;

(iv) $v : [0, 1] \to V$ *is Borel measurable.*

The set of space–time controls will be denoted by $\Gamma(\bar{t}, \bar{u})$. *A solution of the space–time control system* (2.4) *will be called a* space–time trajectory.

We remark again—see the introduction—that a mere interpretation of the original system (2.1) as an equation in measure would lead to an ill-posed problem, for the dependence of $g_1, \ldots, g_m$ on x and u makes it impossible to define a concept of (univalued) trajectory as a map of the original parameter t.

We refer to the appendix for some basic facts concerning the concept of *canonical parametrization* and the related topology on the set of space–time controls. Briefly, the parametrization of a space–time control (t, u, v) is called canonical if the norm $|(t', u')|$ is constant almost everywhere in $[0, 1]$. Any space–time control can be reparametrized in such a way that the resulting space–time control turns out to be canonical. And, up to reparametrization, the corresponding trajectories coincide (see Proposition A.2).

Observe that after introducing new equations we regard t and u both as state variables and as control variables. This allows us to embed problem $\mathcal{P}_{(\bar{t}, \bar{x}, \bar{u}, \bar{k})}$ into the

extended problem

$$(\mathcal{P}^e_{(\bar t,\bar x,\bar u,\bar k)}) \qquad \underset{(t,u,v)\in\Gamma_{K-\bar k}(\bar t,\bar u)}{\text{minimize}} \Phi\big(y[\bar t,\bar x,\bar u;t,u,v](1)\big),$$

where

$$\Gamma_{K-\bar k}(\bar t,\bar u) \doteq \big\{(t,u,v)\in\Gamma(\bar t,\bar u)\ :\ V_0^1(u)\le K-\bar k\big\} \tag{2.6}$$

is the set of admissible space–time controls and $y[\bar t,\bar x,\bar u;t,u,v](\cdot)$ denotes the solution of (2.4), (2.5) corresponding to the space–time control (t,u,v).

Remark 2.2. We point out that Φ is a function of the only variables (x,u). With abuse of notation we write $\Phi(y)$, where $y=(t,x,u)$, instead of $\Phi(x,u)$, just to remind the reader that we are now referring to the space–time extension (2.4).

It is clear that in the space–time setting the original set $W_{K-\bar k}(\bar t,\bar u)$ of admissible controls has to be identified with the subset $\Gamma^+_{K-\bar k}(\bar t,\bar u)\subset\Gamma_{K-\bar k}(\bar t,\bar u)$ formed by the Lipschitz continuous reparametrizations of the graphs of the elements belonging to $W_{K-\bar k}(\bar t,\bar u)$.

The subset $\Gamma^+_{K-\bar k}(\bar t,\bar u)$ turns out to be dense—see [32] and the Appendix—in the set $\Gamma_{K-\bar k}(\bar t,\bar u)$ of space–time controls.

We now prove that the infimum of the extended problem $\mathcal{P}^e_{(\bar t,\bar x,\bar u,\bar k)}$ coincides with the infimum of the original problem $\mathcal{P}_{(\bar t,\bar x,\bar u,\bar k)}$.

THEOREM 2.1. *For every initial condition* $(\bar t,\bar x,\bar u,\bar k)\in[0,T)\times\mathbb{R}^n\times U\times[0,K]$ *one has*

$$\inf_{(t,u,v)\in\Gamma_{K-\bar k}(\bar t,\bar u)}\Phi\big(y[\bar t,\bar x,\bar u;t,u,v](1)\big)=\inf_{(u,v)\in W_{K-\bar k}(\bar t,\bar u)}\Phi\big(x[\bar t,\bar x,\bar u;u,v](T),u(T)\big). \tag{2.7}$$

Proof. Let $(\bar t,\bar x,\bar u,\bar k)$ be a fixed initial datum and let us observe that Gronwall's lemma, together with the bound on the total variation of u, guarantees that there is some positive constant M' such that

$$\begin{aligned} &|y[\bar t,\bar x,\bar u;t,u,v](s)|\le M',\\ &|g_i\big(y[\bar t,\bar x,\bar u;t,u,v](s),v(s)\big)|\le M'\quad (i=0,\dots,m)\quad\text{for a.e. } s\in[0,1]\end{aligned} \tag{2.8}$$

for all $(t,u,v)\in\Gamma_{K-\bar k}(\bar t,\bar u)$. Hence, setting $Q\doteq B_{n+m}[0,M']\cap\mathbb{R}^n\times U$ (where $B_{n+m}[0,M']$ denotes the closed ball of center 0 and radius M' in $\mathbb{R}^{n+m}$), we can identify the vector fields $g_0,\dots,g_m$ and the function Φ with their restrictions to the compact sets $[0,T]\times Q\times V$ and Q, respectively.

By the definition of $\Gamma^+_{K-\bar k}(\bar t,\bar u)$, proving (2.7) is equivalent to checking that the identity

$$\inf_{(t,u,v)\in\Gamma^+_{K-\bar k}(\bar t,\bar u)}\Phi\big(y[\bar t,\bar x,\bar u;t,u,v](1)\big)=\inf_{(t,u,v)\in\Gamma_{K-\bar k}(\bar t,\bar u)}\Phi\big(y[\bar t,\bar x,\bar u;t,u,v](1)\big)$$

holds true. Hence it suffices to show that

$$\inf_{(t,u,v)\in\Gamma^+_{K-\bar k}(\bar t,\bar u)}\Phi\big(y[\bar t,\bar x,\bar u;t,u,v](1)\big)\le\inf_{(t,u,v)\in\Gamma_{K-\bar k}(\bar t,\bar u)}\Phi\big(y[\bar t,\bar x,\bar u;t,u,v](1)\big). \tag{2.9}$$

Since these infima are bounded, for any $\varepsilon>0$ there is a space–time control $(t,u,v)\in\Gamma_{K-\bar k}(\bar t,\bar u)$ verifying

$$\inf_{(\tilde t,\tilde u,\tilde v)\in\Gamma_{K-\bar k}(\bar t,\bar u)}\Phi\big(y[\bar t,\bar x,\bar u;\tilde t,\tilde u,\tilde v](1)\big)\ge\Phi\big(y[\bar t,\bar x,\bar u;t,u,v](1)\big)-\varepsilon/2. \tag{2.10}$$

Note that on the basis of Proposition A.2 it is not restrictive to assume that the norm $|(t', u')|$ is constant almost everywhere in $[0, 1]$, this constant being less than $K + T$. Then, by setting

$$t_\varepsilon(s) \doteq \bar{t} + (T - \bar{t})\frac{(t(s) - \bar{t}) + s\rho_\varepsilon}{(T - \bar{t} + \rho_\varepsilon)} \qquad \forall s \in [0, 1]$$

for a $\rho_\varepsilon \in (0, (T - \bar{t})/2]$ to be chosen, we obtain a space–time control $(t_\varepsilon, u, v) \in \Gamma^+_{K-\bar{k}}(\bar{t}, \bar{u})$ such that

$$|t_\varepsilon(s) - t(s)| \leq 2\rho_\varepsilon \qquad \forall s \in [0, 1],$$

and the corresponding trajectory $x_\varepsilon \doteq x[\bar{t}, \bar{x}, \bar{u}; t_\varepsilon, u, v]$ satisfies

$$\begin{aligned}
|x_\varepsilon(s) - x(s)| &\leq \int_0^s \big|g_0\big(t_\varepsilon(\sigma), x_\varepsilon(\sigma), u(\sigma), v(\sigma)\big) - g_0\big(t(\sigma), x(\sigma), u(\sigma), v(\sigma)\big)\big| t'(\sigma)\, d\sigma \\
&\quad + \int_0^s \sum_{i=1}^m \big|g_i\big(t_\varepsilon(\sigma), x_\varepsilon(\sigma), u(\sigma), v(\sigma)\big) - g_i\big(t(\sigma), x(\sigma), u(\sigma), v(\sigma)\big)\big|\, |u_i'(\sigma)|\, d\sigma \\
&\qquad + \frac{\rho_\varepsilon}{T - \bar{t} + \rho_\varepsilon} \int_0^s \big|g_0\big(t_\varepsilon(\sigma), x_\varepsilon(\sigma), u(\sigma), v(\sigma)\big)\big| [(T - \bar{t}) + t'(\sigma)]\, d\sigma \\
&\leq (K+T) \sum_{i=0}^m \omega_{g_i}(2\rho_\varepsilon) + (m+1)(K+T)L \int_0^s |x_\varepsilon(\sigma) - x(\sigma)|\, d\sigma + 2\frac{T - \bar{t}}{T - \bar{t} + \rho_\varepsilon} M' \rho_\varepsilon.
\end{aligned}$$

By Gronwall's lemma it follows that

$$\big|\Phi\big(x_\varepsilon(1), u(1)\big) - \Phi\big(x(1), u(1)\big)\big| \leq \omega_\Phi\left((2M'\rho_\varepsilon + (K+T)\sum_{i=0}^m \omega_{g_i}(2\rho_\varepsilon))e^{(m+1)(K+T)L}\right).$$

Hence for a ρ_ε small enough from (2.10) we have

$$\inf_{(\tilde{t}, \tilde{u}, \tilde{v}) \in \Gamma_{K-\bar{k}}(\bar{t}, \bar{u})} \Phi\big(y[\bar{t}, \bar{x}, \bar{u}; \tilde{t}, \tilde{u}, \tilde{v}](1)\big) \geq \Phi\big(x_\varepsilon(1), u(1)\big) - \varepsilon,$$

which by the arbitrariness of $\varepsilon > 0$ yields (2.9). □

3. The value function. In this section we introduce the so-called value function for the problem $\mathcal{P}_{(\bar{t}, \bar{x}, \bar{u}, \bar{k})}$ and study its regularity properties.

DEFINITION 3.1. *The map*

$$\mathcal{F}(\bar{t}, \bar{x}, \bar{u}, \bar{k}) \doteq \inf_{(u,v) \in W_{K-\bar{k}}(\bar{t}, \bar{u})} \Phi\big(x[\bar{t}, \bar{x}, \bar{u}; u, v](T), u(T)\big) \tag{3.1}$$

from $[0, T) \times \mathbb{R}^n \times U \times [0, K]$ *into* $\mathbb{R}$ *is called the* value function of the original minimum problem.

DEFINITION 3.2. *The map*

$$\mathcal{V}(\bar{t}, \bar{x}, \bar{u}, \bar{k}) \doteq \inf_{(t,u,v) \in \Gamma_{K-\bar{k}}(\bar{t}, \bar{u})} \Phi\big(y[\bar{t}, \bar{x}, \bar{u}; t, u, v](1)\big) \tag{3.2}$$

from $[0, T] \times \mathbb{R}^n \times U \times [0, K]$ *into* $\mathbb{R}$ *is called the* value function of the extended minimum problem.

The following result follows from Theorem 2.1.

COROLLARY 3.1. *The value function $\mathcal{F}$ of the original minimum problem is bounded and coincides with the value function $\mathcal{V}$ of the extended minimum problem.*

Let us observe that the value function $\mathcal{V}$ of the extended problem is defined even at time $\bar{t} = T$. Furthermore, in Theorem 3.1 below we show that $\mathcal{V}$ is continuous

provided that one of the following two hypotheses on the cone C and the closed set U holds:

(H2)$_C$ The set U coincides with the whole $\mathbb{R}^m$.

(H2)$_U$ the cone C coincides with the whole $\mathbb{R}^m$; moreover for any $\varepsilon > 0$ and $u_1 \in U$ there exists a $\delta > 0$ such that for each $u_2 \in U \cap B(u_1, \delta)$, there is a path $\gamma_{12} \in AC([0,1], U)$ satisfying $\gamma_{12}(0) = u_1$, $\gamma_{12}(1) = u_2$, and

$$\int_0^1 |\gamma'_{12}(s)|\, ds \leq \varepsilon.$$

THEOREM 3.1. *Let $Q_x \subset \mathbb{R}^n$, $Q_u \subset U$ be compact subsets. Then for every $(\bar{t}, \bar{x}, \bar{u}, \bar{k}) \in [0,T] \times Q_x \times Q_u \times [0,K]$ one has the following:*

i) *the functions $x \mapsto \mathcal{V}(\bar{t}, x, \bar{u}, \bar{k})$, $t \mapsto \mathcal{V}(t, \bar{x}, \bar{u}, \bar{k})$, and $k \mapsto \mathcal{V}(\bar{t}, \bar{x}, \bar{u}, k)$ are continuous on Q_x, $[0,T]$, and $[0,K]$, respectively, uniformly with respect to the remaining variables on $[0,T] \times Q_x \times Q_u \times [0,K]$; furthermore, $k \mapsto \mathcal{V}(\bar{t}, \bar{x}, \bar{u}, k)$ is nondecreasing*;

ii) *in addition, if either hypothesis* **(H2)**$_C$ *or hypothesis* **(H2)**$_U$ *is assumed, then the function $u \mapsto \mathcal{V}(\bar{t}, \bar{x}, u, \bar{k})$ is continuous on Q_u, uniformly with respect to the remaining variables on $[0,T] \times Q_x \times Q_u \times [0,K]$. In particular the value function $\mathcal{V}$ is continuous on its domain.*

Proof. By (2.8) the trajectories starting from points of $[0,T] \times Q_x \times Q_u$ lie in the compact set $[0,T] \times B_{n+m}[Q_x \times Q_u; M'] \cap (\mathbb{R}^n \times U)$. Let ω_{g_i} denote the modulus of uniform continuity of g_i $(i = 0, \ldots, m)$ on $[0,T] \times B_{n+m}[Q_x \times Q_u; M'] \cap (\mathbb{R}^n \times U) \times V$, and let ω_Φ be the modulus of continuity of Φ on $B_{n+m}[Q_x \times Q_u; M'] \cap (\mathbb{R}^n \times U)$.

Let $x_1, x_2 \in Q_x$, and consider the difference

$$\mathcal{V}(\bar{t}, x_2, \bar{u}, \bar{k}) - \mathcal{V}(\bar{t}, x_1, \bar{u}, \bar{k}),$$

which can be assumed nonnegative. For any $\varepsilon > 0$ let $(t, u, v) \in \Gamma_{K-\bar{k}}(\bar{t}, \bar{u})$ be a space–time control satisfying

$$\mathcal{V}(\bar{t}, x_1, \bar{u}, \bar{k}) \geq \Phi\big(y[\bar{t}, x_1, \bar{u}; t, u, v](1)\big) - \varepsilon. \tag{3.3}$$

Thus by the definition of $\mathcal{V}$ we have

$$\mathcal{V}(\bar{t}, x_2, \bar{u}, \bar{k}) - \mathcal{V}(\bar{t}, x_1, \bar{u}, \bar{k}) \leq \Phi\big(y[\bar{t}, x_2, \bar{u}; t, u, v](1)\big) - \Phi\big(y[\bar{t}, x_1, \bar{u}; t, u, v](1)\big) + \varepsilon. \tag{3.4}$$

Furthermore standard estimates for the trajectories of (2.4) yield

$$\big|x[\bar{t}, x_2, \bar{u}; t, u, v](s) - x[\bar{t}, x_1, \bar{u}; t, u, v](s)\big| \leq |x_2 - x_1| e^{L(1+m)(K+T)s} \tag{3.5}$$

for all $s \in [0,1]$. Hence (3.4) and (3.5) imply

$$\mathcal{V}(\bar{t}, x_2, \bar{u}, \bar{k}) - \mathcal{V}(\bar{t}, x_1, \bar{u}, \bar{k}) \leq \omega_\Phi\big(e^{L(1+m)(K+T)} |x_2 - x_1|\big) + \varepsilon,$$

which, by the arbitrariness of $\varepsilon > 0$, proves that $x \mapsto \mathcal{V}(\bar{t}, x, \bar{u}, \bar{k})$ is continuous uniformly with respect to $(\bar{t}, \bar{u}, \bar{k})$.

Now let $t_1, t_2 \in [0,T]$, $t_1 \neq t_2$, and consider the difference

$$\mathcal{V}(t_2, \bar{x}, \bar{u}, \bar{k}) - \mathcal{V}(t_1, \bar{x}, \bar{u}, \bar{k}),$$

which can be assumed nonnegative. For any $\varepsilon > 0$, let (t, u, v) be a space–time control for $(t_1, \bar{x}, \bar{u}, \bar{k})$ satisfying

$$\mathcal{V}(t_1, \bar{x}, \bar{u}, \bar{k}) \geq \Phi\big(y[t_1, \bar{x}, \bar{u}; t, u, v](1)\big) - \varepsilon,$$

and consider the space–time control $(\tilde{t}, u, v) \in \Gamma_{K-\bar{k}}(t_2, \bar{u})$, where $\tilde{t}$ is defined as follows:
if $t_1 < t_2$, set

$$\tilde{t}(s) = \begin{cases} t_2, & s \in [0, \bar{s}], \\ t(s), & s \in [\bar{s}, 1], \end{cases}$$

where

$$\bar{s} \doteq \min\{s \in [0,1] \ : \ t(s) = t_2\};$$

if $t_1 \geq t_2$, set

$$\tilde{t}(s) = t(s) - (t_1 - t_2)(1 - s), \qquad s \in [0,1].$$

In both cases the definition of $\tilde{t}$ and Gronwall's lemma imply
(3.6)

$$|\tilde{t}(s) - t(s)| \leq |t_2 - t_1|,$$

$$\big|\tilde{x}(s) - x(s)\big| \leq \big[M'|t_2 - t_1| + (K+T)\sum_{i=0}^{m} \omega_{g_i}(|t_2 - t_1|)\big] e^{L(1+m)(K+T)} \quad \forall s \in [0,1],$$

where we have set $\tilde{x}(\cdot) \doteq x[t_2, \bar{x}, \bar{u}; \tilde{t}, u, v](\cdot)$, $x(\cdot) \doteq x[t_1, \bar{x}, \bar{u}; t, u, v](\cdot)$. It follows that

$$\begin{aligned} &\mathcal{V}(t_2, \bar{x}, \bar{u}, \bar{k}) - \mathcal{V}(t_1, \bar{x}, \bar{u}, \bar{k}) \\ &\quad \leq \omega_\Phi\big(\big[M'|t_2 - t_1| + (K+T)\sum_{i=0}^{m} \omega_{g_i}(|t_2 - t_1|)\big] e^{L(1+m)(K+T)}\big) + \varepsilon, \end{aligned}$$

which, by the arbitrariness of ε, implies that $t \mapsto \mathcal{V}(t, \bar{x}, \bar{u}, \bar{k})$ is continuous uniformly with respect to the remaining variables $(\bar{x}, \bar{u}, \bar{k})$.

Now let $k_1, k_2 \in [0, K]$, with $k_1 \neq k_2$. Since the map $k \mapsto \mathcal{V}(\bar{t}, \bar{x}, \bar{u}, k)$ is nondecreasing, it is not restrictive to consider only the case $k_2 > k_1$. Choose a space–time control for $(\bar{t}, \bar{x}, \bar{u}, k_1)$ satisfying

$$\mathcal{V}(\bar{t}, \bar{x}, \bar{u}, k_1) \geq \Phi\big(y[\bar{t}, \bar{x}, \bar{u}; t, u, v](1)\big) - \varepsilon,$$

and set

$$\tilde{u}(s) = \begin{cases} u(s), & s \in [0, \bar{s}], \\ u(\bar{s}), & s \in [\bar{s}, 1], \end{cases}$$

where

$$\bar{s} \doteq \max\{s \in [0,1] \ : \ V_0^s(u) \leq K - k_2\}.$$

Observe that either $K - k_2 < V_0^1(u) \leq K - k_1$ and $V_0^{\bar{s}}(u) = K - k_2$ with $\bar{s} < 1$, or $\bar{s} = 1$; furthermore, $V_0^1(\tilde{u}) = V_0^{\bar{s}}(u) \leq K - k_2$ so that $(t, \tilde{u}, v) \in \Gamma_{K-k_2}(\bar{t}, \bar{u})$. For every $s \in [\bar{s}, 1)$ one has

$$V_{\bar{s}}^s(u) \leq V_{\bar{s}}^1(u) = V_0^1(u) - V_0^{\bar{s}}(u) = V_0^1(u) - K + k_2 \leq k_2 - k_1.$$

Hence, from the definition of $\tilde{u}$ and applying Gronwall's lemma one obtains

$$\begin{aligned} |\tilde{u}(s) - u(s)| &\leq V_{\bar{s}}^1(u) \leq k_2 - k_1, \\ \big|x[\bar{t}, \bar{x}, \bar{u}; t, \tilde{u}, v](s) - x[\bar{t}, \bar{x}, \bar{u}; t, u, v](s)\big| &\leq \big[mM' V_{\bar{s}}^s(u) + \omega_{g_0}(T|k_2 - k_1|)\big] e^{L(K+T)} \\ &\leq \big[mM'|k_2 - k_1| + T\omega_{g_0}(|k_2 - k_1|)\big] e^{L(K+T)} \quad \forall s \in [0,1]. \end{aligned}$$

This implies

(3.7)

$$\mathcal{V}(\bar t,\bar x,\bar u,k_2)-\mathcal{V}(\bar t,\bar x,\bar u,k_1)\le \omega_\Phi\big(\big[(1+mM')|k_2-k_1|+T\omega_{g_0}(|k_2-k_1|)\big]e^{L(K+T)}\big)+\varepsilon;$$

thence $k\mapsto \mathcal{V}(\bar t,\bar x,\bar u,k)$ is continuous uniformly with respect to the variables $(\bar t,\bar x,\bar u)$. Thus thesis i) of the theorem is proved.

In order to prove ii), let $\varepsilon>0$ and, for a $\delta>0$ to be determined later, let $u_1,u_2\in Q_u$ satisfy $|u_i-\bar u|\le\delta$, $i=1,2$. Let us consider the difference

$$\mathcal{V}(\bar t,\bar x,u_2,\bar k)-\mathcal{V}(\bar t,\bar x,u_1,\bar k),$$

which it is not restrictive to assume is nonnegative. Let $(t,u,v)\in\Gamma_{K-\bar k}(\bar t,u_1)$ be a control satisfying

$$\mathcal{V}(\bar t,\bar x,u_1,\bar k)\ge\Phi\big(y[\bar t,\bar x,u_1;t,u,v](1)\big)-\varepsilon/2.$$

If $\mathbf{(H2)}_C$ is assumed, then the control $(t,\tilde u,v)\doteq(t,u_2-u_1+u(s),v)$ is in $\Gamma_{K-\bar k}(\bar t,u_2)$. The definitions of $\omega_{g_0},\dots,\omega_{g_m}$ and ω_Φ together with standard estimates for the trajectories of (2.4) imply

$$\begin{aligned}\mathcal{V}(\bar t,\bar x,u_2,\bar k)&-\mathcal{V}(\bar t,\bar x,u_1,\bar k)\\&\le\Phi\big(y[\bar t,\bar x,u_2;t,\tilde u,v](1)\big)-\Phi\big(y[\bar t,\bar x,u_1;t,u,v](1)\big)+\varepsilon/2\\&\le\omega_\Phi\big(|u_2-u_1|+(K+T)\sum_{i=0}^m\omega_{g_i}(|u_2-u_1|)e^{L(1+m)(K+T)}\big)+\varepsilon/2.\end{aligned}$$

This yields the continuity of the map $u\mapsto\mathcal{V}(\bar t,\bar x,u,\bar k)$ uniformly with respect to the remaining variables.

We conclude by proving ii) under hypothesis $\mathbf{(H2)}_U$. Let $\rho_\varepsilon\in(0,1)$. If $K-\bar k\le\rho_\varepsilon$, by setting $\tilde u(s)=u_2\ \ \forall s\in[0,1]$ we obtain

(3.8)
$$\begin{aligned}\mathcal{V}(\bar t,\bar x,u_2,\bar k)&-\mathcal{V}(\bar t,\bar x,u_1,\bar k)\\&\le\Phi\big(y[\bar t,\bar x,u_2;t,u_2,v](1)\big)-\Phi\big(y[\bar t,\bar x,u_1;t,u,v](1)\big)+\varepsilon/2\\&\le\omega_\Phi\big(|u_2-u_1|+\rho_\varepsilon+(T\omega_{g_0}(|u_2-u_1|+\rho_\varepsilon)+mM'\rho_\varepsilon)e^{L(1+m)(K+T)}\big)+\varepsilon/2.\end{aligned}$$

Suppose on the contrary that $K-\bar k>\rho_\varepsilon$. Then by $\mathbf{(H2)}_U$ there exists a $\bar\delta>0$ such that if $|u_1-\bar u|<\bar\delta$, $|u_2-\bar u|<\bar\delta$ one has

$$V_0^1(\gamma_{21})\le\rho_\varepsilon/2\qquad(<1)$$

for some path $\gamma_{21}:[0,1]\to U$ such that $\gamma_{21}(0)=u_2$, $\gamma_{21}(1)=u_1$. We set

$$u_{\bar s}(s)=\begin{cases}u(s), & s\in[0,\bar s],\\ u(\bar s), & s\in[\bar s,1],\end{cases}$$

where

$$\bar s\doteq\max\{s\in[0,1]\ :\ V_0^s(u)\le K-\bar k-V_0^1(\gamma_{21})\}.$$

Hence for any $\nu\in V$ the control defined by

$$(\tilde t,\tilde u,\tilde v)=\begin{cases}(\bar t,\gamma_{21}(s/\sigma),\nu), & s\in[0,\sigma],\\ (t,u_{\bar s},v)((s-\sigma)/(1-\sigma)), & s\in[\sigma,1],\end{cases}$$

where $\sigma\doteq V_0^1(\gamma_{21})$ is in $\Gamma_{K-\bar k}(\bar t,u_2)$. Standard estimates yield

$$|\tilde u(s)-u(s)|\le|u_2-u_1|+(3+K+T)V_0^1(\gamma_{21}),$$

from which proceeding as in the previous case one obtains an inequality similar to (3.8). Hence, by choosing $\delta \doteq \min\{\rho_\varepsilon/2, \bar{\delta}\}$, there exists some $\rho_\varepsilon > 0$ such that we have

$$\mathcal{V}(\bar{t}, \bar{x}, u_2, \bar{k}) - \mathcal{V}(\bar{t}, \bar{x}, u_1, \bar{k}) \leq \varepsilon,$$

which implies the continuity of $u \to \mathcal{V}(\bar{t}, \bar{x}, u, \bar{k})$ on Q_u uniformly with respect to the variables $(\bar{t}, \bar{x}, \bar{k})$.

Thus the continuity of the value function $\mathcal{V}$ is proved. □

4. Dynamic programming principle and dynamic programming equation. Let us define the Hamiltonian function $H : [0,T] \times \mathbb{R}^n \times U \times \mathbb{R}^{1+n+m+1} \to \mathbb{R}$ by setting

$$(4.1)\quad H(t,x,u,p_0,p_1,\dots,p_n,p_{n+1},\dots,p_{n+m},p_\infty) \doteq \min_{\substack{v\in V\\(w_0,w)\in S^m_+}} \mathcal{H}(t,x,u,p_0,p_1,\dots,p_n,p_{n+1},\dots,p_{n+m},p_\infty,w_0,w,v),$$

where $\mathcal{H}$ denotes the unminimized Hamiltonian

$$(4.2)\quad \mathcal{H}(t,x,u,p_0,p_1,\dots,p_n,p_{n+1},\dots,p_{n+m},p_\infty,w_0,w,v) \doteq \left\{ \Big(p_0 + \sum_{i=1}^{n} p_i g_0^i(t,x,u,v)\Big) w_0 + \sum_{\substack{i=1,\dots,n\\j=1,\dots,m}} \big(p_i g_j^i(t,x,u,v) + p_{j+n}\big) w^j + p_\infty |w| \right\},$$

while S^m_+ is the intersection of $[0,+\infty[\times C$ and the unit sphere $S^m = \{(w_0,w) \in \mathbb{R}^{1+m} : |(w_0,w)| = 1\}$.

We shall prove that $\mathcal{V}$ solves the dynamic programming equation

$$(\text{DPE})\qquad -H\big(t,x,u,\nabla\mathcal{V}\big) = 0,$$

where $\nabla\mathcal{V}$ stands for $(\nabla_t\mathcal{V}, \nabla_x\mathcal{V}, \nabla_u\mathcal{V}, \nabla_k\mathcal{V})$, and $\nabla_t\mathcal{V}$, $\nabla_x\mathcal{V}$, $\nabla_u\mathcal{V}$, and $\nabla_k\mathcal{V}$ denote the gradients of $\mathcal{V}$ with respect to t, x, u, and k, respectively. The presence of the minus sign in (DPE) is motivated by the fact that we wish to be consistent with the terminology of the theory of viscosity solutions. In fact, like in the nonimpulsive case, the value function $\mathcal{V}$ fails in general to be continuously differentiable, so it can satisfy (DPE) only in a generalized sense. Aiming at self–consistency we recall the definition of viscosity solution of a first-order partial differential equation; see, e.g., [18].

DEFINITION 4.1. *Let E be a subset of $\mathbb{R}^N$. A function $\mathcal{V} \in C^0(E)$ is a* viscosity subsolution *of* (DPE) *at $(t,x,u,k) \in E$ if for any $\lambda \in C^\infty(\mathbb{R}^N)$ such that (t,x,u,k) is a local maximum point of $\mathcal{V} - \lambda$ on E one has*

$$-H\big(t,x,u,\nabla\lambda(t,x,u,k)\big) \leq 0.$$

$\mathcal{V} \in C^0(E)$ is a viscosity supersolution *of* (DPE) *at $(t,x,u,k) \in E$ if for any $\lambda \in C^\infty(\mathbb{R}^N)$ such that (t,x,u,k) is a local minimum point of $\mathcal{V} - \lambda$ on E one has*

$$-H\big(t,x,u,\nabla\lambda(t,x,u,k)\big) \geq 0.$$

$\mathcal{V} \in C^0(E)$ is a viscosity solution *of* (DPE) *at (t,x,u,k) if it is both a viscosity subsolution and a viscosity supersolution.*

In order to state Theorem 4.1 below, let us introduce the domain

$$\Omega \doteq [0,T) \times \mathbb{R}^n \times \overset{\circ}{U} \times [0,K)$$

and the boundary's subsets

$$\begin{aligned} \partial_T\Omega &\doteq \{T\}\times\mathbb{R}^n\times U\times[0,K], \\ \partial'\Omega &\doteq \partial\Omega\setminus\partial_T\Omega. \end{aligned} \tag{4.3}$$

THEOREM 4.1 (dynamic programming equation and boundary conditions). *Assume either hypothesis* $\mathbf{(H2)}_C$ *or hypothesis* $\mathbf{(H2)}_U$. *Then*

a) $\mathcal{V}$ *is a viscosity solution on* Ω *of the dynamic programming equation* (DPE);

b) $\mathcal{V}$ *satisfies*

$$\mathcal{V}(T,x,u,k)\le\Phi(x,u)\qquad\forall(T,x,u,k)\in\partial_T\Omega; \tag{4.4}$$

c) $\mathcal{V}$ *is a viscosity supersolution of* (DPE) *on* $\partial'\Omega$ *and at any point* $(T,x,u,k)\in\partial_T\Omega$ *such that* $\mathcal{V}(T,x,u,k)<\Phi(x,u)$.

Remark 4.1. Note that although the cone $[0,+\infty[\times(T_uU\cap C)$ (where T_uU denotes the contingent cone to U at u; see, e.g., [1]) could be considered the natural range of the control's derivative $(t',u_1',\dots,u_m')$, the minimum in (4.1) is searched over the *compact* set S_+^m. This is due essentially to the bound on the variation of u and to the possibility of replacing any space–time control with its canonical parametrization (see the appendix). On the other hand, the positive homogenity of $\mathcal{H}$ in the variable (w_0,w) allows us to use S_+^m in the definition of H instead of $B_+^{1+m}\doteq[0,+\infty)\times C\cap\{(w_0,w):|(w_0,w)|\le 1\}$. Actually by allowing the elements $(w_0,w,v)\equiv(0,0,v)$ in the domain of minimization of $\mathcal{H}$, we would obtain an equation lacking uniqueness properties; see §5. As a direct consequence of having replaced the unbounded set $[0,+\infty)\times(T_uU\cap C)$ with a compact set, we achieve the continuity of the Hamiltonian H. Incidentally we observe that this approach presents some analogies with the one adopted by G. Barles [5] in an infinite horizon problem.

Remark 4.2. The fact that the domain of minimization of (w_0,w) is independent of u is strictly related to the very definition of viscosity supersolution on a closed set. Indeed it is well known (see, e.g., [40]) that the supersolution condition together with the subsolution condition on the interior accounts for a constraint on the state variables. Actually, in our case the situation is slightly different, since at the boundary points we have an alternative between supersolution condition and an inequality condition; a similar situation is encountered, e.g., in [2], [16], [24].

The proof of Theorem 4.1 will be based on the following dynamic programming principle, whose proof is an obvious adaptation to the parameter–free extended problem $(\mathcal{P}^e_{(\bar t,\bar x,\bar u,\bar k)})$ of the standard reasonings which yield to the dynamic programming principle in the ordinary case.

PROPOSITION 4.1 (dynamic programming principle). *The value function* $\mathcal{V}$ *has the following properties:*

i) *For an initial condition* $(\bar t,\bar x,\bar u,\bar k)\in[0,T]\times\mathbb{R}^n\times U\times[0,K]$ *and an admissible control* $(t,u,v)\in\Gamma_{K-\bar k}(\bar t,\bar u)$, *let* $y\doteq y[\bar t,\bar x,\bar u;t,u,v]$ *be the corresponding trajectory of the extended system* (2.4), (2.5). *Then the map*

$$s\mapsto\mathcal{V}\big(y(s),\bar k+V_0^s(u)\big) \tag{4.5}$$

is nondecreasing.

ii) *If in* i) *the control* (t,u,v) *is optimal, then the map* (4.5) *is constant.*

Proof. Assume by contradiction that there exist s_1, s_2, $0\le s_1<s_2\le 1$, and $\varepsilon>0$ such that

$$\mathcal{V}\big(y(s_2),\bar k+V_0^{s_2}(u)\big)=\mathcal{V}\big(y(s_1),\bar k+V_0^{s_1}(u)\big)-\varepsilon. \tag{4.6}$$

By the definition of $\mathcal{V}$ there is a space–time control $(\check{t}, \check{u}, \check{v}) \in \Gamma_{K-(\bar{k}+V_0^{s_2}(u))}\big((t,u)(s_2)\big)$ satisfying

$$\Phi\big(y[y(s_2);\check{t},\check{u},\check{v}](1)\big) \le \mathcal{V}\big(y(s_2), \bar{k} + V_0^{s_2}(u)\big) + \varepsilon/2. \tag{4.7}$$

Define the space–time control $(\hat{t}, \hat{u}, \hat{v})$ by

$$(\hat{t},\hat{u},\hat{v})(s) \doteq \begin{cases} (t,u,v)(s_1 + 2s(s_2 - s_1)), & s \in [0, 1/2], \\ (\check{t},\check{u},\check{v})(2(s-1/2)), & s \in (1/2, 1], \end{cases}$$

and set $\hat{y} \doteq y[y(s_1);\hat{t},\hat{u},\hat{v}]$. Note that $(\hat{t},\hat{u},\hat{v}) \in \Gamma_{K-(\bar{k}+V_0^{s_1}(u))}\big(t,u)(s_1)\big)$, for we have

$$V_0^1(\hat{u}) = V_{s_1}^{s_2}(u) + V_0^1(\check{u}) \le V_{s_1}^{s_2}(u) + K - \bar{k} - V_0^{s_2}(u) = K - (\bar{k} + V_0^{s_1}(u)).$$

Moreover, by the parameter–free character of the extended system (2.4)—see Proposition A.2—we have

$$\begin{aligned} \hat{y}(1/2) &= y(s_2), \\ \hat{y}(1) &= y[y(s_2);\check{t},\check{u},\check{v}](1). \end{aligned}$$

Hence, by (4.6) and (4.7), we obtain

$$\begin{aligned} \mathcal{V}\big(y(s_1), \bar{k} + V_0^{s_1}(u)\big) &\le \Phi\big(\hat{y}(1)\big) = \Phi\big(y[y(s_2);\check{t},\check{u},\check{v}](1)\big) \\ &\le \mathcal{V}\big(y(s_2), \bar{k} + V_0^{s_2}(u)\big) + \varepsilon/2 = \mathcal{V}\big(y(s_1), \bar{k} + V_0^{s_1}(u)\big) - \varepsilon/2. \end{aligned}$$

Since $\varepsilon > 0$, this proves i).

To prove ii) it is enough to observe that whenever the control (t,u,v) is optimal, on the basis of i) one has

$$\mathcal{V}(\bar{t},\bar{x},\bar{u},\bar{k}) \le \mathcal{V}\big(y(s), \bar{k} + V_0^s(u)\big) \le \Phi\big(y(1)\big) = \mathcal{V}(\bar{t},\bar{x},\bar{u},\bar{k})$$

for every $s \in [0,1]$. □

Proof of Theorem 4.1. We begin by proving that $\mathcal{V}$ is a viscosity subsolution of (DPE) on Ω. Fix a point $(\bar{y},\bar{k}) = (\bar{t},\bar{x},\bar{u},\bar{k}) \in \Omega$ and consider a map $\lambda \in C^\infty(\mathbb{R}^{1+n+m+1})$ such that $\mathcal{V}(\bar{y},\bar{k}) = \lambda(\bar{y},\bar{k})$ and $\mathcal{V} - \lambda$ has a local maximum at $(\bar{y},\bar{k})$. Then

$$\mathcal{V}(t,x,u,k) \le \lambda(t,x,u,k) \qquad \forall (t,x,u,k) \in \Omega \cap B\big((\bar{y},\bar{k}), r\big)$$

for a sufficiently small $r > 0$. Choose $v \in V$ and $w = (w_1,\dots,w_m) \in B^m[0,1] \cap C$, where $B^m[0,1] = \{w \in \mathbb{R}^m : |w| \le 1\}$, and set $w_0 \doteq \sqrt{1-|w|^2}$. Since $\bar{t} < T$, $\bar{u} \in \overset{\circ}{U}$, and $\bar{k} < K$, there exists some $\varepsilon \in (0,1)$ such that the control (t,u,v) defined by

$$(t,u,v)(s) \doteq \begin{cases} (\bar{t} + sw_0, \bar{u} + sw, v), & s \in [0,\varepsilon], \\ (\bar{t} + \varepsilon w_0 + (T - \bar{t} - \varepsilon w_0)(s-\varepsilon)/(1-\varepsilon), \bar{u} + \varepsilon w, v), & s \in (\varepsilon, 1], \end{cases}$$

is in $\Gamma_{K-\bar{k}}(\bar{t},\bar{u})$. Then by the dynamic programming principle one has

$$\lambda(\bar{y},\bar{k}) = \mathcal{V}(\bar{y},\bar{k}) \le \mathcal{V}\big(y[\bar{t},\bar{x},\bar{u};t,u,v](s), \bar{k}+V_0^s(u)\big) \le \lambda\big(y[\bar{t},\bar{x},\bar{u};t,u,v](s), \bar{k}+V_0^s(u)\big),$$

provided $0 < s \le \rho$, with ρ small enough. Dividing the last inequality by s one has

$$\frac{\lambda\big(y[\bar{t},\bar{x},\bar{u};t,u,v](s), \bar{k} + V_0^s(u)\big) - \lambda(\bar{y},\bar{k})}{s} \ge 0 \tag{4.8}$$

for every $s \in (0, \rho]$. Passing to the limit in (4.8) as $s \to 0^+$, we obtain

$$\big(\nabla_t\lambda(\bar t,\bar x,\bar u,\bar k)+\nabla_x\lambda(\bar t,\bar x,\bar u,\bar k)g_0(\bar t,\bar x,\bar u,v)\big)w_0+\nabla_x\lambda(\bar t,\bar x,\bar u,\bar k)\sum_{i=1}^m g_i(\bar t,\bar x,\bar u,v)w^i + \nabla_u\lambda(\bar t,\bar x,\bar u,\bar k)w+\nabla_k\lambda(\bar t,\bar x,\bar u,\bar k)|w|\ge 0.$$

Since w and v are arbitrary in $B^m[0,1]\cap C$ and V, respectively, it follows that

$$-H\big(\bar t,\bar x,\bar u,\nabla\lambda(\bar t,\bar x,\bar u,\bar k)\big)\le 0.$$

Hence $\mathcal{V}$ is a subsolution of (DPE) on Ω.

Let us prove that $\mathcal{V}$ is a supersolution of (DPE) on $\Omega\cup\partial'\Omega$ and at any point $(t,x,u,k)\in\partial_T\Omega$ where $\mathcal{V}(t,x,u,k)<\Phi(x,u)$. Let $(\bar y,\bar k)=(\bar t,\bar x,\bar u,\bar k)\in\overline{\Omega}$ and consider a function $\lambda\in C^\infty(\mathbb{R}^{1+n+m+1})$ such that $\mathcal{V}-\lambda$ has a local minimum on $\overline{\Omega}$ at $(\bar y,\bar k)$ and $\mathcal{V}(\bar y,\bar k)=\lambda(\bar y,\bar k)$. Then

$$\mathcal{V}(t,x,u,k)\ge\lambda(t,x,u,k)\qquad\forall(t,x,u,k)\in\overline{\Omega}\cap B\big((\bar y,\bar k),r\big)$$

for a sufficiently small $r>0$. For any $n\in\mathbb{N}\setminus\{0\}$ let $(t_n,u_n,v_n)\in\Gamma_{K-\bar k}(\bar t,\bar u)$ be a space–time control such that the corresponding trajectory y_n satisfies

$$\Phi\big(y_n(1)\big)\le\mathcal{V}(\bar y,\bar k)+1/n^2. \tag{4.9}$$

The dynamic programming principle yields

$$\lambda\big(y_n(s),\bar k+V_0^s(u_n)\big)\le\mathcal{V}\big(y_n(s),\bar k+V_0^s(u_n)\big)\le\mathcal{V}(\bar y,\bar k)+1/n^2=\lambda(\bar y,\bar k)+1/n^2,$$

provided $0<s\le\rho$, with ρ small enough. By choosing $s=1/n$ and dividing by $1/n$ we obtain

$$n\int_0^{1/n}\mathcal{H}\big(y_n,\nabla\lambda\big(y_n,\bar k+V_0^s(u_n)\big),{t_n}',{u_n}',v_n\big)\,ds\le 1/n \tag{4.10}$$

for every n sufficiently large. Since it is not restrictive to assume that the controls (t_n,u_n,v_n) coincide with their canonical parametrizations, we have $|({t_n}',{u_n}')|(s)=V_0^1(t_n,u_n)$ for almost every $s\in[0,1]$. Now if $(\bar y,\bar k)\in\Omega\cup\partial'\Omega$, one has $V_0^1(t_n,u_n)\ge V_0^1(t_n)\ge T-\bar t>0$. Hence by the continuity —on the bounded set $\overline{\Omega}\cap B\big((\bar t,\bar x,\bar u,\bar k),r\big)$— of all the considered functions, there exists a map $\varepsilon:\mathbb{R}^+\to\mathbb{R}^+$ such that $\lim_{n\to\infty}\varepsilon(n)=0$ and

$$\begin{aligned}\varepsilon(n)&\ge n\int_0^{1/n}\mathcal{H}\big(\bar t,\bar x,\bar u,\bar k,\nabla\lambda(\bar t,\bar x,\bar u,\bar k),{t_n}',{u_n}',v_n\big)\,ds\\ &\ge nV_0^1(t_n,u_n)\int_0^{1/n}\min_{\substack{v\in V\\(w_0,w)\in S^m_+}}\mathcal{H}\big(\bar t,\bar x,\bar u,\bar k,\nabla\lambda(\bar t,\bar x,\bar u,\bar k),w_0,w,v\big)\,ds\\ &\ge(T-\bar t)H\big(\bar t,\bar x,\bar u,\nabla\lambda(\bar t,\bar x,\bar u,\bar k)\big).\end{aligned} \tag{4.11}$$

Therefore, as n tends to infinity, one has

$$-H\big(\bar t,\bar x,\bar u,\nabla\lambda(\bar t,\bar x,\bar u,\bar k)\big)\ge 0,$$

i.e., $\mathcal{V}$ is a viscosity supersolution of (DPE) at $(\bar t,\bar x,\bar u,\bar k)$.

Now let $(\bar y,\bar k)=(T,\bar x,\bar u,\bar k)\in\partial_T\Omega$ and observe that any space–time control $(t,u,v)\in\Gamma(T,\bar u)$ having components $t(s)$,$u(s)$ coinciding with T,$\bar u$, respectively, gives

rise to the constant trajectory $y \equiv (T, \bar{x}, \bar{u})$. Hence thesis b) holds true by the very definition of $\mathcal{V}$; in particular the condition

$$\mathcal{V}(T, \bar{x}, \bar{u}, \bar{k}) = \Phi(\bar{x}, \bar{u}) \tag{4.12}$$

is equivalent to the optimality of any space–time control (t, u, v) with $t(s) \equiv T$ and $u(s) \equiv \bar{u}$. On the contrary, if (4.4) is satisfied as a strict inequality, set

$$\eta = \Phi(\bar{x}, \bar{u}) - \mathcal{V}(T, \bar{x}, \bar{u}, \bar{k}). \tag{4.13}$$

In order to show that $\mathcal{V}$ is a viscosity supersolution of (DPE) at $(T, \bar{x}, \bar{u}, \bar{k})$ we claim the existence of a sequence of controls $(t_n, u_n, v_n) \equiv (T, u_n, v_n)$ enjoying the following properties: i) there exists two positive constants δ, $\bar{n}$ such that

$$V_0^1(u_n) \geq \delta \qquad \forall n \geq \bar{n}; \tag{4.14}$$

ii) the trajectories $y_n \doteq y[\bar{t}, \bar{x}, \bar{u}; t_n, u_n, v_n]$ satisfy (4.9).

In order to prove this claim, assume by contradiction that for any minimizing sequence $\big((T, u_n, v_n)\big)_{n\in\mathbb{N}}$ whose corresponding solutions satisfy (4.9) and for any $\delta > 0$, $\bar{n} > 0$ there exists a $n > \bar{n}$ such that

$$V_0^1(u_n) < \delta.$$

Then one can determine a subsequence, still denoted by $\big((T, u_n, v_n)\big)_{n\in\mathbb{N}}$, such that the corresponding trajectories y_n satisfy

$$|y_n(s) - (T, \bar{x}, \bar{u})| \leq \sum_{i=1}^{m} \int_0^s |\hat{g}_i\big(y_n(s), v_n(s)\big)||u_{in}'(s)|\, ds \leq M' m V_0^s(u_n) < M' m \delta.$$

Then, choosing δ such that $\omega_\Phi(mM'\delta) \leq \eta/4$, for any $n \geq 2/\sqrt{\eta}$ we obtain

$$\big|\Phi(y_n(1)) - \Phi(\bar{x}, \bar{u})\big| \leq \eta/4, \qquad 1/n^2 \leq \eta/4.$$

These inequalities and (4.9) provide a contradiction, for

$$\Phi(\bar{x}, \bar{u}) - \eta/2 \leq \Phi(y_n(1)) - 1/n^2 \leq \mathcal{V}(T, \bar{x}, \bar{u}, \bar{k}) = \Phi(\bar{x}, \bar{u}) - \eta.$$

Hence a sequence of controls (T, u_n, v_n) satisfying (4.14) exists, and the proof is completed by replacing $T - \bar{t}$ with δ in (4.11). □

We conclude this section by showing that (DPE) can be replaced by a quasi–variational inequality. We point out that the latter can be regarded as a generalization of the dynamic programming equation which was obtained in [8] in the special case where m is equal to 1, g_1 is independent of x and u, and C coincides with $[0, +\infty)$. Set

$$\tilde{H}(t, x, u, p) \doteq \min\big\{H_1(t, x, u, p), H_2(t, x, u, p)\big\}, \tag{4.15}$$

where H_1, H_2 are defined by

$$\begin{aligned} H_1(t, x, u, p) &\doteq \min_{v\in V}\Big\{p_0 + \sum_{i=1}^{n} p_i g_0^i(t, x, u, v)\Big\}, \\ H_2(t, x, u, p) &\doteq \min_{\substack{v\in V \\ |w|=1, w\in C}} \Big\{p_\infty + \sum_{\substack{i=1,\dots,n \\ j=1,\dots,m}} p_i g_j^i(t, x, u, v) w^j + \sum_{j=1}^{m} p_{n+j} w^j\Big\}. \end{aligned} \tag{4.16}$$

THEOREM 4.2 (dynamic programming equation in the form of quasi-variational inequality). *Assume either hypothesis* $\mathbf{(H2)}_C$ *or hypothesis* $\mathbf{(H2)}_U$. *Then the following hold:*

a) $\mathcal{V}$ *is a viscosity solution of*

$$-\tilde{H}(t,x,u,\nabla\mathcal{V}) = 0, \tag{DPE}_{(\mathrm{QVI})}$$

on Ω;

b) $\mathcal{V}$ *satisfies*

$$\mathcal{V}(t,x,u,k) \le \Phi(x,u) \qquad \forall (t,x,u,k) \in \partial_T\Omega;$$

c) $\mathcal{V}$ *is a viscosity supersolution of* $(\mathrm{DPE})_{(\mathrm{QVI})}$ *on* $\partial'\Omega$ *and at any point* $(t,x,u,k) \in \partial_T\Omega$ *such that* $\mathcal{V}(t,x,u,k) < \Phi(x,u)$.

Proof. Since $H(t,x,u,p) \le \tilde{H}(t,x,u,p)$ for all $(t,x,u,p) \in \Omega \times \mathbb{R}^{1+n+m+1}$, by the fact that $\mathcal{V}$ is a viscosity subsolution of (DPE) on Ω it follows straightforwardly that $\mathcal{V}$ is a viscosity subsolution of $(\mathrm{DPE})_{(\mathrm{QVI})}$ on Ω.

Now suppose that either $(\bar{y},\bar{k}) \doteq (\bar{t},\bar{x},\bar{u},\bar{k})$ belongs to $\Omega \cup \partial'\Omega$ or it belongs to $\partial_T\Omega$, and assume that $\mathcal{V}(\bar{t},\bar{x},\bar{u},\bar{k}) < \Phi(\bar{x},\bar{u})$. By Theorem 4.1 it follows that for any $\lambda \in C^\infty(\mathbb{R}^{1+n+m+1})$ such that $\mathcal{V} - \lambda$ has a local minimum on $\overline{\Omega}$ at $(\bar{y},\bar{k})$ and $\mathcal{V}(\bar{y},\bar{k}) = \lambda(\bar{y},\bar{k})$, there is a pair $(v,w) \in V \times B^m[0,1] \cap C$ satisfying

$$\begin{aligned}(4.17)\quad &\big(\nabla_t\lambda(\bar{y},\bar{k}) + \nabla_x\lambda(\bar{y},\bar{k})g_0(\bar{t},\bar{x},\bar{u},v)\big)w_0 + \nabla_x\lambda(\bar{y},\bar{k})\sum_{i=1}^m g_i(\bar{t},\bar{x},\bar{u},v)w^i \\ &\qquad + \nabla_u\lambda(\bar{y},\bar{k})w + \nabla_k\lambda(\bar{y},\bar{k})|w| \le 0,\end{aligned}$$

where $w_0 \doteq \sqrt{1-|w|^2}$. If $w = 0$ or $|w| = 1$, then $\mathcal{V}$ is a supersolution of $(\mathrm{DPE})_{(\mathrm{QVI})}$. Otherwise, i.e., if $0 < |w| < 1$, divide (4.17) by $|w|$ and observe that either the first term or the sum of the remaining terms must be nonpositive. Hence $\mathcal{V}$ is a supersolution of $(\mathrm{DPE})_{(\mathrm{QVI})}$. □

Remark 4.3. Theorem 4.2 exhibits a certain analogy of the considered problem with standard impulse control problems; see, e.g., [3], [8]. Indeed the value functions of the latter satisfy certain quasi-variational inequalities, which replace the usual Bellman equation. Actually, the dynamics considered in standard impulse theory can be considered as the simplest case of the dynamics considered in this paper, namely, the case where the vector fields $g_1,\dots,g_m$ are constant. Yet the comparison between the two approaches cannot be pushed further, for the two corresponding minimum problems are not equivalent. Instead, a more strict relation can be recognized between the problems considered here and the questions adressed in E. N. Barron and R. Jensen's paper [6], where (nonimpulsive) controls $\eta(\cdot)$ with bounded variation are considered. Indeed by adding the trivial (impulsive) equation $\dot{z} = \dot{\eta}$ the control system studied in [6] will be reduced to the form considered in this paper.

5. Uniqueness of the solution of (DPE) and verification theorem. In this section we prove a comparison result for viscosity solutions of (DPE). As a consequence we obtain a uniqueness result and a verification theorem for the extended problem $\mathcal{P}^e_{(t,x,u,k)}$.

We assume hypothesis **(H3)** below on the boundary of U. Hypothesis **(H3)**, which excludes the presence of zero-amplitude corners in ∂U, is quite standard in problems involving state constraints; see, e.g., [40].

(H3) There exist a map $\eta \in BUC(U, \mathbb{R}^m)$ and two positive numbers q, r such that

$$B\big(u + t\eta(u), rt\big) \subset \overset{\circ}{U} \qquad \text{for } u \in \partial U \text{ and } 0 < t \leq q.$$

THEOREM 5.1. *Assume hypothesis* **(H3)** *and either* $\mathbf{(H2)}_C$ *or* $\mathbf{(H2)}_U$. *Let* $\mathcal{V}_1$ *be a bounded continuous viscosity subsolution of* (DPE) *in* Ω *which satisfies*

$$\mathcal{V}_1(t, x, u, k) \leq \Phi(x, u) \qquad \forall (t, x, u, k) \in \partial_T \Omega. \tag{5.1}$$

Let $\mathcal{V}_2$ *be a bounded continuous viscosity supersolution of* (DPE) *in* $\Omega \cup \partial'\Omega$ *such that for any* $(t, x, u, k) \in \partial_T \Omega$ *either* $\mathcal{V}_2$ *satisfies the inequality*

$$\mathcal{V}_2(t, x, u, k) \geq \Phi(x, u) \tag{5.2}$$

or it is a viscosity supersolution of (DPE).

Then

$$\mathcal{V}_1 \leq \mathcal{V}_2 \qquad \text{on } \overline{\Omega}. \tag{5.3}$$

Proof. For every $(t, k) \in [0, T] \times [0, K]$ let us define the map $\mathcal{T}_{t,k} : \mathbb{R}^+ \to \mathbb{R}$ by setting

$$\mathcal{T}_{t,k}(r) \doteq \frac{\log r}{1 + t + k}.$$

Let M be a lower bound for the maps $\Phi, \mathcal{V}_1$, and $\mathcal{V}_2$, and let us set

$$Z_i(t, x, u, k) \doteq \mathcal{T}_{t,k}\big(\mathcal{V}_i(T - t, x, u, K - k) - M + 1\big), \qquad i = 1, 2,$$
$$\psi(t, x, u, k) \doteq \mathcal{T}_{t,k}(\Phi(x, u) - M + 1).$$

Then, on the one hand, the map Z_1 turns out to be a bounded continuous subsolution of

(TDPE)

$$Z + \max_{(v, w_0, w) \in V \times S^m_+} \left\{ \frac{1 + t + k}{w_0 + |w|} \mathcal{H}(T - t, x, u, \nabla_t Z, -\nabla_x Z, -\nabla_u Z, \nabla_k Z, v, w_0, w) \right\} = 0$$

in Ω, where $\mathcal{H}$ is the unminimized Hamiltonian defined in (4.2); moreover Z_1 satisfies

$$Z_1(t, x, u, k) \leq \psi(t, x, u, k)$$

on $\partial_0 \Omega \doteq \{0\} \times \mathbb{R}^n \times U \times [0, K]$.

On the other hand, Z_2 is a bounded continuous supersolution of (TDPE) on $\overline{\Omega} \backslash \partial_0 \Omega$. Furthermore, at each point $(0, x, u, k) \in \partial_0 \Omega$, Z_2 either satisfies the inequality $Z_2(0, x, u, k) \geq \psi(0, x, u, k)$ or is a viscosity supersolution of (TDPE). Hence, a straightforward application of Theorem 1.1 in [2] implies that

$$Z_1 \leq Z_2$$

on $\overline{\Omega}$, which in turn yields the thesis. □

THEOREM 5.2 (uniqueness). *Assume hypothesis* **(H3)** *and either* $\mathbf{(H2)}_C$ *or* $\mathbf{(H2)}_U$. *Then the value function* $\mathcal{V}$ *is the unique bounded continuous viscosity solution of* (DPE) *on* Ω *which satisfies the following boundary conditions:*

$(BC)'_1$ $\mathcal{V}$ *is a viscosity supersolution of* (DPE) *at all points of* $[0, T[\times \mathbb{R}^n \times \partial U \times [0, K[\cup [0, T[\times \mathbb{R}^n \times U \times \{K\}$;

$(BC)'_2$ *at each boundary point* (T, x, u, k) *one has* $\mathcal{V}(T, x, u, k) \leq \Phi(x, u)$ *and, moreover, either* $\mathcal{V}$ *is a supersolution of* (DPE) *or it satisfies the relation* $\mathcal{V}(T, x, u, k) = \Phi(x, u)$.

Remark 5.1. By using the same arguments as in the previous theorem a uniqueness result for $(\mathrm{DPE})_{(\mathrm{QVI})}$ can be proved as well. Hence (DPE) and $(\mathrm{DPE})_{(\mathrm{QVI})}$ turn out to be equivalent as soon as one assumes the boundary conditions $(BC)'_1$, $(BC)'_2$. It is worthwhile comparing the latter conditions with the boundary conditions of Dirichlet type assumed by Barron, Jensen, and Menaldi [8] in the particular case when $m = 1$, $u \equiv k \in [0, K]$, g_1 is independent of (x, u), and $C = [0, +\infty)$. Barron–Jensen–Menaldi's conditions can be stated as follows:

$(BC)_1$ the map $\mathcal{V}$ coincides with the value function

$$h_T(\bar{x}, \bar{k}) \doteq \inf_{(T,u,v)\in\Gamma_{K-\bar{k}}(T,\bar{k})} \Phi\big(z[\bar{x}, \bar{k}; u, v](1)\big)$$

on the strip $\{T\}\times\mathbb{R}^n\times[0, K)$, where $z[\bar{x}, \bar{k}; u, v](\cdot)$ is the solution of the *purely impulsive* (integrable) Cauchy problem

$$\begin{cases} z' = \hat{g}_1(T, v(s))u'(s), \\ z(0) = (\bar{x}, \bar{k}); \end{cases}$$

$(BC)_2$ the map $\mathcal{V}$ coincides with the value function

$$h_K(\bar{t}, \bar{x}) \doteq \inf_{v\in\mathcal{B}([\bar{t},T],V)} \Phi\big(x[\bar{t}, \bar{x}; v](T), K\big)$$

on the strip $[0, T) \times \mathbb{R}^n \times \{K\}$, where $x[\bar{t}, \bar{x}; v](\cdot)$ is the solution of the *nonimpulsive* Cauchy problem

$$\begin{cases} \dot{x} = g_0(t, x(t), K, v(t)), \\ x(\bar{t}) = \bar{x}. \end{cases}$$

In particular, in order to construct the maps h_T and h_K one needs solving a class of auxiliary optimization problems whose difficulty is often comparable to the difficulty of the original problem. Instead, conditions $(BC)'_1$, $(BC)'_2$ of Theorem 5.2 refer only to equation (DPE) and to the known function Φ (see also the example in §7).

We conclude this section with a verification theorem, which incidentally provides an answer—in the present, more general, framework—to the question posed by Barron, Jensen, and Menaldi [8] (see the introduction, question b)) about the relationship between optimal controls and dynamic programming equation.

THEOREM 5.3 (verification theorem). *Let $Z \in C(\overline{\Omega})$ be a bounded viscosity subsolution of* (DPE) *in Ω which satisfies the condition $Z \le \Phi$ on $\partial_T\Omega$. Then*

$$Z \le \mathcal{V} \qquad \text{on } \overline{\Omega}. \tag{5.5}$$

Moreover, if for a given $(\bar{t}, \bar{x}, \bar{u}, \bar{k}) \in \overline{\Omega}$ there exists a space–time control $(t, u, v) \in \Gamma_{K-\bar{k}}(\bar{t}, \bar{u})$ such that

$$\Phi\big(y[\bar{t}, \bar{x}, \bar{u}; t, u, v](1)\big) \le Z(\bar{t}, \bar{x}, \bar{u}, \bar{k}),$$

then the control (t, u, v) is optimal and

$$Z(\bar{t}, \bar{x}, \bar{u}, \bar{k}) = \mathcal{V}(\bar{t}, \bar{x}, \bar{u}, \bar{k}).$$

6. Costate, maximum principle, and gradient of the value function. In ordinary control theory it is well known that the costate involved in the Pontryagin's maximum principle coincides—when no endpoint constraints are imposed—with the gradient of the value function evaluated along an optimal trajectory. More generally,

if the value function is not differentiable at some point, the costate belongs to the supergradient of the value function; see, e.g., [7], [22].

It is clear that in order to prove an analogous result for an impulsive system we need to understand the behaviour of the costate in the presence of spatial jumps of the trajectory.

In the special case where $m = 1$ and g_1 is independent of (x, u) (and $C = [0, +\infty)$) the question is posed as an open problem in [8] (see the introduction, question a)). Since the problem with impulses has been reduced to a standard nonimpulsive control problem, under hypothesis $\mathbf{(H2)}_C$ and by simply applying standard arguments (see [7], [22]), it is now easy to provide an answer to the above question in the general case treated in the present paper.

Throughout this section we assume that the vector fields $g_0, \dots, g_m$ and the map Φ are continuously differentiable with respect to the variables t, x and u.

We recall that the space–time Hamiltonian equations in the variables (y, k) and (p, p_k) have the form

$$\begin{aligned} y' &= \nabla_p \mathcal{H}(y, p, p_k, t', u', v), \\ k' &= \nabla_{p_k} \mathcal{H}(y, p, p_k, t', u', v), \\ p' &= -\nabla_y \mathcal{H}(y, p, p_k, t', u', v), \\ p'_k &= -\nabla_k \mathcal{H}(y, p, p_k, t', u', v), \end{aligned} \tag{6.1}$$

where $\mathcal{H}$ is the unminimized Hamiltonian introduced in §4. In components we have

$$\begin{cases} t' = t', \\ x' = g_0(t, x, u, v)t' + \sum_{j=1}^{m} g_j(t, x, u, v)u'_j, \\ u' = u', \\ k' = |u'|, \end{cases} \tag{6.2}$$

$$\begin{cases} p'_0 = -\langle p_x, \nabla_t g_0(t, x, u, v)t' \rangle - \sum_{j=1}^{m} \langle p_x, \nabla_t g_j(t, x, u, v)u'_j \rangle, \\ p'_x = -\langle p_x, \nabla_x g_0(t, x, u, v)t' \rangle - \sum_{j=1}^{m} \langle p_x, \nabla_x g_j(t, x, u, v)u'_j \rangle, \\ p'_u = -\langle p_x, \nabla_u g_0(t, x, u, v)t' \rangle - \sum_{j=1}^{m} \langle p_x, \nabla_u g_j(t, x, u, v)u'_j \rangle, \\ p'_k = 0. \end{cases} \tag{6.3}$$

Note that (6.2) is nothing but the control system (2.4) supplemented with the equation $k' = |u'|$.

By saying that a control $(t(\cdot), u(\cdot), v(\cdot))$ evolves instantaneously at a time $\bar{t} \in [0, T]$ we mean that the preimage $t^{\leftarrow}(\bar{t})$ is a nondegenerate interval $[s_1, s_2]$ on which the component $u(\cdot)$ is not constant. Accordingly, one can compute the *jumps* at time $\bar{t}$ of both the state (y, k) and the costate (p, p_k) by solving the Hamiltonian equations (6.2), (6.3) on the interval $[s_1, s_2]$.

In order to state a maximum principle for the extended problem $\mathcal{P}^e_{(\bar{t}, \bar{x}, \bar{u}, \bar{k})}$ in the

unconstrained case defined by hypothesis $\mathbf{(H2)}_C$ we recall that it is not restrictive to assume that the norm of the derivative $(\hat{t}'(s), \hat{u}'(s))$ is equal to the constant value $V_0^1(\hat{t}, \hat{u}) = \int_0^1 |(\hat{t}'(s), \hat{u}'(s))| ds$ almost everywhere in $[0,1]$; see the appendix.

THEOREM 6.1 (maximum principle). *Let us assume* $\mathbf{(H2)}_C$, *i.e.,* $U = \mathbb{R}^m$. *Fix* $(\bar{t}, \bar{x}, \bar{u}, \bar{k}) \in [0,T] \times \mathbb{R}^{n+m} \times [0,K]$ *and let* $(\hat{t}, \hat{u}, \hat{v})$ *be an optimal control for the extended problem* $\mathcal{P}^e_{(\bar{t},\bar{x},\bar{u},\bar{k})}$, *with* $|(\hat{t}', \hat{u}')| = L$ *almost everywhere in* $[0,1]$ *for some positive constant* L. *Moreover denote the corresponding optimal trajectory by* $(\hat{y}, \hat{k}) = (\hat{t}, \hat{x}, \hat{u}, \hat{k})$.

Then there exists a costate map $(\hat{p}, \hat{p}_k) = (\hat{p}_0, \hat{p}_x, \hat{p}_u, \hat{p}_k) : [0,1] \to \mathbb{R}^{1+n+m+1}$ *such that*

i) $(\hat{y}, \hat{k}, \hat{p}, \hat{p}_k)$ *is a solution of the Hamiltonian equations* (6.1) *corresponding to the control* $(\hat{t}, \hat{u}, \hat{v})$ *and satisfies the boundary conditions*

$$
\begin{aligned}
&(\hat{t}(0), \hat{x}(0), \hat{u}(0), \hat{k}(0)) = (\bar{t}, \bar{x}, \bar{u}, \bar{k}),\\
&\hat{p}_x(1) = \nabla_x \Phi\big(\hat{x}(1), \hat{u}(1)\big), \qquad \hat{p}_u(1) = \nabla_u \Phi\big(\hat{x}(1), \hat{u}(1)\big),\\
&H\big(T, \hat{x}(1), \hat{u}(1), \hat{p}_0(1), \hat{p}_x(1), \hat{p}_u(1), \hat{p}_k(1)\big) = 0,
\end{aligned}
\tag{6.4}
$$

where the Hamiltonian H *is defined as in* §2;

ii) *the minimum relation*

$$
\begin{aligned}
&\mathcal{H}\left(\hat{t}(s), \hat{x}(s), \hat{u}(s), \hat{p}_0(s), \hat{p}_x(s), \hat{p}_u(s), \hat{p}_k(s), \frac{\hat{t}'(s)}{L}, \frac{\hat{u}'(s)}{L}, v(s)\right)\\
&\quad = H\left(\hat{t}(s), \hat{x}(s), \hat{u}(s), \hat{p}_0(s), \hat{p}_x(s), \hat{p}_u(s), \hat{p}_k(s)\right)
\end{aligned}
\tag{6.5}
$$

holds for almost every $s \in [0,1]$, *and the equality*

$$
H\left(\hat{t}(s), \hat{x}(s), \hat{u}(s), \hat{p}_0(s), \hat{p}_x(s), \hat{p}_u(s), \hat{p}_k(s)\right) = 0 \tag{6.6}
$$

holds for all $s \in [0,1]$;

iii) *if* $\hat{u}$ *verifies* $V_0^1(\hat{u}) < K - \bar{k}$, *one has* $\hat{p}_k = 0$ *identically on* $[0,1]$.

This theorem, whose proof will be given after the statement of Theorem 6.2, is a straightforward consequence of the Pontryagin maximum principle when the latter is applied to the extended (nonimpulsive) control problem.

We point out that in the case when the fields g_i are independent of (x,u), some versions of the maximum principle already exist in the literature; see, e.g., [33], [36], [38], [42]. What is more, up to some formal changes a maximum principle for the general case considered here can be already found in [31]. Yet since our main goal is to establish a relationship between the costate and the value function, we prefer to give here a statement and a proof of the maximum principle in the theoretical framework introduced in the previous sections.

In order to state a relationship between the costate and the value function let us introduce the family of maps defined by

$$
\begin{aligned}
&\mathcal{V}^{t,k} : \mathbb{R}^{n+m} \to \mathbb{R}, \qquad (t,k) \in [0,T] \times [0,K],\\
&\mathcal{V}^{t,k}(x,u) \doteq \mathcal{V}(t,x,u,k) \qquad \forall (x,u) \in \mathbb{R}^{n+m}.
\end{aligned}
$$

Moreover, let us recall the definition of superdifferential of a continuous map.

DEFINITION 6.1. *Let* f *be a function in* $C(\mathbb{R}^N)$. *Then for every* $x \in \mathbb{R}^N$ *the subset* $\nabla^+ f(x) \subset \mathbb{R}^N$ *defined by*

$$
\nabla^+ f(x) \doteq \left\{ p \in \mathbb{R}^N : \limsup_{y \to x} \frac{f(y) - f(x) - \langle p\ ,\ y - x \rangle}{|y - x|} \leq 0 \right\}
$$

is called the superdifferential *of* f *at* x.

In the statement of the following theorem $(\hat{t}, \hat{x}, \hat{u}, \hat{k})$ and $(\hat{p}, \hat{p}_k) = (\hat{p}_0, \hat{p}_x, \hat{p}_u, \hat{p}_k)$ have the same meaning as in Theorem 6.1.

THEOREM 6.2 (costate and value function). *For every* $s \in [0,1]$ *one has*

$$(\hat{p}_x(s), \hat{p}_u(s)) \in \nabla^+ \mathcal{V}^{\hat{t}(s),\hat{k}(s)}\big(\hat{x}(s), \hat{u}(s)\big). \tag{6.7}$$

Proof of Theorem 6.1. This theorem follows straightforwardly from the Pontryagin maximum principle for nonimpulsive control systems. Indeed, thanks to the fact that we can restrict the family of controls to the subfamily formed by canonically parametrized control strategies, a control $(\hat{t}, \hat{u}, \hat{v})$ is optimal for problem $\mathcal{P}^e_{(\bar{t},\bar{x},\bar{u},\bar{k})}$ if and only if the control $(\hat{w}_0, \hat{w}, \hat{v}) \doteq (\hat{t}', \hat{u}', \hat{v})$ is optimal for the following ordinary control problem with endpoint constraints:

$$\text{minimize } \Phi\big(z[\bar{t}, \bar{x}, \bar{u}, \bar{k}; w_0, w, v](1)\big)$$

over the trajectories $(z[\bar{t}, \bar{x}, \bar{u}, \bar{k}; (w_0, w, v)](\cdot)$ of

$$\begin{aligned} z' &= \hat{g}_0(z(s), v(s))w_0 + \sum_{i=1}^{m} \hat{g}_i(z(s), v(s))w_i(s), \\ z_k' &= |w|, \\ (z(0), z_k(0)) &= (\bar{t}, \bar{x}, \bar{u}, \bar{k}), \end{aligned}$$

satisfying the endpoint constraints

$$z_0(1) = T, \qquad z_k(1) \le K$$

and corresponding to measurable control maps (w_0, w, v) from $[0,1]$ into

$$\big\{(w_0, w, v) \in [0, +\infty) \times C \times V : \quad |(w_0, w)| \le T + K\big\}.$$

By applying the Pontryagin's maximum principle (see [26]) to this problem we obtain a statement which is equivalent to Theorem 6.1 except that relation (6.5) must be replaced by

(6.8)

$$\begin{aligned} &\mathcal{H}\left(\hat{t}(s), \hat{x}(s), \hat{u}(s), \hat{p}_0(s), \hat{p}_x(s), \hat{p}_u(s), \hat{p}_k(s), \frac{\hat{t}'(s)}{L}, \frac{\hat{u}'(s)}{L}, v(s)\right) \\ &= \min_{(w_0,w,v) \in B_+^{m+1}\left[\frac{T+K}{L}\right] \times V} \left\{\mathcal{H}\left(\hat{t}(s), \hat{x}(s), \hat{u}(s), \hat{p}_0(s), \hat{p}_x(s), \hat{p}_u(s), \hat{p}_k(s), w_0, w, v\right)\right\}, \end{aligned}$$

where

$$B_+^{m+1}\left[\frac{T+K}{L}\right] \doteq \left\{(w_0, w) \in [0, +\infty) \times C : \ |(w_0, w)| \le \frac{T+K}{L}\right\}.$$

On the other hand, since $|(\hat{t}'(s), \hat{u}'(s))| = L$ almost everywhere, in the minimum relation (6.8) one can replace $B_+^{m+1}\left[\frac{T+K}{L}\right]$ with the set S_+^m. Hence (6.8) reduces to (6.5), and the theorem is proved. □

We observe that the proof of Theorem 6.2 cannot be derived directly from analogous results concerning nonimpulsive systems. Indeed, to our knowledge these results concern problems without endpoint constraints, while the trajectories of system (6.2) are subject to

$$\hat{t}(1) = T, \quad \hat{k}(1) \le K.$$

However, since the coordinates $(\hat{x}(1), \hat{u}(1))$ are not constrained, the arguments we use to prove Theorem 6.2 are substantially the same as the ones used in the nonimpulsive case without endpoint constraints; see, e.g., [7], [22].

Proof of Theorem 6.2. By the definition of optimal trajectory there exists a measurable map $\hat{v}$ such that the solution corresponding to the space–time control $(\hat{t}, \hat{u}, \hat{v})$ coincides with the optimal trajectory $(\hat{t}, \hat{x}, \hat{u}, \hat{k})$. Let $s^* \in [0,1]$ and for every initial point $(\tilde{x}, \tilde{u}) \in \mathbb{R}^n \times \mathbb{R}^m$ define the control map

$$\hat{u}_{\tilde{u}} : [s^*, 1] \to \mathbb{R}^m, \qquad \hat{u}_{\tilde{u}}(s) \doteq \hat{u}(s) + \tilde{u} - \hat{u}(s^*).$$

Next consider the cost functional

$$J(\tilde{x}, \tilde{u}) \doteq \Phi\big(\hat{x}[\tilde{x}, \tilde{u}](1), \hat{u}_{\tilde{u}}(1)\big),$$

where $\hat{x}[\tilde{x}, \tilde{u}](\cdot)$ denotes the solution on the interval $[s^*, 1]$ of the Cauchy problem

$$x' = g_0\left(\hat{t}(s), x(s), \hat{u}_{\tilde{u}}(s), \hat{v}(s)\right)\hat{t}'(s) + \sum_{i=1}^{m} g_i\left(\hat{t}(s), x(s), \hat{u}_{\tilde{u}}(s), \hat{v}(s)\right)\hat{u}_i'(s),$$
$$x(s^*) = \tilde{x}.$$

Up to a reparametrization from the interval $[s^*, 1]$ into the standard interval $[0,1]$, the control $(\hat{t}, \hat{u}_{\tilde{u}}, \hat{v}) : [s^*, 1] \to \mathbb{R}^{1+m} \times V$ is feasible for the initial point $(\hat{t}(s^*), \tilde{x}, \tilde{u}, \hat{k}(s^*))$. Hence one has

$$\mathcal{V}\left(\hat{t}(s^*), \tilde{x}, \tilde{u}, \hat{k}(s^*)\right) = \mathcal{V}^{\hat{t}(s^*), \hat{k}(s^*)}(\tilde{x}, \tilde{u}) \leq J(\tilde{x}, \tilde{u})$$

for every $(\tilde{x}, \tilde{u}) \in \mathbb{R}^{n+m}$, and, by the optimality of $(\hat{t}, \hat{u}, \hat{v})$,

$$J(\hat{x}(s^*), \hat{u}(s^*)) = \mathcal{V}^{\hat{t}(s^*), \hat{k}(s^*)}(\hat{x}(s^*), \hat{u}(s^*)).$$

Therefore by the definition of superdifferential of $\mathcal{V}^{\hat{t}(s^*), \hat{k}(s^*)}$ it is sufficient to prove that $J(\tilde{x}, \tilde{u})$ is differentiable at $(\hat{x}(s^*), \hat{u}(s^*))$ and satisfies

$$(p_x(s^*), p_u(s^*)) = \nabla_{x,u} J(\hat{x}(s^*), \hat{u}(s^*)). \tag{6.9}$$

By standard computations involving the differentiability of the solutions of (6.2) with respect to the initial data we obtain

$$\nabla_{x,u} J(\hat{x}(s^*), \hat{u}(s^*)) = \langle \nabla_{x,u} \Phi(\hat{x}(1), \hat{u}(1)) \ , \ Z(1) \rangle, \tag{6.10}$$

where the $(n+m) \times (n+m)$ matrix $Z(\cdot)$ is the solution in $[s^*, 1]$ of the variational Cauchy problem

$$Z'(s) = \left\langle \left[\nabla_{x,u} g_0\big(\hat{t}(s), \hat{x}(s), \hat{u}(s), \hat{v}(s)\big)\hat{t}'(s) + \sum_{i=1}^{m} \nabla_{x,u} g_i\big(\hat{t}(s), \hat{x}(s), \hat{u}(s), \hat{v}(s)\big)\hat{u}_i'(s) \right], Z(s) \right\rangle,$$
$$Z(s^*) = Id.$$

Since $(\hat{p}_x, \hat{p}_u)$ coincides with the unique solution to the adjoint Cauchy problem

$$\begin{aligned}(p'_x(s), p'_u(s)) = -\Big\langle (p_x(s), p_u(s)), \big[\nabla_{x,u} g_0\big(\hat{t}(s), \hat{x}(s), \hat{u}(s), \hat{v}(s)\big)\hat{t}'(s) + \\ + \sum_{i=1}^{m} \nabla_{x,u} g_i(\hat{t}(s), \hat{x}(s), \hat{u}(s), \hat{v}(s))\hat{u}'_i(s)\big]\Big\rangle, \\ (p_x(1), p_u(1)) = \nabla_{x,u}\Phi\left(\hat{x}(1), \hat{u}(1)\right),\end{aligned}$$

one has

$$\frac{d}{ds}\langle (\hat{p}_x(s), \hat{p}_u(s)), Z(s)\rangle = 0$$

on the whole interval $[s^*, 1]$, from which it follows that

$$(\hat{p}_x(s^*), \hat{p}_u(s^*)) = \langle (\hat{p}_x(s^*), \hat{p}_u(s^*)), Id\rangle = \langle \nabla_{x,u}\Phi\left(\hat{x}(s^*), \hat{u}(s^*)\right), Z(1)\rangle.$$

The latter equality and (6.10) yield (6.9), and the theorem is proved. □

7. Example. We apply the results proved in the previous sections to a simple minimum problem. In particular we check the optimality of a certain feedback control by showing that the corresponding cost function satisfies equation (DPE) and the boundary conditions $(\mathrm{BC})'_1$, $(\mathrm{BC})'_2$.

Let T, K, and c be positive constants, and for any $(\bar{t}, \bar{x}, \bar{k}) \in [0, T) \times \mathbb{R} \times [0, K]$ consider the minimum problem:

$$(\mathcal{P}_{(\bar{t},\bar{x},\bar{k})}) \qquad \text{minimize } \arctan\big(x(T)\big)$$

over all the endpoints of the trajectories of

$$\begin{aligned} &\dot{x} = c + \dot{u}_1(t) + x\dot{u}_2(t) \qquad \forall t \in (\bar{t}, T], \\ &x(\bar{t}) = \bar{x}, \end{aligned} \tag{7.1}$$

corresponding to the absolutely continuous controls (u_1, u_2) satisfying $V_{\bar{t}}^T(u_1, u_2) \leq K - \bar{k}$. Moreover assume that the derivatives $(\dot{u}_1, \dot{u}_2)$ belong (for a.e. $t \in [\bar{t}, T]$) to the closed cone

$$C \doteq \big\{(w_1, w_2) \in \mathbb{R}^2 : \ w_1 \leq 0, \ w_2 \geq 0\big\}. \tag{7.2}$$

Following §2, let us consider the extended system relative to (7.1), (7.2):

$$\begin{aligned} &x' = ct'(s) + u'_1(s) + xu'_2(s) \qquad \forall s \in [0, 1], \\ &(t, x)(0) = (\bar{t}, \bar{x}), \quad (u'_1, u'_2)(s) \in C \ \text{ for a.e. } s \in [0, 1]. \end{aligned} \tag{7.3}$$

The form of the equation in (7.3) implies that the optimal solution of problem $\mathcal{P}_{(\bar{t},\bar{x},\bar{k})}$ —and hence the value function $\mathcal{V}$— is independent of the initial values $u_1(\bar{t})$ and $u_2(\bar{t})$ of the controls. Moreover, by the definition of space–time control one has $t'(s) \geq 0$ for a.e. $s \in [0, 1]$. Since both $\dot{u}_1$ and $x\dot{u}_2$ are negative whenever x is negative, heuristics suggests the following strategy: at the initial time let the state jump to the minimum x reachable by spending all the available variation $K - \bar{k}$. After the jump set $\dot{u}_1 \equiv 0 \equiv \dot{u}_2$ and let the state evolve in time (with constant derivative equal to c). The maximum

principle (6.5) yields

$$(w_0, w_1, w_2)(x, s) = \begin{cases} -(K - \bar{k} + T - \bar{t})(1 + x^2)^{-1/2}(0, 1, x) & \text{if } x < 0, \\ -(K - \bar{k} + T - \bar{t})(0, 1, 0) & \text{if } x \geq 0,\ s \in [0, \frac{K-\bar{k}}{K-\bar{k}+T-\bar{t}}], \\ (K - \bar{k} + T - \bar{t})(1, 0, 0) & \text{if } x \geq 0,\ s \in [\frac{K-\bar{k}}{K-\bar{k}+T-\bar{t}}, 1], \end{cases} \tag{7.4}$$

as a control candidate to be optimal. Accordingly, for each initial condition $(\bar{t}, \bar{x}, \bar{k})$ the corresponding terminal position $\hat{x}(T; \bar{t}, \bar{x}, \bar{k})$ is given by

$$\hat{x}(T; \bar{t}, \bar{x}, \bar{k}) = \begin{cases} \sinh\big(\operatorname{arc\,sinh}(\bar{x}) - (K - \bar{k})\big) + c(T - \bar{t}), & \bar{x} \leq 0,\ 0 \leq \bar{k} \leq K, \\ \sinh\big(\bar{x} - (K - \bar{k})\big) + c(T - \bar{t}), & \bar{x} > 0,\ \bar{x} + \bar{k} < K,\ 0 \leq \bar{k} \leq K, \\ \bar{x} - (K - \bar{k}) + c(T - \bar{t}), & \bar{x} > 0,\ \bar{x} + \bar{k} \geq K,\ 0 \leq \bar{k} \leq K, \end{cases}$$

so that the resulting cost is given by $\mathcal{V}(\bar{t}, \bar{x}, \bar{k}) \doteq \arctan\big(\hat{x}(T; \bar{t}, \bar{x}, \bar{k})\big)$. We claim that $\mathcal{V}(t, x, k)$ is the optimal cost; i.e., it coincides with the value function of problem $\mathcal{P}_{(t,x,k)}$. Since $\mathcal{V}$ is continuously differentiable on $\Omega \doteq (0, T) \times \mathbb{R} \times (0, K)$, on the basis of the uniqueness of Theorem 5.2 it is sufficient to verify that i) $\mathcal{V}$ is a classical solution on Ω of (DPE) equation

$$\min\left\{(\nabla_t \mathcal{V} + \nabla_x \mathcal{V} c) w_0 + \nabla_x \mathcal{V}(w_1 + x w_2) + \mathcal{V}_k \sqrt{w_1^2 + w_2^2} : \ (w_0, w_1, w_2) \in S_+^2 \right\} = 0; \tag{7.5}$$

ii) $\mathcal{V}$ is a viscosity supersolution of (DPE) on $\partial\Omega \setminus \big(\{T\} \times \mathbb{R} \times \{K\}\big)$ and satisfies

$$\mathcal{V}(T, x, K) = \arctan(x) \qquad \forall x \in \mathbb{R}. \tag{7.6}$$

Relations (7.5) and (7.6) can be checked by means of straightforward calculations. Hence it only remains to check the supersolution inequality for every $(t, x, k) \in \partial\Omega \setminus \big(\{T\} \times \mathbb{R} \times \{K\}\big)$. If $\lambda \in C^\infty(\overline{\Omega})$ is a map such that $\mathcal{V} - \lambda$ has a local minimum at (t, x, k), then it satisfies the following relations:

$$\nabla_t \lambda(t, x, k) \geq \nabla_t \mathcal{V}(t, x, k), \quad (\nabla_x \lambda, \nabla_k \lambda)(t, x, k) = (\nabla_x \mathcal{V}, \nabla_k \mathcal{V})(t, x, k) \quad \text{if } t = T$$

and

$$\nabla_k \lambda(t, x, k) \geq \nabla_k \mathcal{V}(t, x, k), \quad (\nabla_t \lambda, \nabla_x \lambda)(t, x, k) = (\nabla_t \mathcal{V}, \nabla_x \mathcal{V})(t, x, k) \quad \text{if } k = K.$$

Moreover observe that relation (7.5) holds at any $(t, x, k) \in \partial\Omega \setminus \big(\{T\} \times \mathbb{R} \times \{K\}\big)$ and the minimum in the right-hand side of (7.5) is achieved by a vector $(\bar{w}_0, \bar{w}_1, \bar{w}_2)$ verifying $\bar{w}_0 = 0$ if $t = T$ and $(\bar{w}_1, \bar{w}_2) = 0$ if $k = K$. It follows that

$$(\nabla_t \lambda + \nabla_x \lambda c) \bar{w}_0 + \nabla_x \lambda(\bar{w}_1 + x \bar{w}_2) + \nabla_k \lambda \sqrt{\bar{w}_1^2 + \bar{w}_2^2} = 0 \qquad \text{on } \partial\Omega \setminus \big(\{T\} \times \mathbb{R} \times \{K\}\big).$$

Hence $\mathcal{V}$ is a viscosity supersolution on $\partial\Omega \setminus \big(\{T\} \times \mathbb{R} \times \{K\}\big)$, so we can conclude that $\mathcal{V}$ coincides with the value function of problem $\mathcal{P}_{(t,x,k)}$ $\forall (t, x, k) \in \overline{\Omega}$. In particular the controls (7.4) are optimal.

Appendix.

Canonical parametrizations. We recall the notion of *canonical parametrization* from [32]. For this purpose, if (t, u) is not identically constant let us introduce the

map σ from $[0,1]$ onto itself defined by

$$\sigma(s) \doteq \frac{V_0^s(t,u)}{V_0^1(t,u)} = \frac{\int_0^s |(t',u')|ds}{\int_0^1 |(t',u')|ds}.$$

If (t,u) is constant on the whole interval $[0,1]$, we set

$$(t^c,u^c,v^c) \doteq (t,u,v).$$

If (t,u) is not constant, we set

$$(t^c,u^c,v^c)(\sigma) \doteq (t,u,v)(s), \qquad \sigma = \sigma(s). \tag{D}$$

In principle (D) defines a multivalued map. Yet (t^c,u^c) turns out to be univalued, while v^c is uniquely determined almost everywhere. More precisely we have the following proposition.

PROPOSITION A.1. *The relation* (D) *defines a Lipschitz-continuous map* (t^c,u^c) *on* $[0,1]$, *and the derivative* $(t^{c\prime},u^{c\prime})$, *which exists almost everywhere, has constant norm equal to* $V_0^1(t,u)$. *Moreover* (D) *defines a univalued Borel-measurable map* v^c *almost everywhere in* $[0,1]$.

Thanks to Proposition 1 we can give the notion of *canonical parametrization.*

DEFINITION A.1. *The space–time control* (t^c,u^c,v^c) *defined by relation* (D) *is called the* canonical parametrization *of* (t,u,v).

DEFINITION A.2. *Let* (t_1,u_1,v_1), (t_2,u_2,v_2) *be two space–time controls and let* (t_1^c,u_1^c,v_1^c), (t_2^c,u_2^c,v_2^c) *be the corresponding canonical parametrizations. The space–time control* (t_1,u_1,v_1) *is called* equivalent to (t_2,u_2,v_2) *if* $(t_1^c,u_1^c)(s) = (t_2^c,u_2^c)(s)$ $\forall s \in [0,1]$ *and* $v_1^c(s) = v_2^c(s)$ *for almost every* s *in* $[0,1]$.

Proposition A.2 below illustrates the relationship between the trajectory $y[t,u,v]$ corresponding to a space–time control (t,u,v) and the trajectory $y[t^c,u^c,v^c]$ corresponding to the canonical parametrization (t^c,u^c,v^c) of (t,u,v).

PROPOSITION A.2. *Fix the initial condition* $y(0) = (t_1,x_1,u_1)$. *Then the trajectories* $y[t,u,v]$, $y[t^c,u^c,v^c]$ *satisfy the relation*

$$y[t^c,u^c,v^c](\xi) = y[t,u,v]\,(\sigma^{\leftarrow}(\{\xi\}))$$

for every $\xi \in [0,1]$.

A pseudometric for space–time controls. The notion of canonical parametrization allows us to introduce a pseudometric δ^c on the space $\Gamma(\bar t,\bar u)$ of space–time controls. For every two space–time controls (t_1,u_1,v_1), (t_2,u_2,v_2) let us set

$$\delta^c\big((t_1,u_1,v_1),(t_2,u_2,v_2)\big) \doteq \|(t^c{}_1,u^c{}_1) - (t^c{}_2,u^c{}_2)\| + \|v_1^c - v_2^c\|_1,$$

where $\|\cdot\|$ and $\|\cdot\|_1$ denote the C^0 norm and the L^1 norm, respectively. In particular two space–time controls have δ^c pseudodistance equal to zero if and only if they are equivalent, so δ^c induces a metric on the quotient space.

The following density result was proved in [32].

PROPOSITION A.3. *Any set* $\Gamma^+_{K-\bar k}(\bar t,\bar u)$ *of regular controls is dense, in the topology induced by* δ^c, *in the corresponding set* $\Gamma_{K-\bar k}(\bar t,\bar u)$ *of space–time controls.*

Acknowledgments. The authors wish to thank Martino Bardi and Pierpaolo Soravia for their very useful bibliographical suggestions.

REFERENCES

[1] J. P. AUBIN AND A. CELLINA, *Differential Inclusion*, Springer-Verlag, Berlin, 1984.

[2] M. BARDI AND P. SORAVIA, *A comparison result for Hamilton–Jacobi equations and applications to some differential games lacking controllability*, Funkcial. Ekvac., 37 (1994), pp. 19–43.

[3] G. BARLES, *Deterministic impulse control problems*, SIAM J. Control Optim., 23 (1985), pp. 419–432.

[4] ———, *Uniqueness and regularity results for first–order Hamilton–Jacobi equations*, Indiana Math. J., 39 (1990), pp. 443–466.

[5] ———, *An approach of deterministic control problems with unbounded data*, Ann. Inst. Henri Poincaré Anal Non Linéaire, 7 (1990), pp. 235–258.

[6] E. N. BARRON AND R.JENSEN, *Optimal control problems with no turning back*, J. Differential Equations, 36 (1980), pp. 223–248.

[7] ———, *The Pontrjagin maximum principle from dynamic programming and viscosity solutions to first order partial differential equations*, Trans. Amer. Math. Soc., 298 (1986), pp. 635–641.

[8] E. N. BARRON, R. JENSEN, AND J. L. MENALDI, *Optimal control and differential games with measures*, Nonlinear Anal., 21 (1993), pp. 241–268.

[9] A. BRESSAN, *On differential systems with impulsive controls*, Rend. Sem. Mat. Univ. Padova, 78 (1987), pp. 227–236.

[10] A. BRESSAN AND F. RAMPAZZO, *On differential systems with vector–valued impulsive controls*, Boll. Un. Mat. Ital. B(7), 2 (1988), pp. 641–656.

[11] ———, *Impulsive control systems with commutative vector fields*, J. Optim. Theory Appl., 71 (1991), pp. 67–83.

[12] ———, *Impulsive control systems without commutativity assumption*, J. Optim. Theory Appl., to appear.

[13] A. BRESSAN, *Hyperimpulsive motions and controllizable coordinates for Lagrangean systems*, Atti Accad. Naz. Lincei, Mem. Cl. Sc. Fis. Mat. Natur., 19 (1991), pp. 195–246.

[14] ———, *On some control problem concerning the ski and the swing*, Atti Accad. Naz. Lincei Mem. Cl. Sc. Fis. Mat. Natur., Series IX, 1 (1991), pp. 149–196.

[15] A. BRESSAN AND M. MOTTA, *Some optimization problems with a monotone impulsive character. Approximation by means of structural discontinuities*, Mem. Mat. Acc. Lincei, Series IX, 2 (1994), pp. 31–52.

[16] I. CAPUZZO-DOLCETTA AND P. L. LIONS, *Hamilton–Jacobi equations and state constrained problems*, Trans. Amer. Math. Soc., 318 (1990), pp. 643–668.

[17] C. W. CLARK, F. H. CLARKE, AND G. R. MUNRO, *The optimal exploitation of renewable resource stocks*, Econometrica, 48 (1979), pp. 25–47.

[18] M. G. CRANDALL, L. C. EVANS, AND P. L. LIONS, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc. (1984), p. 282.

[19] G. DAL MASO AND F. RAMPAZZO, *On systems of ordinary differential equations with measures as controls*, Differential Integral Equations, 4 (1991), pp. 739–765.

[20] J. R. DORROH AND G. FERREYRA, *A multi–state, multi–control problem with unbounded controls*, SIAM J. Control Optim., 32 (1994), pp. 1322–1331.

[21] V. A. DYKHTA, *Impulse trajectory extension of degenerated optimal control problems*, in The Liapunov Functions Methods and Applications, P. Borne and V. Matrosov, eds., J. C. Baltzer AG, 1990, pp. 103–109.

[22] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer–Verlag, New York, 1993.

[23] O. HÁJEK, Book review, Bull. Amer. Math. Soc., 12 (1985), pp. 272–279.

[24] H. ISHII, *A boundary value problem of the Dirichlet type for Hamilton–Jacobi equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., (4) 16 (1989), pp. 105–135.

[25] D. F.LAWDEN, *Optimal Trajectories for Space Navigation*, Butterworth, London, 1963.

[26] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, SIAM Series in Applied Mathematics, John Wiley, New York, London, Sydney, 1977.

[27] P. L. LIONS, *Generalized Solutions of Hamilton–Jacobi Equations*, Pitman, London, 1982.

[28] W. LIU AND H. J. SUSSMANN, *Limit of hight oscillatory controls and the approximation of general paths by admissible trajectories*, Proc. C.D.C. IEEE, 1991.

[29] J. P. MAREC, *Optimal Space Trajectories*, Elsevier, Amsterdam Oxford, 1979.

[30] B. M. MILLER, *Optimization of dynamic systems with a generalized control*, translated from

Avtomat. i Telemekh., 6 (1989), pp. 23–34.

[31] B. M. MILLER, *Conditions for the optimality in problems of generalized control.* I. *Necessary conditions for optimality*, translated from Avtomat. i Telemekh., 3 (1992), pp. 50–58.

[32] M. MOTTA AND F. RAMPAZZO, *Space–time trajectories of nonlinear systems driven by ordinary and impulsive controls*, Differential Integral Equations, 8 (1995), pp. 269–288.

[33] L. W. NEUSTADT, *A general theory of minimum–fuel space trajectories*, SIAM J. Control, 3 (1965), pp. 317–356.

[34] F. RAMPAZZO, *Optimal impulsive controls with a constraint on the total variation*, in New Trends in Systems Theory, Progress in Systems and Control Theory, G. Conte, A. M. Perdon, and B. F. Wyman, eds., Springer-Verlag, Boston, MA, 1990, pp. 606–613.

[35] ———, *On the Riemannian structure of a Lagrangian system and the problem of adding time–dependent constraints as controls*, European J. Mech. A Solids, 10 (1991), pp. 405–431.

[36] R. W. RISHEL, *An extended Pontryagin principle for control systems whose control laws contain measures*, SIAM J. Control, 3 (1965), pp. 191–205.

[37] A. V. SARYCHEV, *Nonlinear systems with impulsive and generalized function controls*, Proc. Conf. on Nonlinear Synthesis, Sopron, Hungary, 1989.

[38] W. W. SCHMAEDEKE, *Optimal control theory for nonlinear differential equations containing measures*, SIAM J. Control, 3 (1965), pp. 231–279.

[39] S. P. SETHI, *Dynamic optimal control problems in advertising: a survey*, SIAM Rev., 19 (1977), pp. 685–725.

[40] H. M. SONER, *Optimal control with state–space constraints*, SIAM J. Control Optim., 24 (1986), pp. 552–561, 1110–1122.

[41] H. J. SUSSMANN, *On the gap between deterministic and stochastic ordinary differential equations*, Ann. Probab., 6 (1978), pp. 17–41.

[42] R. B. VINTER AND M. F. L. PEREIRA, *A maximum principle for optimal processes with discontinuous trajectories*, SIAM J. Control Optim., 26 (1988), pp. 205–229.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 226–243, January 1996

010

ASYMPTOTIC STABILITY OF THE OPTIMAL FILTER WITH RESPECT TO ITS INITIAL CONDITION*

DANIEL OCONE† AND ETIENNE PARDOUX‡

Abstract. Consider the problem of estimation of a diffusion signal observed in additive white noise. If the solution to the filtering equations, initialized with an incorrect prior distribution, approaches the true conditional distribution asymptotically in time, then the filter is said to be asymptotically stable with respect to perturbations of the initial condition. This paper presents asymptotic stability results for linear filtering problems and for signals with limiting ergodic behavior. For the linear case, stability of the Riccati equation of Kalman filtering is used to derive almost sure asymptotic stability of linear filters for possibly non-Gaussian initial conditions. In the nonlinear case, asymptotic stability in a weak convergence sense is shown for filters of signal diffusions which converge in law to an invariant distribution.

Key words. nonlinear filtering, asymptotic stability, ergodicity in filtering, forgetting of initial conditions

AMS subject classifications. 93E11, 60G35

1. Introduction.

Let $X = (X_t)_{t\geq 0}$ be a Markov process taking values in $\mathbb{R}^d$, W be an $\mathbb{R}^p$-valued Brownian motion independent of X, and h be a function $h : \mathbb{R}^d \to \mathbb{R}^p$, and set

$$Y_t = \int_0^t h(X_s)\, ds + W_t, \quad t \geq 0. \tag{1}$$

Supposing that we can observe only Y but wish to know X, we would like to compute the conditional distribution

$$\pi_t(A) = E[\mathbf{1}_A(X_t)/\mathcal{Y}_t]$$

of X_t given $\mathcal{Y}_t = \sigma\{Y_s | 0 \leq s \leq t\}$. This is the classical problem of filtering a signal in independent, additive white noise, and we shall call π_t the *exact* filter.

The computation of $(\pi_t)_{t\geq 0}$ is Bayesian in character. Let π_0 denote the distribution of X_0. Knowledge of π_0 and the transition probability laws of X fixes the prior law of X, while π_t is the a posteriori law of X_t based on observing $\{Y_s,\ s \leq t\}$. In this paper, we discuss the sensitivity of the calculation of π_t to errors in the choice of π_0. Suppose that we mistakenly think that another probability measure $\overline{\pi}_0 \neq \pi_0$ is the initial distribution and that we compute a corresponding filter $\overline{\pi}_t$ using this erroneous prior. We want to know how well the *erroneous filter* $\overline{\pi}_t$ will perform as the time t gets large. In particular, it is interesting to find conditions under which

$$\overline{\pi}_t - \pi_t \text{ tends to zero in some sense as } t \to \infty\ . \tag{2}$$

In other words, for large time intervals, the strength of the observations, or the ergodic properties of X itself, correct for erroneous initial conditions in the computation of

* Received by the editors October 8, 1993; accepted for publication (in revised form) August 30, 1994.

† Department of Mathematics, Rutgers University, New Brunswick, NJ 08903. This author's work was done in part while visiting the Université de Provence, Marseille, France, where the author was supported by a visiting professorship in the Mathematics Department.

‡ LATP-CMI, Centre de Mathématiques et d'Informatique, Université de Provence, 39 rue F. Joliot-Curie, F-13453 Marseille cedex 13, France.

a posteriori distributions. We shall describe a situation in which a property like (2) holds by saying that the filter is asymptotically stable with respect to perturbations of the initial conditions. Of course, this is not a mathematically precise definition, and we shall use the term *asymptotically stable* only in this heuristic sense. The precise sense of convergence that can be achieved in (2) will depend on the filtering model and the techniques used.

The Kushner–Stratonovich equation provides a dynamical system interpretation of our stability question. Let L denote the infinitesimal generator of X, and for a measure μ on $\mathbb{R}^d$ and $\phi : \mathbb{R}^d \to \mathbb{R}$, let $\mu(\phi) = \int \phi(x)\, d\mu(x)$. Let $\mathcal{P}(\mathbb{R}^d)$ denote the space of probability measures on $\mathbb{R}^d$. Then, given a probability measure ν on $\mathbb{R}^d$, a process $(\rho_t)_{t\geq 0}$ taking values in $\mathcal{P}(\mathbb{R}^d)$ satisfies the Kushner–Stratonovich equation with initial condition ν if $(\rho_t)_{t\geq 0}$ is adapted to Y and if it satisfies

$$\rho_t(\phi) = \nu(\phi) + \int_0^t \rho_s(L\phi)\, ds + \int_0^t [\rho_s(h\phi) - \rho_s(h)\rho_s(\phi)]\, [dY_s - \rho_s(h)ds] \tag{3}$$

$$\forall\, \phi \in \text{ Domain } (L).$$

When π_0 is the initial distribution of X, $(\pi_t)_{t\geq 0}$ will be the unique solution of (3) with initial condition π_0 once appropriate regularity assumptions are satisfied (see [9]). Given an erroneous initial condition $\overline{\pi}_0$, the erroneous filter $\overline{\pi}_t$ is the solution of (3) with initial condition $\bar{\pi}_0$. Asymptotic stability means roughly that the solution of (3) forgets its initial condition and acquires an asymptotic behavior determined solely by the observation process Y and the generator L.

When X satisfies a linear stochastic differential equation with a Gaussian initial law and h is linear, the Kalman filtering equations compute the exact filter π_t. The study of the asymptotic stability of the Kalman filter has been of course a central feature of Kalman filtering theory from its inception (Kalman and Bucy [6]), and the insensitivity of Kalman filters to initial conditions, given simple controllability and detectability assumptions on the signal-observation dynamics, is of fundamental applied importance. (See Kwakernaak and Sivan [10] for an expository treatment; we don't even try to reference the huge technical literature on this point.) Usually, this stability is stated in terms of stability in law; that is, the erroneous filter $\overline{\pi}_t$ performs as well in the limit as the exact filter and has the same limiting law. It is interesting to note that these results do not require asymptotic ergodic behavior of the signal itself, which may be transient.

Beyond the linear case, general qualitative results on asymptotic filter stability with respect to initial condition perturbation are few. Stettner [15] provides conditions under which the process $(\overline{\pi}_t,\, X_t)_{t\geq 0}$ is ergodic, from which strong stability properties in the sense of (2) follow immediately, in the special case of a discrete-time, discrete-state-space model. Delyon and Zeitouni [2] prove almost sure (a.s.) exponential decay of the total variation $\|\pi_t - \overline{\pi}_t\|$ for filtering an ergodic, finite-state, continuous-time Markov chain whose observations are sufficiently rich. Closely related and, as we shall see, very relevant studies have been carried out by Kunita [7], [8] and Stettner [14] on the behavior of filters when the Markov transition semigroup of the signal is ergodic so that the signal itself forgets its initial conditions. Roughly, they show that when the signal converges in law to an invariant measure independent of the initial law, the exact filter $(\pi_t)_{t\geq 0}$ inherits a similar property. Namely, $(\pi_t)_{t\geq 0}$, itself a Markov process with state space $\mathcal{P}(\mathbb{R}^d)$, will also converge in law to an invariant distribution $M \in \mathcal{P}(\mathcal{P}(\mathbb{R}^d))$, where M is independent of the initial distribution. The filter $(\overline{\pi}_t)_{t\geq 0}$

computed with an incorrect initial condition is not a Markov process, but the pair $(\overline{\pi}_t, X_t)_{t\geq 0}$ is.

This paper presents several results on the asymptotic stability question. The purpose of §2 is to state the strongest possible results in the case of linear system and observation dynamics so as to establish a standard for evaluating progress in nonlinear problems. The results of §2 seem largely known, but it is difficult to find the precise statements we make, and so we have provided proofs. In particular, we emphasize the a.s. weak convergence of $\overline{\pi}_t - \pi_t$ to 0 and the case of non-Gaussian initial conditions, handled by Makowski's [11] reference probability transformation technique. Makowski and Sowers [12] state discrete-time versions of this result. Delyon and Zeitouni [2] also derive a.s. convergence results using Lyapunov exponent techniques.

In §3, we study the ergodic case of Kunita and Stettner. We show that when the signal tends in law to a unique invariant measure independent of the initial law and when $\pi_0 \ll \overline{\pi}_0$, then

$$\lim_{t\to\infty} E|\bar{\pi}_t(\phi) - \pi_t(\phi)| = 0$$

for any bounded continuous ϕ. Briefly, the Stettner–Kunita theory allows one to choose time lengths T so that the conditional expectation

$$E[\varphi(X_t) \ / \ Y_s - Y_{t-T} \ , \ t - T \leq s \leq t],$$

based on the observation of Y for the past T units of time, approximates $\pi_t(\phi)$ to within arbitrary accuracy, uniformly in t, and similarly for $\bar{\pi}_t(\phi)$. These approximations depend only on the observations and on the prior distribution of X_{t-T} under π_0 and $\overline{\pi_0}$, respectively. Since X_{t-T} converges to an invariant distribution independent of π_0 or $\overline{\pi}_0$ as $t \to \infty$, filter stability follows.

Roughly paraphrased, the result of §3 says that if the signal forgets its initial condition, then so does the filter. But we know from the linear case that ergodicity of the signal is not necessary for asymptotic filter stability. In future work we hope to obtain asymptotic filter stability with respect to initial condition perturbations for signals which have nonlinear dynamics and are not necessarily ergodic.

2. Asymptotic stability of the time-invariant Kalman filter. In this section we shall study the filtering problem defined by the linear system

$$dX_t = BX_t\, dt + F dV_t + G dY_t, \tag{4}$$

$$dY_t = HX_t\, dt + dW_t, \quad Y_0 = 0, \tag{5}$$

where $B \in \mathbb{R}^{d\times d}$, $F \in \mathbb{R}^{d\times q}$, $G \in \mathbb{R}^{d\times p}$, $H \in \mathbb{R}^{p\times d}$, and V and W are independent standard Brownian motions taking values in $\mathbb{R}^q$ and $\mathbb{R}^p$, respectively. X_0 is assumed to be a random vector independent of (V, W). Throughout, we shall let $\hat{X}_t$ denote the conditional expectation $E[X_t/\mathcal{Y}_t]$, where $\mathcal{Y}_t = \sigma\{Y_s,\, 0 \leq s \leq t\}$.

We shall let $(Z_t^{z,R},\, P_t^R)$ denote the solution to the Kalman filtering equations initialized at (z, R), where $z \in \mathbb{R}^d$ and R is a symmetric, nonnegative definite, $d \times d$ matrix:

$$\begin{aligned} dZ_t^{z,R} &= BZ_t^{z,R}\, dt + G dY_t + P_t^R H^*[dY_t - HZ_t^{z,R}\, dt], \\ Z_0^{z,R} &= z, \end{aligned} \tag{6}$$

$$\dot{P}_t^R = BP_t^R + P_t^R B^* + FF^* - P_t^R H^* H P_t^R,$$

(7)

$$P_0^R = R.$$

If X_0 is normal with mean $\bar{x}_0$ and variance R_0,

$$\hat{X}_t = Z_t^{\bar{x}_0, R_0}, \quad P_t^{R_0} = E[(X_t - \hat{X}_t)\,(X_t - \hat{X}_t)^*],$$

as is well known.

We shall be concerned with the asymptotic behavior of $\hat{X}_t - Z_t^{z,R}$ as $t \to \infty$, given an arbitrary, square-integrable, initial state X_0. The study of the asymptotic stability of the Kalman filtering equations (6)–(7) is classical and dates to the origins of the Kalman filter theory. Here we shall state a few complements to the theory relating to the question of a.s. convergence of filters. These seem essentially known, but we have not found a good reference, especially in the case of non-Gaussian initial conditions for continuous-time problems.

The classical stability theory introduces the fundamental assumption:

There exists a solution $P_\infty \geq 0$ to the algebraic Riccati equation

(8)
$$0 = BP_\infty + P_\infty B^* + FF^* - P_\infty H^* H P_\infty$$

such that $B - P_\infty H^* H$ is asymptotically stable.

Remark 2.1. If (8) holds and if $P_t^R \to P_\infty$ as $t \to \infty$, then

(9)
$$P_t^R \to P_\infty \text{ as } t \to \infty \text{ exponentially fast.}$$

In fact, for any $0 < \sigma < \inf\{-Re\lambda;\ \lambda \text{ is an eigenvalue of } B - P_\infty H^* H\}$, there is a constant K_σ such that

$$\|P_t^R - P_\infty\| \leq K_\sigma e^{-\sigma t}.$$

The last fact can be proved by observing that

$$\frac{d}{dt}(P_t^R - P_\infty) = [B - 1/2(P_t^R + P_\infty)H^* H](P_t^R - P_\infty) + (P_t^R - P_\infty)[B - 1/2(P_t^R + P_\infty)H^* H]^*$$

and carrying out an analysis similar to that proving (16) in the proof of Theorem 2.3.

The virtue of (8) is that there is a well-known and simple sufficient condition for (8) to hold (see Kwakernaak and Sivan [10]), and it has strong consequences.

LEMMA 2.2. *If (B, H) is detectable and (B, F) is stabilizable, then* (8) *holds, P_∞ is the unique nonnegative definite solution to the algebraic Riccati equation, and $P_t^R \to P_\infty$ exponentially fast for any initial condition $P_0^R = R \geq 0$.*

Proof. Kwakernaak and Sivan [10, Thm. 4.11] may be referred to for the uniqueness of P_∞ and the convergence of $P_t^R \to \infty$. The exponential speed of convergence is shown above. □

Some classical stability statements that follow from this result are as follows.

(a) If (B, H) is detectable, (B, F) is stabilizable, and X_0 is normal with mean z and variance $\sum_0$, the *steady state filter* Z_t^{z, P_∞} is asymptotically optimal in the sense that

$$\lim_{t\to\infty} E[(X_t - Z_t^{z,P_\infty})^*(X_t - Z_t^{z,P_\infty})] = \lim_{t\to\infty} E[(X_t - \hat{X}_t)^*(X_t - \hat{X}_t)] = \operatorname{tr} P_\infty.$$

(b) If, again, (B, H) is detectable and (B, F) is stabilizable but the initial distribution π_0 of X_0 is arbitrary, then

$$Z_t^{0,R} - X_t \text{ converges weakly to } N(0, P_\infty), \tag{10}$$

where $N(0, P_\infty)$ denotes the normal law on $\mathbb{R}^d$ with mean 0 and covariance matrix P_∞. See Vintner [16] for a proof of (10) in an infinite-dimensional state space.

We shall state here some a.s. limit theorems. The first is a rather immediate consequence of condition (8) and Remark 2.1. Define the constant $\bar{\lambda} = \min\{-\mathrm{Re}\lambda | \lambda$ is an eigenvalue of $B - P_\infty H^* H\}$; condition (8) implies that $\bar{\lambda} > 0$.

THEOREM 2.3. *Assume* (8). *Let* X_0 *be normal with mean* m_0 *and covariance* R_0. *Let* $R_1 \geq 0$, *and assume*

$$\lim_{t\to\infty} P_t^{R_0} = P_\infty = \lim_{t\to\infty} P_t^{R_1}. \tag{11}$$

Then for any $z \in \mathbb{R}^d$ *and any* $0 < \sigma < \bar{\lambda}$

$$\lim_{t\to\infty} (\hat{X}_t - Z_t^{z,R_1})\, e^{\sigma t} = 0 \quad \textit{almost surely.} \tag{12}$$

Proof. Recall that $\hat{X}_t = Z_t^{m_0,R_0}$ and that $P_t^{R_0}$ is the error covariance of the optimal filter. For simplicity of notation, set $P_t^0 = P_t^{R_0}$, $P_t^1 = P_t^{R_1}$, and $Z_t = Z_t^{z,R_1}$. Let $d\nu_t = dY_t - H\hat{X}_t\, dt$ define the innovations process, which is a Brownian motion, and observe that

$$d\hat{X}_t = B\hat{X}_t\, dt + G dY_t + P_t^0 H^* d\nu_t, \quad \hat{X}_0 = m_0. \tag{13}$$

From equation (6) for Z_t and from (13)

$$d(\hat{X}_t - Z_t) = (B - P_t^1 H^* H)(\hat{X}_t - Z_t)\, dt + (P_t^0 - P_t^1) H^*\, d\nu_t,$$

$$\hat{X}_0 - Z_0 = m_0 - z. \tag{14}$$

Let $\Phi(t) \in \mathbb{R}^{d\times d}$ be the state transition matrix associated with $B - P_t^1 H^* H$; that is,

$$\dot{\Phi}(t) = (B - P_t^1 H^* H)\, \Phi(t),\ \Phi(0) = I. \tag{15}$$

The proof of Theorem 2.3 is a consequence of the following facts: for every $0 < \sigma < \bar{\lambda}$ there exist an $M_\sigma < \infty$ and a $t_\sigma < \infty$ such that

$$\|\Phi(t)\, \Phi^{-1}(s)\| \leq M_\sigma e^{-\sigma(t-s)} \quad \text{for } t > s \geq t_\sigma, \tag{16}$$

and

$$\|P_t^0 - P_t^1\| \leq M_\sigma e^{-\sigma t} \quad \text{for all } t. \tag{17}$$

We have already seen (17) in Remark 2.1. (16) is very natural in view of the fact that $B - P_t^1 H^* H \to B - P_\infty H^* H$ as $t \to \infty$, and we return to its proof in a moment.

Now from (14) we have

$$\hat{X}_t - Z_t = \Phi(t)(m_0 - z) + \int_0^t \Phi(t)\, \Phi^{-1}(s)[P_s^0 - P_s^1] H^*\, d\nu_s. \tag{18}$$

Clearly (16) implies

$$\lim_{t\to\infty} e^{\sigma t}\,\Phi(t)(m_0 - z) = 0 \quad \text{if} \quad \sigma < \bar{\lambda}. \tag{19}$$

Because ν is Brownian, if $\sigma < \bar{\lambda}$ there is a constant $K_\sigma < \infty$ such that

$$\begin{aligned} &E\left|\int_0^t \Phi(t)\Phi^{-1}(s)[P_s^0 - P_s^1]H^*\,d\nu_s\right|^2 \\ &= \int_0^t \operatorname{tr}\left[\Phi(t)\Phi^{-1}(s)\,[P_s^0 - P_s^1]\,H^*H\,[P_s^0 - P_s^1]\,\Phi^{-1}(s)^*\Phi(t)^*\right] ds \\ &\le K_\sigma\, e^{-2\sigma t}, \end{aligned} \tag{20}$$

where we have used (16) and (17) to derive the last estimate. Also, (19) and (20) show that for any $0 < \sigma < \bar{\lambda}$ there is a K_σ such that

$$E[|\hat{X}_t - Z_t|^2] \le K_\sigma\, e^{-2\sigma t}. \tag{21}$$

By applying the Borel–Cantelli lemma to $\hat{X}_n - Z_n$ and

$$\sup_{n\le t\le n+1} |\hat{X}_t - Z_t - (\hat{X}_n - Z_n)|,$$

we obtain from (21) that

$$\lim_{t\to\infty} |\hat{X}_t - Z_t|\, e^{\sigma t} = 0 \quad \text{almost surely}$$

for any $0 < \sigma < \hat{\lambda}$, which was to be proved.

Finally, we consider (16). This is a consequence of the estimate

$$\|\psi(t)\,\psi^{-1}(s)\| \le M\, e^{-\beta(t-s)}$$

for the solution to $\dot{\psi}(t) = A(t)\psi(t)$, $\psi(0) = I$, where

$$\begin{aligned} &\|e^{A(t)\tau}\| \le Ke^{\gamma\tau} \quad \forall\, t \ge 0,\ \forall \tau \ge 0, \\ &\|\dot{A}(t)\| \le \delta \quad \forall\, t, \end{aligned}$$

and $0 < \beta = \gamma - (\delta K \log K)^{1/2}$, $M = K^2$; see Harris and Miles [5, Thm. 5.10, p. 146]. Apply this to $A(t) = B - P_t^1 H^*H$, using the fact $\dot{P}_t^1 \to 0$ as $t \to \infty$ and the fact that the eigenvalues of $B - P_t^1 H^*H$ will have real parts less than $-\sigma$ for all large enough t, to derive (16). □

We can deduce from Theorem 2.3 a statement about the convergence of probability measures. For a vector m and matrix $Q \ge 0$ let $N(m, Q)$ denote the Gaussian distribution with mean m and covariance Q. We shall write

$$N(m, Q)(\varphi) := \int \varphi(x)\, N(m, Q)\,(dx).$$

In the context of Theorem 2.3, if X_0 has the distribution $N(m_0, R_0)$, then $\pi_t = N(\hat{X}_t, P_t^{R_0})$ is the conditional distribution of X_t given $\mathcal{Y}_t$. On the other hand

$$\bar{\pi}_t := N(Z_t^{z,R_1}, P_t^{R_1})$$

would be the conditional distribution that we would think we were getting if we had started with the wrong initial conditions z and R_1. $\bar{\pi}_t$ solves the Kushner–Stratonovich equation for the conditional distribution starting from the initial condition $N(z, R_1)$.

For a continuous function f on $\mathbb{R}^d$, let

$$||f||_{BL} := \sup_{x\in\mathbb{R}^d} |f(x)| + \sup_{x,y\in\mathbb{R}^d} |x-y|^{-1}|f(x)-f(y)|,$$

and define the metric

$$\beta(\mu,\nu) := \sup\left\{\int f\,(d\mu - d\nu) \;:\; ||f||_{BL} \le 1\right\}$$

on $\mathcal{P}(\mathbb{R}^d)$. It is well known that β metrizes the topology of weak convergence on $\mathcal{P}(\mathbb{R}^d)$. It is easy to prove the following lemma.

LEMMA 2.4. *If* $m_t - m'_t \to 0$ *as* $t \to \infty$ *and if* $\lim_{t\to\infty} Q_t = Q_\infty = \lim_{t\to\infty} Q'_t$, *then*

$$\lim_{t\to\infty} ||N(m_t, Q_t) - N(m'_t, Q'_t)||_{BL} = 0.$$

Theorem 2.3 and Lemma 2.4 immediately imply the following.

COROLLARY 2.5. *Let the assumptions of Theorem* 2.3 *hold. Then*

$$\lim_{t\to\infty} ||\pi_t - \bar{\pi}_t||_{BL} = 0 \quad \textit{almost surely,}$$

where $\pi_t = N(\hat{X}_t, P_t^{R_0})$ *and* $\bar{\pi}_t = N(Z_t^{z,R_1}, P_t^{R_1})$.

We next want to present a similar result for the case in which the initial condition X_0 is not Gaussian. A discrete time version of this result may be found in Sowers [13] and Makowski and Sowers [12]. We shall present the convergence result under the assumption

(22) $\qquad (B, H)$ is detectable and (B, F) is stabilizable.

THEOREM 2.6. *Let* (X, Y) *denote the solution of* (4)–(5). *Assume that* (22) *holds and* $E[|X_0|^2] < \infty$. *Then*

(23) $$\lim_{t\to\infty} \hat{X}_t - Z_t^{z,R} = 0 \quad \textit{almost surely,}$$

and in the L^2 *sense for* any $z \in \mathbb{R}^d$, $R \ge 0$. *Moreover, if* π_t *denotes the conditional distribution of* X_t *given* $\mathcal{Y}_t$,

(24) $$\lim_{t\to\infty} \pi_t(\varphi) - N(Z_t^{z,R}, P_t^R)(\varphi) = 0 \quad \textit{almost surely}$$

for every bounded, uniformly continuous φ.

Proof. The proof uses a formula, due to Makowski [11] and Beneš and Karatzas [1], which expresses the optimal filter for a linear system with non-Gaussian initial conditions in terms of the solution to a Kalman filtering equation. The idea is to decompose the signal X as

(25) $$X_t = e^{Bt} X_0 + \overline{X}_t,$$

where

$$\overline{X}_t := \int_0^t e^{B(t-s)} F dV_s + \int_0^t e^{B(t-s)} G dY_s,$$

and to introduce the new measure $\overline{\mathbb{P}}$ defined by

$$\frac{d\overline{\mathbb{P}}}{d\mathbb{P}} = \exp\left[\int_0^T -\langle He^{Bs}X_0,\, dW_s\rangle - \frac{1}{2}\int_0^T |He^{Bs}X_0|^2\, ds\right].$$

Then on $(\Omega, \overline{\mathbb{P}})$

$$\overline{W}_t := \int_0^t He^{Bs}X_0\, ds + W_t, \qquad t \le T,$$

is a Brownian motion, and X_0 is independent of $(V_t, \overline{W}_t)_{t\le T}$. Moreover

$$\overline{\mathbb{P}}(X_0 \in A) = \mathbb{P}(X_0 \in A)$$

for all Borel A (i.e., the law of X_0 remains unchanged).

Let

$$L_t := \exp\left[\int_0^t \langle He^{Bs}X_0,\, d\overline{W}_s\rangle - \frac{1}{2}\int_0^t |He^{Bs}X_0|^2\, ds\right]$$

(note that $L_T^{-1} = \frac{d\overline{\mathbb{P}}}{d\mathbb{P}}$). Then for any nonnegative measurable function θ of $(X_0, \overline{X}_t)$ (or equivalently of (X_0, X_t))

$$E[\theta(X_0, \overline{X}_t)/\mathcal{Y}_t] = \frac{\overline{E}[\theta(X_0, \bar{X}_t)L_t/\mathcal{Y}_t]}{\bar{E}[L_t/\mathcal{Y}_t]}. \tag{26}$$

From the above formula for $\overline{X}_t$ and the fact that X_0 and $(\overline{W}, Y)$ are $\overline{\mathbb{P}}$-independent, both the numerator and the denominator may be expressed by an integral involving the Kalman filter for the process $(\overline{X}_t, \overline{W}_t)$, given $\mathcal{Y}$, whose equations we now write. Let (Z_t, P_t^0) denote the solution of (6)–(7) with initial conditions $(Z_0, P_t^0) = (0, 0)$. Define (S_t, Q_t, M_t) and $\widetilde{Z}_t$ by the equations

$$\dot{S}_t = BS_t - P_t^0 H^* H(e^{Bt} + S_t), \quad S_0 = 0,$$

$$\dot{Q}_t = -e^{B^* t}H^* HS_t - S_t^* H^* He^{Bt} - S_t^* H^* HS_t, \quad Q_0 = 0, \tag{27}$$

$$\dot{M}_t = e^{B^* t}H^* He^{Bt}, \quad M_0 = 0, \text{ and}$$

$$d\widetilde{Z}_t = (e^{Bt} + S_t)^* H^*(dY_t - HZ_t\, dt), \quad \widetilde{Z}_0 = 0. \tag{28}$$

We then have that

(29)

$$\overline{E}[\theta(X_0, \overline{X}_t)L_t/\mathcal{Y}_t] = \int_{\mathbb{R}^d} \pi_0(dx)\, e^{-1/2\langle M_t x, x\rangle} \left\{\int_{\mathbb{R}^{2d}} \theta(x, r_1) e^{\langle x, r_2\rangle} n_t(dr_1, dr_2)\right\},$$

where n_t is the Gauss measure with mean

$$\begin{pmatrix} Z_t \\ \widetilde{Z}_t \end{pmatrix}$$

and covariance

$$C_t = \begin{pmatrix} P_t^0 & S_t \\ S_t & Q_t \end{pmatrix}.$$

Note that

$$\int_{\mathbb{R}^{2d}} e^{i\langle\lambda,r_1\rangle+i\langle\mu,r_2\rangle+\langle x,r_2\rangle}\, n_t(dr_1\, dr_2) \tag{30}$$
$$= e^{i\langle\lambda,Z_t\rangle+i\langle\mu-ix,\widetilde{Z}_t\rangle} \times e^{-1/2\,\langle C_t\binom{\lambda}{\mu-ix},\binom{\lambda}{\mu-ix}\rangle}$$

$$= e^{1/2\langle Q_t x,x\rangle+\langle x,\widetilde{Z}_t\rangle} \times \exp\Big[i\langle\lambda, Z_t + S_t x\rangle + i\langle\mu, \widetilde{Z}_t + Q_t x\rangle \tag{31}$$
$$\cdot -1/2\left\langle C_t\begin{pmatrix}\lambda\\ \mu\end{pmatrix},\begin{pmatrix}\lambda\\ \mu\end{pmatrix}\right\rangle\Big].$$

Thus

$$\bar{n}_t(dr_1, dr_2) := e^{-1/2\langle Q_t x,x\rangle-\langle x,\widetilde{Z}_t\rangle} e^{\langle x,r_2\rangle} n_t(dr_1, dr_2)$$

is the normal distribution with mean

$$\begin{pmatrix} Z_t \\ \widetilde{Z}_t \end{pmatrix} + \begin{pmatrix} S_t \\ Q_t \end{pmatrix} x$$

and variance C_t since it has the correct characteristic functional by (30). For each t, let U_t denote a $N(0, P_t^0)$ random variable. Then by (26) and (29)

$$E[\varphi(X_t)/\mathcal{Y}_t] = E[\varphi(e^{Bt}x + \overline{X}_t)/\mathcal{Y}_t]$$

$$= \frac{\int \pi_0(dx)\, e^{1/2\langle(Q_t-M_t)x,x\rangle+\langle x,\widetilde{Z}_t\rangle} \int \varphi(e^{Bt}x + r_1)\bar{n}_t(dr_1, dr_2)}{\int \pi_0(dx)\, e^{1/2\langle(Q_t-M_t)x,x\rangle+\langle x,\widetilde{X}_t\rangle}} \tag{32}$$

$$= \frac{\int \pi_0(dx)\, e^{1/2\langle(Q_t-M_t)x,x\rangle+\langle x,\widetilde{Z}_t\rangle} E[\varphi(Z_t + (e^{Bt} + S_t)x + U_t)]}{\int \pi_0(dx)\, e^{1/2\langle(Q_t-M_t)x,x\rangle+\langle x,\widetilde{Z}_t\rangle}}.$$

In the case $\varphi(x) = x$, (32) gives

$$\hat{X}_t = Z_t + E[(e^{Bt} + S_t)X_0/\mathcal{Y}_t]. \tag{33}$$

Note that

$$\frac{d}{dt}(e^{Bt} + S_t) = (F - P_t^0 H^* H)(e^{Bt} + S_t)$$

as follows directly from equation (27) for S_t. Thus $e^{Bt} + S_t$ is the state transition matrix for $B - P_t^0 H^* H$. Since $P_t^0 \to P_\infty$ where $B - P_\infty H^* H$, the same argument as

in the proof of Theorem 2.3 (see (16)) shows that for any $0 < \sigma < \bar{\lambda} = \min\{|\mathrm{Re}\lambda|\,;\ \lambda$ is an eigenvalue of $B - P_\infty H^* H\}$ there exists M such that

$$\|(e^{Bt} + S_t)\| \le M e^{-\sigma t}. \tag{34}$$

Furthermore $E[X_0|\mathcal{Y}_t]$ is a square-integrable martingale, and hence its sample paths are bounded. Thus it follows easily from (33) and (34) that

$$\lim_{t\to\infty} \hat{X}_t - Z_t^{0,0} = 0 \quad \text{almost surely and in } L^2. \tag{35}$$

(Recall $Z_t := Z_t^{0,0}$.) To complete the proof of (23) it only remains to show that

$$\lim_{t\to\infty} Z_t - Z_t^{z,R} = 0 \quad \text{almost surely} \tag{36}$$

and in the L^2-sense for any $z \in \mathbb{R}$ and $R \ge 0$. To do this, we apply the argument of Theorem 2.3 to

$$\begin{aligned} d(Z_t - Z_t^{z,R}) &= (B - P_t^R H^* H)(Z_t - Z_t^{z,R})\, dt + (P_t^0 - P_t^R) H^*\, d\nu_t \\ &\quad + (P_t^0 - P_t^R) H^* H(\hat{X}_t - Z_t)\, dt, \end{aligned} \tag{37}$$

where ν is the innovations process, $\nu_t = Y_t - \int_0^t H\hat{X}_s\, ds$. The details are completely analogous, and we omit them. We remark that (35) is used to handle the effect of the third term in (36).

Finally we consider the proof (24). Because of (25) and Lemma 2.4 it suffices to prove

$$\lim_{t\to\infty} \pi_t(\varphi) - N_t(Z_t, P_t^0)(\varphi) = 0 \quad \text{almost surely} \tag{38}$$

for any bounded, uniformly continuous φ. By (32) and the definition of U_t.

$$\begin{aligned} &\pi_t(\varphi) - N(Z_t, P_t^0)(\varphi) \\ &\quad = \frac{\int \pi_0(x)\, e^{[1/2\langle (Q_t - M_t)x, x\rangle + \langle x, \widetilde{Z}_t\rangle]}\, E[\varphi(Z_t + U_t) - \varphi(Z_t + U_t + (e^{Bt} + S_t)x)]}{\int \pi_0\,(dx)\, \exp[1/2\langle (Q_t - M_t)x, x\rangle + \langle x, \widetilde{Z}_t\rangle]} \end{aligned} \tag{39}$$

Decompose the integral in the numerator into the sum of the integral over the region $|(e^{Bt} + S_t)x| < \delta$ and the integral over the region $|(e^{Bt} + S_t)x| \ge \delta$. Then from (38)

$$\begin{aligned} |\pi_t(\varphi) - N_t(Z_t, P_t^0)(\varphi)| &\le \sup_{|y-y'|<\delta} |\varphi(y) - \varphi(y')| \\ &\quad + 2\|\varphi\|_\infty\, E\{\mathbf{1}_{\{|(e^{Ft} + S_t)X_0| \ge \delta\}} | \mathcal{Y}_t\} \\ &\le \sup_{|y-y'|<\delta} |\varphi(y) - \varphi(y')| + \frac{2\|\varphi\|_\infty}{\delta^2}\, \|e^{Bt} + S_t\|^2\, E[|X_0|^2 | \mathcal{Y}_t]. \end{aligned}$$

By first letting $t \to \infty$ and then using uniform continuity, we arrive at (38) and thus complete the proof. □

Remark 2.7. By being a bit more careful, we can prove, as in Theorem 2.3, that

$$\lim_{t\to\infty} e^{\sigma t}[\hat{X}_t - Z_t^{z,R}] = 0 \quad \text{almost surely}$$

for any $0 < \sigma < \bar{\lambda}$ under the hypotheses of Theorem 2.6.

3. Stability in the case of signal ergodicity. In this section, we prove stability of filters for a class of signals which themselves forget their initial condition in the sense that they converge in law to a unique invariant measure. A result of Stettner [14] and Kunita [7] shows that the conditional distribution $(\pi_t)_{t\geq 0}$ inherits a similar ergodic property, and we use this to obtain stability.

3.1. Filtering model. Rather than model the signal X directly by a stochastic differential equation, we begin more abstractly with a locally compact, complete separable metric space E and a Markov semigroup $(S_t)_{t\geq 0}$ on $C_b(E)$ specifying the transition laws of X. (Note that S_t was used in the previous section to denote part of a covariance matrix; we change that notation from here on.) More precisely, we suppose throughout that $(S_t)_{t\geq 0}$ is a strongly continuous, positive, and conservative ($S_t 1 = 1$) contraction semigroup on $C_b(E)$. We shall assume that Markov processes associated to $(S_t)_{t\geq 0}$ admit càdlàg sample paths. That is, if X denotes the canonical process on the Skorohod space $D([0,\infty) : E)$, we suppose that for each $x \in E$ there is a probability measure P_x on $D([0,\infty) : E)$ for which X is a Markov process with transition semigroup $(S_t)_{t\geq 0}$ and $P_x[X_0 = x] = 1$. In addition, we assume throughout that $x \to P_x(A)$ is measurable for all Borel sets in $D([0,\infty) : E)$ for the topology of uniform convergence on compact time intervals. The measure

$$P^\nu(A) = \int_E P_x(A)\,\nu(dx), \quad A \in \sigma(X_s,\ 0 \leq s < \infty),$$

where $\nu \in \mathcal{P}(E)$, then defines the law of the process corresponding to $(S_t)_{t\geq 0}$ with initial condition ν.

We note for future use the following consequence of the Feller assumption. Let $\{\nu^n\}$ be a sequence of probability laws on E converging weakly to the law ν and $\{X^n\}$ and X denote the Markov processes associated to $(S_t)_{t\geq 0}$ with respective initial laws $\{\nu^n\}$ and ν. Then

$$X^n \Rightarrow X \tag{40}$$

as $n \to \infty$, where $\Rightarrow$ denotes weak convergence of X^n to X in $D([0,\infty) : E)$. Corollary 3.3.2 in Ethier and Kurtz [3] says that to show this it is enough to show that X^n converges weakly to X in $D([0,\infty); E^\triangle)$, where $E^\triangle$ is the one-point compactification of E. However, Theorems 3.9.4 and 3.9.1 in [3] and the fact that the generator of $(S_t)_{t\geq 0}$ is dense in $C_b(E)$ imply that X^n is relatively compact in $D([0,\infty); E^\triangle)$, and the Feller property implies that $(X^n(t_1),\ldots,X^n(t_k)) \Rightarrow (X(t_1),\ldots,X(t_k))$ for any finite set of nonnegative times $t_1,\ldots,t_k$, just as in the proof of Theorem 4.2.5 of [3]. These two facts imply the desired weak convergence (see Theorem 3.7.8 in [3]).

Our filtering model is then specified by the signal-observation pair (X, Y) defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as follows:

$$X = (X_t)_{t\geq 0} \text{ is a càdlàg, } E\text{-valued Markov process with law } P^{\pi_0}; \tag{41}$$

$$Y_t = \int_0^t h(X_s)\,ds + W_t, \tag{42}$$

where W is an $\mathbb{R}^p$-valued Brownian motion independent of X and

$$h : E \to \mathbb{R}^p \text{ is bounded and continuous.} \tag{43}$$

In the model (41)–(42), note that π_0 denotes the initial distribution of X. As usual, π_t shall denote the conditional distribution of X_t given $\mathcal{Y}_t = \sigma(Y_s,\ 0 \le s \le t)$.

It is convenient to work throughout on the canonical probability space $\Omega = D([0,\infty) : E) \times C_0([0,\infty) : \mathbb{R}^p)$ for the signal-observation pair. Henceforth, we let (X, Y) denote the canonical process on Ω. For a probability measure $\nu \in \mathcal{P}(E)$, let Q^ν be the measure on Ω corresponding to the filtering model (41)–(42) when ν is the probability measure of X_0; that is, the marginal of Q^ν on $D([0,\infty) : E)$ is P^ν and, on $(\Omega,\ Q^\nu)$, $Y_t - \int_0^t h(X_s)\,ds$ is a Brownian motion independent of X. Sometimes, to emphasize that we are working on $(\Omega,\ Q^\nu)$, we shall write $(X^\nu,\ Y^\nu)$ for the canonical process. We use $E^\nu[\cdot]$ to denote expectations with respect to Q^ν and $\pi^\nu = (\pi_t^\nu)_{t\ge 0}$ to denote the conditional distribution of X_t given $\sigma\{Y_s,\ s \le t\}$ on $(\Omega,\ Q^\nu)$.

This canonical formulation is useful because we shall need to consider (41)–(42) for arbitrary initial laws ν for X_0. In particular, it allows us to define a semigroup for the conditional law process. Endow $\mathcal{P}(E)$ with the topology of weak convergence. For $F \in C_b(\mathcal{P}(E))$, define

$$(M_t F)(\nu) = E^\nu[F(\pi_t^\nu)], \quad t \ge 0,\ \nu \in \mathcal{P}(E).$$

Given the assumptions above made on $(S_t)_{t\ge 0}$ and in (41)–(42), $(M_t)_{t\ge 0}$ defines a Feller, Markov transition semigroup on $C_b(\mathcal{P}(E))$; see Stettner [14].

3.2. Ergodicity assumptions. We formulate next the precise ergodicity assumptions that we require of the signal semigroup.

$(S_t)_{t\ge 0}$ admits a unique invariant measure μ and

$$\text{(H1)} \qquad \limsup_{t\to 0} \int_E |S_t f(x) - \mu(f)|\,\mu(dx) = 0 \quad \forall\, f \in C_b(E).$$

For $\nu \in \mathcal{P}(E)$, let νS_t denote the law of X_t^ν. We say that S_t forgets ν for μ if

$$\text{(H2)} \qquad \nu S_t \to \mu \text{ weakly as } t \to \infty.$$

The main point of (H1) and (H2) is Stettner's result, Theorem 3 in [14], lifting ergodic properties of $(S_t)_{t\ge 0}$ to $(M_t)_{t\ge 0}$.

LEMMA 3.1. a) *If $(S_t)_{t\ge 0}$ satisfies* (H1), *there is a unique measure M on $\mathcal{P}(E)$ such that M is $(M_t)_{t\ge 0}$-invariant and*

$$\int_{\mathcal{P}(E)} \eta(\phi)\, M(d\eta) = \mu(\phi) \quad \forall \phi \in C_b(E).$$

b) *If, in addition, ν satisfies* (H2), *then for every $F \in C_b(\mathcal{P}(E))$*

$$\lim_{t\to\infty} (M_t F)(\nu) = \int F(\eta)\, M(d\eta) =: M(F).$$

3.3. Stability result. Our stability theorem compares the correct estimate $\pi_t(\phi)$ for a fixed function ϕ to a filter $\overline{\pi}_t(\phi)$ computed with an erroneous initial condition $\overline{\pi}_0$. Rather than work with the Kushner–Stratonovich equation, we construct π_t and $\overline{\pi}_t$ by means of the Kallianpur–Striebel formula. It is convenient to do this on the canonical space $\Omega = D([0,\infty) : E) \times C_0([0,\infty) : \mathbb{R}^p)$ with the canonical process (X, Y). Let λ denote Wiener measure on $C_0([0,\infty)\,;\,\mathbb{R}^p)$ and y denote an element of $C_0([0,\infty)\,;\,\mathbb{R}^p)$.

Let

$$L_t(X,Y) = \exp\left\{\int_0^t h(X_s)\cdot dY_s - \frac{1}{2}\int_0^t |h(X_s)|^2\, ds\right\}.$$

Recall the definition of P^ν on $D([0,\infty)\,;\,E)$ from §3.1. Then we may define

$$\rho_t(\phi)(\nu\,;\,y) := \frac{E^{P^\nu\times\lambda}[\phi(X_t)\,L_t(X,Y)\;/\;Y_s,\ s\le t](y)}{E^{P^\nu\times\lambda}[L_t(X,Y)\;/\;Y_s,\ s\le t](y)},$$

where the dependence on $y\in C_0([0,\infty)\,:\,\mathbb{R}^p)$ is explicitly indicated. Then for the filtering model (X,Y) specified in (41), (42) with initial condition π_0

$$\pi_t(\phi) = \rho_t(\phi)(\pi_0\,;\,Y). \tag{44}$$

In other words, on (Ω, Q^{π_0}), $\rho_t(\phi)(\pi_0, Y)$ defines a version of the optimal filter $E^{\pi_0}[\phi(X_t)|\mathcal{Y}_t]$. Now fix $\overline{\pi}_0\in\mathcal{P}(E)$, $\overline{\pi}_0\neq\pi_0$, and define

$$\overline{\pi}_t(\phi) = \rho_t(\phi)(\overline{\pi}_0\,;\,Y). \tag{45}$$

On $(\Omega, Q^{\overline{\pi}_0})$, $\overline{\pi}_t(\phi)$ computes the optimal filter, but on (Ω, Q^{π_0}) it corresponds to a filter computed with the incorrect initial condition $\overline{\pi}_0$. (The perturbed filter $\overline{\pi}_t(\phi)$ will satisfy the Kushner–Stratonovich equation on $(\Omega, Q^{\overline{\pi}_0})$ and hence also on (Ω, Q^{π_0}) because the laws of Y^{π_0} and $Y^{\overline{\pi}_0}$ are mutually absolutely continuous up to any finite time.) We wish to assess the performance of $\overline{\pi}_t(\phi)$ on (Ω, Q^{π_0}), that is, when π_0 is the true initial measure.

Given an initial distribution ν for X_0, let R^ν denote the marginal of Q^ν on $C_0([0,\infty)\,:\,\mathbb{R}^p)$; that is, R^ν is the law of the observation process when X_0 has distribution ν.

THEOREM 3.2. *Assume*

(i) $(S_t)_{t\ge0}$ *satifies* (H1);

(ii) π_0 *and* $\overline{\pi}_0$ *both satisfy* (H2) (*that is,* (H2) *is true when* ν *is replaced by* π_0 *and* $\overline{\pi}_0$);

(iii) $R^{\pi_0}\ll R^{\overline{\pi}_0}$.

Then for every bounded, continuous $\varphi : E\to\mathbb{R}$

$$\lim_{t\to\infty} E^{\pi_0}[(\pi_t(\varphi)-\overline{\pi}_t(\varphi))^2] = 0.$$

Remark 3.3. 1. Condition (iii) says roughly that on supp R^{π_0}, it is impossible to distinguish with certainty from the entire history of the observations, whether the initial condition is π_0 or $\overline{\pi}_0$. Condition (iii) is certainly satisfied if $\pi_0\ll\overline{\pi}_0$ as one may easily see by conditioning Y^ν on $\sigma(X_0^\nu)$.

2. The choice of $E^{\overline{\pi}_0}(\pi_t(\phi)-\overline{\pi}_t(\phi))^2$ to measure the difference of $\pi_t(\phi)$ and $\overline{\pi}_t(\phi)$ is made for convenience of calculation. In fact, because $|\pi_t(\phi)|$, $|\overline{\pi}_t(\phi)|\le\|\phi\|_\infty$ almost surely for all t, $\pi_t(\phi)-\overline{\pi}_t(\phi)\to0$ in $L^p(Q^{\overline{\pi}_0})$ for all $p\ge1$.

Proof of Theorem 3.2. The proof takes several steps. The first step is to establish a uniform finite memory approximation of $\pi_t(\phi)$ and $\overline{\pi}_t(\phi)$.

Let $\mathcal{Y}_t^s = \sigma\{Y_r - Y_s\,;\,s\le r\le t\}$ be the σ-algebra of the increments of the observations on $[s,t]$. We shall call

$$\pi_{t-T,t}(\phi) := E^{\pi_0}[\phi(X_t)/\mathcal{Y}_t^{t-T}]$$

the (exact) filter of memory length T. Because of the time-homogeneity of X, computing $\pi_{t-T,t}(\phi)$ is the same as computing the filter on $[0,T]$ when the initial condition is $\pi_0 S_{t-T}$ since $\pi_0 S_{t-T}$ is the law of X_{t-T} when X_0 has distribution π_0. In other words,

$$\pi_{t-T,t}(\phi) = \rho_T(\phi)(\pi_0 S_{t-T},\, Y_{t-T+\cdot} - Y_{t-T}) \tag{46}$$

(see (44)). Likewise, we define the finite memory filter computed with the wrong initial condition:

$$\overline{\pi}_{t-T,t}(\phi) = \rho_T(\phi)(\overline{\pi}_0 S_{t-T},\, Y_{t-T+\cdot} - Y_{t-T}). \tag{47}$$

LEMMA 3.4. *Let hypotheses* (i)–(iii) *of Theorem* 3.2 *be satisfied. Then for every* $\varepsilon > 0$ *there is a* T_ε *and a* t_ε *such that*

$$E[(\pi_t(\phi) - \pi_{t-T_\varepsilon,t}(\phi))^2] < \varepsilon \quad \forall\, t \geq t_\varepsilon \tag{48}$$

and

$$E[(\overline{\pi}_t(\phi) - \overline{\pi}_{t-T_\varepsilon,t}(\phi))^2] < \varepsilon \quad \forall\, t \geq t_\varepsilon. \tag{49}$$

Remark 3.5. We obtain in (48) and (49) an estimate for a fixed-length finite memory filter uniform in t for all large t. The expectations in (48) and (49) are both evaluated assuming that π_0 is the initial law of X.

Proof. We shall first establish (48). (49) will follow, using a similar argument and the absolute continuity of R^{π_0} with respect to $R^{\overline{\pi}_0}$.

Let the continuous bounded function $F_\phi : \mathcal{P}(E) \to \mathbb{R}$ be given by $F_\phi(\eta) = \eta^2(\phi)$. Observe that

$$\begin{aligned} &E^{\pi_0}[(\pi_t(\phi) - \pi_{t-T,t}(\phi))^2] \\ &\quad = E^{\pi_0}[(\pi_t(\phi))^2] - 2E^{\pi_0}\{E^{\pi_0}[\phi(X_t)/\mathcal{Y}_t]\, E^{\pi_0}[\phi(X_t)/\mathcal{Y}_t^{t-T}]\} \\ &\qquad + E^{\pi_0}[(\pi_{t-T,t}(\phi))^2] \\ &\quad = E^{\pi_0}(\pi_t(\phi))^2 - E^{\pi_0}(\pi_{t-T,t}(\phi))^2 \\ &\quad = (M_t F_\phi)(\pi_0) - (M_T F_\phi)(\pi_0 S_{t-T}), \end{aligned} \tag{50}$$

where the last line follows from (46). Lemma 3.1 implies that there exists a T_ε such that

$$|(M_t F_\phi)(\mu) - M(F_\phi)| < \varepsilon/3 \quad \forall t \geq T_\varepsilon \tag{51}$$

and

$$|(M_t F_\phi)(\pi_0) - M(F_\phi)| < \varepsilon/3 \quad \forall t \geq T_\varepsilon. \tag{52}$$

Since $(M_t)_{t\geq 0}$ is Feller, so that $\eta \to M_t F_\phi(\eta)$ is continuous in η in the weak topology (see Stettner [14]), and since $\pi_0 S_t \to \mu$ weakly as $t \to \infty$, we may choose t_ε so that, with T_ε already fixed,

$$|M_{T_\varepsilon} F_\phi(\mu) - M_{T_\varepsilon} F_\phi(\pi_0 S_{t-T_\varepsilon})| < \varepsilon/3 \quad \forall\, t \geq t_\varepsilon. \tag{53}$$

Combining (51)–(53) gives the result (48).

To establish (49), we work on the canonical space for the observations. By (45) and (47)

$$(\overline{\pi}_t(\phi) - \overline{\pi}_{t-T,t}(\phi))^2 = \psi(t-T,t\,;\,Y),$$

where

$$\psi(t-T,t\,;\,y) := (\rho_t(\phi)(\overline{\pi}_0\,;\,y) - \rho_T(\phi)(\overline{\pi}_0 S_{t-T}\,;\,y_{t-T+\cdot} - y_{t-T}))^2$$

for $y \in C_0([0,\infty)\,:\,\mathbb{R}^p)$.

Thus, since the law of Y is R^{π_0} and by hypothesis (iii) $R^{\pi_0} \ll R^{\overline{\pi}_0}$,

$$E^{\pi_0}[(\overline{\pi}_t(\phi) - \overline{\pi}_{t-T,t}(\phi))^2] = \int_{C_0([0,\infty):\mathbb{R}^p)} \psi(T-t,t\,;\,y)\, R^{\pi_0}(dy)$$

(54)

$$= \int \psi(T-t,t\,;\,y) \frac{dR^{\pi_0}}{dR^{\overline{\pi}_0}}(y)\, R^{\overline{\pi}_0}(dy).$$

However,

$$\int \psi(T-t,t\,;\,y)\, R^{\overline{\pi}_0}(dy) = E^{\overline{\pi}_0}(\pi_t^{\overline{\pi}_0}(\phi) - \pi_{t-T,t}^{\overline{\pi}_0}(\phi))^2,$$

where in the right-hand side $E^{\overline{\pi}_0}$ denotes expectation on a space where $\overline{\pi}_0$ is the true distribution of X_0 and $\pi_t^{\overline{\pi}_0}(\phi)$ is the true conditional estimate. By the same argument as above, for each $\delta > 0$ we can choose $\bar{t}_\delta$ and $\overline{T}_\delta$ so that $E^{\overline{\pi}_0}[(\pi_t^{\overline{\pi}_0}(\phi) - \pi_{t-T_\delta,t}^{\overline{\pi}_0}(\phi))^2] < \delta$ for all $t \geq t_\delta$. Now fix $\varepsilon > 0$ and choose $K \geq 1$ so large that

$$\int \frac{dR^{\pi_0}}{dR^{\overline{\pi}_0}}(y)\, \mathbf{1}_{\left\{\frac{dR^{\pi_0}}{dR^{\overline{\pi}_0}}(y) \geq K\right\}} dR^{\overline{\pi}_0}(y) < \frac{\varepsilon}{8\|\phi\|_\infty^2}. \tag{55}$$

Then let $T_\varepsilon = \overline{T}_{\varepsilon/2K}$, $t_\varepsilon = \bar{t}_{\varepsilon/2K}$. Using (54), (55), and the fact that

$$|\psi(T-t,t\,;\,y)| \leq 4\|\phi\|_\infty^2, \quad R^{\overline{\pi}_0} \text{ almost surely},$$

we get

$$E^{\pi_0}[(\overline{\pi}_t(\phi) - \overline{\pi}_{t-T_\varepsilon,t}(\phi))^2] < \varepsilon \quad \forall\, t \geq t_\varepsilon,$$

thereby proving (49). It is clear that we can find T_ε and t_ε that work simultaneously for (48) and (49) as in the statement of the lemma. □

The second step of the proof will be to demonstrate the following lemma.

LEMMA 3.6. *For any fixed* T

$$\lim_{t\to\infty} E[(\pi_{t-T,t}(\phi) - \overline{\pi}_{t-T,t}(\phi))^2] = 0. \tag{56}$$

Before giving the proof of (56) we show how to use it to complete the proof of the stability result Theorem 3.2. The idea is simply to write for $t \geq t_\varepsilon$, where t_ε and T_ε are from Lemma 3.4,

$$\begin{aligned}
&E^{\pi_0}[(\pi_t(\phi) - \overline{\pi}_t(\phi))^2] \\
&\quad \leq 3\{E^{\pi_0}[(\pi_t(\phi) - \pi_{t-T_\varepsilon,t}(\phi))^2] + E^{\pi_0}[(\overline{\pi}_t(\phi) - \overline{\pi}_{t-T_\varepsilon,t}(\phi))^2] \\
&\qquad + E^{\pi_0}[(\pi_{t-T_\varepsilon,t}(\phi) - \overline{\pi}_{t-T_\varepsilon,t}(\phi))^2]\} \\
&\quad \leq 3\{\varepsilon + E^{\pi_0}[(\pi_{t-T_\varepsilon,t}(\phi) - \overline{\pi}_{t-T_\varepsilon,t}(\phi))^2]\} \quad \forall\, t \geq T_\varepsilon.
\end{aligned}$$

By first taking $t \to \infty$ and using Lemma 3.6 and then taking $\varepsilon \downarrow 0$, we obtain the result

$$\lim_{t\to\infty} E[(\pi_t(\phi) - \overline{\pi}_t(\phi))^2] = 0,$$

and that proves Theorem 3.2.

Proof of Lemma 3.6. The idea of the proof is that as t increases, $\pi_0 S_{t-T}$ and $\overline{\pi}_0 S_{t-T}$ converge weakly to the same limit, and hence, using (45) and (47), $\overline{\pi}_{t-T,t}(\phi)$ and $\pi_{t-T,t}(\phi)$ also converge. Indeed, if for every path y the function taking the path x to $\phi(X_T)L_T(x,y)$ were continuous in x, then the weak convergence of $\pi_0 S_{t-T}$ and $\overline{\pi}_0 S_{t-T}$ would imply that $\overline{\pi}_{t-T,t}(\phi)(y)$ converges to $\pi_{t-T,t}(\phi)(y)$. This is not the case, but by using Skorohod's representation theorem we can complete the argument. The arguments we use very much follow the techniques of Goggin [4], who establishes conditions for weak convergence of filters. As in [4], we shall use the following inequality, here stated abstractly.

LEMMA 3.7. *Let Z, Z' be nonnegative, integrable random variables, U be a bounded random variable, and $\mathcal{G}$ be a sub-σ-algebra. Then*

$$E\left[Z\left|\frac{E[UZ/\mathcal{G}]}{E[Z/\mathcal{G}]} - \frac{E[UZ'/\mathcal{G}]}{E[Z'/\mathcal{G}]}\right|\right] \le 2\|U\|_\infty E[|Z - Z'|],$$

where $\|U\|_\infty = \operatorname{ess\,sup}|U|$ and, by convention,

$$(E[Z/\mathcal{G}])^{-1} \equiv 0 \text{ on } \{E[Z/\mathcal{G}] = 0\},$$

and similarly for Z'.

Proof of Lemma 3.7.

$$\begin{aligned}
&E\left[Z\left|\frac{E[UZ/\mathcal{G}]}{E[Z/\mathcal{G}]} - \frac{E[UZ'/\mathcal{G}]}{E[Z'/\mathcal{G}]}\right|\right] \\
&= E\left[Z\left|\frac{E[U(Z-Z')/\mathcal{G}]}{E[Z/\mathcal{G}]} + \frac{E[UZ'/\mathcal{G}]}{E[Z/\mathcal{G}]} - \frac{E[UZ'/\mathcal{G}]}{E[Z'/\mathcal{G}]}\right|\right] \\
&= E\left[Z\left|\frac{E[U(Z-Z')/\mathcal{G}]}{E[Z/\mathcal{G}]} + \frac{E[UZ'/\mathcal{G}]}{E[Z'/\mathcal{G}]} \times \frac{E[Z'-Z/\mathcal{G}]}{E[Z/\mathcal{G}]}\right|\right] \\
&\le \|U\|_\infty E\left[Z\frac{E[|Z-Z'|/\mathcal{G}]}{E[Z/\mathcal{G}]}\right] + \|U\|_\infty E\left[Z\frac{E[|Z'-Z|/\mathcal{G}]}{E[Z/\mathcal{G}]}\right] \\
&= 2\|U\|_\infty E|Z'-Z|. \qquad \square
\end{aligned}$$

T being given, we proceed by constructing for each $t > T$ a measure $\tilde{P}_t$ on a probability space $\tilde{\Omega}$ with processes $(X, \bar{X}, W)$ such that

$$\begin{cases} W & \text{is a Brownian motion independent of } (X, \overline{X}), \\ X & \text{is a Markov process with initial condition } \pi_0 S_{t-T} \text{ and semigroup } (S_s)_{s\ge 0}, \\ \overline{X} & \text{is a Markov process with initial condition } \overline{\pi}_0 S_{t-T} \text{ and semigroup } (S_s)_{s\ge 0}. \end{cases} \tag{57}$$

The purpose of this construction is to have a common probability space for evaluating the expectations defining $\rho_T(\phi)(\overline{\pi}_0 S_{t-T}, Y)$ and $\rho_T(\phi)(\pi_0 S_{t-T}, Y)$. To explain, we define

$$Y_s = \int_0^s h(X_u)\, du + W_s.$$

If

$$\frac{d\tilde{Q}_T}{d\tilde{P}_t} := L_T^{-1}(X,Y)$$

it is clear by the usual Girsanov argument that, under $\overline{Q}_t$, Y is a Brownian motion independent of $(X,\overline{X})$. Thus, referring to the definition of $\rho_t(\phi)(\nu\,;\,y)$,

$$\rho_T(\phi)(\overline{\pi}_0\,S_{t-T},\,Y) = \frac{E^{\tilde{Q}_T}[\phi(\overline{X}_T)\,L_T\,(\overline{X},Y)/Y_s,\,0\le s\le T]}{E^{\tilde{Q}_T}[L_T(\overline{X},Y)/Y_s,\,0\le s\le T]},\ \tilde{P}_t \text{ almost surely,} \tag{58}$$

because under $\tilde{Q}_T$ the law of $(\overline{X},Y)$ is $P_{\overline{\pi}_0 S_{t-T}}\times\lambda$. Likewise

$$\rho_T(\phi)(\pi_0\,S_{t-T},Y) = \frac{E^{\tilde{Q}_T}[\phi(X_T)\,L_T\,(X,Y)/Y_s,\,0\le s\le T]}{E^{\tilde{Q}_T}[L_T(X,Y)/Y_s,\,0\le s\le T]}. \tag{59}$$

Since $|\rho_T(\phi)(\nu,\cdot)|\le\|\phi\|_\infty$ almost surely for any ν, it follows that

$$\begin{aligned}
&E[(\pi_{t-T,t}(\phi)-\overline{\pi}_{t-T,t}(\phi))^2]\\
&\quad\le 2\|\phi\|_\infty E^{\tilde{P}_t}[|\rho_T(\phi)(\pi_0\,S_{t-T},\,Y)-\rho_T(\phi)(\overline{\pi}_0\,S_{t-T},Y)|]\\
&\quad= 2\|\phi\|_\infty\,E^{\tilde{Q}_t}[L_T(X,Y)|\,\rho_T(\phi)(\pi_0\,S_{t-T},\,Y)-\rho_T(\phi)(\overline{\pi}_0\,S_{t-T},\,Y)|]
\end{aligned}$$

Now set $\mathcal{G}=\sigma\{Y_s,\,0\le s\le T\}$, write (59) as

$$\begin{aligned}
&\frac{E^{\tilde{Q}_T}[\phi(\overline{X}_T)\,L_T\,(X,Y)/\mathcal{G}]}{E^{\tilde{Q}_T}[L_T(X,Y)/\mathcal{G}]}+\frac{E^{\tilde{Q}_T}[(\phi(X_T)-\phi(\overline{X}_T))L_T(X,Y)/\mathcal{G}]}{E^{\tilde{Q}_T}[L_T(X,Y)/\mathcal{G}]}\\
&=\frac{E^{\tilde{Q}_T}[\phi(\overline{X}_T)\,L_T\,(X,Y)/\mathcal{G}]}{E^{\tilde{Q}_T}[L_T(X,Y)/\mathcal{G}]}+E^{\tilde{P}_T}[\phi(X_T)-\phi(\overline{X}_T)]/\mathcal{G}],
\end{aligned}$$

and use Lemma 3.7 to derive

$$\begin{aligned}
&E[(\pi_{t-T,t}(\phi)-\overline{\pi}_{t-T,t}(\phi))^2]\\
&\quad\le 2\|\phi\|_\infty\{2\|\phi\|_\infty\,E^{\tilde{Q}_t}|L_T(X,Y)-L_T(\overline{X},Y)|\\
&\qquad\qquad+E^{\tilde{P}_t}|\phi(X_T)-\phi(\overline{X}_T)|\}.
\end{aligned} \tag{60}$$

Note that (60) is true regardless of the joint law of X and $\overline{X}$.

To complete the proof it is necessary to show that the right-hand side of (60) approaches 0 along any sequence of times $\{t_n\}$ such that $t_n\to 0$. Let X^n denote the sequence of $D([0,\infty):E)$-valued Markov processes corresponding to the semigroup $(S)_{t\ge0}$ and initial laws $\pi_0\,S_{t_k-T}$, and let $\overline{X}^k$ denote that sequence corresponding to initial laws $\overline{\pi}_0\,S_{t_k-T}$. Since $\pi_0\,S_{t_k-T}$ and $\overline{\pi}_0\,S_{t_k-T}$ both converge weakly to the invariant measure μ as $k\to\infty$ by assumption (ii) of Theorem 2.3, it follows from the discussion at (40) that $X^k\Rightarrow\hat{X}$ and $\overline{X}^k\Rightarrow\tilde{X}$, where $\tilde{X}$ is the stationary process with initial law μ and transition semigroup $(S_t)_{t\ge0}$. By Skorohod's representation theorem, as stated and proved in Theorem 3.1.8 in [3], we may assume that there is a common probability space $(\tilde{\Omega}.\tilde{P})$ on which $\tilde{X}$ and the sequences X^k and $\overline{X}^k$ are defined such that the convergence is almost sure in $D([0,\infty):E)$. Let this probability

space also support a Brownian motion Y independent of the other processes. On this probability space, (60) translates into

$$
\begin{aligned}
&E[(\pi_{t-T,t}(\phi) - \overline{\pi}_{t-T,t}(\phi))^2] \\
&\quad \le 2\|\phi\|_\infty \{2\|\phi\|_\infty E^{\tilde{P}} |L_T(X^n, Y) - L_T(\overline{X}^n, Y)| \\
&\qquad + E^{\tilde{P}} |\phi(X_T^n) - \phi(\overline{X}_T^n)|\}.
\end{aligned} \tag{61}
$$

It is immediate that the second term on the right-hand side converges to 0 as $n \to \infty$. As for the first term, note that $\{L_T(X^n, Y) - L_T(\overline{X}^n, Y\}$ converges to 0 in probability and, because h is bounded, is uniformly integrable. Hence the first term also converges to 0, which completes the proof. □

Acknowledgments. We wish to thank Professor H. J. Kushner for bringing the asymptotic stability question for filters to our attention. We also thank the anonymous reviewers for their comments, which helped to generalize and simplify the proof of Theorem 3.2, and for other improvements.

REFERENCES

[1] V. E. Beneš and I. Karatzas, *Estimation and control for linear, partially observable systems with non-Gaussian initial distribution*, Stochastic Process. Appl., 14 (1983), pp. 233–248.

[2] B. Delyon and O. Zeitouni, *Lyapunov exponents for filtering problems*, in Applied Stochastic Analysis, M.H.A. Davis and R.J. Elliot, eds., Gordon and Breach, New York, 1991, pp. 531–535.

[3] S. Ethier and T. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.

[4] E. Goggin, *Convergence in distribution of conditional expectations*, Ann. Probab., 22 (1994), pp. 1097–1114.

[5] C. J. Harris and J. F. Miles, *Stability of Linear Systems: Some Aspects of Kinematic Similarity*, Academic Press, New York, 1980.

[6] R. E. Kalman and R. S. Bucy, *New results in linear filtering and prediction theory*, ASME Trans. Part D, 83 (1961), pp. 95–108.

[7] H. Kunita, *Asymptotic behavior of the nonlinear filtering errors of Markov processes*, J. Multivariate Anal., 1 (1971), pp. 365–393.

[8] ———, *Ergodic properties of nonlinear filtering processes*, in Spatial Stochastic Processes, K. S. Alexander and J. C. Watkins, eds., Birkhäuser, Boston, 1991.

[9] T. Kurtz and D. Ocone, *Unique characterization of conditional distributions in nonlinear filtering*, Ann. Probab., 16 (1988), pp. 80–107.

[10] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

[11] A. M. Makowski, *Filtering formulae for partially observed linear systems with non-Gaussian initial conditions*, Stochastics, 16 (1986), pp. 1–24.

[12] A. M. Makowski and R. B. Sowers, *Discrete-time filtering for linear systems with non-Gaussian initial conditions: Asymptotoic behavior of the difference between the MMSE and the LMSE estimates*, IEEE Trans. Automat. Control, 37 (1992), pp. 114–121.

[13] R. B. Sowers, *New discrete-time filtering results*, M.S. thesis, Dep. Elec. Eng., Univ. Maryland, College Park, MD, Aug. 1988, also in Systems Research Center, Tech. Rep. TR 88-85, 1988.

[14] L. Stettner, *On invariant measure of filtering processes*, in Stochastic Differential Systems, Lecture Notes in Control and Information Sciences 126, Springer-Verlag, Berlin, 1989.

[15] ———, *Invariant measures of the pair: State, approximate filtering process*, Colloq. Math., LXII (1991), pp. 347–351.

[16] R. Vintner, *Filter stability for stochastic evolution equations*, SIAM J. Control Optim., 15 (1977), pp. 465–485.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 244–263, January 1996

011

NONDEGENERATE SOLUTIONS AND RELATED CONCEPTS IN AFFINE VARIATIONAL INEQUALITIES*

M. C. FERRIS† AND J. S. PANG‡

This paper is dedicated to Professor O. L. Mangasarian on the occasion of his 60th birthday (January 12, 1994). We submitted this paper for publication on Professor Mangasarian's birthday to a journal that he has been associated with for many years, the SIAM Journal on Control and Optimization. *He has made many significant contributions to the topics addressed in this paper, namely, error bounds, weak sharp minima, minimum principle sufficiency, and complementarity problems. We are both indebted to him for his constant encouragement, advice, and fruitful collaborations over many years. Without his help and guidance, this paper would not have been possible.*

Abstract. The notion of a strictly complementary solution for complementarity problems is extended to that of a nondegenerate solution of variational inequalities. Several equivalent formulations of nondegeneracy are given. In the affine case, an existence theorem for a nondegenerate solution is given in terms of several related concepts which are shown to be equivalent in this context. These include a weak sharp minimum, the minimum principle sufficiency, and error bounds. The gap function associated with the variational inequality plays a central role in this existence theorem.

Key words. Variational inequalities, nondegenerate solutions, weak sharp minima, minimum principle, error bounds

AMS subject classifications. 90C33, 65K05, 90C25

1. Introduction.

Strict complementarity is a familiar notion in the context of optimization problems and complementarity theory. A classical result proved in [17, Cor. 2A] shows that a solvable linear complementarity problem defined by a skew-symmetric matrix must possess a strictly complementary solution. In general, the property of strict complementarity of a solution to an optimization or a complementarity problem plays an important role in many aspects of such a problem. Historically, Fiacco and McCormick [14] used this property to develop the first sensitivity theory of nonlinear programs under perturbation. Robinson [40] has introduced a generalized notion of strict complementarity and considered its role in parametric nonlinear programming.

In recent years, the strict complementarity property was given a renewed emphasis in the analysis of many iterative algorithms for solving linear and nonlinear programs and complementarity problems. Dunn [10] and Burke and Moré [6] used a geometric definition of a strictly complementary solution to a nonlinear program and showed how such a solution was essential for the successful identification of active constraints in a broad class of gradient based methods for solving constrained optimization problems. Güler and Ye [19] showed that many interior-point algorithms for linear programs generated a sequence of iterates whose limit points satisfied the

* Received by the editors January 19, 1994; accepted for publication (in revised form) September 12, 1994.

† Computer Sciences Department, University of Wisconsin, Madison, WI 53706 (ferris@cs.wisc.edu). The work of this author was based on research supported by National Science Foundation grant CCR-9157632 and Air Force Office of Scientific Research grant F49620-94-1-0036.

‡ Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD 21218-2689 (jsp@vicp.mts.jhu.edu). The work of this author was based on research supported by National Science Foundation grants DDM-9104078 and CCR-9213739 and by Office of Naval Research grant N00014-93-1-0228.

strict complementarity condition; they also extended the result to a monotone linear complementarity problem having a strictly complementary solution. Monteiro and Wright [36] demonstrated that the existence of a strictly complementary solution was essential for the fast convergence of these interior-point algorithms for a monotone linear complementarity problem.

The theory of error bounds for inequality systems has in recent years become an active area of research within the field of mathematical programming. In this regard, Hoffman [21] obtained the first error bound for a system of finitely many linear inequalities. The generalizations of Hoffman's result are too numerous to be mentioned here. There are several factors that have motivated this proliferation of activities. In general, an error bound is an inequality that bounds the distance function from a test vector to the solutions of a system of inequalities in terms of a residual function. Part of the importance of an error bound is that it provides the foundation for exact penalization of mathematical programs [24], [30]; this in turn is strongly connected to the theory of optimality conditions for nonlinear programs [4]. Error bounds play an important role in the convergence analysis (particularly in establishing the convergence rates) of many iterative algorithms for solving various mathematical programs. These include the the matrix splitting methods for linear complementarity problems [8, Chap. 5] and affine variational inequalities [25], various descent methods for convex minimization problems [26]–[28], and interior-point methods for linear programs and extensions [23], [35], [42]. Error bounds can also be used to design inexact iterative methods [37], [16].

The concept of a weak sharp minimum for a constrained optimization problem was introduced in [11]. The usefulness of this concept in establishing the finite convergence of various iterative algorithms was discussed in several subsequent papers [12], [5], [1]. Among the classes of optimization problems that possess weak sharp minima are linear programs [32] and certain convex quadratic programs and monotone linear complementarity problems [5].

Finally, the minimum principle [29] is a well-known set of conditions that must be satisfied by any local minimum of a nonlinear program with a convex feasible region. One way to state this principle is in terms of the gap function [20] of the nonlinear program; informally, this principle states that a local minimum of a nonlinear program must be a global minimizer of the gap function over the same convex feasible region of the program. In [13], Ferris and Mangasarian studied the "converse" of this principle for the class of convex programs and coined the term *minimum principle sufficiency* when this converse was valid. They also showed (Theorem 6 in [13]) that for a convex quadratic program, the minimum principle sufficiency is equivalent to the existence of weak sharp minima of the program and that of a nondegenerate solution in the primal-dual linear complementarity formulation of the quadratic program. This somewhat unexpected result therefore links up the various concepts that we have discussed so far.

The present research is motivated by the desire to gain a better understanding of the concepts of strict complementarity, error bounds, weak sharp minima, and minimum principle sufficiency for various mathematical programs and how these concepts are related. The results in [13], [31] suggest that for a monotone linear complementarity problem and its "natural" convex quadratic program [8, Chap. 3], all these concepts are equivalent (to be made precise later). In this paper, we shall extend the equivalences to a monotone affine variational inequality.

By adding appropriate multipliers to the constraints of an affine variational in-

equality, this problem becomes equivalent to a linear complementarity problem [38]. In view of the results available for the linear complementarity problem [13], [31], this transformation therefore raises the question of whether the intended generalized equivalences for the affine variational inequality are of any significant interest. We shall argue that the results derived herein are potentially useful for two reasons: (i) they do not rely on the multipliers of the constraints and hence are independent of the representation of the defining set of the affine variational inequality; and (ii) as it turns out, we shall use a nondifferentiable optimization problem as the bridge to connect the various concepts in question. The latter approach raises the issue of the extent to which these equivalences will remain valid for more general nondifferentiable optimization problems. The full treatment of this last issue is, regrettably, beyond the scope of the present work.

2. Definitions and review. For a given mapping $F : R^n \to R^n$, the nonlinear complementarity problem, which we shall denote NCP (F), is to find a vector $x \in R^n$ such that

$$x \geq 0, \quad F(x) \geq 0, \quad x^T F(x) = 0.$$

A solution $\hat{x}$ of this problem is said to be *strictly complementary*, or *nondegenerate*, if $\hat{x} + F(\hat{x}) > 0$. For an optimization problem of the form

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && x \in C, \end{aligned} \tag{2.1}$$

where $f : R^n \to R$ is continuous and $C \subseteq R^n$ is convex, different forms of nondegeneracy abound in the literature. Dunn [10] and Burke and Moré [6] use the relative interior condition

$$-\nabla f(\hat{x}) \in \operatorname{ri} N_C(\hat{x}) \tag{2.2}$$

to define an optimal solution $\hat{x}$ of (2.1) as being nondegenerate. Here $\operatorname{ri} S$ denotes the relative interior of the convex set S and $N_C(x)$ denotes the normal cone to the convex set C at the point $x \in R^n$, which is defined by

$$N_C(x) \equiv \begin{cases} \{y \in R^n \mid y^T(c - x) \leq 0 \text{ for all } c \in C\} & \text{if } x \in C, \\ \emptyset & \text{otherwise.} \end{cases}$$

Robinson [40] uses the dual form: $T_C(\hat{x}) \cap \nabla f(\hat{x})^{\perp}$ is a subspace where the tangent cone, $T_C(x)$, to C at x is the polar of the normal cone at x; i.e.,

$$T_C(x) \equiv \{z \in R^n \mid z^T y \leq 0 \text{ for all } y \in N_C(x)\}.$$

It is easy to show (see [40, Lem. 2.1] for a proof) that the definition (2.2) is equivalent to the subspace definition. In general, for a convex set $S \subseteq R^n$, the negative of the polar of S is the dual cone of S, which is denoted by S^*.

It is not difficult to extend the notion of strict complementarity to the context of a variational inequality (VI) of the following form: find $x \in C$ such that

$$F(x)^T(y - x) \geq 0 \quad \text{for all } y \in C, \tag{2.3}$$

where $C \subseteq R^n$ is a nonempty closed convex set and $F : R^n \to R^n$ is a continuous mapping. We shall denote this problem by VI (F, C); its (possibly empty) solution

set is denoted $\mathrm{SOL}(F, C)$. When F is affine and given by $F(x) \equiv q + Mx$ for some vector $q \in R^n$, some matrix $M \in R^{n \times n}$, and all vectors $x \in R^n$, we shall append the word "affine" to describe this VI and denote it by AVI (q, M, C); the notation $\mathrm{SOL}(q, M, C)$ will be used to denote the solution set of this AVI.

Given a vector $\hat{x} \in \mathrm{SOL}(F, C)$ by simply replacing $\nabla f(\hat{x})$ by $F(\hat{x})$ in either (2.2) or in Robinson's dual definition, we obtain a definition for $\hat{x}$ to be a nondegenerate solution of the VI (F, C). Thus, $\hat{x}$ is a nondegenerate solution of the VI (F, C) if

$$-F(\hat{x}) \in \operatorname{ri} N_C(\hat{x}). \tag{2.4}$$

A justification for this definition of nondegeneracy for the VI is the well-known fact that the VI (F, C) is equivalent to the generalized equation

$$0 \in F(x) + N_C(x),$$

or equivalently

$$-F(x) \in N_C(x),$$

which easily leads to the generalized definition.

When C is a polyhedron, it is possible to give some further characterizations for the nondegeneracy of a solution $\hat{x} \in \mathrm{SOL}(F, C)$. We shall summarize these characterizations in Proposition 2.2 below. The additional characterizations rely heavily on the face structure of a polyhedral convex set. It is well known that the relative interiors of the faces of a convex set C form a partition of C [41, Thm. 18.2]. Throughout this paper, we will use the notation $\mathcal{F}(x)$ to denote the face of C which contains a vector $x \in C$ in its relative interior. According to [41, Thm. 18.1], $\mathcal{F}(x)$ is the "minimal" face of C containing $x \in C$, minimal in terms of set inclusion. The following result was established in [6].

LEMMA 2.1. *The normal cone to a polyhedral convex set C is constant for all $x \in \operatorname{ri} \mathcal{F}$, where $\mathcal{F}$ is a face of C, henceforth labeled $\mathcal{N}_{\mathcal{F}}$. Furthermore,*

$$\operatorname{aff} \mathcal{F} - x = \operatorname{lin} T_C(x) = (\operatorname{aff} \mathcal{N}_{\mathcal{F}})^{\perp}.$$

As a consequence of this lemma, it follows that $\mathcal{F} - \mathcal{N}_{\mathcal{F}}$ has full dimension and hence has a nonempty interior. This observation will be used in the proof of the following proposition which gives a number of equivalent conditions for a given solution of the VI (F, C) to be nondegenerate. Among these conditions, condition (iv) has been used by Reinoza [39].

PROPOSITION 2.2. *Suppose $\hat{x}$ solves VI (F, C) and C is polyhedral. Let $\mathcal{F} = \mathcal{F}(\hat{x})$ so that $-F(\hat{x}) \in \mathcal{N}_{\mathcal{F}}$. The following statements are equivalent:*

(i) $\hat{x} + F(\hat{x}) \in \operatorname{int}(\mathcal{F} - \mathcal{N}_{\mathcal{F}})$,

(ii) $-F(\hat{x}) \in \operatorname{ri} \mathcal{N}_{\mathcal{F}}$,

(iii) $T_C(\hat{x}) \cap F(\hat{x})^{\perp}$ *is a subspace,*

(iv) $\hat{x}$ *is in the relative interior of the face of C exposed by $-F(\hat{x})$.*

If F is monotone and any one of the above four conditions holds, then $\mathrm{SOL}(F, C) \subseteq \mathcal{F}(\hat{x})$.

Proof. The equivalence of (ii) and (iii) has been noted before. The equivalence of (ii) and (iv) is by [7, Thm. 2.4]. Since $\hat{x} \in \operatorname{ri} \mathcal{F}$ and $-F(\hat{x}) \in \operatorname{ri} \mathcal{N}_{\mathcal{F}}$, it follows that $\hat{x} + F(\hat{x}) \in \operatorname{ri} \mathcal{F} + \operatorname{ri}(-\mathcal{N}_{\mathcal{F}}) = \operatorname{ri}(\mathcal{F} - \mathcal{N}_{\mathcal{F}})$ which, as we have noted, has a nonempty

interior. Thus (ii) implies (i). We now show that (i) implies (ii). First note that $\hat{x} + F(\hat{x}) \in \operatorname{ri} \mathcal{F} + \operatorname{ri}(-\mathcal{N}_{\mathcal{F}})$, so suppose

$$\hat{x} + F(\hat{x}) = y + z$$

with $y \in \operatorname{ri} \mathcal{F}$ and $z \in \operatorname{ri}(-\mathcal{N}_{\mathcal{F}})$. Then $y - \hat{x} \in \operatorname{aff} \mathcal{F} - \hat{x}$, $F(\hat{x}) - z \in \operatorname{aff} \mathcal{N}_{\mathcal{F}}$, and these two subspaces are orthogonal. Hence $y - \hat{x} = 0 = F(\hat{x}) - z$ as required.

For the final statement of the proposition, let $z \in \operatorname{SOL}(F, C)$ be arbitrary. Since F is monotone, it follows that (see, e.g., [3])

$$F(c)^T(c - z) \geq 0 \quad \text{for all } c \in C,$$

which implies, since $\hat{x} \in C$, that $F(\hat{x})^T(\hat{x} - z) \geq 0$. However, $\hat{x}$ also solves the VI (F, C), so

$$F(\hat{x})^T(c - \hat{x}) \geq 0 \quad \text{for all } c \in C,$$

implying $F(\hat{x})^T(\hat{x} - z) = 0$. Hence,

$$z \in \{c \in C \mid F(\hat{x})^T(c - \hat{x}) = 0\},$$

which is $\mathcal{F}(\hat{x})$ by [7, Thm. 2.4]. □

In the remainder of this paper, we shall focus on the AVI (q, M, C). As stated before, our goal is to establish the equivalence of the existence of a nondegenerate solution to this problem and a number of related concepts. In what follows, we shall describe each of these concepts more formally.

The notion of a weak sharp minimum was introduced in [11] and extensively analyzed in [5], [13]. The formal definition is as follows.

DEFINITION 2.3. *Let $f : R^n \to R \cup \{\infty\}$ and $C \subseteq R^n$. A nonempty subset $S \subseteq C$ is a set of* weak sharp minima *for the problem* (2.1) *if there is a scalar $\alpha > 0$ such that for all $x \in C$ and all $y \in S$*

$$f(x) \geq f(y) + \alpha \operatorname{dist}(x \mid S), \tag{2.5}$$

where

$$\operatorname{dist}(x \mid S) \equiv \inf\{\|z - x\| : z \in S\}$$

is the distance from the point x to S measured by any norm.

Note that a set of weak sharp minima for (2.1), if it exists, must be equal to the set of global minimizers of f over C. In general, for the problem (2.1), it would be useful to know when a set of weak sharp minima exists. As mentioned in the introduction, an affirmative answer to this question is known for a linear program and certain convex quadratic programs.

Observe that if the problem (2.1) has a weak sharp minimum, then the inequality (2.5), which is equivalent to

$$\operatorname{dist}(x \mid S) \leq \alpha^{-1}(f(x) - f_{\min}) \quad \text{for all } x \in C, \tag{2.6}$$

where $f_{\min}$ is the minimum value of f on C, can be interpreted as providing an *error bound* for an arbitrary feasible point x to the set of minimizers of (2.1), with the residual given by the deviation of the objective value $f(x)$ from its minimum value. Consequently, a necessary and sufficient condition for the existence of a weak sharp

minimum for the problem (2.1) is the existence of an error bound of the type (2.6) where S is the set of minimizers of (2.1).

The notion of minimum principle sufficiency was introduced in [13]. The minimum principle is a well-known necessary optimality condition for a program of the form (2.1), where C is convex; this principle states that, for a continuously differentiable function f, if $\bar{x}$ solves (2.1) then $\bar{x} \in \mathrm{SOL}(\nabla f, C)$. Roughly speaking, minimum principle sufficiency is the converse assumption; nevertheless, in order to make this precise, it will be necessary for us to introduce the gap function associated with the VI (F, C). Specifically, the gap function for the latter problem is the extended-valued function $g : R^n \to R \cup \{\infty\}$ given by

$$g(x) \equiv x^T F(x) - \omega(x) \quad \text{for all } x \in R^n, \tag{2.7}$$

where

$$\omega(x) \equiv \inf\{z^T F(x) : z \in C\}. \tag{2.8}$$

The function ω was introduced in [18], where it was used for stability analysis of the AVI. Let

$$\Omega(x) \equiv \operatorname{argmin}\{z^T F(x) : z \in C\};$$

it is understood that if the minimum value in $\omega(x)$ is not attained, then $\Omega(x)$ is defined to be the empty set. We note that if C is polyhedral, then $\omega(x)$ is the optimum objective value of a linear program.

The following proposition summarizes some important properties of the two functions g and ω. No proof is needed for these properties.

PROPOSITION 2.4. *Let $F : R^n \to R^n$ be a mapping and C be a closed convex subset of R^n. The following statements are valid.*

(i) *The function $\omega : R^n \to R \cup \{-\infty\}$ is concave and extended-valued; if F is a monotone affine function, then g is convex.*

(ii) *The function g is nonnegative on C.*

(iii) *A vector $x \in \mathrm{SOL}(F, C)$ if and only if $x \in \Omega(x)$ or, equivalently, $x \in C$ and $g(x) = 0$.*

(iv) *If C is polyhedral, then*

$$\begin{aligned} \operatorname{dom}\omega &\equiv \{x \in R^n \mid \omega(x) > -\infty\} \\ &= \{x \in R^n \mid F(x) \in (\operatorname{rec} C)^*\}, \end{aligned}$$

where $(\operatorname{rec} C)^*$ is the dual of the recession cone of C.

(v) *If C is polyhedral and F is affine, then ω is piecewise linear and g is piecewise quadratic.*

Returning to the problem (2.1) and letting $F \equiv \nabla f$, we see that the minimum principle for this problem can be stated simply as: if x is a local minimizer of (2.1), then $x \in \Omega(x)$. Obviously, if f is a convex function, then every vector $x \in C$ with the property that $x \in \Omega(x)$ must be a global minimizer of (2.1). For a convex function f, the minimum principle sufficiency stipulates that for all optimal solutions x of (2.1), or equivalently, for all x such that $x \in \Omega(x)$, if $x' \in \Omega(x)$, then x' is also a global minimizer of (2.1). In what follows, we shall give several equivalent formulations for this sufficiency property, one of which will be the basis for generalization to a nondifferentiable function f.

PROPOSITION 2.5. *Let $f : R^n \to R$ be a continuously differentiable convex function and $C \subseteq R^n$ be a closed convex set. Assume that $\mathcal{S} \equiv \operatorname{argmin}\{f(x) : x \in C\} \neq \emptyset$. The following statements are equivalent.*

(a) *The minimum principle sufficiency holds for the minimization problem* (2.1).

(b) *For all $x \in \mathcal{S}$, $\mathcal{S} = \Omega(x)$, where $\Omega(x) \equiv \operatorname{argmin}\{z^T \nabla f(x) : z \in C\}$.*

(c) *For all $x \in \mathcal{S}$,*

$$[z \in C,\ \nabla f(x)^T(z - x) = 0] \Rightarrow z \in \mathcal{S}.$$

If in addition, C is polyhedral and $\mathcal{S} \neq \emptyset$, then any one of the above statements is further equivalent to

(d) *$\mathcal{S}$ is a set of weak sharp minima for* (2.1).

Proof. Since $\mathcal{S} \subseteq \Omega(x)$ for all optimal solutions x of (2.1), the equivalence of (a) and (b) is obvious. That (c) is also equivalent to (a) or (b) is equally obvious because x solves (2.1) if and only if $x \in \Omega(x)$. Finally, the equivalence of (d) and the above statements was proved in [5, Thm. 4.2]. □

Remark. Theorem 4.2 in [5] shows that in the above proposition, (d) always implies (a) for an arbitrary closed convex set C; nevertheless, Example 4.3 in [5] shows that the polyhedrality of C is needed for the reverse implication.

3. Miscellaneous preliminary results. We have now defined all the concepts we shall deal with in this paper. Our ultimate goal is to link them together for the monotone AVI (q, M, C), where M is assumed to be positive semidefinite and C is polyhedral. The linkage is via the gap function g for this AVI. Motivation for using this function g stems partly from statement (iii) in Proposition 2.4, which suggests that g is a likely candidate for a residual function for the AVI. This choice is also supported by some error bound results in [18] which are derived with the aid of some additional properties of the monotone AVI. In what follows, we shall summarize the relevant results for later use. Throughout the rest of this paper, we shall fix the vector $q \in R^n$, the matrix $M \in R^{n \times n}$, and the set $C \subseteq R^n$. We shall assume that M is positive semidefinite and C is a polyhedral. We shall further assume that $\mathrm{SOL}(q, M, C) \neq \emptyset$.

There are two important constants associated with the solution set of the monotone AVI (q, M, C). Indeed, by results in [18], there exist a vector $d \in R^n$ and a scalar $\sigma \in R_+$, both dependent on the data (q, M, C), such that

$$d = (M + M^T)x, \quad \sigma = x^T M x \tag{3.1}$$

for all $x \in \mathrm{SOL}(q, M, C)$. Furthermore, $\mathrm{SOL}(q, M, C)$ can be characterized, using these constants, as

$$\mathrm{SOL}(q, M, C) = \{x \in C \mid \omega(x) - (q^T x + \sigma) \geq 0, (M + M^T)x = d\}.$$

Since, for every $x \in \mathrm{SOL}(q, M, C), \omega(x) \leq (q + Mx)^T x = (q + d - M^T x)^T x = (q + d)^T x - \sigma$, simple algebra gives the alternative characterization:

$$\begin{aligned} \mathrm{SOL}(q, M, C) &= \{x \in C \mid \omega(x) - (q + d)^T x + \sigma \geq 0, (M + M^T)x = d\} \\ &= \{x \in C \mid \omega(x) - (q + d)^T x + \sigma = 0, (M + M^T)x = d\}. \end{aligned}$$

For a given polyhedral cone $K \subseteq R^n$, the AVI (q, M, K) is equivalent to a generalized linear complementarity problem which is to find a vector $y \in R^n$ such that

$$y \in K, \quad q + My \in K^*, \quad \text{and} \quad y^T(q + My) = 0,$$

where

$$K^* \equiv \{y \in R^n \mid y^T x \geq 0 \ \forall x \in K\}$$

is the dual cone of K. In this case, we shall use the prefix GLCP instead of AVI to describe the problem. The feasible region of GLCP (q, M, K) is given by

$$\text{FEA}(q, M, K) \equiv \{y \in K \mid q + My \in K^*\}. \tag{3.2}$$

Since $\mathcal{F} - \mathcal{N}_{\mathcal{F}} \subseteq K + K^*$ and both have full dimension, it follows from Proposition 2.2 that if $\hat{y}$ is a strictly complementary solution of the GLCP (q, M, K), then

$$\hat{y} + q + M\hat{y} \in \text{int}(K + K^*).$$

It is known [38] that the AVI (q, M, C) is equivalent to a mixed linear complementarity problem in higher dimensions. In what follows, we shall establish a connection between the nondegenerate solutions of these two problems. For this purpose, we shall represent C as

$$C = \{x \in R^n \mid Ax \geq b\} \tag{3.3}$$

for some matrix $A \in R^{m \times n}$ and vector $b \in R^m$. Then a vector $x \in C$ is a solution of AVI (q, M, C) if and only if there exists a vector $\lambda \in R^m$ such that the following conditions hold:

$$\begin{aligned} 0 &= q + Mx - A^T\lambda, \\ w &= Ax - b, \\ w \geq 0, &\quad \lambda \geq 0, \quad w^T\lambda = 0. \end{aligned}$$

These conditions define the GLCP (p, N, K) where the variable y and the data (p, N, K) are given by

$$y \equiv \begin{pmatrix} x \\ \lambda \end{pmatrix}, \quad p \equiv \begin{pmatrix} q \\ -b \end{pmatrix}, \quad N \equiv \begin{bmatrix} M & -A^T \\ A & 0 \end{bmatrix}, \tag{3.4}$$

and $K \equiv R^n \times R^m_+$. Specializing Proposition 2.2 to the latter GLCP, we can show that a solution $(\hat{x}, \hat{\lambda})$ of GLCP (p, N, K) is nondegenerate if and only if $\hat{w} + \hat{\lambda} > 0$, where $\hat{w} \equiv A\hat{x} - b$. Based on this observation, the following result is easy to prove.

PROPOSITION 3.1. *Let C be given by (3.3). A solution $\hat{x}$ of the AVI (q, M, C) is nondegenerate if and only if for some $\hat{\lambda}$, $(\hat{x}, \hat{\lambda})$ is a nondegenerate solution of the GLCP (p, N, K).*

Proof. Let

$$\mathcal{I} \equiv \{i \mid (A\hat{x} = b)_i\}$$

be the index set of active constraints at $\hat{x}$. By the definition of $\mathcal{F} = \mathcal{F}(\hat{x})$, we have

$$\mathcal{F} = \{x \in C \mid (Ax = b)_i \text{ for all } i \in \mathcal{I}\}$$

and $\hat{x} \in \text{ri}\,\mathcal{F}$. Hence,

$$\mathcal{N}_{\mathcal{F}} = \{A^T\lambda \mid \lambda \in R^m_-, \lambda_i = 0, \text{ for all } i \notin \mathcal{I}\}.$$

From the theory of convex polyhedra, particularly [41, Thm. 6.6], we have

$$\operatorname{ri}\mathcal{F} = \{x \in \mathcal{F} \mid (Ax > b)_i \text{ for all } i \notin \mathcal{I}\}$$
$$\operatorname{ri}\mathcal{N}_{\mathcal{F}} = \{A^T\lambda \mid \lambda_i < 0 \text{ for all } i \in \mathcal{I};\ \lambda_i = 0 \text{ for all } i \notin \mathcal{I}\}.$$

Hence, according to Proposition 2.2, $\hat{x}$ is nondegenerate if and only if $\hat{x} \in \operatorname{ri}\mathcal{F}$ and $-(q + M\hat{x}) \in \operatorname{ri}\mathcal{N}_{\mathcal{F}}$. From this, the existence of the desired $\hat{\lambda}$ is obvious. □

The GLCP (p, N, K) defined above is related to the linear program defining the function $\omega(x)$, which is given by

$$\omega(x) \equiv \min\{z^T(q + Mx) : z \in C\};$$

see (2.8). The dual of this linear program, denoted $\Delta(x)$, is

$$\begin{aligned} &\text{maximize} && b^T\lambda \\ &\text{subject to} && q + Mx - A^T\lambda = 0, \quad \lambda \geq 0. \end{aligned}$$

We shall let $\Lambda(x)$ denote the (possibly empty) optimal solution set of $\Delta(x)$. The following result summarizes an important relation between the dual program $\Delta(x)$ and the GLCP (p, N, K) as well as two properties of $\Delta(x)$ as a parametric linear program with a changing right-hand side in the constraints.

PROPOSITION 3.2. *The following three statements hold:*

(a) *if $\hat{x} \in \mathrm{SOL}(q, M, C)$, then a pair $(\hat{x}, \hat{\lambda})$ solves the GLCP (p, N, K) if and only if $\hat{\lambda} \in \Lambda(\hat{x})$;*

(b) *there exists a constant $\alpha > 0$ such that for all $x \in R^n$ with $\Lambda(x) \neq \emptyset$ and all λ feasible to $\Delta(x)$,*

$$-b^T\lambda + \omega(x) \geq \alpha \operatorname{dist}(\lambda \mid \Lambda(x)); \tag{3.5}$$

(c) *there exists a constant $\beta > 0$ such that for all x and x' in R^n with $\Lambda(x) \neq \emptyset$ and $\Lambda(x') \neq \emptyset$,*

$$\Lambda(x) \subseteq \Lambda(x') + \beta\|x - x'\|\mathcal{B}(0, 1),$$

where $\mathcal{B}(0, 1)$ is the unit Euclidean ball in R^m.

Proof. Statement (a) is obvious. For statement (b), observe that if $\Lambda(x) \neq \emptyset$ for some x, then $\omega(x)$ is finite and equal to the optimal objective value of $\Delta(x)$. By [32, Lem. A.1], every solvable linear program has a nonempty set of weak sharp minima. A careful look at the proof of this result reveals that the constant associated with such a set of weak sharp minima is independent of the right-hand side in the constraints of the program. Thus (b) follows. Statement (c) follows from the Lipschitzian property of the solutions to a parametric right-hand sided linear program as proved in [33, Thm. 2.4]. □

We shall associate the following optimization problem with the AVI (q, M, C):

$$\begin{aligned} &\text{minimize} && g(x) \\ &\text{subject to} && x \in C, \end{aligned} \tag{3.6}$$

where g is the gap function defined in (2.7) with $F(x) \equiv q + Mx$. By Proposition 2.4, the function g is convex, piecewise quadratic, and possibly extended-valued; it is in general not Fréchet differentiable. We should mention that recently there have been

several differentiable optimization problems introduced for the study of a monotone VI [2], [15], [34]; since the objective functions of the latter optimization problems are not known to be convex even for a monotone AVI, it is therefore not clear whether our results can be extended to these other (possibly nonconvex) optimization formulations of the AVI.

Since C is polyhedral, it can be represented as

$$C = \operatorname{conv} G + \operatorname{rec} C \tag{3.7}$$

for some finite point set $G \subseteq R^n$, where $\operatorname{conv} G$ denotes the convex hull of G and $\operatorname{rec} C$ denotes the recession cone of C. When C is a cone, we have $G = \{0\}$ and $C = \operatorname{rec} C$. Clearly, the problem (3.6) can be equivalently stated as

$$\begin{array}{ll} \text{minimize} & x^T(q + Mx) - \tilde{\omega}(x) \\ \text{subject to} & x \in C, \quad q + Mx \in (\operatorname{rec} C)^*, \end{array} \tag{3.8}$$

where

$$\tilde{\omega}(x) \equiv \min\{z^T(q + Mx) : z \in G\}. \tag{3.9}$$

When C is a cone, the latter formulation reduces to

$$\begin{array}{ll} \text{minimize} & x^T(q + Mx) \\ \text{subject to} & x \in \operatorname{FEA}(q, M, C), \end{array}$$

since $C = \operatorname{rec} C$ and $\tilde{\omega}(x)$ is identically equal to zero in this case; see (3.2) for the definition of $\operatorname{FEA}(q, M, C)$.

Unlike the function $\omega(x)$, $\tilde{\omega}(x)$ is finite valued for all $x \in R^n$, and it is dependent on the point set G (in particular, on the representation of C). Nevertheless, $\omega(x) = \tilde{\omega}(x)$ for all $x \in \operatorname{dom} \omega$; recall that by Proposition 2.4, $\operatorname{dom} \omega$ consists of all vectors x satisfying $q + Mx \in (\operatorname{rec} C)^*$. The function $\tilde{\omega}(x)$ will play an important part in the proofs (but not the statements) of the results involving the AVI (q, M, C). We shall let $\operatorname{FEA}(q, M, C)$ denote the feasible region of the problem (3.8). This coincides with the previous definition (3.2) when C is a cone. Trivially, we have $\operatorname{SOL}(q, M, C) \subseteq \operatorname{FEA}(q, M, C)$. Moreover, the problem (3.6) is equivalent to

$$\begin{array}{ll} \text{minimize} & g(x) \\ \text{subject to} & x \in \operatorname{FEA}(q, M, C). \end{array} \tag{3.10}$$

Although the function g is not Fréchet differentiable, it is directionally differentiable at every vector in $\operatorname{FEA}(q, M, C)$ along all feasible directions. This fact is made precise in the following result.

PROPOSITION 3.3. *Let $q \in R^n$ and $M \in R^{n \times n}$ be arbitrary; let $C \subseteq R^n$ be a polyhedral set. For any vectors $\bar{x}$ and x in $\operatorname{FEA}(q, M, C)$, the directional derivative*

$$\omega'(\bar{x}; x - \bar{x}) \equiv \lim_{\tau \downarrow 0} \frac{\omega(\bar{x} + \tau(x - \bar{x})) - \omega(\bar{x})}{\tau}$$

exists, is finite, and is equal to

$$\min\{u^T M(x - \bar{x}) : u \in \Omega(\bar{x})\},$$

where $\Omega(\bar{x}) \equiv \operatorname{argmin}\{z^T(q + M\bar{x}) : z \in C\}$; *hence,* $g'(\bar{x}; x - \bar{x})$ *exists and is equal to*

$$(x - \bar{x})^T(q + (M + M^T)\bar{x}) - \omega'(\bar{x}; x - \bar{x}).$$

Proof. Since $q + M\bar{x} \in (\operatorname{rec} C)^*$, $\Omega(\bar{x}) \neq \emptyset$. It suffices to verify that

$$\omega'(\bar{x}; x - \bar{x}) = \min\{u^T M(x - \bar{x}) : u \in \Omega(\bar{x})\},$$

and that this derivative is finite. Since both $\bar{x}$ and x are in $\operatorname{FEA}(q, M, C)$, it follows that

$$\omega(\bar{x} + \tau(x - \bar{x})) = \tilde{\omega}(\bar{x} + \tau(x - \bar{x}))$$

for all $\tau \in [0, 1]$. Hence, we have

$$\omega'(\bar{x}; x - \bar{x}) = \tilde{\omega}'(\bar{x}; x - \bar{x}) = \min\{u^T M(x - \bar{x}) : u \in \tilde{\Omega}(\bar{x})\},$$

where

$$\tilde{\Omega}(\bar{x}) \equiv \operatorname{argmin}\{z^T(q + M\bar{x}) : z \in G\}$$

is a nonempty, finite subset of $\Omega(\bar{x})$. Since $\tilde{\omega}'(\bar{x}; x - \bar{x})$ is finite, thus so is $\omega'(\bar{x}; x - \bar{x})$. Moreover, we have

$$\omega'(\bar{x}; x - \bar{x}) \geq \min\{u^T M(x - \bar{x}) : u \in \Omega(\bar{x})\}. \tag{3.11}$$

Since

$$\begin{aligned} \omega(\bar{x} + \tau(x - \bar{x})) &= \min\{z^T(q + M\bar{x}) + \tau z^T M(x - \bar{x}) : z \in C\} \\ &\leq \min\{z^T(q + M\bar{x}) + \tau z^T M(x - \bar{x}) : z \in \Omega(\bar{x})\} \\ &= \omega(\bar{x}) + \tau \min\{u^T M(x - \bar{x}) : u \in \Omega(\bar{x})\}, \end{aligned}$$

it follows that the reverse inequality in (3.11) also holds. Consequently, equality holds in (3.11). □

Note that if $\bar{x} \in \operatorname{SOL}(q, M, C)$ and M is positive semidefinite, then Proposition 3.3 yields

$$g'(\bar{x}; x - \bar{x}) = (x - \bar{x})^T(q + d) - \omega'(\bar{x}; x - \bar{x}) \tag{3.12}$$

for all $x \in \operatorname{FEA}(q, M, C)$, where $d = (M + M^T)\bar{x}$ is one of the two constants associated with the solutions of the AVI (q, M, C). With the above proposition, we can now discuss the extension of the minimum principle sufficiency to the nondifferentiable gap minimization problem (3.6) or equivalently to (3.10). Some related work on error bounds for convex, piecewise quadratic minimization problems, of which (3.6) is a special case, can be found in [22]. The following result establishes two properties of solutions to the AVI (q, M, C).

PROPOSITION 3.4. *Let* $q \in R^n$ *be arbitrary,* $M \in R^{n \times n}$ *be positive semidefinite, and* $C \subseteq R^n$ *be a polyhedral set. If* x *and* $\bar{x}$ *are any two vectors in* $\operatorname{SOL}(q, M, C)$, *then* $g'(\bar{x}; x - \bar{x}) = 0$ *and*

$$\omega(x) = \omega(\bar{x}) + \omega'(\bar{x}; x - \bar{x}). \tag{3.13}$$

Proof. Since $\mathrm{SOL}(q, M, C)$ is convex, $\bar{x}+\tau(x-\bar{x}) \in \mathrm{SOL}(q, M, C)$ for all $\tau \in [0, 1]$. Hence for all such τ,

$$g(\bar{x} + \tau(x - \bar{x})) = 0,$$

which easily implies $g'(\bar{x}; x - \bar{x}) = 0$.

Since x and $\bar{x}$ belong to $\mathrm{SOL}(q, M, C)$, we have

$$\begin{aligned}
&\omega(x) = x^T(q + Mx) \\
&= x^T(q + M\bar{x}) + x^T M(x - \bar{x}) \\
&= \bar{x}^T(q + M\bar{x}) + (x - \bar{x})^T(q + M\bar{x}) + \bar{x}^T M(x - \bar{x}) + (x - \bar{x})^T M(x - \bar{x}) \\
&\geq \omega(\bar{x}) + \min\{u^T M(x - \bar{x}) : u \in \Omega(\bar{x})\} \\
&= \omega(\bar{x}) + \omega'(\bar{x}; x - \bar{x}) \geq \omega(x),
\end{aligned}$$

where the last inequality follows from the concavity of ω. □

Alternatively stated, Proposition 3.4 says that for a monotone AVI (q, M, C) and any $\bar{x} \in \mathrm{SOL}(q, M, C)$, we have

$$\begin{aligned}
&\mathrm{SOL}(q, M, C) \\
&\subseteq \{x \in \mathrm{FEA}(q, M, C) \mid g'(\bar{x}; x - \bar{x}) = 0, \omega(x) = \omega(\bar{x}) + \omega'(\bar{x}; x - \bar{x})\}.
\end{aligned} \tag{3.14}$$

We say that the *restricted minimum principle sufficiency* (RMPS) holds for the problem (3.10) if for any $\bar{x} \in \mathrm{SOL}(q, M, C)$, equality holds in (3.14); or equivalently, the implication holds:

$$\left.\begin{aligned} x \in \mathrm{FEA}(q, M, C), g'(\bar{x}; x - \bar{x}) = 0 \\ \omega(x) = \omega(\bar{x}) + \omega'(\bar{x}; x - \bar{x}) \end{aligned}\right\} \Rightarrow x \in \mathrm{SOL}(q, M, C). \tag{3.15}$$

The word *restricted* that describes this property reflects the additional restriction—equation (3.13)—that the vector x has to satisfy in order for it to be a solution of AVI (q, M, C). If ω is a smooth (linear) function on $\mathrm{FEA}(q, M, C)$ (instead of a piecewise linear function), the latter restriction is redundant. In particular, this is the case when C is a cone.

The following two results give some necessary and sufficient conditions for the two conditions, $g'(\bar{x}; x - \bar{x}) = 0$ and (3.13), to hold separately. Although these results are not needed in the proof of the main theorem in the next section, they give some insights into the RMPS property of the AVI.

PROPOSITION 3.5. *Let $q \in R^n$ and $M \in R^{n \times n}$ be arbitrary; let $C \subseteq R^n$ be a polyhedral set. Let $\bar{x} \in \mathrm{SOL}(q, M, C)$ and $x \in \mathrm{FEA}(q, M, C)$ be given. Then $g'(\bar{x}; x - \bar{x}) = 0$ if and only if $x \in \Omega(\bar{x})$ and*

$$(u - \bar{x})^T(q + Mx) \geq 0 \quad \textit{for all } u \in \Omega(\bar{x}). \tag{3.16}$$

Proof. Indeed, by Proposition 3.3, we have $g'(\bar{x}; x - \bar{x}) = 0$ if and only if

$$(x - \bar{x})^T \left(q + (M + M^T)\bar{x}\right) = \min\{u^T M(x - \bar{x}) : u \in \Omega(\bar{x})\},$$

or equivalently

$$(x - \bar{x})^T(q + M\bar{x}) = \min\{(u - \bar{x})^T M(x - \bar{x}) : u \in \Omega(\bar{x})\}.$$

Since $\bar{x} \in \mathrm{SOL}(q, M, C)$ and $x \in C$, the left-hand side is nonnegative, whereas the right-hand side is nonpositive because $\bar{x} \in \Omega(\bar{x})$. Consequently, $g'(\bar{x}; x - \bar{x}) = 0$ if and only if

$$0 = (x - \bar{x})^T (q + M\bar{x}) = \min\{(u - \bar{x})^T M(x - \bar{x}) : u \in \Omega(\bar{x})\}.$$

The first equality is equivalent to $x \in \Omega(\bar{x})$. Moreover, for all $u \in \Omega(\bar{x})$, we have $(u - \bar{x})^T (q + M\bar{x}) = 0$; hence,

$$(u - \bar{x})^T M(x - \bar{x}) = (u - \bar{x})^T (q + Mx).$$

Consequently,

$$\min\{(u - \bar{x})^T M(x - \bar{x}) : u \in \Omega(\bar{x})\} = 0$$

if and only if (3.16) holds. □

PROPOSITION 3.6. *Let $q \in R^n$ and $M \in R^{n \times n}$ be arbitrary; let $C \subseteq R^n$ be a polyhedral set. Let $\bar{x}$ and x be any two vectors in* $\mathrm{FEA}(q, M, C)$. *Then the following are equivalent:*

(i) (3.13) *holds,*
(ii) $\Omega(x) \cap \Omega(\bar{x}) \neq \emptyset$,
(iii) *for all* $\lambda \in (0, 1)$,

$$\omega(\lambda x + (1 - \lambda)\bar{x}) = \lambda\omega(x) + (1 - \lambda)\omega(\bar{x}).$$

Proof. Suppose (3.13) holds. Then for any $u \in \Omega(\bar{x})$ such that $u^T M(x - \bar{x}) = \omega'(\bar{x}; x - \bar{x})$, we have

$$\begin{aligned} \omega(x) &\leq u^T (q + Mx) \\ &= u^T (q + M\bar{x}) + u^T M(x - \bar{x}) \\ &= \omega(\bar{x}) + \omega'(\bar{x}; x - \bar{x}). \end{aligned}$$

Hence $\omega(x) = u^T (q + Mx)$, which implies $u \in \Omega(x) \cap \Omega(\bar{x})$. Conversely, if $u \in \Omega(x) \cap \Omega(\bar{x})$, then

$$\begin{aligned} \omega(x) &= u^T (q + Mx) \\ &= u^T (q + M\bar{x}) + u^T M(x - \bar{x}) \\ &\geq \omega(\bar{x}) + \omega'(\bar{x}; x - \bar{x}). \end{aligned}$$

By the concavity of ω, we have

$$\omega(x) \leq \omega(\bar{x}) + \omega'(\bar{x}; x - \bar{x}).$$

Thus (i) is equivalent to (ii). The equivalence of (ii) and (iii) follows from the fact that $\Omega(x)$ is the subdifferential of the support function of C at $-(Mx + q)$ and [9, Lem. 5.3]. □

4. The main result. We are now ready to state the main result of this paper. This result gives various necessary and sufficient conditions for the existence of a nondegenerate solution for a monotone AVI.

THEOREM 4.1. *Let $q \in R^n$ be arbitrary, $M \in R^{n \times n}$ be positive semidefinite, and $C \subseteq R^n$ be a polyhedral set. Suppose* $\mathrm{SOL}(q, M, C) \neq \emptyset$. *Let $d \in R^n$ and $\sigma \in R_+$ be the two constants associated with the AVI (q, M, C); see* (3.1). *The following statements are equivalent:*

(a) *The AVI* (q, M, C) *has a nondegenerate solution; that is,* (2.4) *holds.*
(b) *The set* $\mathrm{SOL}(q, M, C)$ *is a set of weak sharp minima for the problem* (3.6).
(c) *There exists a constant* $\gamma > 0$ *such that for all* $x \in C$

$$\mathrm{dist}\,(x \mid \mathrm{SOL}(q, M, C)) \leq \gamma g(x). \tag{4.1}$$

(d) *The representation*

$$\mathrm{SOL}(q, M, C) = \{x \in C \mid \omega(x) - (q + d)^T x + \sigma \geq 0\} \tag{4.2}$$

holds.
(e) *The restricted minimum principle sufficiency holds for the problem* (3.10); *i.e., the implication* (3.15) *holds.*

As it turns out, the proof of this theorem, except for the equivalence of (b) and (c), is rather complicated. We shall divide the entire proof into several parts. Throughout the proof we will assume, if necessary, that C is written in the form (3.3) or (3.7). Note that since the function ω is in general not differentiable, the equivalence of (b) and (e) does not follow from Proposition 2.5.

The easiest part is the equivalence of (b) and (c); this follows from the remark made after Definition 2.3 and the observation that $g_{\min} = 0$. Note that effectively, the inequality (4.1) concerns only those vectors $x \in \mathrm{FEA}(q, M, C)$; indeed, since $g(x) = \infty$ for all $x \in C \setminus \mathrm{FEA}(q, M, C)$, (4.1) trivially holds for the latter vectors x.

The following lemma establishes (a) $\Rightarrow$ (d).

LEMMA 4.2. *Under the assumptions of Theorem* 4.1, *statement* (a) *implies statement* (d).

Proof. Let S denote the right-hand set in (4.2). It suffices to verify

$$S \subseteq \mathrm{SOL}(q, M, C),$$

because the reverse inclusion is always valid. Let $x \in S$ and let $\hat{x}$ be a nondegenerate solution of AVI (q, M, C). Since $\omega(x)$ is finite, its dual program $\Delta(x)$ has an optimal solution λ that satisfies

$$b^T \lambda = \omega(x).$$

Since $\hat{x} \in \mathrm{SOL}(q, M, C)$ is nondegenerate, by Propositions 3.1 and 3.2 there exists a $\hat{\lambda} \in \Lambda(\hat{x})$ satisfying

$$\hat{\lambda}^T (A\hat{x} - b) = 0 \quad \text{and} \quad \hat{\lambda} + A\hat{x} - b > 0.$$

We have

$$\begin{aligned} \omega(x) &\geq (q + d)^T x - \sigma \\ &= (q + (M + M^T)\hat{x})^T x - \hat{x}^T M \hat{x} \\ &= (q + M\hat{x})^T x + \hat{x}^T M (x - \hat{x}) \\ &= \hat{\lambda}^T A x + (\lambda - \hat{\lambda})^T A \hat{x} \\ &= \hat{\lambda}^T (Ax - b) + \lambda^T (A\hat{x} - b) + \lambda^T b, \end{aligned}$$

which yields

$$0 \geq \hat{\lambda}^T (Ax - b) + \lambda^T (A\hat{x} - b) \geq 0.$$

Since $\hat{\lambda} + A\hat{x} - b > 0$ and λ and $Ax - b$ are both nonnegative, it follows easily that $\lambda^T(Ax - b) = 0$. Thus $x \in \mathrm{SOL}(q, M, C)$ as desired. □

Next we prove (d) $\Rightarrow$ (c). The proof of this implication uses the following consequence of the famous Hoffman error bound for systems of linear inequalities [21]. Let P be a polyhedral set in R^n, and let E and f be, respectively, a matrix and vector of compatible dimensions. If the polyhedron

$$S \equiv \{x \in P : Ex \geq f\}$$

is nonempty, then there exists a constant $c > 0$ such that

$$\mathrm{dist}\,(x \mid S) \leq c\|(Ex - f)_-\|_\infty \quad \text{for all } x \in P,$$

where the subscript $_-$ denotes the nonpositive part of a vector.

LEMMA 4.3. *Under the assumptions of Theorem* 4.1, *statement* (*d*) *implies statement* (*c*).

Proof. Invoking the function $\tilde{\omega}(x)$ defined in (3.9), we can express (4.2) equivalently as

$$\begin{aligned}&\mathrm{SOL}(q, M, C)\\ &\quad = \{x \in \mathrm{FEA}(q, M, C) \mid z^T(q + Mx) - (q + d)^T x + \sigma \geq 0, \forall z \in G\}.\end{aligned}$$

By the aforementioned consequence of Hoffman's result, we deduce the existence of a constant $\gamma > 0$ such that for all $x \in \mathrm{FEA}(q, M, C)$,

$$\mathrm{dist}\,(x \mid \mathrm{SOL}(q, M, C)) \leq \gamma \max_{z \in G} \left(z^T(q + Mx) - (q + d)^T x + \sigma\right)_- .$$

To complete the proof, it remains to verify that for all $x \in \mathrm{FEA}(q, M, C)$ and all $z \in G$

$$\left(z^T(q + Mx) - (q + d)^T x + \sigma\right)_- \leq x^T(q + Mx) - \omega(x).$$

Since $x^T(q+Mx) \geq \omega(x)$ for all $x \in C$, it suffices to show that for all $x \in \mathrm{FEA}(q, M, C)$ and $z \in G$

$$(q + d)^T x - \sigma - z^T(q + Mx) \leq x^T(q + Mx) - \omega(x);$$

in turn, since $z^T(q + Mx) \geq \omega(x)$, it suffices to verify

$$(q + d)^T x - \sigma \leq x^T(q + Mx).$$

For some $\bar{x} \in \mathrm{SOL}(q, M, C)$, the left-hand side of the above inequality is equal to

$$\begin{aligned}&(q + (M + M^T)\bar{x})^T x - \bar{x}^T M \bar{x}\\ &\quad = (q + Mx)^T x - (\bar{x} - x)^T M(\bar{x} - x) \leq (q + Mx)^T x,\end{aligned}$$

where the last inequality follows by the positive semidefiniteness of M. □

We next show that (d) and (e) are equivalent. The proof of this equivalence is based on the following lemma which shows that the two sets on the right-hand sides of (3.14) and (4.2) are equal.

LEMMA 4.4. *Under the assumptions of Theorem* 4.1,

$$\begin{aligned}&\{x \in \mathrm{FEA}(q, M, C) \mid \omega(x) - (q + d)^T x + \sigma \geq 0\}\\ &\quad = \{x \in \mathrm{FEA}(q, M, C) \mid g'(\bar{x}; x - \bar{x}) = 0, \omega(x) = \omega(\bar{x}) + \omega'(\bar{x}; x - \bar{x})\}\end{aligned} \tag{4.3}$$

for any $\bar{x} \in \mathrm{SOL}(q, M, C)$; *hence statements* (d) *and* (e) *are equivalent.*

Proof. Let x be any vector belonging to the right-hand set in (4.3). Combining (3.12) and (3.13), we deduce

$$\omega(x) = \omega(\bar{x}) + (x - \bar{x})^T(q + d).$$

Thus

$$\omega(x) - (q + d)^T x + \sigma = \omega(\bar{x}) - (q + d)^T \bar{x} + \sigma = 0,$$

where the last equality holds because $\bar{x} \in \mathrm{SOL}(q, M, C)$. This establishes one inclusion in (4.3). To show the reverse inclusion, let x belong to the left-hand set in (4.3). By the concavity of ω, we have

$$\begin{aligned} 0 &\leq \omega(x) - (q + d)^T x + \sigma \\ &\leq \omega(\bar{x}) - (q + d)^T \bar{x} + \sigma - (q + d)^T (x - \bar{x}) + \omega'(\bar{x}; x - \bar{x}) \\ &= -g'(\bar{x}; x - \bar{x}) \leq 0. \end{aligned}$$

Thus equality holds throughout and (4.3) follows. The equivalence of statements (d) and (e) is now obvious. □

Finally, we show that (c) ⇒ (a). Before presenting the details of the proof, we explain the key steps involved. First, we recall the GLCP (p, N, K) that is equivalent to the AVI (q, M, C); see (3.4) for the definition of this GLCP. Consider the convex quadratic program in the variable (x, λ):

$$\begin{aligned} &\text{minimize} && x^T(q + Mx) - b^T\lambda \\ &\text{subject to} && 0 = q + Mx - A^T\lambda \\ &&& Ax - b \geq 0, \quad \lambda \geq 0; \end{aligned} \tag{4.4}$$

this is the "natural" quadratic program associated with the GLCP (q, N, K). We will show that condition (c) in Theorem 4.1 implies that this program has a nonempty set of weak sharp minima; the proof of this implication will use Proposition 3.2. Thus by Proposition 2.5, the minimum principle sufficiency holds for (4.4). Next by using a similar proof technique as in [13, Thm. 13], we will establish that the GLCP (p, N, K) has a nondegenerate solution. Proposition 3.1 will then imply that the AVI (q, M, C) has a nondegenerate solution.

In what follows, let $y \equiv (x, \lambda)$; also let $f(y)$ denote the objective function of (4.4). Note that $f(y) = y^T(p + Ny)$ and the matrix N is positive semidefinite; moreover, the feasible region of (4.4) is precisely $\mathrm{FEA}(p, N, K)$.

LEMMA 4.5. *Under the assumptions of Theorem* 4.1, *statement* (c) *implies that*

$$\mathrm{SOL}(p, N, K) = \{y \in \mathrm{FEA}(p, N, K) \mid \nabla f(\bar{y})^T(y - \bar{y}) \leq 0\} \tag{4.5}$$

for any $\bar{y} \in \mathrm{SOL}(p, N, K)$.

Proof. Since $\mathrm{SOL}(q, M, C) \neq \emptyset$, it follows that $\mathrm{SOL}(p, N, K) \neq \emptyset$; moreover, the optimal solution set of (4.4) is equal to $\mathrm{SOL}(p, N, K)$. The claimed equation (4.5) is a consequence of the minimum principle sufficiency holding for (4.4); see Proposition 2.5. Thus by the analysis made above, it suffices to show that condition (c) in Theorem 4.1 implies that there exists a constant $\gamma' > 0$ such that

$$x^T(q + Mx) - b^T\lambda \geq \gamma' \mathrm{dist}\,(y \mid \mathrm{SOL}(p, N, K)) \tag{4.6}$$

for all $y \equiv (x, \lambda) \in \mathrm{FEA}(p, N, K)$. Let y be any such vector. Then $x \in \mathrm{FEA}(q, M, C)$ and λ is feasible to $\Delta(x)$. Thus $\Lambda(x) \neq \emptyset$ and the inequality (3.5) is valid for this pair (x, λ). We have

$$\begin{aligned} x^T(q + Mx) - b^T\lambda &= g(x) + \omega(x) - b^T\lambda \\ &\geq \gamma^{-1}\mathrm{dist}\,(x \mid \mathrm{SOL}(q, M, C)) + \alpha\mathrm{dist}\,(\lambda \mid \Lambda(x))\,, \end{aligned}$$

where the last inequality follows from (3.5) and (4.1). Pick $(x', \lambda') \in \mathrm{SOL}(q, M, C) \times \Lambda(x)$ such that

$$\|x - x'\| = \mathrm{dist}\,(x \mid \mathrm{SOL}(q, M, C)) \quad \text{and} \quad \|\lambda - \lambda'\| = \mathrm{dist}\,(\lambda \mid \Lambda(x))\,.$$

Since $x' \in \mathrm{SOL}(q, M, C)$, it follows that $\omega(x')$ is finite and thus $\Lambda(x') \neq \emptyset$. By part (c) of Proposition 3.2, there exists $\tilde{\lambda} \in \Lambda(x')$ satisfying

$$\|\lambda' - \tilde{\lambda}\| \leq \beta\|x - x'\|.$$

By part (a) of the same proposition, the pair $(x', \tilde{\lambda}) \in \mathrm{SOL}(p, N, K)$. Consequently, we have

$$\begin{aligned} &\mathrm{dist}\,(y \mid \mathrm{SOL}(q, N, K)) \\ &\leq \|x - x'\| + \|\lambda - \tilde{\lambda}\| \\ &\leq \mathrm{dist}\,(x \mid \mathrm{SOL}(q, M, C)) + \mathrm{dist}\,(\lambda \mid \Lambda(x)) + \|\lambda' - \tilde{\lambda}\| \\ &\leq (1 + \beta)\mathrm{dist}\,(x \mid \mathrm{SOL}(q, M, C)) + \mathrm{dist}\,(\lambda \mid \Lambda(x))\,. \end{aligned}$$

Thus by letting

$$\gamma' \equiv \min\left(\frac{1}{\gamma(1+\beta)}, \alpha\right),$$

it is easy to see that (4.6) must hold. □

LEMMA 4.6. *Under the assumptions of Theorem* 4.1, *statement* (c) *implies statement* (a).

Proof. It suffices to show that the GLCP (p, N, K) has a nondegenerate solution. By the expression of $\mathrm{SOL}(p, N, K)$ given in Lemma 4.5 and by expanding $\nabla f(\bar{y})^T(y - \bar{y})$, such a solution exists if and only if the following linear program in the variables $(x, \lambda, \varepsilon)$ has a feasible solution with a negative objective value:

$$\begin{aligned} \text{minimize} \quad & -\varepsilon \\ \text{subject to} \quad & 0 = q + Mx - A^T\lambda, \\ & Ax - b \geq 0, \quad \lambda \geq 0, \\ & \left(q + (M + M^T)\bar{x}\right)^T (x - \bar{x}) - b^T(\lambda - \bar{\lambda}) \leq 0, \\ & \lambda + Ax - b \geq \varepsilon e, \end{aligned}$$

where e is the vector of all ones and $(\bar{x}, \bar{\lambda})$ is an arbitrary solution of the GLCP (p, N, K). Assume that the GLCP (p, N, K) does not have a nondegenerate solution. Since the above linear program is feasible, with $(x, \lambda, \varepsilon) \equiv (\bar{x}, \bar{\lambda}, 0)$ as a feasible solution, the assumption implies that the program has an optimal solution with zero

objective value. By letting (u, v, ζ, w) be an optimal dual solution, we have

$$\begin{aligned}
&M^T u + A^T(v+w) - \zeta\left(q + (M+M^T)\bar{x}\right) = 0,\\
&-Au + b\zeta + w \le 0,\\
&e^T w = 1,\\
&v, \zeta, w \ge 0,\\
&-q^T u + b^T(v+w) + \zeta\left(b^T\bar{\lambda} - \bar{x}^T(q + (M+M^T)\bar{x})\right) = 0.
\end{aligned}$$

Premultiplying the first equation by u^T, the second constraint by $(v+w)^T$, and the last equation by $-\zeta$, adding the resulting constraints, using the fact that $b^T\bar{\lambda} - \bar{x}^T(q + M\bar{x}) = 0$, and simplifying, we deduce

$$(u - \zeta\bar{x})^T M(u - \zeta\bar{x}) + (v+w)^T w \le 0.$$

Since M is positive semidefinite and both w and v are nonnegative, the last inequality implies that $w = 0$, which contradicts the equation $e^T w = 1$. □

Combining the above lemmas, we have the following proof of Theorem 4.1.

Proof of main theorem. From Lemmas 4.2–4.4 and 4.6, as well as the previously mentioned equivalence of (b) and (c), we see that the following implications are valid:

$$\begin{array}{ccccc}
\text{(a)} & \Rightarrow & \text{(d)} & \Leftrightarrow & \text{(e)}\\
 & & \Downarrow & & \\
 & & \text{(c)} & \Rightarrow & \text{(a)}\\
 & & \Updownarrow & & \\
 & & \text{(b)} & &
\end{array}$$

Consequently, all five statements (a)–(e) are equivalent. □

In summary, Theorem 4.1 has shown that for a monotone AVI, the following five properties are equivalent: (a) existence of a nondegenerate solution, (b) existence of a nonempty set of weak sharp minima for the gap minimization problem, (c) validity of an error bound in terms of the gap function alone, (d) a simplified representation of the solution set, and (e) validity of the restricted minimum principle sufficiency for the gap minimization problem.

We conclude this paper by giving an application of Theorem 4.1 that generalizes the classical result of Goldman and Tucker [17] mentioned in the beginning of this paper.

COROLLARY 4.7. *Let $q \in R^n$ be arbitrary, $M \in R^{n\times n}$ be positive semidefinite, and $C \subseteq R^n$ be a polyhedral set. Suppose* $\mathrm{SOL}(q, M, C) \neq \emptyset$ *and* $\mathrm{FEA}(q, M, C)$ *is contained in the null space of $M + M^T$. Then the AVI (q, M, C) has a nondegenerate solution.*

Proof. Since $\mathrm{FEA}(q, M, C)$ is contained in the null space of $M + M^T$, it follows that $x^T M x = 0$ for all $x \in \mathrm{FEA}(q, M, C)$. Thus the two constants, d and σ, of the AVI (q, M, C) are both equal to zero. Moreover, it is easy to verify that the right-hand set in (4.2) reduces to

$$\{x \in C \mid u^T(q + Mx) - x^T(q + Mx) \ge 0 \text{ for all } u \in C\},$$

which is exactly $\mathrm{SOL}(q, M, C)$. Thus property (d) of Theorem 4.1 holds, and the corollary is established. □

REFERENCES

[1] F. A. Al-Khayyal and J. Kyparisis, *Finite convergence of algorithms for nonlinear programs and variational inequalities*, J. Optim. Theory Appl., 70 (1991), pp. 319–332.

[2] G. Auchmuty, *Variational principles for variational inequalities*, Numer. Funct. Anal. Optim., 10 (1989), pp. 863–874.

[3] A. Auslender, *Optimization Méthods Numériques*, Masson, Paris, 1976.

[4] J. V. Burke, *An exact penalization viewpoint of constrained optimization*, SIAM J. Control Optim., 29 (1991), pp. 968–998.

[5] J. V. Burke and M. C. Ferris, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.

[6] J. V. Burke and J. J. Moré, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197–1211.

[7] ———, *Exposing constraints*, SIAM J. Optim., 4 (1994), pp. 573–595.

[8] R. W. Cottle, J. S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, New York, 1992.

[9] V. Demýanov and L. Vasilév, *Nondifferentiable Optimization*, Optimization Software, Inc., New York, 1985.

[10] J. C. Dunn, *On the convergence of projected gradient processes to singular critical points*, J. Optim. Theory Applications, 55 (1987), pp. 203–216.

[11] M. C. Ferris, *Weak Sharp Minima and Penalty Functions in Mathematical Programming*, Ph.D. thesis, University of Cambridge, England, 1988.

[12] ———, *Finite termination of the proximal point algorithm*, Math. Programming, 50 (1991), pp. 359–366.

[13] M. C. Ferris and O. L. Mangasarian, *Minimum principle sufficiency*, Math. Programming, 57 (1992), pp. 1–14.

[14] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley & Sons, New York, 1968.

[15] M. Fukushima, *Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems*, Math. Programming, 53 (1992), pp. 99–110.

[16] S. A. Gabriel and J.-S. Pang, *An inexact NE/SQP method for solving the nonlinear complementarity problem*, Computational Optim. Appl., 1 (1992), pp. 67–91.

[17] A. J. Goldman and A. W. Tucker, *Theory of linear programming*, in Linear Inequalities and Related Systems, H. W. Kuhn and A. W. Tucker, eds., Princeton University Press, Princeton, NJ, 1956, pp. 53–97.

[18] M. S. Gowda and J. S. Pang, *On the boundedness and stability of solutions to the affine variational inequality problem*, SIAM J. Control Optim., 32 (1994), pp. 421–441.

[19] O. Güler and Y. Ye, *Convergence behavior of interior-point algorithms*, Math. Programming, 60 (1993), pp. 215–228.

[20] D. W. Hearn, *The gap function of a convex program*, Oper. Res. Lett., 1 (1982), pp. 67–71.

[21] A. J. Hoffman, *On approximate solutions of systems of linear inequalities*, Journal of Research of the National Bureau of Standards, 49 (1952), pp. 263–265.

[22] W. Li, *Error bounds for piecewise convex quadratic programs and applications*, SIAM J. Control Optim., 33 (1995), pp. 1510–1529.

[23] Z.-Q. Luo, *Convergence Analysis of Primal-Dual Interior-Point Algorithms for Convex Quadratic Programs*, Tech. report, Communications Research Laboratory, McMaster University, Hamilton, Ontario L8S 4K1, Canada, 1992.

[24] Z.-Q. Luo, J.-S. Pang, D. Ralph, and S.-Q. Wu, *Exact penalization and stationarity conditions of mathematical programs with equilibrium constraints*, Tech. report 275, Communications Research Laboratory, McMaster University, Hamilton, Ontario, L8S 4K1, Canada, 1993.

[25] Z.-Q. Luo and P. Tseng, *Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem*, SIAM J. Optim., 2 (1992), pp. 43–54.

[26] ———, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM J. Control Optim., 30 (1992), pp. 408–425.

[27] ———, *Error bounds and convergence analysis of feasible descent methods: A general approach*, Ann. Oper. Res., 47 (1993), pp. 157–178.

[28] ———, *On the convergence rate of dual ascent methods for strictly convex minimization*, Math. Oper. Res., 18 (1993), pp. 846–867.

[29] O. L. Mangasarian, *Nonlinear Programming*, McGraw–Hill, New York, 1969.

[30] ———, *Sufficiency of exact penalty minimization*, SIAM J. Control Optim., 23 (1985), pp. 30–37.

[31] O. L. MANGASARIAN, *Error bounds for nondegenerate monotone linear complementarity problems*, Math. Programming, 48 (1990), pp. 437–445.

[32] O. L. MANGASARIAN AND R. R. MEYER, *Nonlinear perturbation of linear programs*, SIAM J. Control Optim., 17 (1979), pp. 745–752.

[33] O. L. MANGASARIAN AND T.-H. SHIAU, *Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems*, SIAM J. Control Optim., 25 (1987), pp. 583–595.

[34] O. L. MANGASARIAN AND M. V. SOLODOV, *Nonlinear complementarity as unconstrained and constrained minimization*, Math. Programming, 62 (1993), pp. 277–298.

[35] R. D. C. MONTEIRO AND T. TSUCHIYA, *Limiting behavior of the derivatives of certain trajectories associated with a monotone horizontal linear complementarity problem*, Tech. report, Department of Systems and Industrial Engineering, University of Arizona, Tucson, 1992. Mathematics of Operations Research, to appear.

[36] R. D. C. MONTEIRO AND S. J. WRIGHT, *Local convergence of interior–point algorithms for degenerate LCP*, Comput. Optim. Appl., 3 (1994), pp. 131–155.

[37] J. S. PANG, *Inexact Newton methods for the nonlinear complementarity problem*, Math. Programming, 36 (1986), pp. 54–71.

[38] ———, *Complementarity problems*, in Handbook in Global Optimization, R. Horst and P. Pardalos, eds., Kluwer Academic Publishers, Boston, 1994.

[39] J. A. REINOZA, *A Degree for Generalized Equations*, Ph.D. thesis, University of Wisconsin, Madison, Wisconsin, 1979.

[40] S. M. ROBINSON, *Local structure of feasible sets in nonlinear programming, Part III: Stability and sensitivity*, Mathematical Programming Study, 30 (1987), pp. 45–66.

[41] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[42] S. J. WRIGHT, *A Path–Following Infeasible–Interior–Point Algorithm for Linear and Quadratic Problems*, Tech. Report MCS–P401–1293, Argonne National Laboratory, Argonne, IL, 1993.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 264–294, January 1996

012

CONSTRAINED LQR PROBLEMS IN ELLIPTIC DISTRIBUTED CONTROL SYSTEMS WITH POINT OBSERVATIONS*

ZHONGHAI DING†, LINK JI‡, AND JIANXIN ZHOU†

Abstract. In this paper, we study (bound constrained) LQR problems in distributed control systems governed by the elliptic equation with point observations, which are motivated by problems in corrosion engineering and contemporary "smart materials." Several regularity and characterization theorems have been established. In particular, three decomposition formulas are obtained to characterize the optimal control and optimal layer density, and are used to direct the numerical computations. These results cannot be obtained by the traditional Galerkin variational method. In the process, several useful lemmas are established, which are of independent interest. We point out that the classical Lagrangian multiplier method (LMM) may fail to provide a reliable numerical algorithm. Based on our characterization results and the boundary element method, two algorithms are proposed to carry out numerical computations. It has been shown by our numerical experiments that both algorithms are efficient and insensitive to the partition number of the boundary. An adaptive local refinement scheme has also been designed to handle the rough behavior of the optimal solution around sensor locations.

Key words. linear-quadratic regulator, distributed boundary control, point observation, potential theory, boundary element method, numerical method

AMS subject classifications. 49N10, 49J20, 93C20, 31A10, 65N38

1. Introduction.

Let Ω be an (interior or exterior) open domain in $\Re^{\mathcal{N}}$ ($\mathcal{N} = 2, 3$) with bounded smooth boundary $\Gamma = \Gamma_0 \cup \Gamma_c$ (i.e., Γ is a C^∞ surface). We study the following linear-quadratic regulator (LQR) problem in a distributed parameter control system governed by the elliptic equation

$$\text{LQR} \left\{ \begin{array}{l} \min\limits_{u\in\mathcal{U}} J(u) = \sum\limits_{k=1}^{M} \mu_k |w(P_k) - Z_k|^2 + \gamma \int_{\Gamma_c} u^2(x) d\sigma_x, \\ \text{subject to (1.1)} \left\{ \begin{array}{rcll} \Delta w(x) & = & f(x), & \text{in } \Omega, \\ \dfrac{\partial w(x)}{\partial n} & = & g(x), & \text{on } \Gamma_0, \\ \dfrac{\partial w(x)}{\partial n} & = & u(x), & \text{on } \Gamma_c, \\ \int_{\Gamma_c} u(x) d\sigma_x & = & -\int_{\Gamma_0} g(x) d\sigma_x & + \int_{\Omega} f(x) dx, \end{array} \right. \end{array} \right.$$

where

$f(x)$ is a given (loading) function in Ω,
$\frac{\partial}{\partial n}$ is the outward normal derivative,
$g(x)$ is a given Neumann type boundary data on Γ_0,
$u \in \mathcal{U}$, is a Neumann type boundary control on Γ_c,
$\mathcal{U} \subset L^2(\Gamma_c)$, is the admissible control set,
$\gamma, \mu_k > 0, 1 \le k \le M$, are given weighting factors,
$P_k \in \partial\Omega, 1 \le k \le M$, are prescribed "sensor locations,"
$Z_k \in \Re, 1 \le k \le M$, are prescribed "target" values at P_k.

* Received by the editors August 13, 1993; accepted for publication (in revised form) September 16, 1994.

† Department of Mathematics, Texas A&M University, College Station, TX 77843. The research of these authors was supported in part by AFOSR grant 91-0097.

‡ Department of Oceanography, Texas A&M University, College Station, TX 77843.

In the above setting of the LQR problem, the loading function $f(x)$, Neumann boundary data $g(x)$ and the admissible control set $\mathcal{U}$ will be chosen such that the solution $w(x)$ of Neumann problem (1.1) is continuous on $\overline{\Omega}$, otherwise point observations $w(P_k)$ $(1 \leq k \leq M)$ will be meaningless. In fact, f can be any given function in $H^r(\Omega)$, $r > -1$ when $\mathcal{N} = 2$ and $r > -\frac{1}{2}$ when $\mathcal{N} = 3$; g can be any given function in $L^2(\Gamma_0)$ when $\mathcal{N} = 2$ and in $L^p(\Gamma_0)$ $(p > 2)$ when $\mathcal{N} = 3$. The admissible control set will be specified later (see (2.3) and (3.4)).

The study of the above system is motivated by problems in cathodic protection systems in corrosion engineering (see [13] and [14]). Cathodic protection systems have been employed extensively in ships, offshore structures and pipeline networks, and other structures in a corrosive environment. The physical domain Ω occupied by a "corrosive fluid" can be either finite, as in the case of electrolyte container protection, or infinite with a bounded boundary, as in the case of ship propeller protection where Ω is the sea surrounding the ship. The boundary Γ is the walls of the container or the surface of the ship, and Γ_0 is the insulated (or painted) part of the boundary. The only control is the current $(\frac{\partial w}{\partial n} = u)$ on the part Γ_c of the boundary. So the LQR problem is to obtain an optimal current $u(x)$ on Γ_c (anodes) that produces a desired potential distribution $w(x)$ in a certain interested area (cathode) which is to be protected. For contemporary "smart materials," $P_k, 1 \leq k \leq M$, are the locations of piezoelectric sensors to measure the deformation at these points and $w(P_k), 1 \leq k \leq M$, are called point observations. We wish to find the values of $u(x)$ on Γ_c such that at sensor locations $P_k, 1 \leq k \leq M$, the observation values $w(P_k)$ are as close as possible to the target values Z_k with least possible control cost $\int_{\Gamma_c} u^2(x) d\sigma_x$. The formulation of LQR can also be adjusted to meet other physical interest.

General LQR problems governed by elliptic equations on smooth domains were first studied by J. L. Lions in Chapter 2 of [9] and were based essentially upon the Galerkin variational method, which leads to a characterization formula of optimal control coupled with an adjoint system. Unfortunately it is not directly applicable to our LQR problem. Recently, Ji and Chen [7] studied the above LQR problem by using the potential theory and boundary element method (BEM). Their approach has been shown to provide certain important advantages over the traditional Galerkin variational approach. It can provide rather explicit information about the control and state, and it is amenable to direct numerical computation through BEM. Motivated by the observation of their numerical results and computer graphics, Ji and Chen [7] proved some regularity results. They show that for $\mathcal{N} = 3$, LQR has no nontrivial solution and for $\mathcal{N} = 2$, LQR has a unique solution that may contain certain singularities at sensor locations. At the end of their paper, they proposed to study the LQR problems with a bound constraint on the control. Once an inequality constraint is added to the LQR problems, numerically it becomes very difficult to handle. The optimal control behaves roughly, especially around the sensor locations. The unknown singularities around sensor locations may result in divergence and instability in numerical computations.

Motivated by Ji and Chen's results, we will use the potential theory to study the constrained and unconstrained LQR problems. We first prove certain regularity and characterization results for unconstrained LQR. From our characterization theorem and the singularity decomposition formula of optimal control, we point out that the classical Lagrangian multiplier method (LMM) is not reliable to provide numerical solution for unconstrained LQR. Then we establish several regularity and characterization theorems for the constrained LQR. Our singularity decomposition formulas

play key roles in characterizing the singularities in optimal controls and optimal layer densities. In the course of the proofs, several useful lemmas have been established, which are of independent interest. Based upon our (decomposed) characterization results for the constrained LQR, a gradient truncation method and an iterative truncation method have been developed to carry out numerical computations on several test problems. In both methods, truncation techniques have been proposed to handle the bound constraints. An adaptive local refinement scheme is proposed in the iterative truncation method to enhance the convergence and stability in numerical computations. Our test problems show that both methods are efficient and insensitive to the partition number of the boundary of domain. This is a significant advantage of our methods over other numerical methods. Since the optimal control problem under consideration is governed by a partial differential equation, the partition number of the boundary can be very large and any numerical method sensitive to the partition number of boundary may fail to carry out numerical computations.

The results in this paper are derived for interior domain problems. For exterior domain problems, parallel results can also be obtained with suitable modifications.

Before the discussion of LQR problems, let us briefly recall the potential theory, BEM, and Ji and Chen's results.

Let $E(x,\xi)$ be the fundamental solution of the Laplacian, i.e.,

$$\Delta_\xi E(x,\xi) = -\delta(x-\xi), \quad x,\xi \in \Re^{\mathcal{N}}. \tag{1.2}$$

It is well known [1, p. 214] that

$$E(x,\xi) = \begin{cases} -\dfrac{1}{2\pi}\ln|x-\xi|, & x,\xi \in \Re^2, \\ \dfrac{1}{4\pi}\dfrac{1}{|x-\xi|}, & x,\xi \in \Re^3. \end{cases} \tag{1.3}$$

According to [1, Chap. 6], any solution w of (1.1) can be represented as a sum of a volume potential and a simple-layer-potential:

$$w(x) = \left[-\int_\Omega E(x,\xi)f(\xi)d\xi\right] + \int_\Gamma E(x,\xi)\eta(\xi)d\sigma_\xi, \quad x \in \Omega, \tag{1.4}$$

where η is called a layer density, to be determined from the boundary integral equation (BIE)

$$\frac{1}{2}\eta(x) + \int_\Gamma \frac{\partial E(x,\xi)}{\partial n_x}\eta(\xi)d\sigma_\xi = \frac{\partial w(x)}{\partial n} + \frac{\partial}{\partial n_x}\int_\Omega E(x,\xi)f(\xi)d\xi, \quad x \in \Gamma. \tag{1.5}$$

Since the last term of (1.5) is known and fixed once the inhomogeneous term $f(x)$ is given, without loss of generality, from now on we assume $f(x) \equiv 0$. Then the BIE for the layer density $\eta(x)$ becomes

$$\begin{cases} \dfrac{1}{2}\eta(x) + \displaystyle\int_\Gamma \frac{\partial E(x,\xi)}{\partial n_x}\eta(\xi)d\sigma_\xi = g(x), & x \in \Gamma_0, \\ \dfrac{1}{2}\eta(x) + \displaystyle\int_\Gamma \frac{\partial E(x,\xi)}{\partial n_x}\eta(x)d\sigma_\xi = u(x), & x \in \Gamma_c. \end{cases} \tag{1.6}$$

Once the layer density η is found, the solution $w(x)$ of (1.1) can be computed from (1.4). Because BIE (1.6) is linear in terms of η, g, and u, for simplicity, throughout of this paper we assume $g(x) \equiv 0$.

In BEM, the boundary $\Gamma = \Gamma_0 \cup \Gamma_c$ is divided into N pieces (elements). N is called the partition number of the boundary. Assume that the layer density $\eta(x)$ is piecewise smooth, e.g., piecewise constant, piecewise linear, ..., etc.; then the BIE (1.6) becomes a linear algebraic system of order N. This system can be solved for $\eta(x_i)$ and then $w(x)$ can be computed from a discretized version of (1.4) for any $x \in \overline{\Omega}$.

The following example for LQR was considered by Ji and Chen in [7].

EXAMPLE 1. *For LQR, let* $\mathcal{N} = 2$ *and* Ω *be the unit circle centered at the origin. Assume that* $M = 3$, $P_k = k\frac{\pi}{2}, 1 \le k \le 3, Z_{(1,2,3)} = (1, 0, 1)$, *and* $\Gamma_0 = \emptyset, \Gamma_c = \Gamma$.

They have applied the LMM to reformulate the optimization problem of Example 1 and implemented numerically by BEM. Since only linear equality constraints are involved in the LQR problem, LMM leads to solving a linear system. They obtained Fig. 1 for the optimal controls with different boundary partition number N.

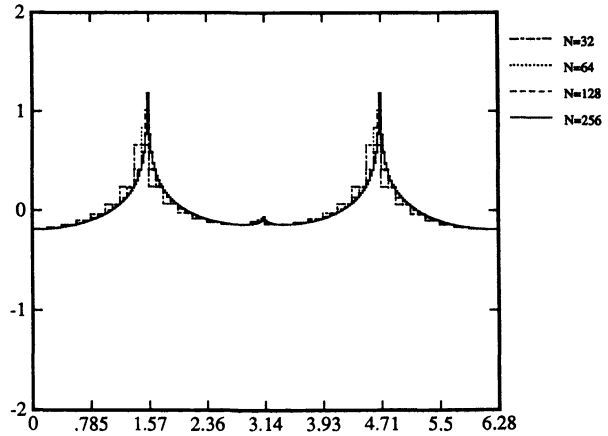


FIG. 1. *Optimal controls* $u(x)$ *of the LQR in Example* 1, *computed by LMM, for different boundary partition number* N.

In observing Fig. 1, they found that when the boundary partition number N increases, the magnitudes of the optimal control $u(x)$ at sensor locations P_k increase without bound. Motivated by this observation, Ji and Chen proved the following theorem.

THEOREM 1.1 (see [7]). (i) *For* $\mathcal{N} = 3$ *and* $\Gamma_0 = \emptyset$, *LQR does not have any nontrivial optimal control* u^* *in* $L^2(\Gamma_c)$;

(ii) *For* $\mathcal{N} = 2$, *LQR has a unique optimal control* u^* *that is differentiable at every point* $x \in \Gamma_c, x \neq P_k, 1 \le k \le M$. *While around* $P_k, 1 \le k \le M, u^*$ *has at worst a logarithmic singularity of magnitude* $\mathcal{O}(\ln|x - P_k|)$.

They were unable to characterize singularities in the optimal control of LQR. In this paper, the singularities in optimal controls of constrained and unconstrained LQR problems are displayed explicitly (see Theorems 3.9 and 4.4).

Naturally, for practical and numerical purpose, a bound constraint of control, $|u(x)| \le B$, should be added to LQR.

2. Constrained LQR problems and gradient projection method. Consider the following constrained LQR problem:

$$\text{Constrained LQR} \begin{cases} \min J(u) = \sum_{k=1}^{M} \mu_k |w(P_k) - Z_k|^2 + \gamma \int_{\Gamma_c} u^2(x) d\sigma_x, \\ \text{subject to (2.2)} \begin{cases} \Delta w(x) = 0, & \text{in } \Omega, \\ \dfrac{\partial w(x)}{\partial n} = 0, & \text{on } \Gamma_0, \\ \dfrac{\partial w(x)}{\partial n} = u(x), & \text{on } \Gamma_c, \end{cases} \\ u \in \mathcal{U}, \end{cases} \tag{2.1}$$

where

$$\mathcal{U} = \left\{ v \in L^2(\Gamma_c) \mid \int_{\Gamma_c} u(x) d\sigma_x = 0, \ \mid v(x) \mid \leq B \ \text{ on } \ \Gamma_c \right\}. \tag{2.3}$$

Let

$$L_0^p(\Gamma_c) = \left\{ f \in L^p(\Gamma_c) \mid \int_{\Gamma_c} f(\xi) d\sigma_\xi = 0 \right\}, \quad 1 \leq p < \infty.$$

When $\mathcal{N} = 2$, $\mathcal{U}$ is a bounded, closed, and convex subset of $L_0^2(\Gamma_c)$. By [10] and the Sobolev imbedding theorem, (2.1) admits a solution $w(x)$, unique up to a constant, in $H^{\frac{3}{2}}(\Omega) \subset C^{0,\alpha}(\overline{\Omega})$ where $0 < \alpha < \frac{1}{2}$ for any $u \in L_0^2(\Gamma_c)$. When $\mathcal{N} = 3$, $\mathcal{U}$ is a bounded, closed and convex subset of $L_0^p(\Gamma_c)$ for any $p > 2$. By [4], [5], and the Sobolev imbedding theorem, (2.1) admits a solution $w(x)$, unique up to a constant, in $W^{1+\frac{1}{p},p}(\Omega) \subset C^{0,\alpha}(\overline{\Omega})$ where $0 < \alpha < \frac{p-2}{p}$ for any $u \in \mathcal{U}$. Thus the point observation makes sense. For each given $u \in \mathcal{U}$, (2.1) has a unique solution $w \in C(\overline{\Omega})$ such that

$$\sum_{k=1}^{M} \mu_k |w(P_k) - Z_k|^2 = \min_{c \in \Re} \sum_{k=1}^{M} \mu_k |w(P_k) + c - Z_k|^2.$$

A calculation shows that $w(x)$ must satisfy

$$\sum_{k=1}^{M} \mu_k (w(P_k) - Z_k) = 0. \tag{2.4}$$

Therefore the constrained LQR problem is well posed. Once an optimal control is found, the optimal state can be obtained by solving (2.1) and (2.4).

By [4], [5], [10], the solution Tu of (2.1) satisfying

$$\sum_{k=1}^{M} \mu_k w(P_k) = 0$$

defines a linear bounded operator T from $L_0^2(\Gamma_c)$ to $C^{0,\alpha}(\overline{\Omega})$ with $0 < \alpha < \frac{1}{2}$ for $\mathcal{N} = 2$ and from $L_0^p(\Gamma_c)$ to $C^{0,\alpha}(\overline{\Omega})$ with $0 < \alpha < \frac{p-2}{p}$ and $p > 2$ for $\mathcal{N} = 3$. Thus

the solution $w(x)$ of (2.1) and (2.4) can be expressed as

$$w(x) = Tu(x) + \frac{1}{\sum_{k=1}^{M} \mu_k} \sum_{k=1}^{M} \mu_k Z_k.$$

THEOREM 2.1. *The constrained LQR problem has a unique optimal control* $u^* \in \mathcal{U}$ *and a unique optimal state* $w^* \in C(\overline{\Omega})$ *satisfying* (2.1) *and* (2.4), *such that*

$$J(v) \geq J(u^*), \quad \forall v \in \mathcal{U}. \tag{2.5}$$

Proof. When $\mathcal{N} = 2$, the constrained LQR problem is well posed in Hilbert space $L^2(\Gamma_c)$. By applying Theorem 1.1 of Chapter I of [9], we obtain that the constrained LQR problem admits a unique optimal control $u^* \in \mathcal{U}$ and a unique optimal state $w^* \in C(\overline{\Omega})$ satisfying (2.1) and (2.4) such that (2.5) holds. When $\mathcal{N} = 3$, the constrained LQR problem is no longer in Hilbert space. Hence Theorem 1.1 of Chapter I of [9] cannot be applied directly to the above constrained LQR problem. Let $p > 2$ and $L = \inf\{J(u) : u \in \mathcal{U}\}$. Assume $\{u_n\} \subset \mathcal{U}$ and $\{w_n(x)\}$, the solution of (2.1) and (2.4) with $u(x) = u_n(x)$, such that

$$L \leq J(u_n) \leq L + \frac{1}{n}.$$

Since $\mathcal{U}$ is bounded, closed, and convex (therefore weakly closed) in $L^p(\Gamma_c)$, there exists a subsequence $\{u_{n_m}\} \subset \{u_n\}$ and $u^* \in \mathcal{U}$ such that

$$\lim_{m \to \infty} u_{n_m} = u^* \quad \text{weakly in } L^p(\Gamma_c). \tag{2.6}$$

Hence

$$\lim_{k \to \infty} Tu_{n_m} = Tu^* \quad \text{weakly in } C^{0,\alpha}(\overline{\Omega}), \quad 0 < \alpha < \frac{p-2}{p}.$$

Let $w^*(x)$ be the solution of (2.1) and (2.4) with $u(x) = u^*(x)$. Then

$$\lim_{k \to \infty} w_{n_m} = w^* \quad \text{weakly in } C^{0,\alpha}(\overline{\Omega}), \quad 0 < \alpha < \frac{p-2}{p}.$$

Noticing that the injection from $C^{0,\alpha_1}(\overline{\Omega})$ to $C^{0,\alpha_2}(\overline{\Omega})$ is compact for any $0 < \alpha_2 < \alpha_1 < 1$, we obtain

$$\lim_{m \to \infty} w_{n_m} = w^* \quad \text{strongly in } C^{0,\alpha}(\overline{\Omega}), \quad 0 < \alpha < \frac{p-2}{p}.$$

Hence

$$\lim_{m \to \infty} \sum_{k=1}^{M} \mu_k \left| w_{n_m}(P_k) - Z_k \right|^2 = \sum_{k=1}^{M} \mu_k \left| w^*(P_k) - Z_k \right|^2.$$

By (2.6), we have

$$\lim_{m \to \infty} \gamma \int_{\Gamma_c} (u_{n_m}(x))^2 d\sigma_x \geq \gamma \int_{\Gamma_c} (u^*(x))^2 d\sigma_x.$$

Thus

$$L = \lim_{m\to\infty} J(u_{n_m}) \geq J(u^*) \geq L.$$

Therefore the existence of optimal control of the constrained LQR problem is proved when $\mathcal{N} = 3$. The uniqueness of optimal control can be easily obtained by using the strict convexity of $J(u)$ on $\mathcal{U}$. □

Thus the existence and uniqueness of the optimal control of the constrained LQR problem is established. But with this constraint—an inequality constraint, numerically the problem becomes very tough to handle. There are three major difficulties in numerical computations:

(1) The consistency condition $\int_{\Gamma_c} u(x)d\sigma_x = 0$ — a linear equality constraint;

(2) The Neumann boundary condition

$$\begin{cases} \dfrac{\partial w(x)}{\partial n} &= 0, & x \in \Gamma_0, \\ \dfrac{\partial w(x)}{\partial n} &= u(x), & x \in \Gamma_c. \end{cases}$$

For each $u(x)$, equation (2.1) has many solutions $w(x)$, which is unique up to a constant;

(3) The bound constraint on control, $|u(x)| \leq B, \quad x \in \Gamma_c$—an inequality constraint.

Basically there are two types of approaches in numerical algorithm design for the (constrained) LQR problems. The first is to use the layer density η as the variable of the LQR problem. The state variable w and control variable u are expressed as functions of η. Then difficulties (1) and (2) are automatically resolved. But difficulty (3) becomes much more complicated. The second is to use the control u as the variable of the LQR problem. Then difficulty (3) becomes more direct but difficulty (2) remains and difficulties (1) and (3) are mixed up.

Due to (1.4) and (1.6), the gradient ∇J of the objective functional J with respect to either the control variable u or the layer density variable η can be computed directly without invoking any adjoint systems.

Using the layer density as the variable, we apply the gradient projection method (GPM), a classical method, to solve the constrained LQR problem. So the state, control, and objective functional are all represented as functions of the layer density. Therefore the only difficulty is to deal with the inequality constraint. This method can be briefly stated as following:

Step 0: Given initial guess η_0;
Step 1: Using (1.4) and (1.6) to solve for state variable w_0 and control variable u_0;
Step 2: Compute $\nabla J(\eta_0)$;
Step 3: If some constraints are active, project $\nabla J(\eta_0)$ to the constraint surface;
Step 4: Find $\alpha_0 = \arg\min_{\alpha\in\Re} J(\eta_0 - \alpha\nabla J(\eta_0))$ subjected to $|u(\eta_0 - \alpha\nabla J(\eta_0))| \leq B$ and set $\eta = \eta_0 - \alpha_0\nabla J(\eta_0)$;
Step 5: If $|\eta - \eta_0|_{L^2(\Gamma)} < \varepsilon_\eta$ then compute $u(\eta_0 - \alpha_0\nabla J(\eta_0))$, output, and stop else set $\eta_0(x) = \eta(x) \quad x \in \Gamma$, goto Step 1.

REMARK 1. (1) Step 1 can be performed numerically by BEM [1].

(2) In Step 4, the linear search is still an inequality constrained minimization problem.

(3) More technical details have been omitted in Step 3 because this method is not important in this paper.

The GPM has been used to solve the test problem stated in Example 1. The numerical values of the cost $J_{\min}$ corresponding to different partition numbers N of the boundary and different bounds B are presented in the first column of Table 2. The optimal controls are shown in Fig. 2.

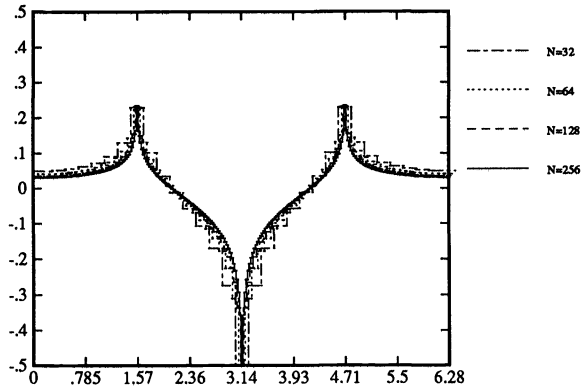


FIG. 2. *Optimal controls $u(x)$ of the LQR in Example 1, computed by GPM, for different boundary partition number N with $B = 0.5$.*

From Table 2, we found that, when the bound B is fixed and the partition number N of the boundary increases, the value of $J_{\min}$ increases. This is certainly not what we expect. The reason can be explained as follows. In the above GPM, the inequality constraint is treated *pointwisely*. When the partition number of the boundary increases, more inequality constraints are involved, hence the value of $J_{\min}$ may increase. So this method is *sensitive* to the partition number of the boundary. And compared to the two numerical algorithms developed in this paper, this method is also rather slow.

Observe that the profiles of Figs. 1 and 2 are quite different. Since the control bound used in Fig. 2 is considerably large and is active only on a very small part of the boundary, the profile of Fig. 2 should be quite close to that of Fig. 1. What is wrong? To answer this question, we need more information about the optimal control.

3. Characterizations of optimal control for LQR problems. Before giving the main results of characterization of optimal control, let us first introduce the simple layer potential $\mathcal{S}$, boundary operators $\mathcal{K}$ and $\mathcal{K}^*$, and several basic properties of these operators.

For any $f \in L^2(\Gamma)$ and $x \in \Re^{\mathcal{N}}$, define the simple layer potential by

$$\mathcal{S}(f)(x) = \int_\Gamma E(x,\xi) f(\xi) d\sigma_\xi = \begin{cases} \dfrac{1}{4\pi} \displaystyle\int_\Gamma \frac{1}{|x-\xi|} f(\xi) d\sigma_\xi, & \mathcal{N} = 3, \\ -\dfrac{1}{2\pi} \displaystyle\int_\Gamma \ln|x-\xi|\, f(\xi) d\sigma_\xi, & \mathcal{N} = 2. \end{cases} \tag{3.1}$$

The simple layer potential $\mathcal{S}$ is well defined and is continuous across the boundary Γ (see [1, pp. 223–225]). For any $f \in L^2(\Gamma)$ and $x \in \Gamma$, define the boundary operator

$\mathcal{K}$ and $\mathcal{K}^*$ by

$$
\begin{aligned}
(\mathcal{K}f)(x) &= p.v.\int_\Gamma \frac{\partial}{\partial n_\xi}E(x,\xi)f(\xi)d\sigma_\xi \\
&\equiv \lim_{\epsilon\to 0+}\int_{\Gamma\bigcap\{|x-\xi|>\epsilon\}} \frac{\partial}{\partial n_\xi}E(x,\xi)f(\xi)d\sigma_\xi \\
&= \begin{cases} \dfrac{1}{4\pi}\lim\limits_{\epsilon\to 0+}\displaystyle\int_{\Gamma\bigcap\{|x-\xi|>\epsilon\}} \dfrac{< n_\xi, x-\xi >}{|x-\xi|^3}f(\xi)d\sigma_\xi, & \mathcal{N}=3, \\ \dfrac{1}{2\pi}\lim\limits_{\epsilon\to 0+}\displaystyle\int_{\Gamma\bigcap\{|x-\xi|>\epsilon\}} \dfrac{< n_\xi, x-\xi >}{|x-\xi|^2}f(\xi)d\sigma_\xi, & \mathcal{N}=2, \end{cases}
\end{aligned} \tag{3.2}
$$

$$
\begin{aligned}
(\mathcal{K}^*f)(x) &= p.v.\int_\Gamma \frac{\partial}{\partial n_x}E(x,\xi)f(\xi)d\sigma_\xi \\
&\equiv \lim_{\epsilon\to 0+}\int_{\Gamma\bigcap\{|x-\xi|>\epsilon\}} \frac{\partial}{\partial n_x}E(x,\xi)f(\xi)d\sigma_\xi \\
&= \begin{cases} -\dfrac{1}{4\pi}\lim\limits_{\epsilon\to 0+}\displaystyle\int_{\Gamma\bigcap\{|x-\xi|>\epsilon\}} \dfrac{< n_x, x-\xi >}{|x-\xi|^3}f(\xi)d\sigma_\xi, & \mathcal{N}=3, \\ -\dfrac{1}{2\pi}\lim\limits_{\epsilon\to 0+}\displaystyle\int_{\Gamma\bigcap\{|x-\xi|>\epsilon\}} \dfrac{< n_x, x-\xi >}{|x-\xi|^2}f(\xi)d\sigma_\xi, & \mathcal{N}=2. \end{cases}
\end{aligned} \tag{3.3}
$$

For the operators defined above, the following results are well known.

THEOREM 3.1 (see [1], [3]). *Let $1 \le p < +\infty$ and $0 \le s \le 1$; the following operators are continuous:*

$$
\begin{aligned}
\mathcal{K} &: W^{s,p}(\Gamma) \longmapsto W^{s,p}(\Gamma); \\
\mathcal{K}^* &: W^{s,p}(\Gamma) \longmapsto W^{s,p}(\Gamma).
\end{aligned}
$$

In addition, $\mathcal{K}$ ($\mathcal{K}^*$, resp.) is the adjoint operator of $\mathcal{K}^*$ ($\mathcal{K}$, resp.) in $L^2(\Gamma)$. Here $W^{s,p}(\Gamma)$ denotes the usual Sobolev space.

THEOREM 3.2 (see [1]). *Let $s \ge 0$ and $f \in L^2(\Gamma)$; then*

$$
\mathcal{S} : H^s(\Gamma) \longmapsto H^{s+\frac{3}{2}}(\Omega)
$$

is a linear bounded operator, hence a linear bounded operator from $H^s(\Gamma)$ to $H^{s+1}(\Gamma)$, and

$$
\lim_{x\in\Omega, x\to y}\frac{\partial}{\partial n_x}\mathcal{S}(f)(x) = \frac{1}{2}f(y) + \mathcal{K}^*(f)(y), \quad a.e. \quad on\ \Gamma,
$$

where $H^s(\Gamma) = W^{s,2}(\Gamma)$.

THEOREM 3.3 ([8], [11]). *There exists an absolute constant $\varepsilon = \varepsilon(\Gamma) > 0$ such that for $2-\varepsilon \le p < +\infty$, the operator $\frac{1}{2}I + \mathcal{K}^*$ is invertible in $L_0^p(\Gamma)$, where $L_0^p(\Gamma) = \{f \in L^p(\Gamma) | \int_\Gamma f(\xi)d\sigma_\xi = 0\}$.*

Notice the fact that the operator $\frac{1}{2}I + \mathcal{K}^*$ is a linear and bounded operator from $L^p(\Gamma)$ into $L_0^p(\Gamma)$ for $1 < p < +\infty$. Then there exists a unique $f_0 \in L^p(\Gamma)$ such that

$$
\left(\frac{1}{2}I + \mathcal{K}^*\right)f_0 = 0, \qquad \int_\Gamma f_0(\xi)d\sigma_\xi = 1.
$$

It is known (see pp. 286–287 in [1]) that $\mathcal{S}f_0 = \text{constant} \neq 0$ for any $\Omega \subset \Re^3$ and $\mathcal{S}f_0 = \text{constant} \neq 0$ for "most" domains $\Omega \subset \Re^2$. Indeed, $\mathcal{S}f_0 \neq 0$ for any domain $\Omega \subset \Re^2$ with diameter $= \sup\{|x_1 - x_2| \mid x_1, x_2 \in \Omega\} < 1$ [6], and any bounded domain $\Omega \subset \Re^2$ can be transformed into a domain with diameter less than 1 by a scale transform $x \to x/k$, $x \in \Re^2$, with a sufficiently large $k > 0$. Thus, throughout this paper we assume $\mathcal{S}f_0 \neq 0$ for $\Omega \subset \Re^2$. Define

$$L^p_{\perp f_0}(\Gamma) = \left\{ g \in L^p(\Gamma) \Big| \int_\Gamma g(\xi) f_0(\xi) d\sigma_\xi = 0 \right\}.$$

Applying the interpolation theorem [10] and the results in [8], [11], we can easily prove the following theorem.

THEOREM 3.4. *For $0 \leq s \leq 1$ and $1 < p \leq 2 + \varepsilon$, where ε is an absolute constant that depends on Γ, $\frac{1}{2}I + \mathcal{K} : H^s(\Gamma) \bigcap L^2_0(\Gamma) \longmapsto H^s(\Gamma) \bigcap L^2_{\perp f_0}(\Gamma)$ is invertible and $\frac{1}{2}I + \mathcal{K} : L^p_0(\Gamma) \longmapsto L^p_{\perp f_0}(\Gamma)$ is invertible.*

With the above theorems, we can prove the first main theorem, a characterization result of optimal controls for LQR problems in $\Re^2$ with the admissible control set

$$\mathcal{U} = L^2_0(\Gamma_c) = \left\{ u \in L^2(\Gamma_c) \Big| \int_{\Gamma_c} u(x) d\sigma_x = 0 \right\}. \tag{3.4}$$

THEOREM 3.5. *Let $\Omega \subset \Re^2$. The LQR problem admits a unique optimal control $u^* \in \mathcal{U} = L^2_0(\Gamma_c)$ and a unique optimal state $w^* \in C(\overline{\Omega})$ satisfying (2.1) and (2.4) such that*

$$J(v) \geq J(u^*), \quad \forall v \in L^2_0(\Gamma_c). \tag{3.5}$$

Furthermore the optimal control is characterized by

$$u^*(x) = -\frac{1}{\gamma} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{M} \mu_k (w^*(P_k) - Z_k) E(P_k, \cdot)(x) + C_0, \quad x \in \Gamma_c, \tag{3.6}$$

where

$$C_0 = -\frac{1}{\gamma |\Gamma_c|} \int_{\Gamma_0} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{M} \mu_k (w^*(P_k) - Z_k) E(P_k, \cdot)(x) d\sigma_x \tag{3.7}$$

and

$$u^* \in H^{\frac{1}{2}-\varepsilon}(\Gamma_c), \quad w^* \in H^{2-\varepsilon}(\Omega) \quad \forall\, 0 < \varepsilon \leq \frac{1}{2}. \tag{3.8}$$

Proof. Since the LQR problem is well posed in Hilbert space $L^2(\Gamma_c)$, the existence and uniqueness of the optimal control $u^* \in L^2_0(\Gamma_c)$ and optimal state w^*, as well as (3.5) can be obtained by applying Theorem 1.1 in [9] and the Sobolev imbedding theorem. We need only prove (3.6)–(3.8). From (1.4) and (1.6), $w(x) = \mathcal{S} \circ (\frac{1}{2} + \mathcal{K}^*)^{-1} \tilde{u}(x) + C$ is the solution of equation (2.1), where C is an arbitrary constant and

$$\tilde{u}(x) = \begin{cases} 0, & x \in \Gamma_0, \\ u(x) & x \in \Gamma_c \end{cases} \tag{3.9}$$

as well as $u \in L_0^2(\Gamma_c)$. Hence

$$w^*(x) = \mathcal{S} \circ \left(\frac{1}{2}I + \mathcal{K}^*\right)^{-1} \tilde{u}^*(x) + C^*, \tag{3.10}$$

where C^* is determined from (2.4), i.e.,

$$C^* = -\frac{1}{\sum_{k=1}^M \mu_k} \sum_{k=1}^M \mu_k \left(\mathcal{S} \circ \left(\frac{1}{2}I + \mathcal{K}^*\right)^{-1} \tilde{u}^*(P_k) - Z_k\right). \tag{3.11}$$

Therefore, from (3.5) we have

$$\tilde{J}(u) \geq \tilde{J}(u^*), \qquad \forall u \in L_0^2(\Gamma_c), \tag{3.12}$$

where

$$\tilde{J}(u) = \sum_{k=1}^M \mu_k \left| \left(\mathcal{S} \circ \left(\frac{1}{2}I + \mathcal{K}^*\right)^{-1} \tilde{u}\right)(P_k) - Z_k + C^* \right|^2 + \gamma \int_{\Gamma_c} u^2(x) d\sigma_x.$$

For any $u \in L_0^2(\Gamma_c)$, $u^* + \theta(u - u^*) \in L_0^2(\Gamma_c)$ $\forall \theta \in [0,1]$,

$$\lim_{\theta \to 0+} \frac{1}{\theta}[\tilde{J}(u^* + \theta(u - u^*)) - \tilde{J}(u^*)] \geq 0.$$

Denote

$$\beta_k = \mu_k(w^*(P_k) - Z_k).$$

We obtain

$$\sum_{k=1}^M \beta_k \mathcal{S} \circ \left(\frac{1}{2}I + \mathcal{K}^*\right)^{-1} (\tilde{u} - \tilde{u}^*)(P_k) + \gamma \int_{\Gamma_c} u^*(u - u^*) d\sigma_\xi \geq 0, \quad \forall u \in L_0^2(\Gamma_c).$$

Using the definition of the simple layer $\mathcal{S}$, we have

$$\int_\Gamma \left\{\sum_{k=1}^M \beta_k E(P_k, \xi)\right\} \cdot \left(\frac{1}{2}I + \mathcal{K}^*\right)^{-1} (\tilde{u} - \tilde{u}^*)(\xi) d\sigma_\xi + \gamma \int_{\Gamma_c} u^*(\xi)(u(\xi) - u^*(\xi)) d\sigma_\xi \geq 0, \quad \forall u \in L_0^2(\Gamma_c). \tag{3.13}$$

Note that $E(P_k, \xi) = -\frac{1}{2\pi} \ln |P_k - \xi|$ and $E(P_k, \cdot) \in H^s(\Gamma)$ for $0 \leq s < \frac{1}{2}$ ([7]). It is easy to verify that

$$\sum_{k=1}^M \beta_k E(P_k, \cdot) \in H^s(\Gamma) \bigcap L^2_{\perp f_0}(\Gamma), \quad 0 \leq s < \frac{1}{2}. \tag{3.14}$$

Thus, applying Theorems 3.1 and 3.4, we have

$$\int_{\Gamma_c} \left[\frac{1}{\gamma}\left(\frac{1}{2}I + \mathcal{K}\right)^{-1} \sum_{k=1}^M \beta_k E(P_k, \xi) + u^*(\xi)\right](u(\xi) - u^*(\xi)) d\sigma_\xi \geq 0, \quad \forall u \in L_0^2(\Gamma_c). \tag{3.15}$$

Since $L_0^2(\Gamma_c)$ is a linear space, we obtain

$$\int_{\Gamma_c} \left[\frac{1}{\gamma} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{M} \beta_k E(P_k, \xi) + u^*(\xi) \right] u(\xi) d\sigma_\xi = 0, \quad \forall u \in L_0^2(\Gamma_c). \tag{3.16}$$

For any $u \in L^2(\Gamma_c)$,

$$u - \frac{1}{|\Gamma_c|} \int_{\Gamma_c} u(\xi) d\sigma_\xi \in L_0^2(\Gamma_c).$$

Thus, by the Fubini theorem, we have

$$\int_{\Gamma_c} \left[\frac{1}{\gamma} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{M} \beta_k E(P_k, \xi) + C + u^*(\xi) \right] u(\xi) d\sigma_\xi = 0, \quad \forall u \in L^2(\Gamma_c), \tag{3.17}$$

where

$$C = -\frac{1}{\gamma |\Gamma_c|} \int_{\Gamma_c} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{M} \beta_k E(P_k, \cdot)(x) d\sigma_x. \tag{3.18}$$

By Theorem 3.4, $(\frac{1}{2} I + \mathcal{K})^{-1} \sum_{k=1}^{M} \beta_k E(P_k, \cdot) \in L_0^2(\Gamma)$, we have

$$C = \frac{1}{\gamma |\Gamma_c|} \int_{\Gamma_0} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{M} \beta_k E(P_k, \cdot)(x) d\sigma_x. \tag{3.19}$$

It follows from (3.17) that

$$\frac{1}{\gamma} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{M} \beta_k E(P_k, \xi) + C + u^*(\xi) = 0, \quad \text{a.e. on } \Gamma_c.$$

Therefore

$$u^*(x) = -\frac{1}{\gamma} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{M} \beta_k E(P_k, \cdot)(x) - C, \quad \text{a.e. on } \Gamma_c, \tag{3.20}$$

where C is determined by (3.18) or (3.19). By (3.14) and Theorem 3.4,

$$u^* \in H^{\frac{1}{2}-\epsilon}(\Gamma_c), \qquad 0 < \epsilon \leq \frac{1}{2}. \tag{3.21}$$

By Theorems 3.2 and 3.3, it follows from (3.11) and (3.22) that

$$w^* \in H^{2-\epsilon}(\Omega), \qquad 0 < \epsilon \leq \frac{1}{2}.$$

Thus we have proved (3.6)–(3.8). □

To see the profile of the optimal control u^*, we need to know more information (i.e., singular behavior) of $u^*(x)$ at each sensor location $P_k, 1 \leq k \leq M$. The singularities of the optimal control u^* around sensor locations are displayed through Theorem 3.9. Let us first establish the following two lemmas.

LEMMA 3.6. *For $\Omega \subset \Re^{\mathcal{N}}$ $(\mathcal{N} = 2, 3)$ and $P \in \Gamma$, we have*

$$\mathcal{K}E(P,\cdot)(x) = \mathcal{S}\left(\frac{\partial}{\partial n_\xi}E(P,\xi)\right)(x), \quad \forall x \in \Gamma. \tag{3.22}$$

Proof. For $x \in \Gamma$ and $x \neq P$, let $\Omega_\epsilon = \Omega \setminus \{B(x,\epsilon) \bigcup B(P,\epsilon)\}$ and $\Gamma_\epsilon = \partial\Omega_\epsilon$ for any sufficiently small $\epsilon > 0$, where $B(x,\epsilon)$ is the ball of radius ϵ centered at x. Let n_ξ always denote the unit outward normal along Γ_ϵ for $\xi \in \Gamma_\epsilon$. It follows from the divergence theorem that

$$\int_{\Gamma_\epsilon}\left[\frac{\partial}{\partial n_\xi}E(x,\xi)\right]E(P,\xi)d\sigma_\xi = \int_{\Gamma_\epsilon}\left[\frac{\partial}{\partial n_\xi}E(P,\xi)\right]E(x,\xi)d\sigma_\xi. \tag{3.23}$$

Thus we have

$$\begin{aligned}
&\mathcal{K}E(P,\cdot)(x)\\
&\quad = p.v.\int_\Gamma \frac{\partial}{\partial n_\xi}E(x,\xi)E(P,\xi)d\sigma_\xi\\
&\quad \equiv \lim_{\epsilon\to 0+}\int_{\Gamma\bigcap\{|x-\xi|>\epsilon\}\bigcap\{|P-\xi|>\epsilon\}} \frac{\partial}{\partial n_\xi}E(x,\xi)E(P,\xi)d\sigma_\xi\\
&\quad = \lim_{\epsilon\to 0+}\left\{\int_{\Gamma_\epsilon}\frac{\partial}{\partial n_\xi}E(x,\xi)E(P,\xi)d\sigma_\xi - \int_{\Omega\bigcap\partial B(x,\epsilon)}\frac{\partial}{\partial n_\xi}E(x,\xi)E(P,\xi)d\sigma_\xi\right.\\
&\qquad\qquad \left. - \int_{\Omega\bigcap\partial B(P,\epsilon)}\frac{\partial}{\partial n_\xi}E(x,\xi)E(P,\xi)d\sigma_\xi\right\}\\
&\quad = \lim_{\epsilon\to 0+}\left\{\int_{\Gamma_\epsilon}\frac{\partial}{\partial n_\xi}E(P,\xi)E(x,\xi)d\sigma_\xi - \int_{\Omega\bigcap\partial B(x,\epsilon)}\frac{\partial}{\partial n_\xi}E(x,\xi)E(P,\xi)d\sigma_\xi\right.\\
&\qquad\qquad \left. - \int_{\Omega\bigcap\partial B(P,\epsilon)}\frac{\partial}{\partial n_\xi}E(x,\xi)E(P,\xi)d\sigma_\xi\right\}\\
&\quad = \lim_{\epsilon\to 0+}\left\{\int_{\Gamma\bigcap\{|x-P|>\epsilon\}\bigcap\{|P-\xi|>\epsilon\}}\frac{\partial}{\partial n_\xi}E(P,\xi)E(x,\xi)d\sigma_\xi\right.\\
&\qquad\qquad + \int_{\Omega\bigcap\partial B(x,\epsilon)}\left(\frac{\partial}{\partial n_\xi}E(P,\xi)E(x,\xi) - E(P,\xi)\frac{\partial}{\partial n_\xi}E(x,\xi)\right)d\sigma_\xi\\
&\qquad\qquad \left. + \int_{\Omega\bigcap\partial B(P,\epsilon)}\left(\frac{\partial}{\partial n_\xi}E(P,\xi)E(x,\xi) - E(P,\xi)\frac{\partial}{\partial n_\xi}E(x,\xi)\right)d\sigma_\xi\right\}.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\mathcal{K}E(P,\cdot)(x)\\
&\quad = \int_\Gamma \frac{\partial}{\partial n_\xi}E(P,\xi)E(x,\xi)d\sigma_\xi\\
&\qquad + \lim_{\epsilon\to 0+}\int_{\Omega\bigcap\partial B(x,\epsilon)}\left(\frac{\partial}{\partial n_\xi}E(P,\xi)E(x,\xi) - E(P,\xi)\frac{\partial}{\partial n_\xi}E(x,\xi)\right)d\sigma_\xi\\
&\qquad + \lim_{\epsilon\to 0+}\int_{\Omega\bigcap\partial B(P,\epsilon)}\left(\frac{\partial}{\partial n_\xi}E(P,\xi)E(x,\xi) - E(P,\xi)\frac{\partial}{\partial n_\xi}E(x,\xi)\right)d\sigma_\xi
\end{aligned}$$

$$= \int_\Gamma \frac{\partial}{\partial n_\xi} E(P,\xi)E(x,\xi)d\sigma_\xi - \frac{1}{2}E(P,x) + \frac{1}{2}E(P,x)$$
$$= \int_\Gamma E(x,\xi)\frac{\partial}{\partial n_\xi} E(P,\xi)d\sigma_\xi.$$

Hence the lemma is proved. □

LEMMA 3.7. *Let $\Omega \subset \Re^{\mathcal{N}}$ ($\mathcal{N} = 2, 3$) be a bounded domain with smooth boundary Γ. Then there exists a constant $C = C(\Gamma)$ such that*

$$|\langle n_\xi, x - \xi\rangle| \le C(\Gamma)|x - \xi|^2, \quad \forall x, \xi \in \Gamma.$$

Proof. See [1, p. 222]. □

LEMMA 3.8. *For $\Omega \subset \Re^2$, let $\{P_k\}_{k=1}^M \subset \Gamma$ and $\{\beta_k\}_{k=1}^M \subset \Re$ satisfying*

$$\sum_{k=1}^M \beta_k = 0.$$

Then

$$\left(\frac{1}{2}I + \mathcal{K}\right)^{-1} \sum_{k=1}^M \beta_k E(P_k,\cdot)(x)$$
$$= 2\sum_{k=1}^M \beta_k E(P_k,x) - 2\left(\frac{1}{2}I + \mathcal{K}\right)^{-1} \sum_{k=1}^M \beta_k \mathcal{S}\left(\frac{\partial}{\partial n_\xi} E(P_k,\xi)\right)(x) \tag{3.24}$$
$$- \sum_{k=1}^M \frac{2\beta_k}{|\Gamma|} \int_\Gamma E(P_k,\xi)d\sigma_\xi$$

and

$$\left(\frac{1}{2}I + \mathcal{K}\right)^{-1} \sum_{k=1}^M \beta_k \mathcal{S}\left(\frac{\partial}{\partial n_\xi} E(P_k,\xi)\right) \in C(\Gamma).$$

Proof. From Lemma 3.6, we can see that

$$\sum_{k=1}^M \beta_k E(P_k,x) = \left(\frac{1}{2}I + \mathcal{K}\right)(2\sum_{k=1}^m \beta_k E(P_k,\xi) + C_0)(x) - 2\mathcal{S}(\eta)(x), \tag{3.25}$$

where

$$\eta(\xi) = \sum_{k=1}^M \beta_k \frac{\partial}{\partial n_\xi} E(P_k,\xi), \tag{3.26}$$

$$C_0 = -\frac{2}{|\Gamma|}\int_\Gamma \sum_{k=1}^M \beta_k E(P_k,x)d\sigma_x. \tag{3.27}$$

Applying Theorem 3.4, we obtain that

$$\left(\frac{1}{2}I + \mathcal{K}\right)^{-1} \sum_{k=1}^M \beta_k E(P_k,\cdot)(x)$$

$$= 2\sum_{k=1}^{M}\beta_k E(P_k,x) - 2\left(\frac{1}{2}I+\mathcal{K}\right)^{-1}\sum_{k=1}^{M}\beta_k\mathcal{S}\left(\frac{\partial}{\partial n_\xi}E(P_k,\xi)\right)(x)$$
$$-\sum_{k=1}^{M}\frac{2\beta_k}{|\Gamma|}\int_\Gamma E(P_k,\xi)d\sigma_\xi.$$

Since

$$\frac{\partial}{\partial n_\xi}E(P_k,\xi) = \frac{1}{2\pi}\cdot\frac{< n_\xi, P_k-\xi >}{|P_k-\xi|^2}$$

and Γ is smooth, we obtain

$$\frac{\partial}{\partial n_\xi}E(P_k,\xi)\in L^\infty(\Gamma) \tag{3.28}$$

by applying Lemma 3.7. Hence by Theorem 3.2, we have

$$\mathcal{S}(\frac{\partial}{\partial n_\xi}E(P_k,\xi))\in H^1(\Gamma),$$

and by Theorem 3.4 and the Sobolev imbedding theorem

$$\left(\frac{1}{2}I+\mathcal{K}\right)^{-1}\sum_{k=1}^{M}\beta_k\mathcal{S}\left(\frac{\partial}{\partial n_\xi}E(P_k,\xi)\right)\in C(\Gamma). \qquad \square$$

Applying Lemma 3.8, we obtain our second main theorem—the singularity decomposition formula of the optimal control.

THEOREM 3.9. *Under the assumptions of Theorem* 3.5, *the optimal control can be decomposed as*

$$u^*(x) = -\frac{1}{\gamma}\left(\frac{1}{2}I+\mathcal{K}\right)^{-1}\sum_{k=1}^{M}\mu_k(w^*(P_k)-Z_k)E(P_k,\cdot)(x)+C_0$$
$$= -\frac{2}{\gamma}\sum_{k=1}^{M}\mu_k(w^*(P_k)-Z_K)E(P_k,x)$$
$$+\frac{2}{\gamma}\left(\frac{1}{2}I+\mathcal{K}\right)^{-1}\sum_{k=1}^{M}\mu_k(w^*(P_k)-Z_k)\mathcal{S}\left(\frac{\partial}{\partial n_\xi}E(P_k,\xi)\right)(x)+C \tag{3.29}$$

and

$$\left(\frac{1}{2}I+\mathcal{K}\right)^{-1}\sum_{k=1}^{M}\mu_k(w^*(P_k)-Z_k)\mathcal{S}\left(\frac{\partial}{\partial n_\xi}E(P_k,\xi)\right)\in C(\Gamma),$$

where

$$C = C_0+\frac{1}{\gamma}\sum_{k=1}^{M}\frac{2\mu_k}{|\Gamma|}\int_\Gamma(w^*(P_k)-Z_k)E(P_k,\xi)d\sigma_\xi.$$

TABLE 1
The values of optimal control at sensor locations.

	LMM	GPM	(3.29)
Z_k	$w^*(P_k)$	$w^*(P_k)$	$u^*(P_k)$
1	0.7562	0.7799	$+\infty$
0	0.4876	0.4402	$-\infty$
1	0.7562	0.7799	$+\infty$

From Theorem 3.9, we can see that the first term of the right-hand side of (3.29) contains all possible singular terms and that at sensor location $P_k, 1 \le k \le M$, the sign of the singular term is completely determined by the sign of the term $(w^*(P_k) - Z_k)$. Now we return to Example 1 and utilize (3.29) to answer the question asked at the end of §2. Let us examine the values of the optimal controls u^* of the LQR and the constrained LQR at three sensor locations, which are computed by LMM and by our GPM.

From Table 1, we can see that the profiles of u^*'s in Fig. 2 computed by GPM are correct and the profiles of u^*'s in Fig. 1 computed by LMM are not correct because the middle cusp, according to (3.29), should be pointed to $-\infty$.

Next we analyze the failure of LMM. Notice that in the numerical computation of LQR problems, LMM is implemented by BEM with the layer density η as the variable of the problems. The next theorem indicates that the optimal layer density η^* behaves just like the optimal control u^*. It can also be decomposed into a singular part and a bounded part.

THEOREM 3.10. *Under the assumption of Theorem* 3.5 *and* $\Gamma_0 = \emptyset$, *the optimal layer density* η^* *has the same possible singularities as that of the optimal control* u^* *and can be decomposed into following singular part and bounded part:*

$$\eta^*(x) = \left(\frac{1}{2}I + \mathcal{K}^*\right)^{-1} u^*(x)$$

$$= -\frac{4}{\gamma}\sum_{k=1}^{M} \mu_k(w^*(P_k) - Z_k)E(P_k, x) \tag{3.30}$$

$$+\frac{4}{\gamma}\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\sum_{k=1}^{M} \mu_k(w^*(P_k) - Z_k)\mathcal{S}\left(\frac{\partial}{\partial n_\xi}E(P_k, \xi)\right)(x) + 2C \tag{3.31}$$

$$-2\left(\frac{1}{2}I + \mathcal{K}^*\right)^{-1} \circ \mathcal{K}^* u^*(x), \tag{3.32}$$

where C *is defined in Theorem* 3.9.

Proof. From Theorem 3.9, we already knew that

$$\frac{4}{\gamma}\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\sum_{k=1}^{M} \mu_k(w^*(P_k) - Z_k)\mathcal{S}\left(\frac{\partial}{\partial n_\xi}E(P_k, \xi)\right) + 2C \in C(\Gamma).$$

It then follows from Theorem 3.5 that $u^* \in H^s(\Gamma), 0 \le s < \frac{1}{2}$. Since $\Omega \subset \Re^2$ has bounded smooth boundary Γ, $\mathcal{K}^*$ is a linear bounded operator from $H^s(\Gamma)$ to $C^\infty(\Gamma)$ [1, pp. 249–250]. Therefore

$$\mathcal{K}^* u^* \in C^\infty(\Gamma).$$

By Theorem 3.3, we have

$$\left(\frac{1}{2}I + \mathcal{K}^*\right)^{-1} \circ \mathcal{K}^* u^* \in C(\Gamma).$$

Hence (3.31) and (3.32) are bounded. To prove the theorem, we only need to note that

$$u^*(x) = 2\left(\frac{1}{2}I + \mathcal{K}^*\right) u^*(x) - 2\mathcal{K}^* u^*(x).$$

Thus

$$\begin{aligned}\eta^*(x) &= \left(\frac{1}{2}I + \mathcal{K}^*\right)^{-1} u^*(x) \\ &= 2u^*(x) - 2\left(\frac{1}{2}I + \mathcal{K}^*\right)^{-1} \mathcal{K}^* u^*(x). \qquad \square\end{aligned}$$

In BEM, the layer density η is approximated by piecewise smooth elements. Since the optimal layer density has the same possible (logarithmic) singularities at sensor locations as the optimal control, many elements are required around each sensor location for a good approximation. But this will cause the system size to be too large to handle. The following facts show how poor the approximation is.

$$\left[\int_{-\delta}^{\delta} \left(\ln|x| - (\ln\delta - 1)\right)^2 dx\right]^{\frac{1}{2}} = \min_{y\in\Re}\left[\int_{-\delta}^{\delta} \left(\ln|x| - y\right)^2 dx\right]^{\frac{1}{2}} = (2\delta)^{\frac{1}{2}}, \tag{3.33}$$

$$\left\|\ln x - \frac{3}{\delta}x - \left(\ln\delta - \frac{5}{2}\right)\right\|_{L^2(0,\delta)} = \min_{a,b\in\Re} \|\ln x - (ax+b)\|_{L^2(0,\delta)} = \frac{\sqrt{3}}{2}\delta^{\frac{1}{2}}, \tag{3.34}$$

$$\begin{aligned}&\left\|\ln x - \left(\frac{3}{\delta^2}x^2 - \frac{4}{\delta}x + \ln\delta - \frac{1}{3}\right)\right\|_{L^2(0,\delta)} \\ &\quad = \min_{a,b,c\in\Re} \|\ln x - (ax^2 + bx + c)\|_{L^2(0,\delta)} \\ &\quad = \left[2\delta\left(\frac{7}{3} - 2\ln\delta\right)\right]^{\frac{1}{2}} + \text{higher-order terms.}\end{aligned} \tag{3.35}$$

In other words, to have an accuracy of order 10^{-4} for the optimal layer density, one must let the length 2δ of element be of order 10^{-8}. So LMM cannot provide us a reliable numerical solution.

Since the GPM is sensitive to the partition number of the boundary and rather slow, we wish to develop some fast and reliable numerical methods for the constrained LQR problem. The next section will be devoted to this purpose and the derivation of some characterizations of optimal control of constrained LQR problems.

4. Characterizations of optimal control for constrained LQR problems. We first establish the following important lemma.

LEMMA 4.1. *Let $-\infty < a < b < +\infty$, $f(x)$ be a Lebesgue measurable function on (a,b), and $B > 0$ be given. Then there is a $C \in \Re$ such that*

$$\int_a^b [f(x)+C]_B dx = 0, \tag{4.1}$$

where

$$[f(x)+C]_B = \begin{cases} B & \text{if } f(x)+C > B, \\ f(x)+C & \text{if } -B \le f(x)+C \le B, \\ -B & \text{if } f(x)+C < -B, \end{cases} \tag{4.2}$$

is the truncation of the function $f(x)+C$ by the bound B, and the map $\phi : \Re \to \Re$ defined by

$$\phi(\lambda) = \int_a^b [f(x)+\lambda]_B dx, \quad \lambda \in \Re$$

is continuous and increasing.

Proof. The function $\phi(\lambda)$ is obviously well-defined on $\Re$ and

$$-B(b-a) \le \phi(\lambda) \le B(b-a).$$

First we will prove that $\phi(\lambda)$ is an increasing continuous function. Suppose λ_1, $\lambda_2 \in \Re$, $\lambda_1 \le \lambda_2$, and $\lambda_2 - \lambda_1 < 2B$, and let

$$\begin{aligned} I_1 &= \{x \in (a,b) |\ f(x) > B - \lambda_1\}, \\ I_2 &= \{x \in (a,b) |\ B - \lambda_2 < f(x) \le B - \lambda_1\}, \\ I_3 &= \{x \in (a,b) |\ -B - \lambda_1 \le f(x) \le B - \lambda_2\}, \\ I_4 &= \{x \in (a,b) |\ -B - \lambda_2 \le f(x) < -B - \lambda_1\}, \\ I_5 &= \{x \in (a,b) |\ f(x) < -B - \lambda_2\}. \end{aligned}$$

Then

$$I_1 \bigcup I_2 \bigcup I_3 \bigcup I_4 \bigcup I_5 = (a,b) \text{ and } I_i \bigcap I_j = \emptyset, \ \forall i \ne j.$$

Therefore

$$\begin{aligned} \phi(\lambda_2) - \phi(\lambda_1) &= \int_a^b ([f(x)+\lambda_2]_B - [f(x)+\lambda_1]_B) dx \\ &= \int_{I_1} ([f(x)+\lambda_2]_B - [f(x)+\lambda_1]_B) dx \\ &\quad + \int_{I_2} ([f(x)+\lambda_2]_B - [f(x)+\lambda_1]_B) dx \\ &\quad + \int_{I_3} ([f(x)+\lambda_2]_B - [f(x)+\lambda_1]_B) dx \\ &\quad + \int_{I_4} ([f(x)+\lambda_2]_B - [f(x)+\lambda_1]_B) dx \\ &\quad + \int_{I_5} ([f(x)+\lambda_2]_B - [f(x)+\lambda_1]_B) dx \\ &= \int_{I_2} (B - [f(x)+\lambda_1]_B) dx + \int_{I_3} (\lambda_2 - \lambda_1) dx \\ &\quad + \int_{I_4} ([f(x)+\lambda_2]_B + B) dx \ge 0. \end{aligned}$$

Hence

$$\phi(\lambda_1) \leq \phi(\lambda_2), \ \forall \lambda_1 \leq \lambda_2,$$

and

$$\begin{aligned} 0 &\leq \phi(\lambda_2) - \phi(\lambda_1) \\ &= \int_{I_2} (B - [f(x) + \lambda_1]_B) dx + \int_{I_3} (\lambda_2 - \lambda_1) dx + \int_{I_4} ([f(x) + \lambda_2]_B + B) dx \\ &\leq 2B\text{mes}(\{x \in (a,b)| \ -B - \lambda_2 \leq f(x) < -B - \lambda_1\}) \\ &\quad +(b-a)(\lambda_2 - \lambda_1) + 2B\text{mes}(\{x \in (a,b)| \ B - \lambda_2 < f(x) \leq B - \lambda_1\}). \end{aligned}$$

Applying Lusin's theorem, we have that for any given $\epsilon > 0$, $\exists \delta > 0$ such that if $|\lambda_2 - \lambda_1| < \delta$, then

$$|\phi(\lambda_2) - \phi(\lambda_1)| < \epsilon. \tag{4.3}$$

Thus $\phi(\lambda)$ is an increasing continuous function. Note

$$\lim_{\lambda \to +\infty} \text{mes}(\{x \in (a,b)| \ f(x) > B - \lambda\}) = b - a$$

and

$$\lim_{\lambda \to -\infty} \text{mes}(\{x \in (a,b)| \ f(x) < -B - \lambda\}) = b - a.$$

Thus

$$\lim_{\lambda \to +\infty} \phi(\lambda) = B(b-a),$$

$$\lim_{\lambda \to -\infty} \phi(\lambda) = -B(b-a).$$

Therefore there exists a $C \in \Re$ such that

$$\phi(C) = 0. \qquad \square$$

Using the same method, we can prove the following lemma.

LEMMA 4.2. *Let $f(x)$ be a Lebesgue measurable function on Γ_c and let $B > 0$ be given. Then there exists a $C \in \Re$ such that*

$$\int_{\Gamma_c} [f(x) + C]_B d\sigma_x = 0, \tag{4.4}$$

where

$$[f(x) + C]_B = \begin{cases} B, & \text{if } f(x) + C > B, \\ f(x) + C, & \text{if } -B \leq f(x) + C \leq B, \\ -B, & \text{if } f(x) + C < -B. \end{cases} \tag{4.5}$$

Denote

$$\mathcal{U} = \{v \in L_0^2(\Gamma_c)| \ |v(\xi)| \leq B, \qquad \text{a.e. on } \Gamma_c\}. \tag{4.6}$$

THEOREM 4.3. *The constrained LQR problem has a unique optimal control* $u^* \in \mathcal{U}$ *and a unique optimal state* $w^* \in C(\overline{\Omega})$ *satisfying* (2.1) *and* (2.4). *The optimal control is characterized by*

$$u^*(\xi) = -\left[\frac{1}{\gamma}\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\sum_{k=1}^{M}\mu_k(w^*(P_k) - Z_k)E(P_k,\xi) + C\right]_B, \quad \forall x \in \Gamma_c, \tag{4.7}$$

where C *is a constant such that*

$$\int_{\Gamma_c}\left[\frac{1}{\gamma}\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\sum_{k=1}^{M}\mu_k(w^*(P_k) - Z_k)E(P_k,\xi) + C\right]_B d\sigma_\xi = 0. \tag{4.8}$$

Proof. The existence and uniqueness of the optimal control $u^* \in \mathcal{U}$ and the optimal state w^* follow from Theorem 2.1. So we only have to prove (4.7) and (4.8). Denote

$$\beta_k = \mu_k(w^*(P_k) - Z_k).$$

For $\Omega \subset \Re^2$, repeating the same argument as (3.9)–(3.15) in the proof of Theorem 3.5, we have

$$\int_{\Gamma_c}\left[\frac{1}{\gamma}\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\sum_{k=1}^{M}\beta_k E(P_k,\xi) + u^*(\xi)\right](u(\xi) - u^*(\xi))d\sigma_\xi \geq 0, \quad \forall u \in \mathcal{U}. \tag{4.9}$$

Since $u,\ u^* \in \mathcal{U}$, we obtain

$$\int_{\Gamma_c}\left[\frac{1}{\gamma}\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\sum_{k=1}^{M}\beta_k E(P_k,\xi) + C + u^*(\xi)\right](u(\xi) - u^*(\xi))d\sigma_\xi \geq 0, \forall u \in \mathcal{U}, \tag{4.10}$$

where $C \in \Re$ is a constant in Lemma 4.2 such that

$$\int_{\Gamma_c}\left[\frac{1}{\gamma}\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\sum_{k=1}^{M}\beta_k E(P_k,\xi) + C\right]_B d\sigma_\xi = 0. \tag{4.11}$$

Let

$$G(\xi) = \frac{1}{\gamma}\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\sum_{k=1}^{M}\beta_k E(P_k,\xi) + C, \tag{4.12}$$

$$\tilde{u}(\xi) = -[G(\xi)]_B = -\left[\frac{1}{\gamma}\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\sum_{k=1}^{M}\beta_k E(P_k,\xi) + C\right]_B; \tag{4.13}$$

then $\tilde{u} \in \mathcal{U}$. For any $v \in \mathcal{U}$ we have

$$\begin{aligned}
&\int_{\Gamma_c}\left\{\frac{1}{\gamma}\left(\frac{1}{2}I + \mathcal{K}\right)^{-1}\sum_{k=1}^{M}\beta_k E(P_k,\xi) + C + \tilde{u}(\xi)\right\}(v(\xi) - \tilde{u}(\xi))d\sigma_\xi \\
&= \int_{\Gamma_c}(G(\xi) + \tilde{u}(\xi))(v(\xi) - \tilde{u}(\xi))d\sigma_\xi \qquad (4.14)\\
&= \int_{\Gamma_c^+}(G(\xi) - B)(v(\xi) + B)d\sigma_\xi + \int_{\Gamma_c^-}(G(\xi) + B)(v(\xi) - B)d\sigma_\xi \\
&\geq 0,
\end{aligned}$$

where

$$\Gamma_c^+ = \{x \in \Gamma_c | G(x) > B\} \quad \text{and} \quad \Gamma_c^- = \{x \in \Gamma_c | G(x) < -B\}.$$

Let $u(\xi) = \tilde{u}(\xi)$ in (4.10) and $v(\xi) = u^*(\xi)$ in (4.14). We have

$$\int_{\Gamma_c} \left\{ \frac{1}{\gamma} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{M} \beta_k E(P_k, \xi) + C + u^*(\xi) \right\} (\tilde{u}(\xi) - u^*(\xi)) d\sigma_\xi \geq 0, \tag{4.15}$$

and

$$\int_{\Gamma_c} \left\{ \frac{1}{\gamma} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{M} \beta_k E(P_k, \xi) + C + \tilde{u}(\xi) \right\} (u^*(\xi) - \tilde{u}(\xi)) d\sigma_\xi \geq 0. \tag{4.16}$$

Adding (4.15) to (4.16), we obtain

$$\int_{\Gamma_c} (u^*(\xi) - \tilde{u}(\xi))^2 d\sigma_\xi \leq 0.$$

Therefore

$$u^*(\xi) = \tilde{u}(\xi) \qquad \text{a.e. on } \Gamma_c.$$

Thus we have proved (4.7) and (4.8). For $\Omega \subset \Re^3$, using Theorems 3.3 and 3.4, the proof of (4.7) and (4.8) is exactly the same as above. □

REMARK 2. The constant C in (4.7), the expression of optimal control, is uniquely determined from (4.8), if Γ_c is connected. When Γ_c is not connected, the constant C may be not unique. In this case, the set of all such C's is a bounded closed interval. For all C in this interval, the value of the optimal control remains the same.

For $\Omega \subset \Re^2$, coupled with the decomposition formula (3.29), the characterization formula (4.7) can be applied in numerical algorithm to compute the exact value of the optimal control at any point $x \in \Gamma$. Unfortunately, the decomposition formula (3.29) is only valid for $\Omega \subset \Re^2$. For the same purpose, we will establish a decomposition formula for $\Omega \subset \Re^3$. The proof is different and much harder.

THEOREM 4.4. *Let $\Omega \subset \Re^3$ be a bounded domain with smooth boundary Γ. Let $\{\beta_k\}_{k=1}^M \subset \Re$ satisfy*

$$\sum_{k=1}^{M} \beta_k = 0.$$

Then the following decomposition formula holds:

$$\left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{M} \beta_k E(P_k, \cdot)(x)$$

$$= 2 \sum_{k=1}^{M} \beta_k E(P_k, x) - 4 \sum_{k=1}^{M} \beta_k \mathcal{S} \left(\frac{\partial E(P_k, \xi)}{\partial n_\xi} \right)(x) \tag{4.17}$$

$$+4 \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{M} \beta_k \mathcal{S} \circ \mathcal{K}^* \left(\frac{\partial E(P_k, \xi)}{\partial n_\xi} \right)(x) + C_0 \tag{4.18}$$

with

$$C_0 = -\frac{1}{|\Gamma|}\int_\Gamma \left[2\sum_{k=1}^{M}\beta_k E(P_k,\xi) - 4\sum_{k=1}^{M}\beta_k \mathcal{S}\left(\frac{\partial E(P_k,\xi)}{\partial n_\xi}\right)\right] d\sigma_\xi,$$

where (4.17) *is the singular part with a dominant term* $2\sum_{k=1}^{M}\beta_k E(P_k,x)$ *and* (4.18) *is the bounded part.*

To prove the theorem, we need several lemmas.

LEMMA 4.5. *Let* $\Omega \subset \Re^3$ *be a bounded domain with smooth boundary* Γ. *For any* $P \in \Gamma$,

$$E(P,\cdot) \in L^\alpha(\Gamma), \quad 1 \le \alpha < 2,$$

and

$$E(P,\cdot) \notin L^2(\Gamma).$$

Proof. See [2]. □

By Theorem 3.4 and Lemma 4.5, the function $(\frac{1}{2}I+\mathcal{K})^{-1}\sum_{k=1}^{M}\beta_k E(P_k,\cdot)$ is well defined only in $L^\alpha(\Gamma)$ with $2-\varepsilon < \alpha < 2$, where ε is defined in Theorem 3.4.

LEMMA 4.6. *For* $x, y > 0$ *and* $0 < \alpha < 1$, *we have the following inequality:*

$$\left|\frac{1}{x} - \frac{1}{y}\right| \le M(\alpha)|x-y|^\alpha \max\left\{\frac{1}{x^{1+\alpha}}, \frac{1}{y^{1+\alpha}}\right\},$$

where $M(\alpha) = (\frac{1-\alpha}{\alpha})^{1-\alpha}$.

Proof. See [2]. □

LEMMA 4.7. *Let* $\Omega \subset \Re^3$ *be a bounded domain with smooth boundary* Γ, *and* $0 < \alpha < 2$. *There exists a constant* $C = C(\Gamma,\alpha)$ *such that*

$$\int_\Gamma \frac{1}{|x-\xi|^\alpha} d\sigma_\xi \le C(\Gamma,\alpha), \quad \forall x \in \Gamma.$$

Proof. The proof follows from a direct calculation, and thus is omitted. □

LEMMA 4.8. *For any* $x, y \in \Gamma, 0 \le \alpha < 1, \alpha < \beta < \frac{1+\alpha}{2}$ *there exists a constant* $C(\Gamma,\alpha,\beta) > 0$ *such that*

$$\int_\Gamma \frac{1}{|x-\xi|^{1+\alpha}} \cdot \frac{1}{|y-\xi|} d\sigma_\xi \le C(\Gamma,\alpha,\beta) \cdot \frac{1}{|x-y|^\beta}.$$

Proof. Using the inequality $(a+b)^\beta \le a^\beta + b^\beta$, $a > 0$, $b > 0$, $0 < \beta < 1$, we have

$$\begin{aligned}
&\int_\Gamma \frac{|x-y|^\beta}{|x-\xi|^{1+\alpha}|y-\xi|} d\sigma_\xi \\
&\quad \le \int_\Gamma \frac{|x-\xi|^\beta + |y-\xi|^\beta}{|x-\xi|^{1+\alpha}|y-\xi|} d\sigma_\xi \\
&\quad = \int_\Gamma \frac{1}{|x-\xi|^{1-(\beta-\alpha)}} \cdot \frac{1}{|y-\xi|} d\sigma_\xi + \int_\Gamma \frac{1}{|x-\xi|^{1+\alpha}} \cdot \frac{1}{|y-\xi|^{1-\beta}} d\sigma_\xi \\
&\quad \le \left(\int_\Gamma \frac{1}{|x-\xi|^{p_1(1-(\beta-\alpha))}} d\sigma_\xi\right)^{\frac{1}{p_1}} \cdot \left(\int_\Gamma \frac{1}{|y-\xi|^{q_1}} d\sigma_\xi\right)^{\frac{1}{q_1}} \\
&\qquad + \left(\int_\Gamma \frac{1}{|x-\xi|^{p_2(1+\alpha)}} d\sigma_\xi\right)^{\frac{1}{p_2}} \cdot \left(\int_\Gamma \frac{1}{|y-\xi|^{q_2(1-\beta)}} d\sigma_\xi\right)^{\frac{1}{q_2}}.
\end{aligned}$$

The last inequality follows from the Hölder inequality, where $2 < p_1 < \frac{2}{1-(\beta-\alpha)}, q_1 = \frac{p_1}{p_1-1}$ and $\frac{2}{1+\beta} < p_2 < \frac{2}{1+\alpha}, q_2 = \frac{p_2}{p_2-1}$; thus $1 < p_1(1-(\beta-\alpha)) < 2, 1 < q_1 < 2$ and $1 < p_2(1+\alpha) < 2, 1 < q_2(1-\beta) < 2$. Here we have used the fact $0 \le \alpha < 1, \alpha < \beta < \frac{\alpha+1}{2}$. By using Lemma 4.7, we obtain

$$\int_\Gamma \frac{|x-y|^\beta}{|x-\xi|^{1+\alpha}|y-\xi|} d\sigma_\xi \le C(\Gamma,\alpha,\beta),$$

where $C(\Gamma,\alpha,\beta)$ is a constant depending on Γ, α, and β. Thus the lemma is proved. □

LEMMA 4.9. *Let $f \in L^q(\Gamma)$, $\frac{4}{3} < q < 2$. Let $0 < \alpha < \frac{3q-4}{q}$. Then*

$$\mathcal{S} \circ \mathcal{K}^* : L^q(\Gamma) \to C^{0,\alpha}(\Gamma)$$

is a linear continuous mapping and there exists a constant $C = C(\Gamma,\alpha,q)$ such that

$$|\mathcal{S} \circ \mathcal{K}^*(f)(x) - \mathcal{S} \circ \mathcal{K}^*(f)(y)| \le C|x-y|^\alpha \|f\|_{L^q(\Gamma)}, \quad \forall x, y \in \Gamma.$$

Proof. For $x \in \Gamma$, by Fubini's theorem, we have

$$\begin{aligned}\mathcal{S} \circ \mathcal{K}^*(f)(x) &= -\frac{1}{16\pi^2} \int_\Gamma \left[\frac{1}{|x-\xi|} \int_\Gamma \frac{\langle n_\xi, \xi-\eta\rangle}{|\xi-\eta|^3} f(\eta) d\sigma_\eta \right] d\sigma_\xi \\ &= -\frac{1}{16\pi^2} \int_\Gamma \left[\frac{1}{|x-\xi|} \int_\Gamma \frac{\langle n_\xi, \xi-\eta\rangle}{|\xi-\eta|^3} d\sigma_\xi \right] f(\eta) d\sigma_\eta.\end{aligned}$$

It follows from Lemma 3.7 that for any $x, y \in \Gamma$,

$$\begin{aligned}&|\mathcal{S} \circ \mathcal{K}^*(f)(x) - \mathcal{S} \circ \mathcal{K}^*(f)(y)| \\ &\le \frac{1}{16\pi^2} \int_\Gamma \left| \int_\Gamma \left(\frac{1}{|x-\xi|} - \frac{1}{|y-\xi|} \right) \frac{\langle n_\xi, \xi-\eta\rangle}{|\xi-\eta|^3} d\sigma_\xi \right| |f(\eta)| d\sigma_\eta \\ &\qquad \text{(by Lemma 3.7)} \\ &\le \frac{C(\Gamma)}{16\pi^2} \int_\Gamma \left\{ \int_\Gamma \left| \frac{1}{|x-\xi|} - \frac{1}{|y-\eta|} \right| \frac{1}{|\xi-\eta|} d\sigma_\xi \right\} |f(\eta)| d\sigma_\eta \\ &\qquad \text{(by Lemma 4.6)} \\ &\le \frac{C(\Gamma)M(\alpha)}{16\pi^2} \int_\Gamma \left\{ \int_\Gamma |x-y|^\alpha \max\left\{ \frac{1}{|x-\xi|^{1+\alpha}}, \frac{1}{|y-\xi|^{1+\alpha}} \right\} \frac{1}{|\xi-\eta|} d\sigma_\xi \right\} |f(\eta)| d\sigma_\eta \\ &\le \frac{C(\Gamma)M(\alpha)}{16\pi^2} |x-y|^\alpha \int_\Gamma \left\{ \int_\Gamma \left(\frac{1}{|x-\xi|^{1+\alpha}} + \frac{1}{|y-\xi|^{1+\alpha}} \right) \frac{1}{|\xi-\eta|} d\sigma_\xi \right\} |f(\eta)| d\sigma_\eta \\ &\qquad \text{(apply Lemma 4.8 with } \alpha < \beta < \tfrac{1+\alpha}{2}\text{)} \\ &\le \frac{C(\Gamma)M(\alpha)}{16\pi^2} |x-y|^\alpha C(\Gamma,\alpha,\beta) \int_\Gamma \left[\frac{1}{|x-\eta|^\beta} + \frac{1}{|y-\eta|^\beta} \right] |f(\eta)| d\sigma_\eta \\ &\le \frac{C(\Gamma)M(\alpha)}{16\pi^2} C(\Gamma,\alpha,\beta) |x-y|^\alpha \|f\|_{L^q(\Gamma)} \left\{ \left[\int_\Gamma \frac{d\sigma_\eta}{|x-\eta|^{\beta p}} \right]^{\frac{1}{p}} + \left[\int_\Gamma \frac{d\sigma_\eta}{|y-\eta|^{\beta p}} \right]^{\frac{1}{p}} \right\} \\ &\le C|x-y|^\alpha \|f\|_{L^q(\Gamma)},\end{aligned}$$

where the last ineqality follows from Lemma 4.7 with the fact that $\alpha < \beta < \frac{1+\alpha}{2}$, $p = \frac{q}{q-1}$, and $0 < \alpha < \frac{2q-2}{q}$ imply $0 < \beta p < 2$; and where the constant C depends only on q, Γ, α. □

Proof of Theorem 4.4. From [12, p. 56], we have

$$\mathcal{S} \circ \mathcal{K}^* = \mathcal{K} \circ \mathcal{S}.$$

Thus by Lemma 3.6, it is easy to verify that

$$\sum_{k=1}^{M} \beta_k E(P_k, \xi) = \left(\frac{1}{2} I + \mathcal{K}\right) \left(2 \sum_{k=1}^{M} \beta_k E(P_k, \xi) - 4 \sum_{k=1}^{M} \beta_k \mathcal{S} \left(\frac{\partial E(P_k, \xi)}{\partial n_\xi}\right) + C_0\right)$$
$$+4 \sum_{k=1}^{M} \beta_k \mathcal{S} \circ \mathcal{K}^* \left(\frac{\partial E(P_k, \xi)}{\partial n_\xi}\right).$$

Upon using Theorem 3.4, we get

$$\left(\frac{1}{2} I + \mathcal{K}\right)^{-1} \sum_{k=1}^{M} \beta_k E(P_k, \xi) = 2 \sum_{k=1}^{M} \beta_k E(P_k, \xi) - 4 \sum_{k=1}^{M} \beta_k \mathcal{S} \left(\frac{\partial E(P_k, \xi)}{\partial n_\xi}\right) + C_0$$
$$+4 \left(\frac{1}{2} I + \mathcal{K}\right)^{-1} \sum_{k=1}^{M} \beta_k \mathcal{S} \circ \mathcal{K}^* \left(\frac{\partial E(P_k, \xi)}{\partial n_\xi}\right).$$

From Lemmas 3.7 and 4.5, we have

$$\frac{\partial E(P_k, \xi)}{\partial n_\xi} \in L^q(\Gamma), \quad 1 < q < 2$$

and

$$\frac{\partial E(P_k, \xi)}{\partial n_\xi} \notin L^2(\Gamma).$$

Applying Lemma 4.8, for $0 < \beta < 1$, we have

$$\left|\mathcal{S}\left(\frac{\partial E(P_k, \xi)}{\partial n_\xi}\right)(x)\right| \leq C(\Gamma, \beta) \frac{1}{|x - P_k|^\beta}, \quad 1 \leq k \leq M,$$

where $C(\Gamma, \beta)$ depends on Γ and β. Therefore the dominant term of the singular part (4.17) is

$$2 \sum_{k=1}^{M} \beta_k E(P_k, x).$$

By Lemma 4.9, for $0 < \alpha < 1$, we have

$$\mathcal{S} \circ \mathcal{K}^* \left(\frac{\partial E(P_k, \xi)}{\partial n_\xi}\right) \in C^{0,\alpha}(\Gamma), \quad 1 \leq k \leq M.$$

Since

$$\frac{1}{2} \left(\frac{1}{2} I + \mathcal{K}\right)^{-1} \sum_{k=1}^{M} \beta_k \mathcal{S} \circ \mathcal{K}^* \left(\frac{\partial E(P_k, \xi)}{\partial n_\xi}\right)$$
$$= \sum_{k=1}^{M} \beta_k \mathcal{S} \circ \mathcal{K}^* \left(\frac{\partial E(P_k, \xi)}{\partial n_\xi}\right) - \mathcal{K} \circ \left(\frac{1}{2} I + \mathcal{K}\right)^{-1} \sum_{k=1}^{M} \beta_k \mathcal{S} \circ \mathcal{K}^* \left(\frac{\partial E(P_k, \xi)}{\partial n_\xi}\right),$$

by Theorem 3.4, Lemma 4.9, and [1], we have

$$\left(\frac{1}{2}I+\mathcal{K}\right)^{-1}\sum_{k=1}^{M}\beta_k\mathcal{S}\circ\mathcal{K}^*\left(\frac{\partial E(P_k,\xi)}{\partial n_\xi}\right)\in C(\Gamma).$$

Therefore (4.18) is continuous and then bounded on Γ. □

REMARK 3. It is worthwhile to indicate the importance of the decomposed results, Theorems 3.9 and 4.4. When $x\to P_{k_0}$ for some $1\le k_0\le M$, $E(P_{k_0},x)\to+\infty$. The original characterization formulas (3.6) and (4.7) are not computable and fail to provide further information about the optimal control u^* around the singular point P_k. Due to Theorems 3.9 and 4.4, we cannot only determine the sign of u^* at P_k but also compute the exact value of $u^*(P_k)$. These results also make it possible to develop numerical algorithms. We also obtained Theorem 4.3 from Theorem 3.9—a decomposed characterization of the optimal layer density. With this result, we are able to detect the failure of the classical LMM in providing reliable numerical solutions.

5. Numerical algorithms for solving constrained LQR problems. Motivated by characterization results obtained so far, we will develop two numerical algorithms in this section to solve the constrained LQR problems. Observe that due to equation (2.4), the optimal state w^* is uniquely determined by the optimal control. We formulate the constrained LQR problem in terms of the control variable u, so the bound constraint on controls becomes more directly handleable. The first algorithm is called the gradient truncation method (GTM) and the second one is called the iterative truncation method (ITM). Our numerical examples show that these algorithms are efficient and insensitive to the partition number of the boundary, a significant advantage over other methods. Theorems 3.9, 4.3, and 4.4, and Lemma 3.8 are used in the development of these algorithms.

The GTM is constructed using the gradient of cost functional and characterization results in §4.

Step 0: Give initial guess $u_0\in\mathcal{U}$;
Step 1: Use the potential theory and BEM (see §1) to solve $w(x)$ from

$$\begin{cases}\Delta w(x)=0 & \text{in }\Omega,\\ \dfrac{\partial w(x)}{\partial n}=0 & \text{on }\Gamma_0,\\ \dfrac{\partial w(x)}{\partial n}=u_0(x) & \text{on }\Gamma_c;\end{cases}$$

Step 2: Find C_0 such that

$$\sum_{k=1}^{M}\mu_k(w(P_k)+C_0-Z_k)=0$$

and set $w_0(x)=w(x)+C_0\quad x\in\Gamma$;
Step 3: Compute $\nabla J(u_0)$;
Step 4: Find $\alpha_0=\arg\min_\alpha J\left(u_0-\alpha*\nabla J(u_0)\right)$;
Step 5: Find C_1 such that

$$\int_{\Gamma_c}\left[u_0(x)-\alpha_0\nabla J(u_0)+C_1\right]_B d\sigma_x=0$$

and set $u(x)=\left[u_0(x)-\alpha_0\nabla J(u_0)+C_1\right]_B\quad x\in\Gamma_c$;

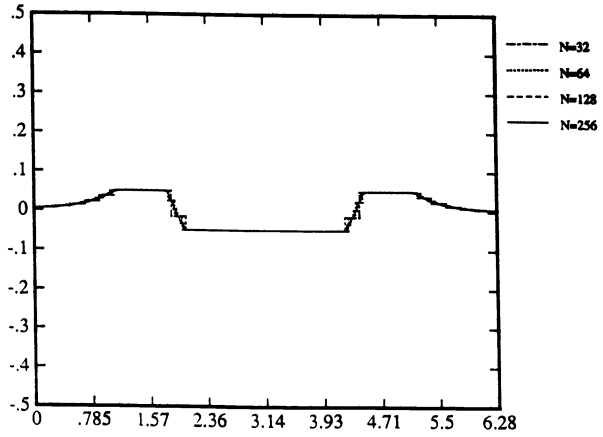


FIG. 3. *Optimal controls* $u^*(x)$ *of the constrained LQR in Example* 1, *computed by GTM with* $B = 0.05$.

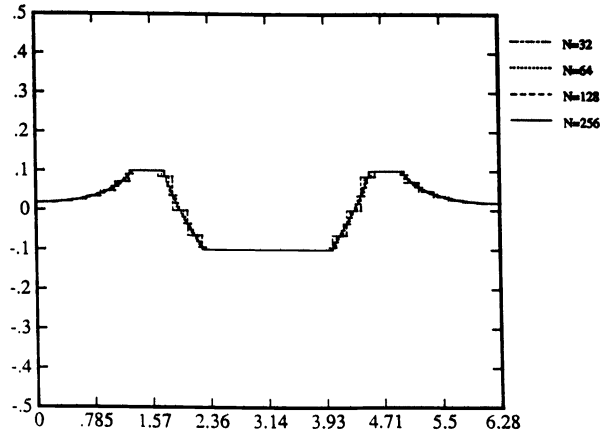


FIG. 4. *Optimal controls* $u^*(x)$ *of the constrained LQR in Example* 1, *computed by GTM with* $B = 0.1$.

Step 6: If $|u - u_0|_{L^2(\Gamma_c)} < \varepsilon_u$ then output and stop
else set $u_0(x) = u(x) \quad x \in \Gamma_c$, goto Step 1.

REMARK 4. (1) Step 4 in GTM is a minimization problem without constraints, while Step 4 of GPM solves a constrained minimization problem;

(2) Step 5 is motivated by our characterization result (4.7) and (4.8). It takes care of the consistency condition and handles the bound constraint uniformly.

Applying the GTM to the constrained LQR problem in Example 1, we obtain the second column in Table 2 and Figs. 3, 4, and 5. From Table 2, it can be seen that this method is efficient in the comparison to GPM. From Figs. 3, 4, and 5, when the control bound B increases, we observed that u^* does not reach the bound B at sensor locations where $w^*(P_k) - Z_k \neq 0$. This is different from our characterization formula (4.7) in Theorem 4.3. The reason is that GTM does not use (4.7) and (4.8) exactly (see the difference between Step 5 in GTM and (4.7)). So if we want to catch the characterization of the optimal control described in (4.7), we have to develop a numerical algorithm based on it. The ITM is constructed this way and considers the

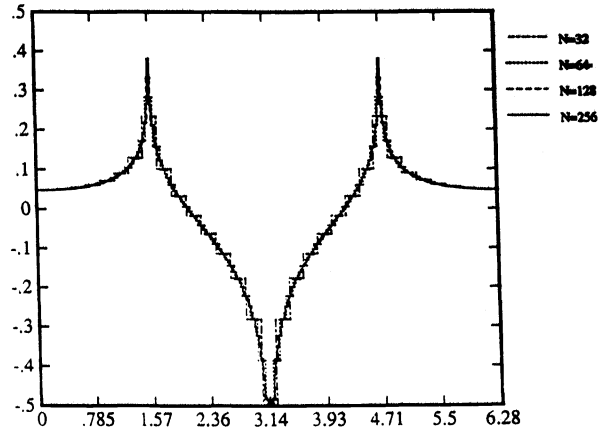


FIG. 5. *Optimal controls $u^*(x)$ of the constrained LQR in Example* 1, *computed by GTM with* $B = 0.5$.

optimal control u^* as a fixed point of (4.7).

Step 0: Given initial guess $u_0 \in \mathcal{U}$;
Step 1: Use the potential theory and BEM (see §1) to solve $w(x)$ from

$$\begin{cases} \Delta w(x) = 0 & \text{in } \Omega, \\ \dfrac{\partial w(x)}{\partial n} = 0 & \text{on } \Gamma_0, \\ \dfrac{\partial w(x)}{\partial n} = u_0(x) & \text{on } \Gamma_c; \end{cases}$$

Step 2: Find C_0 such that

$$\sum_{k=1}^{M} \mu_k (w(P_k) + C_0 - Z_k) = 0,$$

and set $w_0(x) = w(x) + C_0 \quad x \in \Gamma$;
Step 3: To compute $u(x)$, for $\mathcal{N} = 2$, use (4.7) and (3.29), for $\mathcal{N} = 3$ use (4.7), (4.17), and (4.18);
Step 4: If $|u - u_0|_{L^2(\Gamma_c)} < \varepsilon_u$ then output and stop,
else set $u_0(x) = \frac{1}{2}[u_0(x) + u(x)] \quad x \in \Gamma_c$, goto Step 1.

All the formulas in Step 3 of the ITM scheme treat the bound constraint uniformly, so it can handle only uniform bound constraints. Originally the last formula in Step 4 of the ITM scheme was

$$u_0(x) = u(x) \quad x \in \Gamma_c.$$

But this diverges in some of our numerical experiments. The current formula $u_0(x) = \frac{1}{2}[u_0(x) + u(x)] \quad x \in \Gamma_c$, a relaxation formula, is used to enhance the stability of convergence. All of our numerical experiments have confirmed this and have shown that this method is efficient and catches the characterization of the optimal control. Applying ITM and the adaptive local refinement scheme (to be explained next) to the constrained LQR problems in Example 1, we obtain Table 2 and Figs. 6, 7, and 8.

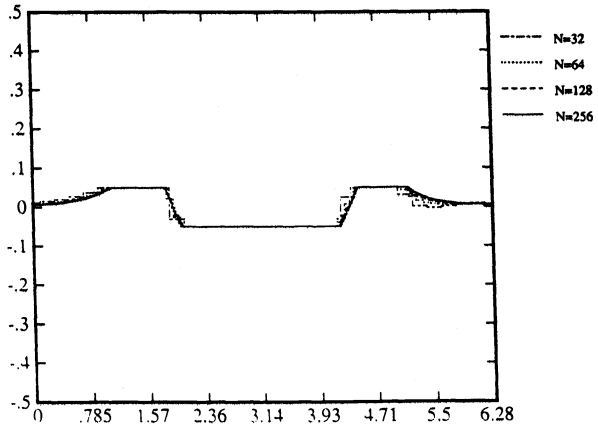


FIG. 6. *Optimal controls* $u^*(x)$ *of the constrained LQR in Example* 1, *computed by ITM with* $B = 0.05$.

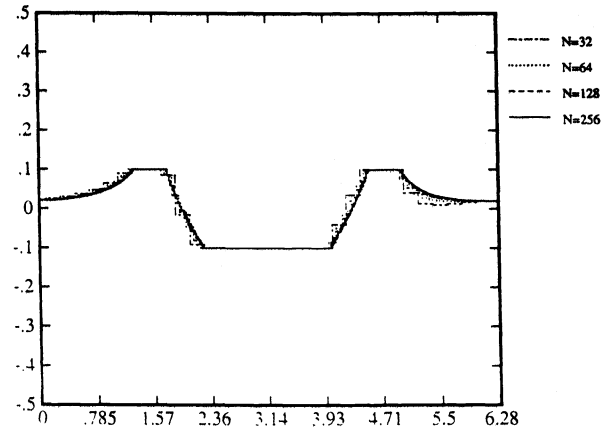


FIG. 7. *Optimal controls* $u^*(x)$ *of the constrained LQR in Example* 1, *computed by ITM with* $B = 0.1$.

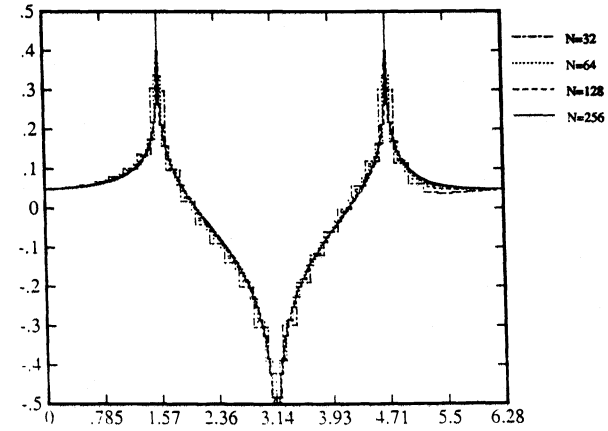


FIG. 8. *Optimal controls* $u^*(x)$ *of the constrained LQR in Example* 1, *computed by ITM with* $B = 0.5$.

TABLE 2
Comparison of convergence and $J_{\min}$.

$B = 0.05$	GPM		GTM		ITM	
N	Itn	$J_{\min}$	Itn	$J_{\min}$	Itn	$J_{\min}$
32	29	0.6261	3	0.5907	3	0.5914
64	25	0.6314	4	0.5888	3	0.5891
128	28	0.6357	5	0.5879	3	0.5880
256	27	0.6391	5	0.5874	3	0.5874
$B = 0.1$	GPM		GTM		ITM	
N	Itn	$J_{\min}$	Itn	$J_{\min}$	Itn	$J_{\min}$
32	29	0.5913	5	0.5396	6	0.5403
64	29	0.5998	6	0.5361	5	0.5363
128	28	0.6072	6	0.5343	5	0.5344
256	26	0.6134	7	0.5335	4	0.5335
$B = 0.5$	GPM		GTM		ITM	
N	Itn	$J_{\min}$	Itn	$J_{\min}$	Itn	$J_{\min}$
32	20	0.4569	5	0.4569	9	0.4609
64	25	0.4542	5	0.4487	10	0.4494
128	31	0.4613	7	0.4448	11	0.4449
256	27	0.4720	7	0.4427	11	0.4428

In Table 2, the error control is $\varepsilon_u = 10^{-4}$. One can see that both GTM and ITM are efficient in the comparison to GPM. If we compare the minimum values $J_{\min}$ w.r.t. the partition number N of the boundary, we find that both the GTM and ITM schemes are insensitive to the partition number of the boundary. This is a significant advantage of GTM and ITM schemes over other algorithms. Since the optimal control problems under consideration are governed by partial differential equations, the partition number of the boundary can be very large, therefore any numerical method sensitive to the partition number may fail to carry out the numerical computation.

Figs. 6, 7, and 8 show that the ITM scheme does catch the characterization of the optimal control at sensor locations $P_k, 1 \leq k \leq M$, no matter how large the bound is. However, this will also cause the divergence of the scheme. Since the optimal control of the LQR problem has singularities around the sensor locations, even after truncation around the sensor locations, the peak in the graph of u^* around the sensor locations may still be very narrow. When the partition number of the boundary is fixed, the length of each element in BEM is fixed. If the control bound B is large, the property that u^* reaches the bound at sensor locations will causes too much control around sensor locations and result in the divergence of the scheme. This is why ITM diverges in computing the constrained LQR problems in Example 1 for $B = 1.5$. Piecewise linear or quadratic elements are not helpful in this situation (see (3.33)–(3.35)). Refining the partition of the boundary can relieve the problem, but this will enlarge the size of the problem. To overcome this difficulty, the following adaptive local refinement scheme is proposed.

For $N = 2$, add two nodal points around each sensor location. Then carry out the ITM scheme. If ITM converges then continue. If ITM still does not converge, move two added nodal points closer to the sensor location and go on.

For $N = 3$, the basic idea is the same. When the adaptive local refinement goes on around each sensor location, the partition number of the boundary is fixed. After we coupled this scheme with the ITM scheme, the algorithm converged in all our numerical examples. The adaptive local refinement scheme can be expected to

be useful in other numerical algorithms where the solution has singularities at some points.

Finally we point out that although the motivations of GTM and ITM are quite different, it can be shown that they are essentially the same,[1] that is, Step 3 of ITM is essentially equal to Step 4 of GTM with $\alpha_0 = \frac{1}{2\gamma}$. Of course, a relaxation formula has to be added in ITM to enhance the convergence stability. Since singularities are treated differently in GTM and ITM, both methods have different advantages and disadvantages. The GTM works for the problem with any bound constraints on the control, but the numerical solution of optimal control may fail to reach the control bound around some sensor locations where, according to characterizations of optimal control, the optimal control should reach the control bound. While the numerical solution of the optimal control computed by ITM does catch the characterization of the optimal control, it may also cause stability problems in the convergence of ITM. To remedy this defect, an adaptive local refinement scheme is proposed to handle the rough behavior of the optimal control around sensor locations without enlarging the size of the problem considered. The ITM coupled with an adaptive local refinement scheme works very well for the LQR problem with any bound constraint on the control; in particular, it has a potential in solving bound-constrained LQR problems on nonsmooth domains where some sensors are placed at vertex points of the boundary (see [2]). Many details are omitted in both numerical algorithms because of space constraints, e.g., most of the integrals involved are singular integrals, and careful treatments are necessary to have a successful computation. The convergence analysis of GTM and ITM will be discussed in a separate paper.

6. Conclusion. In this paper, several regularity and characterization theorems for LQR and constrained LQR problems have been established. In particular, three decomposition formulas are proved to characterize the optimal control and the optimal layer density and to direct numerical computations. These results cannot be obtained by the traditional Galerkin variational method. Along with the proofs, several useful lemmas are established, which are of independent interest. We point out that the classical Lagrangian multiplier method is not reliable to provide numerical algorithms for this kind of problem due to the existence of singularities in the solutions. Based upon the characterization results, two numerical algorithms, GTM and ITM, have been developed to carry out the numerical computations of the constrained LQR problems. Both methods are efficient and insensitive to the partition number of the boundary, a significant advantage over other algorithms. Both algorithms have been carried out for several numerical examples and the numerical results confirm our analysis.

Finally, we point out that all of the above results are valid if the boundary Γ is a piecewise C^2 surface with point sensors placed at smooth points of the boundary. But the constrained LQR problems on nonsmooth domains with some sensors placed at the vertex points of the boundary are quite different. The theoretical analysis is much tougher. The related results are presented in a forthcoming paper [2].

Acknowledgments. The authors thank Professor G. Chen for his comments, suggestions, and constant interest in our work. We also thank two referees for their valuable suggestions and comments.

[1] We thank one of the referees for bringing this to our attention.

REFERENCES

[1] G. CHEN AND J. ZHOU, *Boundary Element Methods*, Academic Press, London, San Diego, 1992.

[2] Z. DING AND J. ZHOU, *Constrained LQR Problems Governed by the Potential Equation on Lipschitz Domain with Point Observations*, J. Math. Pures Appl., 74 (1995), pp. 317–344.

[3] E. B. FABES, M. JODEIT JR., AND N. M. RIVIERE, *Potential techniques for boundary value problems on C^1-domains*, Acta Math., 141 (1978), pp. 165–186.

[4] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1983.

[5] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman Publishing Inc., Boston, 1985.

[6] G. HSIAO AND R. C. MACCAMY, *Solutions of boundary value problems by integral equations of the first kind*, SIAM Rev., 15 (1973), pp. 687–705.

[7] L. JI AND G. CHEN, *Point observation in linear quadratic elliptic distributed control systems*, in Proceedings of American Mathematical Society Summer Conference on Control and Identification of Partial Differential Equations, Society for Industrial and Applied Mathematics, Philadelphia, 1993, pp. 155–170.

[8] C. E. KENIG, *Recent Progress on Boundary-Value Problems on Lipschitz Domains*, A.M.S. Proceedings of Symposia in Pure Mathematics, 43 (1985), pp. 175–205.

[9] J. L. LIONS, *Contrôle Optimal des Systèms Gouvernés par Deséquations aux Dèrivèes Partielles*, Dunod, Gauthier-Villars, Paris, 1968.

[10] J. L. LIONS AND E. MAGENES, *Nonhomogeneous Boundary Value Problems and Applications*, Vol.1, Springer-Verlag, New York, 1970.

[11] G. VERCHOTA, *Layer potentials and regularity for the Dirichlet problems for Laplace's equation in Lipschitz domains*, J. Funct. Anal., 59(1984), pp. 572–611.

[12] ———, *Layer Potentials and Boundary Value Problems for Laplace's Equation on Lipschitz Domains*, Ph.D Thesis, University of Minnesota, Minneapolis, 1982.

[13] N. G. ZAMANI AND J. M. CHUANG, *Optimal control of current in a cathodic protection system: A numerical investigation*, Optimal Control Appl. Methods, 8 (1987), pp. 339–350.

[14] N. G. ZAMANI, J. F. PORTER, AND A. A. MUFTI, *A survey of computational efforts in the field of corrosive engineering*, J. Numer. Methods Engrg., 23(1986), pp. 1295–1311.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 295–310, January 1996

013

AVERAGE OPTIMALITY IN MARKOV CONTROL PROCESSES VIA DISCOUNTED-COST PROBLEMS AND LINEAR PROGRAMMING*

ONÉSIMO HERNÁNDEZ-LERMA† AND JEAN B. LASSERRE‡

Abstract. This paper shows the existence of solutions to the average-cost problem for Markov control processes on Borel spaces, with possibly unbounded costs and noncompact control constraint sets. This is done via a combination of the well-known "vanishing discount" approach and recent results on the linear programming formulation for both discounted- and average-cost Markov control problems.

Key words. (discrete-time) Markov control processes, average-cost criterion, discounted-cost criterion, linear programming (in general vector spaces)

AMS subject classifications. 93E20, 90C40

1. Introduction.

Among the several ways available to analyze average-cost (AC) Markov control processes, two of the most commonly used are the "vanishing discount" and the linear programming (LP) approaches. In the latter case, one introduces a suitable linear program and its dual and gives conditions for their value to coincide and for their common value to equal the value of the AC problem. In the former approach, on the other hand, one defines discounted-cost problems with a discount factor $\alpha \in (0,1)$ and shows that under suitable assumptions, appropriately normalized functions converge to the AC value as $\alpha \uparrow 1$. These approaches are not directly comparable: they require (apparently) independent settings.

This paper presents, for Markov control processes on Borel spaces, a combination of the two approaches: First, for each $\alpha \in (0,1)$, we introduce a "discounted" linear program (P_α) and its dual (P_α^*), and then we give conditions for some modified, equivalent versions, (MP_α) and (MP_α^*), to converge in some sense, as $\alpha \uparrow 1$, to programs (MP_1), (MP_1^*) related to the AC problem. Our main results are that (MP_1) and the AC problem are both solvable, with the same value J^*, and that there is no duality gap, i.e., $J^* = \sup(MP_1^*)$. These conclusions are of course not new. What is indeed new is the way we obtain them here, via the linear programs (MP_α)–(MP_α^*). In particular, in contrast with the usual hypotheses [2], [15], [16], we do not require the discounted differential cost (h_α in Assumption 5.1) to be majorized by a finite function. Thus, on the one hand, we obtain a new set of conditions ensuring solvability of the AC problem, and on the other, we provide a setting for the comparison of the vanishing-discount and the LP approaches.

Related literature. This paper is essentially a sequel to [8] and [7], where we developed the LP formulation of AC and discounted-cost problems, respectively, for Markov control processes with *Borel* state and action spaces, allowing *unbounded* one-stage costs and *noncompact* control constraint sets. [7] and [8]—see also, e.g., [2], [3], [12], [13]—present many related references. For the vanishing-discount approach see, e.g., [2], [6], [15], [16]. The LP terminology we use is based on [1, Chap. 3].

* Received by the editors March 10, 1993; accepted for publication (in revised form) September 22, 1994. This work is part of a joint research project sponsored by CONACYT (México) and CNRS (France).

† Departamento de Matemáticas, CINVESTAV–IPN, Apartado Postal 14–740, 07000 México, D.F., México (ohernand@math.cinvestav.mx). The work of this author was also supported by CONACYT grant 1332-E9206.

‡ LAAS–CNRS, 7 Avenue du Colonel Roche, 31077 Toulouse cédex, France (lasserre@laas.fr).

Organization of the paper. Section 2 introduces the Markov control problems we are interested in. The linear programs $(P_\alpha) - (MP_\alpha)$ and their duals $(P_\alpha^*) - (MP_\alpha^*)$, $0 < \alpha < 1$, are defined in §3, together with conditions for them to be consistent. For $\alpha = 1$, the linear program (MP_1) and the AC problem are both shown to be solvable in §4; it is also shown that they are equivalent in the sense that they have the same optimal value. In §5 it is proved that there is no duality gap for (MP_1), i.e., (MP_1) and its dual (MP_1^*) have the same value. In the last section, §6, we briefly compare the usual vanishing-discount-factor approach with the version presented here, and we also make some comments on how one can get deterministic (as opposed to randomized) policies with optimality properties.

2. Markov control processes.

Notation. We essentially use the same notation as in [7], [8]. In particular, if S is a Borel space, $\mathcal{B}(S)$, $C(S)$, and $\mathcal{P}(S)$ stand for the Borel σ-algebra, the space of bounded and continuous functions, and the set of probability measures on S, respectively. If S and T are Borel spaces, then the family of stochastic kernels on S given T is denoted by $\mathcal{P}(S|T)$.

Let (X, A, Q, c) be a stationary *Markov control model* [2], [4], [5], [10] satisfying the following conditions. The *state space* X and the *action* (or control) *set* A are both Borel spaces. To each $x \in X$ is associated a nonempty set $A(x) \in \mathcal{B}(A)$ whose elements are the feasible control actions when the system is in state x. The set

$$K := \{(x, a) |\ x \in X,\ a \in A(x)\} \tag{2.1}$$

of admissible state-actions pairs is assumed to be a Borel subset of $X \times A$.

The *transition law* Q, or $Q(B|x, a)$ with $B \in \mathcal{B}(X)$ and $(x, a) \in K$, is a stochastic kernel on X given K, which is assumed to be *weakly continuous*, i.e., $\int v(y)Q(dy|.)$ is in $C(K)$ whenever $v \in C(X)$. The *one-stage cost* c is a nonnegative lower semicontinuous function on K.

To ensure that the set $\mathcal{F}$ defined next is nonempty, we will assume that K contains the graph of a measurable map. (Conditions for the latter to hold are given by so-called "measurable selection theorems" [2], [4], [5], [10].)

DEFINITION 2.1. *$\mathcal{F}$ denotes the set of all measurable functions $f : X \to A$ such that $f(x) \in A(x)$ for all $x \in X$, and Φ stands for the set of all stochastic kernels $\varphi \in \mathcal{P}(A|X)$ that satisfy $\varphi(A(x)|x) = 1$ for all $x \in X$.*

DEFINITION 2.2. *Let $H_0 := X$ and $H_t := K \times H_{t-1}$, $t = 1, 2, \ldots$. A (control) policy is a sequence $\delta = \{\delta_t\}$ of stochastic kernels $\delta_t \in \mathcal{P}(A|H_t)$ that satisfy the constraint $\delta_t(A(x_t)|h_t) = 1$ for all $h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t)$ in H_t, $t = 0, 1, \ldots$. The set of all policies is denoted by Δ. A policy δ is said to be* relaxed *(or randomized stationary) if there exists $\varphi \in \Phi$ such that $\delta_t(.|h_t) = \varphi(.|x_t)$ for all $h_t \in H_t$, $t = 0, 1, \ldots$, and it is called* deterministic stationary *if, for some $f \in \mathcal{F}$, $\delta_t(.|h_t)$ is concentrated at $f(x_t)$ for all $h_t \in H_t$, $t = 0, 1, \ldots$.*

We shall identify $\mathcal{F}$ (resp., Φ) with the set of all deterministic stationary (resp., relaxed) policies.

Let $(\Omega, \mathcal{E})$ be the measurable space that consists of the sample space $\Omega := (X \times A)^\infty$ and the corresponding product σ-algebra $\mathcal{E}$. Then for every policy δ and initial distribution $\nu \in \mathcal{P}(X)$ a probability P_ν^δ and a stochastic process $\{(x_t, a_t),\ t = 0, 1, \ldots\}$ are defined on $(\Omega, \mathcal{E})$ in a canonical way (see, e.g., [10, p. 80]), where x_t and a_t denote the state and action at time t, respectively. The expectation operator with respect to

P_ν^δ is denoted by E_ν^δ. If ν is concentrated at some initial state $x_0 = x$, we write P_ν^δ and E_ν^δ as P_x^δ and E_x^δ, respectively.

Performance criteria. For every $\alpha \in (0,1)$, let

$$V_\alpha(\delta,\nu) := E_\nu^\delta\left[\sum_{t=0}^{\infty}\alpha^t c(x_t,a_t)\right] \tag{2.2}$$

be the total expected α-discounted cost when using the policy δ, given the initial distribution ν. We may write

$$V_\alpha(\delta,\nu) := \int_X V_\alpha(\delta,x)\nu(dx), \quad \text{where } V_\alpha(\delta,x) := E_x^\delta\left[\sum_{t=0}^{\infty}\alpha^t c(x_t,a_t)\right].$$

Let us define

$$V_\alpha^*(\nu) := \inf_\delta V_\alpha(\delta,\nu), \quad \nu \in \mathcal{P}(X).$$

For a given ν, δ is said to be *α-discount ν-optimal* if $V_\alpha(\delta,\nu) = V_\alpha^*(\nu)$, whereas if $V_\alpha(\delta,x) = V_\alpha^*(x)$ for all $x \in X$, then δ is said to be *pointwise α-discount optimal.*

Now we define the long-run expected AC when using the policy δ, given the initial distribution ν, as

$$J(\delta,\nu) := \limsup_{n\to\infty}\, n^{-1}E_\nu^\delta\left[\sum_{t=0}^{n-1}c(x_t,a_t)\right], \tag{2.3}$$

and let

$$J^* := \inf_\nu \inf_\delta J(\delta,\nu). \tag{2.4}$$

(In [8], J^* is written as $\inf_\Delta$ AC.) A pair (δ^*,ν^*) consisting of a policy δ^* and an initial distribution ν^* is said to be a *minimum pair* [8], [11] if $J(\delta^*,\nu^*) = J^*$. The problem of finding a minimum pair is sometimes referred to as the *AC problem*, and J^* is called the *value* of the AC problem.

This AC problem has been studied by Kurano [11] using Doeblin's (ergodicity) condition and in [8] using an LP approach. Here we again use the LP approach, but not directly on the AC problem itself as in [8]; we use instead linear programs associated with the α-discount problems [7]. This requires, to begin with, that J^* is finite. We will thus make the following assumption.

Assumption 2.3. $J(\hat\delta,\hat\nu) < \infty$ for some $\hat\delta \in \Delta$ and $\hat\nu \in \mathcal{P}(X)$.

Assumption 2.3 ensures, of course, that J^* is finite, but it also guarantees that $V_\alpha(\hat\delta,\hat\nu)$ is finite for every $\alpha \in (0,1)$, since by a well-known Tauberian theorem (see, e.g., [2], [15], [16])

$$\limsup_{\alpha\uparrow 1}\,(1-\alpha)V_\alpha(\hat\delta,\hat\nu) \le J(\hat\delta,\hat\nu). \tag{2.5}$$

Note that the latter inequality implies

$$\limsup_{\alpha\uparrow 1}\,(1-\alpha)V_\alpha^*(\hat\nu) \le J(\hat\delta,\hat\nu). \tag{2.6}$$

In the next section we consider linear programs associated with the problem of minimizing $V_\alpha(\delta,\hat\nu)$ over Δ, which are then used in §4 to obtain a minimum pair.

3. Linear programs. The plan of this section is as follows. First, for every $\alpha \in (0,1)$, we state linear programs (P_α) and (P_α^*) associated with the α-discounted cost problem, which are then shown to be equivalent to modified linear programs (MP_α), (MP_α^*). For $\alpha = 1$, the latter are programs associated with the AC problem. (The LP terminology that we use is borrowed from [1, Chap. 3].)

We begin with the introduction of two dual pairs of vector spaces, which are the same already used in [7], [8].

Dual pairs. Let K be the set in (2.1) and $b: K \to R$ the function

$$b(x,a) := c_0 + c(x,a), \tag{3.1}$$

where c_0 is a given positive number. (Recall that, by assumption, the one-stage cost is nonnegative; hence $b \geq c_0$.) We define $F(K)$ as the vector space of all real-valued measurable functions v on K such that

$$\|v\|_b := \sup_{(x,a)} |v(x,a)|/b(x,a) \ < \infty. \tag{3.2}$$

Note that c is in $F(K)$. (If necessary, a function $v \in F(K)$ is considered to be extended to all of $X \times A$ in an arbitrary way as long as measurability and (3.2) are preserved.) Now let $M(K)$ be the vector space consisting of all the finite signed measures μ on $X \times A$ concentrated on K and such that

$$\int b d|\mu| \ < \infty,$$

where $|\mu|$ denotes the total variation of μ. Then $(M(K),\ F(K))$ is a dual pair with respect to the bilinear form $\langle \mu, v \rangle := \int v d\mu$.

Similarly, let $b_1 : X \to R$ be a positive measurable function bounded away from zero and define $F(X)$ as the vector space of all the real-valued measurable functions u on X such that

$$\|u\|_{b_1} := \sup_x \ |u(x)|/b_1(x) \ < \infty,$$

and let $M(X)$ be the vector space of all the finite signed measures ν on X for which $\int b_1 d|\nu| \ < \infty$. Then $(M(X),\ F(X))$ is a dual pair with respect to the bilinear form $\langle \nu, u \rangle := \int u d\nu$.

Remark. We adopt here the convention that, unless explicitly stated otherwise, a vector space in a dual pair is *always* endowed with the *weak topology* [1]. Convergence in this topology will be denoted $\xrightarrow{\sigma}$. However, for finite nonnegative—in particular, probability—measures, we will also use the usual notion of "weak convergence," noted $\xrightarrow{w}$ or w-convergence. Thus, e.g., $\mu^n \xrightarrow{w} \mu$ (resp., $\mu^n \xrightarrow{\sigma} \mu$) in $M^+(K) := \{\mu \in M(K) | \mu \geq 0\}$ means that

$$\langle \mu^n, v \rangle \to \langle \mu, v \rangle \quad \forall v \in C(K) \quad (\text{resp., } \forall v \in F(K)).$$

Clearly, σ-convergence implies w-convergence.

To guarantee that the linear programs (P_α) and (P_α^*) below are properly defined, throughout the following we suppose the following assumption.

Assumption 3.1. (a) $\hat{\nu}$ is in $M(X)$, i.e., $\int b_1 d\hat{\nu} \ < \infty$.

(b) $b(x,a) \geq b_1(x) \geq c_1$ for all $(x,a) \in K$ and some positive number c_1.

(c) $\int b_1(y) Q(dy|.)$ is in $F(K)$, i.e., $\sup_{(x,a)} \int b_1(y) Q(dy|x,a)/b(x,a) \ < \infty$.

Linear programs. For each positive $\alpha \leq 1$, let $T_\alpha : M(K) \to M(X)$ and $T_\alpha^* : F(X) \to F(K)$ be the linear operators defined as

$$(T_\alpha \mu)(B) := \mu_1(B) - \alpha \int Q(B|x,a)\mu(d(x,a)) \tag{3.3}$$

for all $\mu \in M(K)$, $B \in \mathcal{B}(X)$, where $\mu_1(B) := \mu(B \times A)$ is the *projection* (or marginal) of μ on X, and

$$(T_\alpha^* u)(x,a) := u(x) - \alpha \int u(y) Q(dy|x,a), \quad u \in F(X), \ (x,a) \in K. \tag{3.4}$$

In view of Assumption 3.1, these linear operators are continuous. Moreover, T_α^* is the *adjoint* of T_α, i.e.,

$$\langle T_\alpha \mu, u \rangle = \langle \mu, T_\alpha^* u \rangle \quad \forall \mu \in M(K), \ u \in F(X). \tag{3.5}$$

Consider the following linear programs:
(P_α) minimize $\langle \mu, c \rangle$ subject to

$$\mu \in M^+(K) \ \text{ and } \ T_\alpha \mu = \hat{\nu}; \tag{3.6}$$

(P_α^*) maximize $\langle \hat{\nu}, u \rangle$ subject to

$$u \in F(X) \ \text{ and } \ T_\alpha^* u \leq c. \tag{3.7}$$

The linear program (P_α^*) is the *dual* of (P_α) [1]. Under Assumptions 2.3 and 3.1, it has been shown in [7, §4] that (P_α) and the problem of minimizing $V_\alpha(., \hat{\nu})$ are equivalent, i.e, $V_\alpha^*(\hat{\nu}) = \inf(P_\alpha)$, and by weak duality,

$$V_\alpha^*(\hat{\nu}) = \inf(P_\alpha) \geq \sup(P_\alpha^*). \tag{3.8}$$

Moreover, under the additional Assumption 3.2(a) below, (P_α) is solvable and, under Assumption 3.2(b), equality holds in (3.8), i.e.,

$$V_\alpha^*(\hat{\nu}) = \min(P_\alpha) = \sup(P_\alpha^*). \tag{3.9}$$

Before we state Assumption 3.2, note that if μ is feasible for (P_α), then $(1-\alpha)\mu(K) = 1$. Thus $\mu'(.) := (1-\alpha)\mu(.)$ is a probability measure on K.

Assumption 3.2. (a) If $\{\mu^n\}$ is a sequence of feasible solutions for (P_α) such that $\sup_n \int c d\mu^n \leq r$ for some $r > 0$, then $\{\mu^n\}$ is tight.

(b)If $\{\mu^n\}$ is a sequence of probability measures in $M(K)$ such that $\sup_n \int c d\mu^n \leq r$ for some $r > 0$, then $\{\mu^n\}$ is tight.

Assumption 3.2 holds, e.g., if the one-stage cost c is a "moment" (see [8, Rem. 5.6 and §6]). Observe also that Assumption 3.2(b) implies 3.2(a)

We now wish to relate (P_α) and (P_α^*) to the minimum-pair problem. To do this, we modify these programs as follows.

Modified linear programs. Consider the dual pair $(R \times M(X), \ R \times F(X))$ with the bilinear form

$$\langle (r,\nu), (\rho, u) \rangle := r\rho + \langle \nu, u \rangle.$$

For every positive $\alpha \leq 1$, let $L_\alpha : M(K) \to R \times M(X)$ be the linear operator defined as

$$L_\alpha \mu := (\bar{\mu}, T_\alpha \mu) \ \text{ with } \ \bar{\mu} := \mu(K), \ \mu \in M(K).$$

Consider the adjoint $L_\alpha^* : R \times F(X) \to F(K)$ of L_α, i.e.,

$$L_\alpha^*(\rho, u) := \rho + T_\alpha^* u.$$

Now, instead of $(P_\alpha) - (P_\alpha^*)$, the corresponding linear programs are
(MP_α) minimize $\langle \mu, c \rangle$ subject to

$$L_\alpha \mu = (1, (1-\alpha)\hat{\nu}), \quad \mu \in M^+(K); \tag{3.10}$$

(MP_α^*) maximize $(\rho + (1-\alpha)\langle \hat{\nu}, u \rangle)$ $[= \langle (1, (1-\alpha)\hat{\nu}), (\rho, u) \rangle]$ subject to

$$L_\alpha^*(\rho, u) \le c, \quad (\rho, u) \in R \times F(X). \tag{3.11}$$

In particular, for $\alpha = 1$ we obtain linear programs associated with the AC problem [8]:
(MP_1) minimize $\langle \mu, c \rangle$ subject to

$$L_1 \mu = (1, 0), \quad \mu \in M^+(K); \tag{3.12}$$

(MP_1^*) maximize ρ $[= \langle (1,0), (\rho, u) \rangle]$ subject to

$$L_1^*(\rho, u) \le c, \quad (\rho, u) \in R \times F(X). \tag{3.13}$$

It turns out that, for positive $\alpha < 1$, the modified linear programs are equivalent to the original programs in the following sense.

PROPOSITION 3.3. *For every $0 < \alpha < 1$, we have the following:*

(a) *if μ is feasible for (P_α), then $\mu'(.) = (1-\alpha)\mu(.)$ is feasible for (MP_α) and*

$$\langle \mu, c \rangle = \langle \mu', c \rangle / (1-\alpha).$$

(b) *Conversely, if μ is feasible for (MP_α), then $\tilde{\mu}(.) := \mu(.)/(1-\alpha)$ is feasible for (P_α) and*

$$\langle \tilde{\mu}, c \rangle = \langle \mu, c \rangle / (1-\alpha).$$

Hence

$$\inf(P_\alpha) = (1-\alpha)^{-1} \inf(MP_\alpha). \tag{3.14}$$

Similarly for the dual problems:

(c) *If u is feasible for (P_α^*), then for any real number m, the pair (ρ, u') defined as $\rho = (1-\alpha)m$ and $u'(.) := u(.) - m$ is feasible for (MP_α^*) and*

$$\langle \hat{\nu}, u \rangle = [\rho + (1-\alpha)\langle \hat{\nu}, u' \rangle] / (1-\alpha).$$

(d) *If (ρ, u) is feasible for (MP_α^*), then $u'(.) := u(.) + \rho/(1-\alpha)$ is feasible for (P_α^*) and*

$$\langle \hat{\nu}, u' \rangle = [\rho + (1-\alpha)\langle \hat{\nu}, u \rangle] / (1-\alpha).$$

Hence

$$\sup(P_\alpha^*) = (1-\alpha)^{-1} \sup(MP_\alpha^*). \tag{3.15}$$

We omit the easy proof.

From (3.8)–(3.9) and (3.14)–(3.15), we obtain the following.

COROLLARY 3.4. *If Assumptions* 2.3 *and* 3.1 *hold, then for every* $0 < \alpha < 1$,

(a) $(1-\alpha)V_\alpha^*(\hat{\nu}) = \inf(MP_\alpha) \geq \sup(MP_\alpha^*)$.

If, in addition, Assumption 3.2 *holds, then*

(b) (MP_α) *is solvable and*

$$(1-\alpha)V_\alpha^*(\hat{\nu}) = \min(MP_\alpha) = \sup(MP_\alpha^*). \tag{3.16}$$

In part (b) of the corollary, solvability of (MP_α), $0 < \alpha < 1$, means, of course, that there is a measure $\mu^\alpha \in M(K)$ that satisfies (3.10) and

$$\min(MP_\alpha) = \int c d\mu^\alpha. \tag{3.17}$$

In the next two sections we show that (3.16) holds "in the limit as $\alpha \uparrow 1$," so that

$$J^* = \min(MP_1) = \sup(MP_1^*), \tag{3.18}$$

with J^* as in (2.4). In particular, the solvability of (MP_1) is equivalent to the existence of a minimum pair. The first equality in (3.18) is proved in §4; the second, in §5.

4. Existence of minimum pairs. To prove the first equality in (3.18), let us first note the following elementary fact. (Recall the notation introduced in the remark preceding Assumption 3.1.)

LEMMA 4.1. *Let* μ *and* μ^α, $0 < \alpha < 1$, *be measures in* $M(K)$ *such that* $\mu^\alpha \xrightarrow{\sigma} \mu$ *as* $\alpha \uparrow 1$, *i.e.,*

$$\lim_{\alpha \uparrow 1} \langle \mu^\alpha, v \rangle = \langle \mu, v \rangle \quad \forall v \in F(K). \tag{4.1}$$

Then $T_\alpha \mu^\alpha \xrightarrow{\sigma} T_1 \mu$ *and* $L_\alpha \mu^\alpha \xrightarrow{\sigma} L_1 \mu$. *If, moreover,* μ^α *is feasible for* (MP_α), $0 < \alpha < 1$, *then* μ *is feasible for* (MP_1). *If we have instead* $\mu^\alpha \xrightarrow{w} \mu$ *in* $M^+(K)$ *as* $\alpha \uparrow 1$, *all of the conclusions hold replacing* $\xrightarrow{\sigma}$ *by* $\xrightarrow{w}$.

Proof. For any $B \in \mathcal{B}(X)$, $T_\alpha \mu^\alpha(B) = T_1 \mu^\alpha(B) - (1-\alpha) \int Q(B|k) \mu^\alpha(dk)$. Therefore, as $\alpha \uparrow 1$,

$$\lim T_\alpha \mu^\alpha = \lim T_1 \mu^\alpha = T_1 \mu, \tag{4.2}$$

where the latter equality is due to (4.1) and the continuity of T_1. Furthermore, letting $v(.) \equiv 1$ in (4.1), we obtain $\overline{\mu^\alpha} \to \overline{\mu}$. Thus $L_\alpha \mu^\alpha \xrightarrow{\sigma} L_1 \mu$, completing the proof of the first statement. Now if each μ^α satisfies (3.10), then $\overline{\mu} = 1$, and from (4.2),

$$T_1 \mu(B) = \lim_{\alpha \uparrow 1} (1-\alpha) \hat{\nu}(B) = 0 \quad \forall B \in \mathcal{B}(X),$$

i.e., μ satisfies (3.12) and therefore is feasible for (MP_1). Finally, to prove the last statement it suffices to note that $\mu^\alpha \xrightarrow{w} \mu$ implies w-convergence of the projections, i.e., $\mu_1^\alpha \xrightarrow{w} \mu_1$, and also

$$\int_K Q(.|k) \mu^\alpha(dk) \xrightarrow{w} \int_K Q(.|k) \mu(dk) \quad \text{in} \quad M^+(X)$$

by the weak continuity of Q. □

THEOREM 4.2. *Suppose that Assumptions* 2.3, 3.1, *and* 3.2(a) *hold, and for every* $\alpha \in (0,1)$, *let* μ^α *be an optimal solution for* (MP_α), *i.e.* (*from the first equality in* (3.16)),

$$(1-\alpha)V_\alpha^*(\hat{\nu}) = \min(MP_\alpha) = \int c d\mu^\alpha, \quad 0 < \alpha < 1. \tag{4.3}$$

Then

(a) *there exist a sequence* $\alpha(n) \uparrow 1$, *a number* $j^* \le J(\hat{\delta}, \hat{\nu})$, *and a measure* μ^* *feasible for* (MP_1) *such that*

$$j^* = \lim_n (1-\alpha(n))V_{\alpha(n)}^*(\hat{\nu}) \ge \int c d\mu^*. \tag{4.4}$$

(b) *If* j^* *is such that, for any initial distribution* ν,

$$j^* \le \lim_n (1-\alpha(n))V_{\alpha(n)}^*(\nu), \tag{4.5}$$

then there exists a minimum pair (φ^*, ν^*), *where* $\varphi^* \in \Phi$ *is a relaxed policy,* $\nu^* = \mu_1^*$ (*the projection of* μ^* *on* X), *and*

$$j^* = J(\varphi^*, \mu_1^*) = J^*. \tag{4.6}$$

Hence

(c) μ^* *is optimal for* (MP_1) *and*

$$J^* = \min(MP_1) = \int c d\mu^*. \tag{4.7}$$

Proof. (a) From (2.6), there is a sequence $\alpha(n) \uparrow 1$ and a number $j^* \le J(\hat{\delta}, \hat{\nu})$ such that

$$j^* = \lim_n (1-\alpha(n))V_{\alpha(n)}^*(\hat{\nu}) = \lim_n \int c d\mu^{\alpha(n)}.$$

By Assumption 3.2(a), with $r = j^* + \epsilon$ for some $\epsilon > 0$, $\{\mu^{\alpha(n)}\}$ has a weakly convergent subsequence. Combining this with Lemma 4.1, there is a subsequence $\{\alpha(n_i)\}$ of $\{\alpha(n)\}$ and a measure μ^* feasible for (MP_1) such that $\mu^{\alpha(n_i)} \xrightarrow{w} \mu^*$; in particular, since c is lower semicontinuous, we obtain (4.4) with $\{\alpha(n_i)\}$ in lieu of $\{\alpha(n)\}$.

(b) By (3.12) and a well-known result ([4, p. 89, Thm. 2]; [10, Cor. 12.7]; [8 Lem. 4.7]), the measure μ^* can be "disintegrated" into a relaxed policy φ^* and a probability measure $\nu^* = \mu_1^*$ so that

$$\mu^*(B \times C) = \int_B \varphi^*(C|x)\mu_1^*(dx) \quad \forall B \in \mathcal{B}(X),\ C \in \mathcal{B}(A),$$

and such that

$$\int c d\mu^* = J(\varphi^*, \mu_1^*). \tag{4.8}$$

Thus, by (4.4) and the definition of J^* in (2.4), $j^* \ge J^*$. The reverse inequality, $j^* \le J^*$, follows from (4.5) and (2.6).

(c) From the "disintegration" result referred to in part (b), for any feasible measure μ for (MP_1) there is a relaxed policy φ and an initial distribution $\nu(=\mu_1)$ such that $\int c d\mu = J(\varphi, \nu)$. Since this is true for any such μ, $\inf(MP_1) \geq J^*$. This inequality, together with (4.6), yields (c). □

Remark 4.3. A sufficient condition for (4.5) is Assumption 5.1(b), as shown in the proof of Theorem 5.3.

Part (c) in Theorem 4.2 shows that solving (MP_1) and finding a minimum pair are "equivalent" problems in the sense that $J^* = \min(MP_1)$; i.e., the first equality in (3.18) holds. In the next section we prove the second equality in (3.18).

5. Absence of duality gap. A standard argument (see, e.g., [8, Lem. 4.5]) shows that if (ρ, u) is feasible for (MP_1^*) and (δ, ν) is any policy-initial-distribution pair with $J(\delta, \nu) < \infty$, then $\rho \leq J(\delta, \nu)$; hence $\rho \leq J^*$. This implies

$$\sup(MP_1^*) \leq J^*. \tag{5.1}$$

To get the equality in (5.1) we shall use Assumption 5.1 below on the pointwise α-discount value function $V_\alpha^*(x)$, $x \in X$, $0 < \alpha < 1$.

Assumption 5.1. For every $\alpha \in (0,1)$ and $x \in X$, $V_\alpha^*(x) < \infty$ and

(a) V_α^* satisfies the dynamic programming equation

$$V_\alpha^*(x) = \min_{a \in A(x)} \left[c(x,a) + \alpha \int V_\alpha^*(y) Q(dy|x,a) \right], \quad x \in X; \tag{5.2}$$

(b) for some state $z \in X$, $N \geq 0$, and $\alpha_0 \in [0,1)$, the function $h_\alpha(x) := V_\alpha^*(x) - V_\alpha^*(z)$, $x \in X$, satisfies $h_\alpha(x) \geq -N \ \ \forall x \in X$, $\alpha \in (\alpha_0, 1)$.

(c) V_α^* belongs to $F(X)$.

Sufficient conditions for (5.2) are well known (see, e.g., [2], [4], [5], [9], [10]). (Observe that, from (2.6), $V_\alpha^*(x) < \infty$ for every $\alpha \in (0,1)$ and $\hat{\nu}$-almost all $x \in X$.)

On the other hand, the conditions (b) and (c) in Assumption 5.1 are model dependent and usually have to be verified directly, which is easily done in some cases: For (b), see [15], [16], [2, §§5.2, 6.2]; and for (c), a sufficient condition is, for example, the existence of a policy $\pi \in \Delta$ with a bounded average cost. In this case, by a well-known Tauberian theorem,

$$\limsup_{\alpha \uparrow 1} (1-\alpha) V_\alpha^*(x) \leq J(x, \pi) \leq g \ \ \forall x \in X,$$

so that, for every fixed discount factor $\alpha \in (0,1)$ and some constant $K > 0$,

$$\sup_x \frac{V_\alpha^*(x)}{b_1(x)} \leq \sup_x \frac{(g+K)}{(1-\alpha) b_1(x)} < \infty$$

since $b_1(.)$ is bounded away from zero. Hence $V_\alpha^* \in F(X)$ for every $\alpha \in (0,1)$.

Note that in contrast to previous work (see, e.g., [2], [15], [16]), *we do not require* h_α *to be majorized* by some function independent of α. The latter condition implies the "unichain" assumption for average optimal policies. We shall return to this point in §6.

LEMMA 5.2. *If* (a) *and* (c) *of Assumption* 5.1 *hold, then* (P_α^*) *and* (MP_α^*) *are solvable for every* $0 < \alpha < 1$, *and*

$$\max(MP_\alpha^*) = (1-\alpha)\max(P_\alpha^*) = (1-\alpha)\langle \hat{\nu}, V_\alpha^* \rangle. \tag{5.3}$$

Proof. If V_α^* is in $F(X)$ and satisfies (5.2), then it obviously satisfies (3.7), i.e., V_α^* is feasible for (P_α^*). Hence, $\int V_\alpha^* d\hat{\nu} \le \sup(P_\alpha^*)$. The reverse inequality is also true, since $u \le V_\alpha^*$ whenever u satisfies (3.7); see, e.g., [9]. Therefore $\max(P_\alpha^*) = \int V_\alpha^* d\hat{\nu}$, which together with (3.15) yields (5.3). □

Throughout the remainder of this section, let $z \in X$ be the fixed state in Assumption 5.1(b), and let $\hat{\nu}$ the initial distribution in Assumption 2.3 be the Dirac measure at z. Then, under the assumptions of Theorem 4.2(a), (4.4) becomes

$$j^* = \lim_n (1 - \alpha(n)) V_{\alpha(n)}^*(z) = \lim \rho_{\alpha(n)} \tag{5.4}$$

for some sequence $\alpha(n) \uparrow 1$, where $\rho_\alpha := (1 - \alpha) V_\alpha^*(z)$. Moreover, it is convenient to rewrite (5.2) in the form

$$\rho_\alpha + h_\alpha(x) = \min_{a \in A(x)} \left[c(x,a) + \alpha \int h_\alpha(y) Q(dy|x,a) \right]. \tag{5.5}$$

Finally, without loss of generality, we may assume that h_α is *nonnegative* (cf. Assumption 5.1(b)), for if (ρ, u) is feasible for (MP_α^*), then so is $(\rho - (1-\alpha)N, u + N)$ for any constant N.

THEOREM 5.3. *If Assumption* 5.1 *and the hypotheses of Theorem* 4.2(a) *hold, then*

$$\sup(MP_1^*) = J^* \quad (= \min(MP_1)). \tag{5.6}$$

Proof. If, as assumed above, $\hat{\nu}$ is the Dirac measure at z, then, by (5.3) and (5.5), (ρ_α, h_α) is optimal for (MP_α^*), $0 < \alpha < 1$, with $h_\alpha(.) \ge 0$ and $\max(MP_\alpha^*) = \rho_\alpha$. We also note that if (ρ, u) is feasible for (MP_α^*) and $u \ge 0$, then (ρ, u) is feasible for (MP_β^*) for all $\beta \in [\alpha, 1]$ since

$$\begin{aligned} \rho + u(x) &\le c(x,a) + \alpha \int u(y) Q(dy|x,a) \\ &\le c(x,a) + \beta \int u(y) Q(dy|x,a) \quad \forall \alpha \le \beta \le 1. \end{aligned}$$

Therefore, from the definition of (MP_α^*), $\forall \alpha_0 < \alpha < 1$,

$$\begin{aligned} \rho_\alpha &= \max(MP_\alpha^*) \\ &= \sup\{\rho + (1-\alpha)u(z) |\ (\rho, u) \ \text{ satisfies (3.11)}\} \\ &= \sup\{\rho + (1-\alpha)u(z) |\ (\rho, u) \ \text{ satisfies (3.11), } u \ge 0\} \\ &\le \sup(MP_1^*) + \sup\{(1-\alpha)u(z) |\ (\rho, u) \text{ satisfies (3.13)}\}. \end{aligned}$$

Thus, from (5.4) and letting $\alpha \uparrow 1$, $j^* \le \sup(MP_1^*)$. To prove the reverse inequality note first that Assumption 5.1(b) implies (4.5) since, for any initial distribution ν,

$$\begin{aligned} (1-\alpha) V_\alpha^*(\nu) &\ge (1-\alpha) \int V_\alpha^*(x) \nu(dx) \\ &= (1-\alpha) \int h_\alpha(x) \nu(dx) + (1-\alpha) V_\alpha^*(z) \\ &\ge -(1-\alpha)N + (1-\alpha) V_\alpha^*(z). \end{aligned}$$

Hence, by weak duality and (4.6)–(4.7),

$$\sup(MP_1^*) \le \inf(MP_1) = J^* = j^*.$$

This completes the proof of (5.6). □

From the proof of Theorem 5.3, observe that the pair (ρ_α, h_α) is maximizing for (MP_1^*) since (ρ_α, h_α) is feasible for that problem and $\rho_\alpha \to J^*$ as $\alpha \uparrow 1$.

6. Existence of pointwise optimal policies. The objectives of this final section are to make a brief comparison between the standard "vanishing discount" approach and the version presented in the previous sections and to comment on the existence of "pointwise" optimal policies—as opposed to minimum pairs—that can be derived from our results.

In the standard vanishing discount approach [2], [15], [16] one tries to obtain the so-called *average-cost optimality equation*

$$J^* + h(x) = \min_{a \in A(x)} \left[c(x,a) + \int h(y) Q(dy|x,a) \right], \quad x \in X, \tag{6.1}$$

or the *optimality inequality*

$$J^* + h(x) \geq \min_{a \in A(x)} \left[c(x,a) + \int h(y) Q(dy|x,a) \right], \quad x \in X, \tag{6.2}$$

starting from the dynamic programming equation (5.2). The basic idea is to rewrite (5.2) in the form (5.5) and give conditions on ρ_α and $h_\alpha(.)$ under which, for some sequence $\alpha \uparrow 1$, $\rho_\alpha \to J^*$ and $h_\alpha(.) \to h(.)$, with $(J^*, h(.))$ satisfying (6.1) or (6.2). Finally, if $h(.)$ is bounded from below and if a deterministic stationary policy $f \in \mathcal{F}$ is such that $f(x) \in A(x)$ attains the minimum in the right-hand side of (6.1) or (6.2) for all $x \in X$, then f satisfies

$$J(f,x) = \inf_\delta J(\delta, x) = J^* \quad \text{for } \textit{all } x \in X \tag{6.3}$$

provided that in Assumption 5.1(b) we impose the *additional* requirement

$$h_\alpha(x) \leq g(x) \quad \forall x \in X, \ \alpha \in (\alpha_0, 1), \tag{6.4}$$

for some *finite-valued* function g [15], [16]. Without (6.4) we cannot guarantee (6.3) for *all* x. To see what can go wrong consider the following elementary "multichain" example.

Example 6.1. $X := \{1,2,3\}$, $A = \{1\}$, and $c(x,1) = c_x > 0$, with $c_1 < c_2, c_3$. The transition law $Q(\{y\}|x,1)$, which we write as $P_{xy}(1)$, is given by

$$P_{xx}(1) = 1 \ \text{ if } x = 1,2; \ P_{31}(1) = \beta = 1 - P_{32}, \ \text{ where } 0 < \beta < 1.$$

Then $V_\alpha^*(x) = c_x/(1-\alpha)$ if $x = 1,2$, and $V_\alpha^*(3) = c_3 + \alpha(1-\alpha)^{-1}[\beta c_1 + (1-\beta)c_2]$, and V_α^* satisfies assumptions 5.1(a) and (c). To verify Assumption 5.1(b), it suffices to take $z = 1$ and $\alpha_0 := c_1/c_2$ so that $c_1 < \alpha c_2$ $(< c_2)$ for all $\alpha \in (\alpha_0, 1)$. This yields $h_\alpha \geq 0$. (If we take $z = 2$ or 3, then h_α does not satisfy Assumption 5.1(b).)

Moreover, $h(x) := \lim_{\alpha \uparrow 1} h_\alpha(x) = 0$ if $x = z = 1$, and $h(x) = +\infty$ if $x = 2,3$; therefore, (6.4) does *not* hold for a *finite*-valued function g. However, the "minimum" cost, given the initial state $x \in X$, is $J^*(x) = c_x$ if $x = 1,2$, and $J^*(3) = \beta c_1 + (1-\beta)c_2$, so that $J^* := \inf_x J^*(x) = c_1$, i.e.,

$$J^* = j^* := \lim_{\alpha \uparrow 1} (1-\alpha) V_\alpha^*(z) = c_1.$$

Thus the pair (J^*, h) trivially satisfies (6.1), but (6.3) does *not* hold for *all* $x \in X$. □

Example 6.1 suggests that a "unichain" assumption is implicit somehow in (6.4). Since we have not made any such assumption here nor assumed (6.4), we would expect to get a pointwise result such as (6.1)–(6.2) or (6.3) only for states in a proper subset of X. In the next theorem we identify such a subset in the context of Theorem 5.3.

THEOREM 6.2. *Suppose that Assumption* 5.1 *and the hypotheses of Theorem* 4.2(a) *hold, and let* $\mu*$ *and* (φ^*, μ_1^*) *be as in* (4.6)–(4.7). *Then*

(a) *the randomized policy* φ^* *is average optimal for* μ_1^**-almost all* (*a.a.*) *initial states, i.e.,*

$$J(\varphi^*, x) = J^* \quad \text{for } \mu_1^*\text{-a.a. } x \in X; \tag{6.5}$$

(b) *there exists a measurable function* h *on* X *bounded from below and such that*

$$J^* + h(x) \geq c(x,a) + \int h(y)Q(dy|x,a) \tag{6.6}$$

for μ^**-a.a.* $(x,a) \in K$;

(c) *if the set* $S := \{x \in X |$ (6.6) *holds and* $h(x) < \infty\}$ *is nonempty, then there exists a deterministic stationary policy* $f \in \mathcal{F}$ *such that*

$$J^* + h(x) \geq c(x, f(x)) + \int h(y)Q(dy|x, f(x)) \quad \forall x \in S, \tag{6.7}$$

and

$$J(f, x) = J^* \quad \forall x \in S. \tag{6.8}$$

For instance, a sufficient condition for S *to be nonempty is that, for some constant* $g < \infty$,

$$\limsup_{\alpha \uparrow 1} |V_\alpha(z) - J^*/(1-\alpha)| < g < \infty$$

(*where* z *is as in Assumption* 5.1).

Remark. The above sufficient condition simply states that the optimal discounted cost in state z converges sufficiently fast to the AC and can be checked in many problems (see, e.g., the linear-quadratic regulator problem [4]). In fact, we may obtain (6.6) directly from (4.6)–(4.7) or the stronger result (6.2) from Assumptions 2.3 and 5.1. However, we chose the above form of Theorem 6.2 because we wish to relate the vanishing-discount-factor approach with the LP results in §§3–5.

Proof of Theorem 6.2. (a) As in the proof of Theorem 4.2(b), disintegrate μ^* as $\mu^*(d(x,a)) = \varphi^*(da|x)\mu_1(dx)$ and note that the condition $T_1\mu^* = 0$ in (3.12) can also be written as

$$\mu_1^*(B) = \int_X Q(B|x, \varphi^*)\mu_1^*(dx) \quad \forall B \in \mathcal{B}(X), \tag{6.9}$$

where $Q(.|x, \varphi^*) := \int_A Q(.|x,a)\varphi^*(da|x)$. In other words, (6.9) says that μ_1^* is an invariant probability measure for the stochastic kernel $Q(.|., \varphi^*)$. We also have that

$$J^* = \int c d\mu^* = \int c(x, \varphi^*)\mu_1^*(dx),$$

where $c(.,\varphi^*) := \int_A c(.,a)\varphi^*(da|.)$. Thus $c(x,\varphi^*)$, the one-stage cost of the randomized policy φ^*, is μ_1^*-integrable, and then the individual ergodic theorem (see, e.g., Theorem 6 in [17, p. 388]) yields that the limit

$$J(\varphi^*,x) := \lim_{n\to\infty} n^{-1}E_x^{\varphi^*}\left[\sum_{t=0}^{n-1} c(x_t,\varphi^*)\right]$$

exists for μ_1^*-a.a. $x \in X$ and satisfies

$$(*) \qquad \int_X c(x,\varphi^*)\mu_1^*(dx) = \int_X J(\varphi^*,x)\mu_1^*(dx) = J^*.$$

Finally, to conclude that (6.5) holds, let $B := \{x|\ J(\varphi^*,x) > J^*\}$ and note that, by (2.4), the complement of B is $B^c := \{x|\ J(\varphi^*,x) = J^*\}$. Hence, from the second equality in (*),

$$J^* = \int_B J(\varphi^*,x)\mu_1^*(dx) + J^*\mu_1^*(B^c),$$

which is equivalent to

$$\int_B J(\varphi^*,x)\mu_1^*(dx) = J^*\mu_1^*(B),$$

that is,

$$\int_B [J(\varphi^*,x) - J^*]\mu_1^*(dx) = 0,$$

which, in view of (2.4), implies $\mu_1^*(B) = 0$. This completes the proof of (6.5).

(b) Let N be as in Assumption 5.1(b) and define

$$\rho_{\alpha,N} := \rho_\alpha - (1-\alpha)N, \quad h_{\alpha,N}(x) := h_\alpha(x) + N, \quad x \in X,\ \alpha \in (\alpha_0,1).$$

Then we may rewrite (5.5) as

$$\rho_{\alpha,N} + h_{\alpha,N}(x) = \min_{a\in A(x)}\left[c(x,a) + \alpha\int h_{\alpha,N}(y)Q(dy|x,a)\right] \quad \forall x \in X,$$

and therefore, since $h_{\alpha,N}(.) \geq 0$ and $\alpha < 1$, the pair $(\rho_{\alpha,N}, h_{\alpha,N})$ is feasible for (MP_1^*), i.e.,

$$\rho_{\alpha,N} + h_{\alpha,N}(x) \leq c(x,a) + \int h_{\alpha,N}(y)Q(dy|x,a) \quad \forall (x,a) \in K.$$

Let $u_\alpha(x,a)$ be a "slack variable," i.e., u_α is a measurable nonnegative function such that, $\forall (x,a) \in K$,

$$(6.10) \qquad \rho_{\alpha,N} + h_{\alpha,N}(x) + u_\alpha(x,a) = c(x,a) + \int h_{\alpha,N}(y)Q(dy|x,a).$$

Thus, defining

$$(6.11) \qquad v(x,\varphi^*) := \int_A v(x,a)\varphi^*(da|x) \quad \text{for all } x \in X,\ v \in F(K),$$

integration of both sides of (6.10) with respect to μ^* yields (by (6.9))

$$\rho_{\alpha,N} + \int h_{\alpha,N} d\mu_1^* + \int u_\alpha d\mu_1^* = \int c d\mu^* + \int h_{\alpha,N} d\mu_1^*,$$

i.e., from (4.7)

$$\int u_\alpha d\mu^* = J^* - \rho_{\alpha,N}.$$

In the last step, note that we may indeed cancel out the integral $\int h_{\alpha,N} d\mu_1^*$ since, by Assumption 5.1(c), this integral is finite. Notice also that with $\alpha(n)$ as in (5.4)

$$\lim_n \rho_{\alpha(n),N} = j^* = J^*,$$

where the last equality comes from (5.6). Therefore, $\lim_n \int u_{\alpha(n)} d\mu^* = 0$, which in turn, using Fatou's lemma (recall that $u_\alpha \geq 0$), yields

$$\liminf_n u_{\alpha(n)}(x,a) = 0 \quad \text{for } \mu^* - a.a.\ (x,a) \in K. \tag{6.12}$$

Finally, define

$$h(x) := \liminf_n h_{\alpha(n)}(x) = \liminf_n h_{\alpha(n),N}(x) - N \geq -N, \qquad x \in X.$$

Let x be an arbitrary state for which (6.12) holds for some $a \in A(x)$, and then take the $\lim_i$ in (6.10) over a subsequence $\alpha(n_i)$ of $\alpha(n)$ for which $h_{\alpha(n_i)}(x) \to h(x)$ and $u_{\alpha(n_i)}(x,a) \to 0$. This yields, by Fatou's lemma again,

$$J^* + h(x) - N \geq c(x,a) + \int h(y) Q(dy|x,a) - N,$$

i.e., $J^* + h(x) \geq c(x,a) + \int h(y) Q(dy|x,a)$. Hence, since $(x,a) \in K$ was an arbitrary pair for which (6.12) holds, we conclude (6.6).

(c) We will first prove the last statement, i.e., the set S is nonempty if

$$\limsup_{\alpha \uparrow 1} |V_\alpha(z) - J^*/(1-\alpha)| < g < \infty.$$

Indeed, from the optimality equation (5.2) we have

$$(1-\alpha) V_\alpha^*(z) + h_\alpha(x) \leq c(x,a) + \alpha \int h_\alpha(y) Q(dy|x,a) \quad \forall x, a$$

so that integrating both sides with respect to μ^* and using

$$\begin{aligned} T_\alpha \mu^*(B) &= T_1 \mu^*(B) + (1-\alpha) \int Q(B|k) \mu^*(dk) \\ &= (1-\alpha) \mu_1^*(B) \quad \forall B \in \mathcal{B}(X) \end{aligned}$$

(see §4) yields

$$(1-\alpha) V_\alpha^*(z) + (1-\alpha) \int h_\alpha d\mu_1^* \leq \int c d\mu^* = J^*$$

and thus

$$0 \leq \int (h_\alpha + N) d\mu_1^* \leq \frac{J^* - (1-\alpha)V_\alpha^*(z)}{1-\alpha} + N \leq 2g + N$$

for all α sufficiently close to 1. Then, by Fatou's lemma,

$$0 \leq \int (h + N) d\mu_1^* < \infty,$$

which in turn implies that h is finite μ_1^*-a.e.

This and (6.6) prove that the set S is nonempty. It contains the support of the invariant probability measure μ_1^*.

Integrating both sides of (6.6) with respect to the measure $\varphi^*(da|x)$ yields

$$J^* + h(x) \geq \int_A \left[c(x,a) + \int h(y) Q(dy|x,a) \right] \varphi^*(da|x) \quad \mu_1^* - \text{a.e.} \tag{6.13}$$

Since $J^* + h(x) < \infty$ for each $x \in S$, the right-hand side of (6.13) is a finite-valued function on S. Therefore, by the measurable selection theorem of Blackwell and Ryll-Nardzewski (see, e.g., [4]) there exists a deterministic stationary policy $f \in \mathcal{F}$ satisfying

$$\int_A \left[c(x,a) + \int h(y) Q(dy|x,a) \right] \varphi^*(da|x) \geq c(x, f(x)) + \int h(y) Q(dy|x, f(x))$$

for all $x \in S$, which together with (6.13) yields (6.7). Moreover, S is "absorbent" in the sense that if $x \in S$, then $Q(S|x, f(x)) = 1$ for μ_1^*-a.a. $x \in S$; otherwise we would get a contradiction to (6.6). This fact and (6.7) yield (6.8) by a standard recursion argument [2], [15], [16].

It is worth noting that the converse of Theorem 6.2(b) holds, which provides a "minimum pair" version of (6.2)–(6.3). That is, if a measure μ^* is feasible for (MP_1) and (6.6) holds μ^*-a.e., then μ^* is optimal for (MP_1), (φ^*, μ_1^*) is a minimum pair, and (4.6)–(4.7) hold.

Theorem 6.2, vis-à-vis (6.1)–(6.3), is perhaps not surprising. Previous works for *countable* (mainly finite) Markov decision processes in the multichain case (e.g., [3], [12]) have shown that what the LP approach does is to identify the set of states of an ergodic chain with minimum cost rate, as in Example 6.1. In our general (Borel) state case, our *guess* is that such a set is precisely the set S in Theorem 6.2(c). However, a precise answer on this issue requires further research on "ergodic decompositions" for *controlled* Markov processes, which for practical purposes is an untouched problem. In fact, to our knowledge, the only related work is Kurano's [11] for the very special case in which the state space is *compact* and *Doeblin's condition* holds. (For noncontrolled Markov chains, see, e.g., [14].)

REFERENCES

[1] E. J. Anderson and P. Nash, *Linear Programming in Infinite-Dimensional Spaces*, John Wiley, Chichester, 1987.

[2] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus, *Discrete-time controlled Markov processes with average cost criterion: A survey*, SIAM J. Control Optim., 31 (1993), pp. 282–344.

[3] E. V. Denardo, *On linear programming in a Markov decision problem*, Manag. Sci., 16 (1970), pp. 281–288.

[4] E. B. DYNKIN AND A. A. YUSHKEVICH, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.
[5] O. HERNÁNDEZ-LERMA, *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.
[6] ———, *Existence of average optimal policies in Markov control processes with strictly unbounded costs*, Kybernetika (Prague), 29 (1993), pp. 1–17.
[7] O. HERNÁNDEZ-LERMA AND D. HERNÁNDEZ-HERNÁNDEZ, *Discounted cost Markov decision processes on Borel spaces: The linear programming formulation*, J. Math. Anal. Appl., 183 (1994), pp. 335–351.
[8] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Linear programming and average optimality of Markov control processes on Borel spaces—unbounded costs*, SIAM J. Control Optim., 32 (1994), pp. 480–500.
[9] O. HERNÁNDEZ-LERMA AND M. MUNOZ DE OZAK, *Discrete-time Markov control processes with discounted unbounded costs: Optimality criteria,* Kybernetika (Prague), 28 (1992), pp. 191–213.
[10] K. HINDERER, *Foundations of Non-Stationary Dynamic Programming with Discrete-Time Parameter*, Lecture Notes Oper. Res. 33, Springer-Verlag, New York, 1970.
[11] M. KURANO, *The existence of a minimum pair of state and policy for Markov decision processes under the hypothesis of Doeblin*, SIAM J. Control Optim., 27 (1989), pp. 296–307.
[12] L. C. M. KALLENBERG, *Linear Programming and Finite Markovian Control Problems*, Math. Centre Tracts 148, Mathematische Centrum, Amsterdam, 1983.
[13] J. B. LASSERRE, *Average optimal stationary policies and linear programming in countable state Markov decision processes*, J. Math. Anal. Appl., 183 (1994), pp. 233–249.
[14] S. P. MEYN AND R. L. TWEEDIE, *Stability of Markovian processes* I: *Criteria for discrete-time chains*, Adv. in Appl. Probab., 24 (1992), pp. 542–574.
[15] R. MONTES-DE-OCA AND O. HERNÁNDEZ-LERMA, *Conditions for average optimality in Markov control processes with unbounded costs and controls*, summary in J. Math. Systems Estim. Control, 4 (1994), pp. 145–148. The full paper is available via anonymous ftp in `trick.ntp.springer.de`, file name `/jmsec/11617.ps`.
[16] L. I. SENNOTT, *Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs*, Oper. Res., 37 (1989), pp. 626–633.
[17] K. YOSIDA, *Functional Analysis*, 6th ed., Springer-Verlag, New York, 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 311–328, January 1996

014

APPROXIMATIONS IN DYNAMIC ZERO-SUM GAMES I*

MABEL M. TIDBALL† AND EITAN ALTMAN‡

Abstract. We develop a unifying approach for approximating a "limit" zero-sum game by a sequence of approximating games. We discuss both the convergence of the values and the convergence of optimal (or "almost" optimal) strategies. Moreover, based on optimal policies for the limit game, we construct policies which are almost optimal for the approximating games. We then apply the general framework to state approximations of stochastic games, to convergence of finite horizon problems to infinite horizon problems, and to convergence in the discount factor and in the immediate reward.

Key words. zero-sum games, approximations, stochastic games

AMS subject classifications. 90D05, 93E05

1. Introduction.

In many cases, one encounters dynamic games for which time and space are continuous and possibly unbounded. In general, numerical solution of such games involve discretization both in time and in space. In pursuit evasion games in particular (see Bardi, Falcone, and Soravia [6] and Pourtallier and Tidball [21]) and in differential games in general (see Pourtallier and Tolwinsky [22], Tidball and González [25]), the time and space discretization often leads to dynamic programming that has a stochastic game interpretation. The numerical solution then typically requires a finite state approximation.

Approximations in dynamic games has therefore been an active area of research for several decades. Several schemes for discretization of time and space and for approximations have been developed for differential games [6], [7], [21], [22], [25]. In stochastic games, much attention was devoted to approximations of infinite-horizon problems by (long) finite-horizon ones, e.g., [14], [19], [24], [27]; discretization of the state space has further been considered by Nowak [20] and Whitt [31], [32] (who also discretizes the actions spaces). Except for [20], [31], all the above references consider approximation of the value function of the games.

The aim of this paper is to study in a systematic way approximations in games, not only of the values but also for the policies. We begin by developing a general framework for establishing the convergence of the upper and lower values of a sequence of games G_n, $n = 1, 2, \ldots$, to a value R (which we assume exists) of a limit game $G = G_\infty$. We are further interested in the following questions: (i) Do (almost) optimal policies converge (in some sense)? (ii) Assume that u and v are (almost) optimal for some approximating game G_n (where n is large enough in some sense). Can we construct from these almost optimal policies for the limit game? (iii) Assume that u and v are (almost) optimal for the limit game. Can we use them to construct almost optimal policies for the approximating game G_n for n large enough?

Problem (ii) above arises in the following situation. Suppose that two players use some approximating numerical schemes to obtain "good" policies, e.g., time discretization. Each player might be using a different discretization scheme. Yet each

* Received by the editors September 14, 1993; accepted for publication (in revised form) September 22, 1994.

† Facultad de Ciencias Exactas, Ingeniera y Agrimensura, Universidad Nacional de Rosario, Pellegrini 250, 2000 Rosario, Argentina. The research of this author was partially performed during a visit to INRIA, Centre Sophia-Antipolis, Sophia-Antipolis, France.

‡ INRIA, Centre Sophia-Antipolis, 2004 Route des Lucioles, B.P. 93, 06902 Sophia-Antipolis cedex, France.

player would like to ensure that regardless of the discretization scheme used by the other player, he or she can guarantee some value, which would be "almost" the value of the nondiscretized game. In fact, since the real game that is played is the nondiscretized one, the desired discretization should perform well even if the adversary uses an optimal policy for the nondiscretized game.

Problem (iii), on the other hand, arises in the opposite situation, and this serves as an additional motivation for studying approximations in games. There are many examples of dynamic games where one can solve an infinite limiting game easier, where problems related to the boundaries are avoided. Indeed, examples are given in §8 of stochastic games with (large) finite state space for which the natural approach for constructing almost optimal policies is to solve a limit game with a countable state space.

After establishing the general theory for approximations in §2, we apply it to several approximation problems in discrete-time stochastic games with discounted reward and denumerable state space. Applications to other dynamic games are the subject of future research. The basic model of the stochastic games is presented in §3. We then present three schemes for state approximation in §4 for the case of infinite horizon. This generalizes many results on the convergence of the optimal value in Markov decision processes (i.e., stochastic games with a single player), e.g., [11], [15], [?], [29], [30]. Other related work on finite state approximations in Markov decision processes are [1], [2], [26]. In §5 we extend the results on state approximations for infinite horizon to the case of finite horizon by a transformation of the state space. In §6 we study the convergence of the finite-horizon problem to the infinite-horizon one, and we combine state approximations with approximation of the horizon. In §7 we study the stability of stochastic games in the discount factor and in the immediate reward. Applications of approximation methods developed in this paper are presented in §8, and some generalizations are finally discussed in §9.

2. Key theorems for approximations. We consider the sequence $G_n = (S_n, U_n, V_n)$, $n = 1, 2, \ldots, \infty$, of generic zero-sum games, where U_n is the set of strategies of player I and V_n is the set of strategies of player II for the nth game. We assume that both U_n and V_n are endowed with some topology. $S_n : U_n \times V_n \to \mathbb{R}$ is a measurable function for all n. We define the upper (lower) value of the game:

$$\overline{R_n} = \inf_{v \in V_n} \sup_{u \in U_n} S_n(u, v) \quad \left(\underline{R_n} = \sup_{u \in U_n} \inf_{v \in V_n} S_n(u, v) \right).$$

$G = (S, U, V) \stackrel{\text{def}}{=} (S_\infty, U_\infty, V_\infty)$ will be called the limit game. It will be assumed that it has a value $R \stackrel{\text{def}}{=} R_\infty$.

An example where G_n does not have a value but G does will be given in §6.3 for computing almost optimal stationary policies for stochastic games with long (but finite) horizon.

A strategy $u^* \in U_n$ is said to be ϵ-optimal for player one in game n if

$$\inf_{v \in V_n} S_n(u^*, v) \geq \inf_{v \in V_n} S_n(u, v) - \epsilon \qquad \forall u \in U_n, \tag{1}$$

which is equivalent to $\inf_{v \in V_n} S_n(u^*, v) \geq \underline{R}_n - \epsilon$. It is said to be strongly ϵ-optimal for player one in game n if it satisfies

$$\inf_{v \in V_n} S_n(u^*, v) \geq \overline{R}_n - \epsilon.$$

A strategy $v^* \in V_n$ is said to be ϵ-optimal for player two in game n if

$$\sup_{u \in U_n} S_n(u, v^*) \leq \sup_{u \in U_n} S_n(u, v) + \epsilon \qquad \forall v \in V_n, \tag{2}$$

which is equivalent to $\sup_{u \in U_n} S_n(u, v^*) \leq \overline{R}_n + \epsilon$. It is said to be strong ϵ-optimal if

$$\sup_{u \in U_n} S_n(u, v^*) \leq \underline{R}_n + \epsilon.$$

Note that strong ϵ-optimality implies ϵ-optimality. If a game has a value $\underline{R}_n = \overline{R}_n$, then strong ϵ-optimality is equivalent to ϵ-optimality.

Assume that (S_n, U_n, V_n) converge (in some sense) to (S, U, V). We are interested in the following questions:

(Q1) Convergence of the values: Does $\underline{R}_n$ (or $\overline{R}_n$) converge to R?

(Q2) Convergence of policies: Fix some $\epsilon \geq 0$. Let ϵ_n be a sequence of positive real numbers such that $\overline{\lim}_{n \to \infty} \epsilon_n \leq \epsilon$. Assume that u_n^* and v_n^* are ϵ_n-optimal policies for the nth game. Are u_n^* and v_n^* "almost" optimal for the limit game for all n large enough?

(Q3) Let $\overline{u} \in U$ (resp., $\overline{v} \in V$) be some limit point of u_n^* (resp., v_n^*), defined above. Is $\overline{u}$ (resp., $\overline{v}$) ϵ-optimal for the limit game?

(Q4) Robustness of the optimal policy: If u^* (resp. v^*) is ϵ-optimal for the limit game, can we derive from it an "almost" (strongly) optimal policy for the nth approximating game for all n large enough?

In most applications that we discuss in this paper, $U_n = U$, $V_n = V$ do not depend on n. However, in several applications this is not the case, e.g., approximations in pursuit evasion games; see Bernhard and Shinar [7]. Another example is given for a state approximation scheme for solving stochastic games; see §4.3.

THEOREM 2.1. *Assume that there exist sequences of functions* $\pi_n^1 : U_n \to U$, $\pi_n^2 : V_n \to V$, $\sigma_n^1 : U \to U_n$, $\sigma_n^2 : V \to V_n$, $n = 1, 2, \ldots$, *such that*

$(\mathcal{A}1)$ $\overline{\lim}_{n \to \infty} [S_n(u, \sigma_n^2(v)) - S(\pi_n^1(u), v)] \leq 0$ *uniformly in* $u \in U_n$ *for each* $v \in V$.

$(\mathcal{A}2)$ $\underline{\lim}_{n \to \infty} [S_n(\sigma_n^1(u), v) - S(u, \pi_n^2(v))] \geq 0$ *uniformly in* $v \in V_n$ *for each* $u \in U$.

Then

(1) $\lim_{n \to \infty} \underline{R}_n = \lim_{n \to \infty} \overline{R}_n = R$.

(2) *For any* $\epsilon' > \epsilon$, *there exists* N *such that* $\pi_n^1(u_n^*)$ *(resp.,* $\pi_n^2(v_n^*)$; *see definitions in* (Q2)) *is* ϵ'*-optimal for the limit game for all* $n \geq N$.

(3) *Let* u^* *(resp.,* v^**) be* ϵ*-optimal for the limit game. Then for all* $\epsilon' > \epsilon$, *there exists* $N(\epsilon')$ *such that* $\sigma_n^1(u^*)$ *(resp.,* $\sigma_n^2(v^*)$*) is strongly* ϵ'*-optimal for the* n*th approximating game for all* $n \geq N(\epsilon')$.

(4) *Suppose*

$(\mathcal{A}3)$ $S(u, v)$ *is a lower semicontinuous function in* u,

$(\mathcal{A}4)$ $S(u, v)$ *is an upper semicontinuous function in* v.

Suppose $\bar{u} \in U$ *(resp.,* $\bar{v} \in V$*) is a limit point of* $\pi_n^1(u_n^*)$ *(resp.,* $\pi_n^2(v_n^*)$*). Then* $\bar{u}$ *(resp.,* $\bar{v}$*) is* ϵ*-optimal for the limit game.*

Remark 2.1. Part (1) of Theorem 2.1 is a generalization of Lemma 1 in [14].

Proof. (1) Choose $\epsilon > 0$. Let $u^* \in U$ and $v^* \in V$ be ϵ-optimal for the limit game G, and choose some sequence $\epsilon(n)$ such that $\lim_{n \to \infty} \epsilon(n) = 0$. Let $u_n \in U_n$

be an $\epsilon(n)$-best response to the policy $\sigma_n^2(v^*)$ in game G_n (i.e., $S_n(u_n, \sigma_n^2(v^*)) \geq S_n(u, \sigma_n^2(v^*)) - \epsilon(n)$ for all $u \in U_n$). Similarly, let $v_n \in V_n$ be an $\epsilon(n)$-best response to the policy $\sigma_n^1(u^*)$ in game G_n.

Choose some $\delta > 0$. By $(\mathcal{A}1)$, there exists N such that for all $n \geq N$ and $u \in U_n$, $S_n(u, \sigma_n^2(v^*)) - S(\pi_n^1(u), v^*) < \delta$. Then for all $n \geq N$

$$\begin{aligned}
\overline{R}_n - R &= \inf_{v \in V_n} \sup_{u \in U_n} S_n(u, v) - \inf_{v \in V} \sup_{u \in U} S(u, v) \\
&\leq \sup_{u \in U_n} S_n(u, \sigma_n^2(v^*)) - \sup_{u \in U} S(u, v^*) + \epsilon \\
&\leq S_n(u_n, \sigma_n^2(v^*)) - S(\pi_n^1(u_n), v^*) + \epsilon + \epsilon(n) \\
&\leq \delta + \epsilon + \epsilon(n).
\end{aligned}$$

Hence $\overline{\lim}_{n\to\infty} \overline{R}_n \leq R + \delta + \epsilon$.

Similarly, by $(\mathcal{A}2)$, one shows that $R \leq \underline{\lim}_{n\to\infty} \underline{R}_n + \delta + \epsilon$. Since $\underline{R}_n \leq \overline{R}_n$, this implies that $\overline{\lim}_{n\to\infty} |R - \underline{R}_n| \leq \delta + \epsilon$ and $\overline{\lim}_{n\to\infty} |R - \overline{R}_n| \leq \delta + \epsilon$. The result follows since ϵ and δ can be chosen arbitrarily small.

(2) Fix some $\delta > 0$. By $(\mathcal{A}1)$, as u_n^* is $\epsilon(n)$-optimal for G_n, and by part (1), there exists $N(\epsilon, \delta)$ such that for $n > N(\epsilon, \delta)$ we have

$$(3) \quad \forall v \in V: \quad S_n(u_n^*, \sigma_n^2(v)) - S(\pi_n^1(u_n^*), v) < \delta, \quad \epsilon(n) < \epsilon + \delta, \quad |\underline{R}_n - R| < \delta,$$

and thus

$$\begin{aligned}
\inf_{v \in V} S(\pi_n^1(u_n^*), v) &\geq \inf_{v \in V} S_n(u_n^*, \sigma_n^2(v)) - \delta \\
&\geq \inf_{v \in V_n} S_n(u_n^*, v) - \delta \geq \underline{R}_n - \delta - \epsilon(n) \geq R - 3\delta - \epsilon.
\end{aligned}$$

So $\pi_n^1(u_n^*)$ is ϵ'-optimal for S with $\epsilon' = 3\delta + \epsilon$.

In the same way, by assumption $(\mathcal{A}2)$ and considering an $\epsilon(n)$-optimal policy v_n^* for G_n, we obtain that $\pi_n^2(v_n^*)$ is ϵ'-optimal for S for all large enough n. The proof follows from the fact that δ was chosen arbitrarily.

(3) Fix some $\delta > 0$. As u^* is an ϵ-optimal strategy in the limit game G and by $(\mathcal{A}2)$, for all n large enough we have

$$\begin{aligned}
\inf_{v \in V_n} S_n(\sigma_n^1(u^*), v) &\geq \inf_{v \in V_n} S(u^*, \pi_n^2(v)) - \delta \\
&\geq \inf_{v \in V} S(u^*, v) - \delta \geq R - \delta - \epsilon \geq \overline{R}_n - 2\delta - \epsilon.
\end{aligned}$$

The proof for v^* is obtained in the same way.

(4) Let $\hat{v} \in V$ be such that $\inf_{v \in V} S(\bar{u}, v) \geq S(\bar{u}, \hat{v}) - \delta$. By $(\mathcal{A}1)$, $(\mathcal{A}3)$, and part (1) of Theorem 2.1, for all $\delta > 0$ there exists $N(\delta)$ such that $n > N(\delta)$ implies $\epsilon(n) < \epsilon + \delta$, and

$$\begin{aligned}
\inf_{v \in V} S(\bar{u}, v) \geq S(\bar{u}, \hat{v}) - \delta &\geq S(\pi_n^1(u_n^*), \hat{v}) - 2\delta \\
&\geq S_n(u_n^*, \sigma_n^2(\hat{v})) - 3\delta \\
&\geq \underline{R}_n - 3\delta - \epsilon(n) \\
&\geq R - 4\delta - \epsilon(n);
\end{aligned}$$

hence, $\bar{u}$ is ϵ'-optimal with $\epsilon' = \epsilon + 5\delta$. In the same way, by $(\mathcal{A}2)$ and $(\mathcal{A}4)$ we prove that $\bar{v}$ is ϵ'-optimal for S. The proof follows from the fact that δ was chosen arbitrarily. □

Remark 2.2. (i) In the rest of the paper, whenever $U_n = U$ and $V_n = V$ do not depend on n, π_n and σ_n will be chosen as the identity maps.

(ii) It follows from the proof of part (1) in the above theorem that if for all G_n, $n = 1, 2, \ldots, \infty$, there exist optimal policies for both players and if $U_n = U$ and $V_n = V$ do not depend on n, then

$$|\overline{R}_n - R| \leq \sup_{u,v} |S_n(u,v) - S(u,v)|, \qquad |\underline{R}_n - R| \leq \sup_{u,v} |S_n(u,v) - S(u,v)|.$$

3. Stochastic games: The model. We will use the results from the previous section to study approximations of zero-sum stochastic games.

- Let $\mathbf{I}$ be a denumerable set of states.
- $\mathbf{A}_i$ ($\mathbf{B}_i$) is a compact metric set of actions for player I (resp., II) at state i. Let $K = \{i, \mathbf{A}_i, \mathbf{B}_i\}_{i \in I}$.
- $r : K \to \mathbb{R}$ is a bounded immediate reward function. (The boundedness condition can be relaxed; see §9.) Let $M \stackrel{\text{def}}{=} \sup_{i,a,b} |r(i,a,b)|$.
- $P(a,b) = [p(i,a,b,E)]_{i,E}$, $a \in \mathbf{A}_i$, $b \in \mathbf{B}_i$ is a (sub) probability transition (from state i to a set $E \subset \mathbf{I}$) when the players use actions a and b.
- β is the discount factor satisfying $0 \leq \beta < 1$.

We shall use the following standard assumption (see, e.g., Nowak [18]):

(M_1) : $r(i, \bullet, \bullet)$ and $p(i, \bullet, \bullet, E)$ are continuous in both actions for any $E \subset \mathbf{I}$.

The game is played in stages $t = 0, 1, 2, \ldots$. If at some stage t the state is i, then the players independently choose actions $a \in \mathbf{A}_i$, $b \in \mathbf{B}_i$. Player II then pays player I the amount $r(i,a,b)$, and at stage $t+1$ the new state is chosen according to the transition probabilities $p(i,a,b,\bullet)$. The game continues at this new state.

Let U and V be the set of behavioral strategies for both players. A strategy $u \in U$ is a sequence $u = (u_0, u_1, \ldots)$, where u_t is a probability measure over the available actions, given the whole history of previous states and of previous actions of both players as well as the current state.

A Markov policy $q = \{q_0, q_1, \ldots\}$ is a policy (for either player I or player II) where q_t is allowed to depend only on t and on the state at time t.

A *stationary* (*mixed*) *policy* g for player one is characterized by a conditional distribution $p^g(\bullet \mid j)$ over $\mathbf{A}_j$, so that $p^g(\mathbf{A}_j \mid j) = 1$, which is interpreted as the distribution over the actions available at state j which player I uses when it is in state j. With some abuse of notation, we shall set $g(\bullet \mid j) = p^g(\bullet \mid j)$ for stationary g. Let S^A be the set of stationary policies for player I, and define similarly the stationary policies S^B for player II. If both players use stationary policies, say u and v, then $\{X_t\}$ becomes a Markov chain with stationary transition probabilities, given by

$$p(j,u,v,k) = \int_{\mathbf{A}_j} \int_{\mathbf{B}_j} p(j,a,b,k) u(da|j) v(db|j). \tag{4}$$

In the following section we are concerned with the infinite-horizon discounted problem. It is known that under (M_1), optimal stationary policies exist for both players; i.e., if u and v are optimal policies for both players when both of them are restricted to stationary policies, then each one of these policies is also optimal against an arbitrary policy of his or her opponent. We shall therefore restrict the game to stationary mixed policies, without loss of generality (see [12], [17]).

Next we introduce a topology on the sets of stationary policies. For any compact metric set Γ, let $M(\Gamma)$ denote the set of probability measures on Γ endowed with the weak topology $\xi(\Gamma)$ (see [18]). The class of stationary policies for player I (and

similarly for player II) can be identified with the set $\prod_{i\in\mathbf{I}} M(\mathbf{A}_i)\times M(\mathbf{B}_i)$; moreover, it is compact with respect to the product topology $\prod_{i\in\mathbf{I}} \xi(\mathbf{A}_i)\times\xi(\mathbf{B}_i)$.

Let (u,v) be a pair of strategies and $i\in\mathbf{I}$ be a fixed initial state. Let I_t, A_t, B_t, $t=0,\dots,$ be the resulting stochastic process of the states and actions of the players. Let $E_i^{u,v}$ denote the expectation with respect to the measure defined by u,v,i. Define the β-discounted game payoff

$$S(i,u,v) = E_i^{u,v}\sum_{t=0}^{\infty}\beta^t r(I_t,A_t,B_t). \tag{5}$$

Let $R(i)$ denote the value of the stochastic game for initial state i. For stationary policies u and v, let the expected current payoff be defined by

$$r(i,u,v) = \int_{\mathbf{A}_i}\int_{\mathbf{B}_i} r(i,a,b)u(da|i)v(db|i). \tag{6}$$

Consider the following (contracting) map:

$$(T_{u,v}f)(i) \stackrel{\text{def}}{=} r(i,u,v) + \beta\sum_{j\in\mathbf{I}} p(i,u,v,j)f(j). \tag{7}$$

Then $S(i,u,v)$ is known to be the unique solution of (7). The value $R(i)$ is the unique solution of

$$R(i) = val\left[r(i,a,b) + \beta\sum_{j\in\mathbf{I}} p(i,a,b,j)R(j)\right]. \tag{8}$$

Moreover, any stationary policies u^* and v^* that choose at any state j the mixed strategies that are optimal for the game

$$\left[r(i,a,b) + \beta\sum_{j\in\mathbf{I}} p(i,a,b,j)R(j)\right]_{a,b}$$

are known to be optimal for the stochastic game S (see [18] for these statements).

Remark 3.1. $S(i,\bullet,\bullet): S^A\times S^B\to\mathbb{R}$ are continuous for all states i. This follows from Corollary 2.2 in Borkar [10]. It will thus follow below that assumptions $(\mathcal{A}3)$ and $(\mathcal{A}4)$ hold.

4. State approximations: Infinite-horizon case. We introduce below several approximating schemes. All of them involve some sequence $\mathbf{I}_n\subset\mathbf{I}$ of sets of states, which are naturally chosen to be increasing. We shall assume

$$\mathbf{I}_n\subset\mathbf{I}_{n+1}, \qquad \bigcup_n \mathbf{I}_n = \mathbf{I}. \tag{$\mathcal{B}1$}$$

The following property will imply conditions $(\mathcal{A}1)$–$(\mathcal{A}2)$ in the various schemes that we consider below:

$$\text{For all integers } r, \qquad \epsilon(r,n) = \sup_{a,b,i\in\mathbf{I}_r}\left\{\sum_{j\notin\mathbf{I}_n} p(i,a,b,j)\right\}\to 0 \text{ as } n\to\infty. \tag{$\mathcal{B}2$}$$

Remark 4.1. Consider the case that the sets $\mathbf{I}_r$ are all finite. Condition (M_1) as well as the compactness of $\mathbf{A}_i$ and $\mathbf{B}_i$ then implies $(\mathcal{B}2)$. Indeed, assume that $(\mathcal{B}2)$ does not hold. Then, there exists some $\alpha > 0$ such that for some i,

$$\overline{\lim_{n\to\infty}} \max_{a,b} \left(\sum_{j\in\mathbf{I}} p(i,a,b,j) 1\{j \notin \mathbf{I}_n\} \right) = \alpha. \tag{9}$$

Let a_n and b_n be some actions achieving the max. (The fact that the max is achieved follows from the compactness and continuity assumption (M_1).) Choose a subsequence $n(\ell), \ell = 1, 2, \ldots$, along which the limsup is obtained and along which a_n and b_n converge to some actions a^* and b^*. Then $p(i, a_{n(\ell)}, b_{n(\ell)}, \bullet)$ converges (pointwise) to the probability $p(i, a^*, b^*, \bullet)$ as $\ell \to \infty$ (by (M_1)). But then it follows from a dominant convergence theorem [23, Chap. 11, §4] and $(\mathcal{B}1)$ that

$$\lim_{\ell\to\infty} \sum_{j\in\mathbf{I}} p(i, a_{n(\ell)}, b_{n(\ell)}, j) 1\{j \notin \mathbf{I}_{n(\ell)}\} = \sum_{j\in\mathbf{I}} p(i, a^*, b^*, j) \cdot 0 = 0,$$

which contradicts (9). Hence $(\mathcal{B}2)$ is established.

For the case of a single player, $(\mathcal{B}2)$ was introduced as an assumption for several approximating schemes by Cavazos-Cadena [11]. Note, however, that in [11], (M_1) and the compactness of the action spaces are not assumed. In order to obtain conditions $(\mathcal{A}1)$ and $(\mathcal{A}2)$ for the approximating schemes below (and hence obtain statements (1), (2), and (3) in Theorem 2.1) one could relax the compactness assumption as well as (M_1); in that case one would indeed need to impose $(\mathcal{B}2)$ as an assumption. The compactness and (M_1) (or other similar assumptions, such as (M_2) or (M_3) from [18]) are required, however, to establish the continuity conditions $(\mathcal{A}3)$ and $(\mathcal{A}4)$ required to establish statement (4) in Theorem 2.1.

Other typical assumptions that imply $(\mathcal{B}2)$ have often been used in the literature; see White [29] and Hernández-Lerma [15] as well as $(\mathcal{B}3)$, introduced in Altman [1], which will be used occasionally below:

$(\mathcal{B}3)$ From any state k, only a finite set of states X_k can be reached.

In all approximations in this section, the approximating games G_n have a value, i.e., $R_n(i) = \underline{R}_n(i) = \overline{R}_n(i)$. Moreover, they will have a saddle point among the stationary policies.

4.1. Approximation scheme I. We define

$$\left(H^1_{u,v} f\right)(i) \stackrel{\text{def}}{=} \begin{cases} r(i,u,v) + \beta \sum_{j\in\mathbf{I}_n} p(i,u,v,j) f(j) & \text{if } i \in \mathbf{I}_n, \\ 0 & \text{if } i \notin \mathbf{I}_n. \end{cases} \tag{10}$$

For this approximating problem we define $S^1_n(i,u,v)$ to be the solution of

$$\left(H^1_{u,v} f\right)(i) = f(i) \qquad \forall i \in \mathbf{I}. \tag{11}$$

S^1_n is thus the total discounted payoff (defined in (5)) for the stochastic game whose transition probabilities are $\bar{p}$ instead of p, where $\bar{p}(i,u,v,j) = p(i,u,v,j)$ if $\{i, j \in \mathbf{I}_n\}$, 0 if $\{i, j \notin \mathbf{I}_n\}$. The value of the game G^1_n is the unique solution of

$$R_n(i) = \begin{cases} val\left[r(i,a,b) + \beta \sum_{j\in\mathbf{I}_n} p(i,a,b,j) R_n(j)\right] & \text{if } i \in \mathbf{I}_n, \\ 0 & \text{if } i \notin \mathbf{I}_n. \end{cases} \tag{12}$$

Moreover, optimal stationary policies u_n^* and v_n^* for the game G_n^1 are obtained by chosing at any state $i \in \mathbf{I}_n$ the mixed strategies that are optimal for the game

$$\left[r(i,a,b) + \beta \sum_{j \in \mathbf{I}_n} p(i,a,b,j) R_n(j) \right]_{a,b}.$$

The proof of the following theorem will enable us to evaluate the precision of the approximation. More precisely, it will enable us to get a bound on $|R_n(\bullet) - R(\bullet)|$ which will be uniform in $i \in J$ where J is an arbitrary fixed subset of $\mathbf{I}$.

THEOREM 4.1. *Assume* (M_1) *and* $(\mathcal{B}2)$. *All statements of Theorem* 2.1 *hold for approximating scheme* I, *where the reward* S *for the limit game* G *is defined in* (5) *and for approximating game* G_n^1 *it is* S_n^1.

Proof. The proof uses an idea by Cavazos-Cadena [11]. We need first to introduce some definitions. Let $\epsilon > 0$. We define

$$g^0(\epsilon, r) = r, \qquad g^k(\epsilon, r) = g(\epsilon, g^{k-1}(\epsilon, r)), \qquad k = 1, 2, \ldots,$$

where

$$g(\epsilon, r) = \min \{m : \epsilon(r, m) \leq \epsilon\}$$

and $\epsilon(r, m)$ is defined in property $(\mathcal{B}2)$. To understand the meaning of $g^s(\epsilon, r)$, we consider first $\epsilon = 0$. Then $\mathbf{I}_{g^{s+1}(0,r)}$ is the set of neighbors of $\mathbf{I}_{g^s(0,r)}$ in the sense that states that are not contained in $\mathbf{I}_{g^{s+1}(0,r)}$ are not reachable from any state in $\mathbf{I}_{g^s(0,r)}$. For $\epsilon > 0$, $\mathbf{I}_{g^{s+1}(\epsilon,r)}$ is the set of "ϵ-neighbors" of $\mathbf{I}_{g^s(\epsilon,r)}$ in the sense that for any state i in $\mathbf{I}_{g^s(\epsilon,r)}$, the states that are not contained in $\mathbf{I}_{g^{s+1}(\epsilon,r)}$ are reachable from i with probability smaller than or equal to ϵ. Note that for any $r \geq 0$

$$\sum_{j \notin \mathbf{I}_{g^{l+1}(\epsilon,r)}} p(i,u,v,j) \leq \epsilon \qquad \forall i \in \mathbf{I}_{g^l(\epsilon,r)},\ \forall l \geq 0,\ \forall u, v. \tag{13}$$

Note also that $g^l(\epsilon, r)$ need not be increasing in l.

Let $J \subset \mathbf{I}$, $\epsilon(J) = \min \{m : J \subset \mathbf{I}_m\}$, and suppose $\epsilon(J) < +\infty$. (This is the case if J is chosen to be finite.) We define

$$m_k(\epsilon, \epsilon(J)) = \max \left\{ \epsilon(J), g(\epsilon, \epsilon(J)), \ldots, g^k(\epsilon, \epsilon(J)) \right\}, \qquad k = 0, 1, 2, \ldots. \tag{14}$$

We show that assumptions $(\mathcal{A}1)$–$(\mathcal{A}4)$ hold for $S_n^1(i,u,v)$ and $S(i,u,v)$ defined above.

Below, ϵ and J are fixed, so for simplicity of notation we shall write g^l instead of $g^l(\epsilon, \epsilon(J))$. Let $n \geq m_k(\epsilon, \epsilon(J))$. Then for all $i \in J$

$$\begin{aligned} |S_n^1(i,u,v) - S(i,u,v)| \leq{}& \beta \sum_{j \in \mathbf{I}_{g^1}} p(i,u,v,j)\, |S_n^1(j,u,v) - S(j,u,v)| \\ &+ \beta \sum_{j \notin \mathbf{I}_{g^1}} p(i,u,v,j)\, |S_n^1(j,u,v) - S(j,u,v)|. \end{aligned} \tag{15}$$

Note that for any state j, $|S_n^1(j,u,v)| \leq M/(1-\beta)$ and $|S(j,u,v)| \leq M/(1-\beta)$. Hence by (13), (15), and $(\mathcal{B}2)$ we obtain

$$|S_n^1(i,u,v) - S(i,u,v)|$$

$$\leq \beta \sum_{j \in \mathbf{I}_{g^1}} p(i,u,v,j)\, |S_n^1(j,u,v) - S(j,u,v)| + \epsilon \frac{2M\beta}{1-\beta}$$

$$\leq \beta \max_{j \in \mathbf{I}_{g^1}} |S_n^1(j,u,v) - S(j,u,v)| + \epsilon \frac{2M\beta}{1-\beta}$$

$$\leq \beta \max_{j \in \mathbf{I}_{g^1}} \left\{ \beta \sum_{\ell \in \mathbf{I}_{g^2}} p(j,u,v,\ell)\, |S_n^1(\ell,u,v) - S(\ell,u,v)| \right.$$

$$\left. + \beta \sum_{\ell \notin \mathbf{I}_{g^2}} p(j,u,v,\ell)\, |S_n^1(\ell,u,v) - S(\ell,u,v)| \right\} + \epsilon \frac{2M\beta}{1-\beta}$$

$$\leq \beta^2 \max_{\ell \in \mathbf{I}_{g^2}} |S_n^1(\ell,u,v) - S(\ell,u,v)| + \epsilon \frac{2M\beta^2}{1-\beta} + \epsilon \frac{2M\beta}{1-\beta} \tag{16}$$

$$\leq \beta^k \max_{\ell \in \mathbf{I}_{g^k}} |S_n^1(\ell,u,v) - S(\ell,u,v)| + 2\epsilon \frac{M\beta}{1-\beta} \sum_{\ell=0}^{k-1} \beta^\ell.$$

The first inequality follows by (13) since $n \geq m_k \geq g^1$ and since $i \in J \subset \mathbf{I}_{g^0}$. Similarly, (16) follows by (13) since $n \geq m_k \geq g^2$ and since $\ell \in \mathbf{I}_{g^1}$. So we have

$$|S_n^1(j,u,v) - S(j,u,v)| \leq 2M \frac{\beta(1-\beta^k)\epsilon/(1-\beta) + \beta^k}{1-\beta}. \tag{17}$$

Hence $(\mathcal{A}1)$ and $(\mathcal{A}2)$ hold true (whereas $(\mathcal{A}3)$–$(\mathcal{A}4)$ are established in Remark 3.1). □

Combining (17) with Remark 2.2 yields the following corollary.

COROLLARY 4.2. *For any $i \in J$, if n is chosen such that $n \geq m_k(\epsilon, \epsilon(J))$, then*

$$|R_n(i) - R(i)| \leq 2M \frac{\beta(1-\beta^k)\epsilon/(1-\beta) + \beta^k}{1-\beta}. \tag{18}$$

In the previous theorem, the sets $\{\mathbf{I}_n\}$ where given a priori. Next we consider a special choice of $\{\mathbf{I}_n\}$ that will be especially useful under assumption $(\mathcal{B}3)$ for a finite set J. This construction will enable us to express in a simple way the sets $\mathbf{I}_n$ needed in (10) in order to approximate R by R_n with a given error. This is especially desirable when it is not easy to compute $m_k(\epsilon, \epsilon(J))$ (and thus Corollary 4.2 cannot be used).

Let J be a given set (for which we would like to get a computable uniform bound on the error of the approximation), and set $Y(i) = \{j : p(i,u,v,j) > 0 \text{ for some } u, v\}$. Then we define $\mathbf{I}_n$ in the following way:

$$\mathbf{I}_0 = J, \qquad \mathbf{I}_{n+1} = \bigcup_{i \in \mathbf{I}_n} Y(i) \bigcup \mathbf{I}_n. \tag{19}$$

Remark 4.2. Note that if J is finite and if $(\mathcal{B}3)$ holds, then all sets $\mathbf{I}_n$ are finite. This construction might be especially useful if the number of states reachable from any given state is small. In that case $\mathbf{I}_n$ do not grow too quickly. We now consider $S_n(i,u,v)$ the solution of (11) with $\mathbf{I}_n$ defined in (19). We have that the following theorem (analogous to (4.1)) holds.

THEOREM 4.3. *Assume* (M_1) *and* $(\mathcal{B}3)$.

(i) *Fix a state $i \in J$. Then all statements of Theorem* 2.1 *hold for the approximating scheme I, where the reward S of the limit game G is defined in* (5) *and the reward S_n for the approximating game G_n is the solution of* (11) *with $\mathbf{I}_n$ defined in* (19).

(ii) *For any $i \in J$ and $n = 0, 1, \ldots$, $|R_n(i) - R(i)| \leq 2M\beta^n/(1-\beta)$.*

Proof. It suffices to prove that $(\mathcal{A}1)$ and $(\mathcal{A}2)$ are satisfied. Let $i \in J$. Then

$$\begin{aligned}
|S_n(i,u,v) - S(i,u,v)| &\leq \beta \sum_{j \in Y(i)} p(i,u,v,j)|S_n(j,u,v) - S(j,u,v)| \\
&\leq \beta \max_{j \in \mathbf{I}_1} |S_n(j,u,v) - S(j,u,v)| \\
&\leq \beta^2 \max_{j \in \mathbf{I}_1} \sum_{k \in Y(j)} p(j,u,v,k)|S_n(k,u,v) - S(k,u,v)| \\
&\quad \cdot \\
&\quad \cdot \\
&\quad \cdot \\
&\leq \beta^n \max_{j \in \mathbf{I}_n} |S_n(j,u,v) - S(j,u,v)| \leq \frac{2M\beta^n}{1-\beta}.
\end{aligned}$$

The proof of (ii) then follows from Remark 2.2. □

As suggested in Remark 4.2, the above method is useful specially when the approximating games have finite states (i.e., $\mathbf{I}_n$ are finite) and the typical number of states (neighbors) is reachable from a state is not too high. If, however, the typical number of neighbors is high, then the sets $\mathbf{I}_n$ become large very rapidly, which suggests that obtaining good estimates of optimal value and policies might require an unacceptably high complexity of computations. We thus present an alternative more general way of constructing finite sets $\mathbf{I}_n$ (even when $(\mathcal{B}3)$ does not hold), which will result in a simple expression for $m_k(\epsilon, \epsilon(J))$ and will thus enable use of Corollary 4.2 to obtain a uniform computable error bound for the approximation for any $i \in J$.

We define a parametrized family $\{\mathbf{I}_n(\epsilon)\}$, where ϵ is a positive real number. Define $\mathbf{I}_0(\epsilon) = J$. $\{\mathbf{I}_n(\epsilon)\}$ are then chosen to be an arbitrary sequence increasing to $\mathbf{I}$ that satisfies the following. If for some $l > 0$, say $l = \hat{l}$,

$$\sup_{a,b,i \in \mathbf{I}_l(\epsilon)} \sum_{j \notin \mathbf{I}_l(\epsilon)} p(i,a,b,j) \leq \epsilon,$$

then $\mathbf{I}_n(\epsilon) = \mathbf{I}$ for all $n > \hat{l}$. Otherwise, $\mathbf{I}_{l+1}$ is chosen such that

$$\sup_{a,b,i \in \mathbf{I}_l(\epsilon)} \sum_{j \notin \mathbf{I}_{l+1}(\epsilon)} p(i,a,b,j) \leq \epsilon.$$

It follows that $\epsilon(J) = 0$, $g^0 = 0$, and hence $g^k = k$ and $m_k(\epsilon, \epsilon(J)) = g^k = k$ for $k \leq \hat{l}$. (The above quantities were defined in the proof of Theorem 4.1.) If J is finite, then it follows from the same arguments as in Remark 4.2 that $\mathbf{I}_n(\epsilon)$ can be chosen to be finite for $n \leq \hat{l}$ and hence in particular $\mathbf{I}_{\hat{l}}(\epsilon)$, which is the truncated state space that should be used to perform approximation scheme I in order to obtain a precision as in Corollary 4.2.

In this setting, Theorem 4.3 (ii) becomes a special case of Corollary 4.2 with $\epsilon = 0$.

4.2. Approximation scheme II. In the previous approximation scheme, the dynamics are seen to be a result of transition probabilities that need not sum to one, even if in the limit game they do sum to one. Indeed, (10) can be considered a stochastic game where we set $p(i,u,v,j) = 0$ for $j \notin \mathbf{I}_n$. In many applications this may be undesirable, and one would like $p(i,u,v,\bullet)$ to remain a probability measure. This is especially the case when we want to learn about the optimal value and (almost) optimal policies for a given specific stochastic games with large finite state space by approximating them through an infinite state game. Indeed, there are cases where one can solve an infinite game more easily, since some boundary problems are avoided. Examples are given in §8.

We assume that $\sum_{j\in\mathbf{I}} p(i,a,b,j) = 1$ for all $a \in \mathbf{A}_i, b \in \mathbf{B}_i$. We define the following sequence of games. We let $\mathbf{I}_n \subset \mathbf{I}$ be an increasing sequence of sets, converging to $\mathbf{I}$, as in the previous section. Define

$$\left(H^2_{u,v}f\right)(i) = \begin{cases} r(i,u,v) + \beta \sum_{j\in\mathbf{I}_n} p^*(i,u,v,j) f(j) & \text{if } i \in \mathbf{I}_n, \\ 0 & \text{if } i \notin \mathbf{I}_n, \end{cases} \tag{20}$$

where

$$p^*(i,u,v,j) = \begin{cases} p(i,u,v,j) + q_n(i,u,v,j) & \text{if } i \in \mathbf{I}_n,\ j \in \mathbf{I}_n, \\ 0 & \text{if } i \notin \mathbf{I}_n \text{ or } j \notin \mathbf{I}_n. \end{cases} \tag{21}$$

$q_n(i,u,v,\bullet)$ is some nonnegative measure satisfying $\sum_{j\in\mathbf{I}_n} (p(i,u,v,j) + q_n(i,u,v,j)) = 1$. Hence,

$$\sum_{j\in\mathbf{I}_n} q_n(i,u,v,j) = \sum_{j\notin\mathbf{I}_n} p(i,u,v,j). \tag{22}$$

We define S^2_n to be the solution of

$$\left(H^2_{u,v}f\right)(i) = f(i). \tag{23}$$

S^2_n is thus the total discounted payoff (defined in (5)) for the stochastic game whose transition probabilities are p^* instead of p.

Assume (M_1) and $(\mathcal{B}2)$. Then all the results of §4.1 still hold. We demonstrate this with the proof of the analogue of Theorem 4.1. It suffices to show that $(\mathcal{A}1)$–$(\mathcal{A}2)$ hold. With the same notation as in §4.1 we obtain

$$\begin{aligned} |S^2_n(i,u,v) - S(i,u,v)| \le{}& \beta \sum_{j\in\mathbf{I}_{g^1}} p(i,u,v,j)\, |S^2_n(j,u,v) - S(j,u,v)| \\ &+ \beta \sum_{j\in\mathbf{I}_{g^1}} q(i,u,v,j)\, |S^2_n(j,u,v) - S(j,u,v)| \\ &+ \beta \sum_{j\notin\mathbf{I}_{g^1}} p(i,u,v,j)\, |S^2_n(j,u,v) - S(j,u,v)|, \end{aligned}$$

and by (22)

$$\begin{aligned} |S^2_n(i,u,v) - S(i,u,v)| &\le \beta \sum_{j\in\mathbf{I}_{g^1}} p(i,u,v,j)\, |S^2_n(j,u,v) - S(j,u,v)| \\ &\quad + 2\beta \sum_{j\notin\mathbf{I}_{g^1}} p(i,u,v,j)\, |S^2_n(j,u,v) - S(j,u,v)| \\ &\le \beta \sum_{j\in\mathbf{I}_{g^1}} p(i,u,v,j)\, |S^2_n(j,u,v) - S(j,u,v)| + 2\epsilon \frac{2M\beta}{1-\beta}. \end{aligned}$$

So continuing as in the proof of Theorem 4.1 we have

$$|S_n^2(j,u,v) - S(j,u,v)| \le 2M\frac{2\beta(1-\beta^k)\epsilon/(1-\beta)+\beta^k}{1-\beta} \tag{24}$$

for $n \ge m_k(\epsilon, \epsilon(J))$.

4.3. Approximation scheme III. The basic idea of the approximation scheme is to fix some stationary policies for both players and use them in all states except for a subset $\mathbf{I}_n$. The problem is then of determining the optimal mixed strategies for both players in the remaining set of states $\mathbf{I}_n$. We are interested in studying the asymptotic behavior of this approach as $\mathbf{I}_n \to \mathbf{I}$. Similar approaches were used in a framework of Markov decision processes (e.g., [2]), where $\mathbf{I}_n$ were assumed finite. We first fix some arbitrary policies $\hat{u} \in U$, $\hat{v} \in V$. We shall now use the framework of Theorem 2.1. Define

$$U_n = \{u \in U : u(i) = \hat{u}(i),\ \forall i \notin \mathbf{I}_n\}, \qquad V_n = \{v \in V : v(i) = \hat{v}(i),\ \forall i \notin \mathbf{I}_n\}.$$

Fix some $i \in \mathbf{I}$. The limit game is defined as $S(u,v) = S(i,u,v)$, where $S(i,u,v)$ is given in (5). For any $u \in U_n, v \in V_n$, define $S_n(u,v) = S(u,v)$ We set π_n^1 and π_n^2 to be the identity mappings and

$$\sigma_n^1(u)(i) = \begin{cases} u(i) & \text{if } i \in \mathbf{I}_n, \\ \hat{u}(i) & \text{if } i \notin \mathbf{I}_n; \end{cases} \qquad \sigma_n^2(v)(i) = \begin{cases} v(i) & \text{if } i \in \mathbf{I}_n, \\ \hat{v}(i) & \text{if } i \notin \mathbf{I}_n. \end{cases}$$

THEOREM 4.4. *Assume* (M_1) *and* ($\mathcal{B}2$), *and fix a state* i. *Then*
(i) *All statements of Theorem* 2.1 *hold for approximation scheme* III.
(ii) R_n *is the unique fixed point of the equation*

$$R_n(k) = \begin{cases} val\left[r(k,a,b) + \beta\sum_{j\in\mathbf{I}} p(k,a,b,j)R_n(j)\right], & k \in \mathbf{I}_n, \\ r(k,\hat{u},\hat{v}) + \beta\sum_{j\in\mathbf{I}} p(k,\hat{u},\hat{v},j)R_n(j), & k \notin \mathbf{I}_n. \end{cases} \tag{25}$$

(iii) *Optimal stationary policies* u_n *and* v_n *for both players are obtained by using at any state* $k \in \mathbf{I}_n$ *mixed strategies that achieve the value in* (25).

The proof of this theorem is similar to the proof of Theorem 4.1.

5. State approximations for the case of finite horizon. Consider the model in §3 with, however, a finite-horizon reward criterion instead of (5):

$$S^{[m]} = E_i^{u,v}\left[\sum_{t=0}^{m} \beta^t r(I_t, A_t, B_t)\right]. \tag{26}$$

It is well known that there exist optimal policies for both players within the class of Markov policies. The value R of $S^{[m]}$ is obtained by the recursion

$$R^{m+1} \stackrel{\text{def}}{=} 0, \tag{27}$$

$$R^k(i) \stackrel{\text{def}}{=} val\left[r(i,u,v) + \beta\sum_{j\in\mathbf{I}} p(i,u,v,j)R^{k+1}(j)\right], \qquad k = 0,\dots,m,$$

$$R \stackrel{\text{def}}{=} R^0.$$

Define $U[m] = (U^0[m], U^1[m], \ldots, U^m[m])$, $V[m] = (V^0[m], V^1[m], \ldots, V^m[m])$, where $U^k[m], V^k[m]$ are the set of mixed strategies which are optimal for the matrix game

$$\left[r(i,a,b) + \beta \sum_{j \in \mathbf{I}} p(i,a,b,j) R^{k+1}(j) \right]_{a,b}, \quad k = 0, \ldots, m, \tag{28}$$

for all $i \in \mathbf{I}$. Then, any Markov policies (u, v) such that $u_t \in U^t$, $v_t \in V^t$, $t = 0, \ldots, m$, are optimal for the stochastic game $S^{[m]}$.

In order to apply the results from §4 to the finite-horizon case we make the following observation. The finite-horizon model is equivalent to the following infinite-horizon model with enlarged state space:

- $\hat{\mathbf{I}} = \mathbf{I} \times \{0, \ldots, m\}$;
- $\hat{\mathbf{A}}_{(i,k)} = \mathbf{A}_i$, $\hat{\mathbf{B}}_{(i,k)} = \mathbf{B}_i$;
- $\hat{r}((i,k), a, b) = r(i, a, b)$;
- $\hat{p}((i,k), a, b, (j,l)) = \begin{cases} p(i,a,b,j) & \text{if } k+1 = l \leq m, \\ 0 & \text{otherwise;} \end{cases}$
- $\hat{\beta} = \beta$.

Define

$$\hat{S}(\hat{i}, \hat{u}, \hat{v}) = E_{\hat{i}}^{\hat{u},\hat{v}} \sum_{t=0}^{\infty} \beta^t r(\hat{I}_t, \hat{A}_t, \hat{B}_t).$$

There is a one-to-one correspondence between stationary policies in the new model and Markov policies in the original one; if $\hat{u}, \hat{v}$ are stationary in the new model, then the corresponding Markov policies in the original model are given by

$$u_t(\bullet|x) = \hat{u}(\bullet|(x,t)), \qquad v_t(\bullet|x) = \hat{v}(\bullet|(x,t)) \tag{29}$$

and vice versa. Moreover, we have

$$S^{[m]}(u, v) = \hat{S}(\hat{u}, \hat{v}).$$

Consequently, the state approximation schemes from the previous section also hold for the case of finite-horizon models. The computation of the (approximating) values and (almost) optimal policies can be done by using the above infinite-horizon model with enlarged state space and then applying (29). $\hat{\mathbf{I}}_n$ may be chosen, for example, as $\hat{\mathbf{I}}_n = \mathbf{I}_n \times \{0, \ldots, m\}$. Note that condition $(\mathcal{B}2)$ will hold for $\hat{\mathbf{I}}_n$ if it holds for $\mathbf{I}_n$.

6. Successive approximations. We study in this section several new aspects of successive approximations. The convergence of the value of successive approximation is already well known [19], [24], [27]. By applying Theorem 2.1, we establish in the following subsection the convergence of (almost) optimal policies. We then study the application of both state approximation and finite horizon approximation. Finally, we discuss the restriction of games with finite horizon to stationary policies.

6.1. Convergence of policies for successive approximations. An interesting application of the results in the previous subsections is the observation that successive approximations (or value iteration) can be viewed as a special case of state approximations. One can define game G_n such that $S_n = S^{[n]}$, where $S^{[n]}$ is given in (26), and consider the limit game $G = G_\infty$ where S is defined in (5). Let u_n^* and v_n^* be

a pair of optimal (or ϵ_n-optimal, where $\lim_{n\to\infty}\epsilon_n = 0$) Markov policies for G_n. Let u^* and v^* be any ϵ-optimal stationary (or Markov) policies for the infinite-horizon game G. If we use for both G_n and G, the equivalent infinite-horizon model with the enlarged state space defined in §5, then u_n^*, v_n^*, u^* and v^* all have an equivalent representation as stationary policies. The problem becomes one of approximating the state space $\mathbf{I}' = (\mathbf{I}\times\mathbb{N})$ by the subsets $\mathbf{I}'_n = (\mathbf{I}\times\{0,1,\dots,n\})$. Note that condition $(\mathcal{B}2)$ holds in this case since the probability of going from any state in $\mathbf{I}'_r$ to any state in $\mathbf{I}'_n$ is zero for $n > r+1$. Then using Theorems 2.1 and 4.1, we conclude with the following theorem.

THEOREM 6.1. *Assume that* $(\mathbf{M_1})$ *holds. Then*

(1) $\lim_{n\to\infty} R_n = R$.

(2) *For any* $\epsilon' > \epsilon$, *there exists* N *such that* u_n^* *(resp.,* v_n^**) is* ϵ'*-optimal for the infinite-horizon game for all* $n \geq N$.

(3) *Let* $\bar{u} \in U$ *(resp.,* $\bar{v} \in V$*) be a limit point of* u_n^* *(resp.,* v_n^**). Then* $\bar{u}$ *(resp.,* $\bar{v}$*) is* ϵ*-optimal for the limit game.*

(4) *For all* $\epsilon' > \epsilon$, *there exists* $N(\epsilon')$ *such that* u^* *is* ϵ'*-optimal for the* n*th approximating game for all* $n \geq N(\epsilon')$.

6.2. Successive approximation and finite state approximation. We use the above approach to combine state approximations with finite horizon reward criterion. Such a combination may be specially useful for computational purposes, where $\mathbf{I}_n$ can be chosen to be finite. We can now compute R_n and (Markov) policies which are optimal for G_n, using approximating schemes introduced in the previous sections, in order to approximate the optimal value R and an almost optimal strategy for the original limit game G. Let again $\mathbf{I}_n \subset \mathbf{I}$ be an increasing sequence of sets of states converging to $\mathbf{I}$. One can repeat the construction of a model with enlarged state space (that includes both the original state space and the time) so that the state space for the nth game G_n is $\hat{\mathbf{I}}_n = (\mathbf{I}_n \times \{0,1,\dots,n\})$ and for the limit game G is $\hat{\mathbf{I}} = (\mathbf{I}\times\mathbb{N})$. This would establish the correctness of approximations based on value iteration for a problem with truncated state space. For example, if we adapt the first approach in §4, we get the approximating values R_n and Markov policies by performing the following iterations:

$$R_n^{n+1}(j) \stackrel{\text{def}}{=} 0,$$

$$R_n^k(j) \stackrel{\text{def}}{=} \begin{cases} val\left[r(i,a,b) + \beta\sum_{j\in\mathbf{I}_n} p(i,a,b,j)R^{k+1}(j)\right] & \text{if } i \in \mathbf{I}_n, \\ 0 & \text{if } i \notin \mathbf{I}_n, \end{cases} \qquad k = 0,\dots,n,$$

$$R_n \stackrel{\text{def}}{=} R_n^0.$$

Define $U_n = (U_n^0, U_n^1, \dots, U_n^n)$, $V_n[m] = (V_n^0, V_n^1, \dots, V_n^n)$, where U_n^k, V_n^k are the set of mixed strategies which are optimal for the game

$$val\left[r(i,a,b) + \beta\sum_{j\in\mathbf{I}_n} p(i,a,b,j)R_n^{k+1}(j)\right]_{a,b}, \qquad k = 0,\dots,m, \tag{30}$$

for any $i \in \mathbf{I}_n$. Then any Markov policies (u,v) such that $u_t \in U_n^t$, $v_t \in V_n^t$, $t = 0,\dots,m$, are optimal for the stochastic game G_n.

6.3. Finite horizon and stationary policies. For simplicity of implementation, one may be interested in restricting to the class of stationary policies in a

stochastic game rather than using Markovian policies (or others). It is well known, however, that finite-horizon games do not have a value within the class of stationary policies. However, it is immediately seen that the conditions $(\mathcal{A}1)$–$(\mathcal{A}4)$ hold when there is restriction to stationary policies, and thus we conclude from Theorem 2.1 that the optimal stationary policies for both players converge (in the sense of Theorem 2.1 (3)) to the strongly optimal policy of the infinite-horizon game as the horizon goes to infinity. Moreover, the lower and upper values converge to the value of the infinite-horizon game.

7. Convergence of the discount factor and immediate reward. We establish in this section the robustness of values and optimal policies with respect to the discount factor and immediate reward. This may be of importance in case that these parameters are not known precisely. One can similarly establish robustness for the random time-varying discount factor and immediate reward. We consider a horizon m which is the same for both the limit and the approximating game and which may be either infinite or finite.

We consider a sequence of stochastic games G_n, $n = 0, 1, 2, \ldots$, where the quantities defining each one of them are as in §3, Eq. (5), if m is infinite or as in §5, Eq. (26), if m is finite. However, the immediate reward and discount factor are replaced by $\beta_n = \beta + \delta_n$, $r_n = r + \rho_n$, where δ_n and ρ_n converge to zero as $n \to \infty$ uniformly in the states and actions. Denote by $S_n(i, u, v)$ the reward for game G_n (as defined either in (5) or in (26)). Then

$$\begin{aligned} &|S(i,u,v) - S_n(i,u,v)| \\ &\leq E_i^{u,v} \sum_{t=0}^{m} \left(|[\beta^t - (\beta+\delta_n)^t]|M + |(\beta+\delta_n)^t \rho_n(I_t, A_t, B_t)| \right), \end{aligned}$$

and we obtain convergence to zero uniformly over all Markovian policies u and v (to which we may restrict ourselves, without loss of generality, as in §5). This implies conditions $(\mathcal{A}1)$ and $(\mathcal{A}2)$, and hence, by Remark 3.1, we see that all statements of Theorem 2.1 hold. This establishes the continuity of the value of the stochastic game as a function of the discount factor β in the open interval $\beta \in (0, 1)$ and as a function of the immediate reward. Moreover, it establishes the convergence of (almost) optimal policies (in the sense of Theorem 2.1).

An especially interesting case is the asymptotics of stochastic games as $\beta \to 1$. We restrict for simplicity to the case of finite state and action spaces. The asymptotic behavior of the value of the game was studied by Bewley and Kohlberg [8]. In fact they establish the convergence of $(1 - \beta)R_\beta(i)$ to the value $R_{average}$ of the expected long run time-average game (where $R_\beta(i)$ is the value of the game with discount factor β and initial state i).

When trying to apply the approximating Theorem 2.1 to the limit as $\beta \to 1$, we are faced with the following problems:

(i) The limit game does not have a value among the stationary (nor even the Markov) policies (see the "big match" by Blackwell and Ferguson [9]).

(ii) The value of the limit game (with the expected average reward) is in general not continuous in the policies (and thus assumptions $(\mathcal{A}3)$ and $(\mathcal{A}4)$ do not hold in general). This is the case even for a single controller, for which it is known that the value may exhibit discontinuity in the parameters (see Gaitsgory and Pervozvanskii [13, p. 407]).

However, both problems are avoided in the case when we restrict ourselves either to games with perfect information or to irreducible games (see Gillette [14]). Games with perfect information (resp., irreducible games) have a saddle point within the class of stationary deterministic policies (resp., stationary mixed policies), and ($\mathcal{A}1$) and ($\mathcal{A}2$) hold; see [14]. Within these classes of policies, ($\mathcal{A}3$) and ($\mathcal{A}4$) also hold; indeed, for the perfect information case this follows from the fact that there is only a finite number of stationary deterministic policies. For the irreducible case, this follows, e.g., from [1].

8. Applications. We present in this section a few problems that motivated our research on approximations in stochastic games. As mentioned in the introduction, many discretization schemes of differential games yield dynamic programming that can be interpreted as representing some stochastic game. In some pursuit evasion games, such as the game of the two cars [22], an additional finite state approximation is then required. The calculations in [22] was done following scheme I (introduced in §4.1). The state transition in the discretized model satisfied property ($\mathcal{B}3$), and, in fact, each state had at most four neighbors (i.e., four states reachable in one transition). This feature motivates the use of Theorem 4.3 for such applications, which not only establishes the convergence but also gives the rate of convergence (or, more precisely, enables computation of n for obtaining any required precision).

Another application of the theory we developed in previous sections are stochastic games appearing in queueing systems. Such problems may serve as models for situations of conflicts between users in telecommunication systems or for worst-case control situations in the presence of some unknown disturbance (in production systems or again in telecommunications applications). An interesting feature in the control of queueing networks is that, often, infinite queues are easier to handle than finite queues, as some boundary problems are avoided. Moreover, the optimal policies for infinite-horizon problems, being stationary, are easier to implement than those for finite-horizon problems. In real applications, however, queues are always finite; moreover, one is often interested in finite-horizon problems (e.g., controlling manufacturing during working hours, etc.). Our results may thus be applied to obtain almost optimal policies for these cases. Here are some examples:

(i) Altman considered in [3] a stochastic game with an infinite state space in order to solve a flow control problem with an infinite buffer. The solution of the problem with a finite buffer [4] seems more involved and was obtained only under an important restriction on the actions of the flow controller (namely, it had to contain an action that corresponds to rejection of arriving customers).

(ii) Altman and Koole [5] solved a game where one or more servers has to be assigned to customers of different classes. In a telecommunication context, the different classes may represent different traffic types, such as voice, video, and data, and the servers may represent a channel through which the traffic has to be transmitted. A controller has to decide a customer of which class will be served next (which traffic will have access to the channel). The input traffic was assumed to be controlled as well, e.g., it may have been the output of some dynamic routing mechanism or dynamic flow control. The problem was posed as a zero-sum stochastic game between the service controller and "nature" which represented the unknown input control mechanism. Simple structural results were obtained for the case of infinite queues. In the case of finite queues, the structure of optimal policies is unknown, even in the case of uncontrolled input. By applying our second state approximation scheme, it follows that the policies obtained for the problem of infinite queues are almost optimal for

the case of finite queues which are large enough.

9. Further generalizations. Although we considered in this paper bounded rewards, it is well known that different sets of conditions exist for which problems with unbounded reward can be transformed into ones with bounded reward. Such transformations have been used in the past for finite state approximations of Markov decision processes; see White [30]. The generalization of such conditions to games are straightforward (see, e.g., Wessels [28]). (For examples of stochastic games with unbounded costs, see [3], [5].)

There are many other useful directions where the general approximation theorems are applicable, on which we continue our investigation. Among these are

(i) Differential games in which a standard problem is to discretize both space and time. Several works have been done in this direction; see [6], [7], [21], [22], [25], where the convergence (and rate of convergence; see [21]) of the values of the approximating games have been established. However, little is known about the convergence of policies. Theorem 2.1 seems to be a suitable tool for approaching these issues.

(ii) Discretization of stochastic games with general state and action spaces. Some results were obtained in the case of a single controller; see §6 of Hernandez-Lerma [15] and references therein. Further results for stochastic games on the convergence of the value and some results on convergence of policies were obtained by Whitt [31] and then generalized to N-person games in [32] and by Nowak [20].

REFERENCES

[1] E. ALTMAN, *Asymptotic properties of constrained Markov decision processes*, Z. Oper. Res., 37 (1993), pp. 151–170.

[2] ———, *Denumerable constrained Markov decision problems and finite approximations*, Math. Oper. Res., 19 (1994), pp. 169–191.

[3] ———, *Flow control using the theory of zero-sum Markov games*, IEEE Trans. Automatic Control, 39 (1994), pp. 814–818.

[4] ———, *Monotonicity of optimal policies in a zero sum game: A flow control model*, in Advances in Dynamic Games and Applications, T. Basar and A. Haurie, eds., Birkhäuser, Boston, 1964, pp. 269–286.

[5] E. ALTMAN AND G. KOOLE, *Stochastic scheduling games with Markov decision arrival processes*, Comput. Math. Appl., 9 (1993), pp. 141–148.

[6] M. BARDI, M. FALCONE, AND P. SORAVIA, *Fully discrete schemes for the value function of pursuit-evasion games*, Annals of the International Society of Dynamic Games, 1 (1993), pp. 89–105.

[7] P. BERNHARD AND J. SHINAR, *On finite approximation of a game solution with mixed strategies*, Appl. Math. Lett., 3 (1990), pp. 1–4.

[8] T. BEWLEY AND E. KOHLBERG, *The asymptotic theory of stochastic games*, Math. Oper. Res., 1 (1976), pp. 197–208.

[9] D. BLACKWELL AND T. FERGUSON, *The big match*, Annals of Mathematical Statistics, 39 (1968), pp. 159–163.

[10] V. S. BORKAR, *A convex analytic approach to Markov decision processes*, Probab. Theory Related Fields, 78 (1988), pp. 583–602.

[11] R. CAVAZOS-CADENA, *Finite-state approximations for denumerable state discounted Markov decision processes*, J. Appl. Math. Optim., 14 (1986), pp. 27–47.

[12] A. FEDERGRUEN, *On N-person stochastic games with denumerable state space*, Adv. in Appl. Prob., 10 (1978), pp. 452–471.

[13] V. A. GAITSGORY AND A. A. PERVOZVANSKII, *Perturbation theory for mathematical programming problems*, J. Optim. Theory Appl., 49 (1986), pp. 389–410.

[14] D. GILLETTE, *Stochastic games with zero stop probabilities*, Ann. Math. Studies 39, M. Dresher, A. W. Tucker, P. Wolfe, eds., Princeton University Press, Princeton, 1957, pp. 179–187.

[15] O. HERNANDEZ-LERMA, *Finite state approximations for denumerable multidimensional-state discounted Markov decision processes*, J. Math. Anal. Appl., 113 (1986), pp. 382–389.

[16] O. HERNANDEZ-LERMA, *Adaptive Control of Markov Processes*, Springer-Verlag, New York, 1989.
[17] A. HORDIJK, O. VRIEZE, AND G. WANDROOIJ, *Semi-Markov strategies in stochastic games*, Mathematical Centre Report BW 68/76, 1976.
[18] A. S. NOWAK, *On zero-sum stochastic games with general state space* I, Probab. Math. Stat., IV (1984), pp. 13-32.
[19] ———, *Approximation Theorems for zero-sum nonstationary stochastic games*, Proc. of the American Math. Soc., 92 (1984), pp. 418–424.
[20] ———, *Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space*, J. Optim. Theory Appl., 45 (1985), pp. 592–602.
[21] O. POURTALLIER AND M. TIDBALL, *A discrete scheme of pursuit-evasion games*, in preparation.
[22] O. POURTALLIER AND B. TOLWINSKY, *Discretization of Isaacs equation: A convergence result*, (1994), in preparation.
[23] H. ROYDEN, *Real Analysis*, Macmillan, New York, 1963.
[24] L. S. SHAPLEY, *Stochastic games*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 1095–1100.
[25] M. TIDBALL AND R. L. V. GONZÁLEZ, *Zero sum differential games with stopping times. Some results about its numerical resolution*, Annals of the International Society of Dynamic Games, 1 (1993), pp. 106–124.
[26] L. C. THOMAS AND D. STENGOS, *Finite state approximation algorithms for average cost denumerable state Markov decision processes*, OR Spektrum, 7 (1985), pp. 27–37.
[27] J. VAN DER WAL, *Successive approximations for average reward Markov games*, Internat. J. Game Theory, 9 (1980), pp. 13–24.
[28] J. WESSELS, *Markov Games with unbounded rewards*, Dynamische Optimierung, M. Schäl, ed., Bonner Mathematische Schriften, 98, Univ. Bonn, Bonn (1977).
[29] D. J. WHITE, *Finite state approximations for denumerable state infinite horizon discounted Markov decision processes*, J. Math. Anal. Appl., 74 (1980), pp. 292–295.
[30] ———, *Finite state approximations for denumerable state infinite horizon discounted Markov decision processes with unbounded rewards*, J. Math. Anal. Appl., 86 (1982), pp. 292–306.
[31] W. WHITT, *Approximations of Dynamic Programs*, I, Math. Oper. Res., 3 (1978), pp. 231–243.
[32] ———, *Representation and approximation of noncooperative sequential games*, SIAM J. Control Optim., 18 (1980), pp. 33–43.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 329–364, January 1996

015

ON AN INVESTMENT-CONSUMPTION MODEL WITH TRANSACTION COSTS*

MARIANNE AKIAN†, JOSÉ LUIS MENALDI‡, AND AGNÈS SULEM§

Abstract. This paper considers the optimal consumption and investment policy for an investor who has available one bank account paying a fixed interest rate and n risky assets whose prices are log-normal diffusions. We suppose that transactions between the assets incur a cost proportional to the size of the transaction. The problem is to maximize the total utility of consumption. Dynamic programming leads to a variational inequality for the value function. Existence and uniqueness of a viscosity solution are proved. The variational inequality is solved by using a numerical algorithm based on policies, iterations, and multigrid methods. Numerical results are displayed for $n = 1$ and $n = 2$.

Key words. portfolio selection, transaction costs, viscosity solution, variational inequality, multigrid methods

AMS subject classifications. 90A09, 93E20, 49L20, 49L25, 65N55, 35R45

1. Introduction.

This paper concerns the theoretical and numerical study of a portfolio selection problem. Consider an investor who has available one riskless bank account paying a fixed rate of interest r and n risky assets modeled by log-normal diffusions with expected rates of return $\alpha_i > r$ and rates of return variation σ_i^2. The investor consumes at rate $c(t)$ from the bank account. Any movement of money between the assets incurs a transaction cost proportional to the size of the transaction, paid from the bank account. The investor is allowed to have a short position in one of the holdings, but his position vector must remain in the closed solvency region $\mathcal{S}$ defined as the set of positions for which the net wealth is nonnegative. The investor's objective is to maximize over an infinite horizon the expected discounted utility of consumption with a HARA (hyperbolic absolute risk aversion)–type utility function.

This problem was formulated for $n = 1$ by Magill and Constantinides [21], who conjectured that the no-transaction region is a cone in the two-dimensional space of position vectors. This fact was proved in a discrete-time setting by Constantinides [8], who proposed an approximate solution based on some assumptions on the consumption process. Davis and Norman proved, in continuous time and without this restriction, that the optimal strategy confines indeed the investor's portfolio to a wedge-shaped region in the portfolio plane [10]. An analysis of the optimal strategy, together with regularity results for the value function, can be found in Fleming and Soner [13, Chap. 8.7] and Shreve and Soner [28]. Taksar, Klass, and Assaf [31] consider a model without consumption and study the problem of maximizing the long-run average growth of wealth. A deterministic model is solved by Shreve, Soner, and Xu [29] with a general utility function which is not necessarily a HARA-type function. A stochastic model driven by a finite-state Markov chain rather than a Brownian motion and with a general but bounded utility function has been investigated in Zariphopoulou [32]. She supposes that the amount of money allocated in

* Received by the editors April 12, 1993; accepted for publication (in revised form) October 5, 1994.

† INRIA, Domaine de Voluceau Rocquencourt, B.P. 105, 78153 Le Chesnay cedex, France (marianne.akian@inria.fr).

‡ Department of Mathematics, Wayne State University, Detroit, MI 48202 (jlm@math.wayne.edu). The research of this author was supported in part by NSF grant DMS-9101360.,

§ INRIA, Domaine de Voluceau Rocquencourt, B.P. 105, 78153 Le Chesnay cedex, France (agnes.sulem@inria.fr).

the assets must remain nonnegative and shows that the value function is the unique constrained viscosity solution of a system of variational inequalities with gradient constraints. Fitzpatrick and Fleming [12] study numerical methods for the optimal investment-consumption model with possible borrowing. They examine a Markov chain discretization of the original continuous problem similar to Kushner's numerical schemes [18]. The convergence arguments rely on viscosity solution techniques.

We consider here Davis and Norman's model [10] in the case where more than one risky asset is allowed. We restrict to power utility functions of the form $\frac{c^\gamma}{\gamma}$ with $0 < \gamma < 1$.

The purpose of the paper is to prove an existence and uniqueness result for the dynamic programming equation associated with this problem and then solve this equation by using an efficient numerical method, the convergence of which is ensured by the uniqueness result.

The mathematical formulation of the problem is given in §2. In §3, we prove that the value function is the unique viscosity solution of a variational inequality. Since the utility and the drift functions are not bounded, uniqueness is not derived from classical results. For the numerical study, an adequate change of variables performed in §4 reduces the dimension of the problem. Then, in §5, the variational inequality is discretized by finite-difference schemes and solved by using an algorithm based on the "Howard algorithm" (policy iteration) and the multigrid method. Numerical results are presented in §6 in the case of one bank account and one or two risky asset(s). They provide the optimal strategy and indicate the shape of the transaction and no-transaction regions. Finally, in §7, a theoretical study of the optimal strategy is done by using properties of the variational inequality; this analysis corroborates the numerical results.

2. Formulation of the problem. Let $(\Omega, \mathcal{F}, P)$ be a fixed complete probability space and $(\mathcal{F}_t)_{t\geq 0}$, a given filtration. We denote by $s_0(t)$ (resp., $s_i(t)$ for $i = 1, \ldots, n$) the amount of money in the bank account (resp., in the ith risky asset) at time t and refer by $s(t) = (s_i(t))_{i=0,\ldots,n}$ the investor position at time t. We suppose that the evolution equations of the investor holdings are

$$(1) \qquad \begin{cases} ds_0(t) = (rs_0(t) - c(t))dt + \displaystyle\sum_{i=1}^{n}(-(1+\lambda_i)d\mathcal{L}_i(t) + (1-\mu_i)d\mathcal{M}_i(t)), \\ ds_i(t) = \alpha_i s_i(t)dt + \sigma_i s_i(t)dW_i(t) + d\mathcal{L}_i(t) - d\mathcal{M}_i(t), \quad i = 1, \ldots, n, \end{cases}$$

with initial values

$$(2) \qquad s_i(0^-) = x_i, \quad i = 0, \ldots, n,$$

where $W_i(t)$, $i = 1, \ldots, n$, are independent Wiener processes, $\mathcal{L}_i(t)$ and $\mathcal{M}_i(t)$ represent cumulative purchase and sale of stock i on $[0, t]$, respectively, and $s(t^-)$ denotes the left-hand limit of the process s at time t. The coefficients λ_i and μ_i represent the proportional transaction costs.

A policy for investment and consumption is a set $(c(t), (\mathcal{L}_i(t), \mathcal{M}_i(t))_{i=1,\ldots,n})$ of adapted processes such that

1. $c(t,\omega) \geq 0$, $\int_0^t c(s,\omega)ds < \infty$ for (t,ω) a.e.,
2. $\mathcal{L}_i(t)$, $\mathcal{M}_i(t)$ are right-continuous and nondecreasing and $\mathcal{L}_i(0^-) = \mathcal{M}_i(0^-) = 0$.

The process $s(t)$ is thus right continuous with the left-hand limit and equations (1) and (2) are equivalent to

$$\begin{cases} s_0(t) &= x_0 + \displaystyle\int_0^t (rs_0(\theta) - c(\theta))d\theta + \sum_{i=1}^n (-(1+\lambda_i)\mathcal{L}_i(t) + (1-\mu_i)\mathcal{M}_i(t)), \\ s_i(t) &= x_i + \displaystyle\int_0^t \alpha_i s_i(\theta)d\theta + \int_0^t \sigma_i s_i(\theta)dW_i(\theta) + \mathcal{L}_i(t) - \mathcal{M}_i(t), \quad i = 1, \ldots, n, \end{cases}$$

for $t \geq 0$.

We define the solvency region as

$$\mathcal{S} = \{x = (x_0, x_1, \ldots, x_n) \in \mathbb{R}^{n+1}, \quad \mathcal{W}(x) \geq 0\},$$

where

$$\mathcal{W}(x) = x_0 + \sum_{i=1}^n \min((1-\mu_i)x_i, (1+\lambda_i)x_i) \tag{3}$$

represents the net wealth, that is, the amount of money in the bank account after performance of the transactions that bring the holdings in the risky assets to zero.

Suppose that the investor is given an initial endowment x in $\mathcal{S}$. A policy is admissible if the bankruptcy time $\overline{\tau}$ defined as

$$\overline{\tau} = \inf\{t \geq 0, s(t) \notin \mathcal{S}\} \tag{4}$$

is infinite. We denote by $\mathcal{U}(x)$ the set of admissible policies. The investor's objective is to maximize over all policies $\mathcal{P}$ in $\mathcal{U}(x)$ the discounted utility of consumption

$$J_x(\mathcal{P}) = E_x \int_0^\infty e^{-\delta t} u(c(t))dt, \tag{5}$$

where E_x denotes expectation given that the initial endowment x, δ is a positive discount factor and $u(c)$ is a utility function defined by

$$u(c) = \frac{c^\gamma}{\gamma}, \quad 0 < \gamma < 1. \tag{6}$$

We define the value function V as

$$V(x) = \sup_{\mathcal{P} \in \mathcal{U}(x)} J_x(\mathcal{P}). \tag{7}$$

We are facing a singular control problem. We refer to Menaldi and Robin [23] and Chow, Menaldi, and Robin [7] for various treatments of singular stochastic control problems.

Remark 2.1. When the process $s(t)$ reaches the boundary $\partial\mathcal{S}$ at time t, i.e., $s(t^-) \in \partial\mathcal{S}$, the only admissible policy is to jump immediately to the origin and remain there with a null consumption (see Shreve, Soner, and Xu [29]). Consequently, if the initial endowment x is on the boundary, then $V(x) = 0$.

Remark 2.2. Let τ denote the exit time of the interior of $\mathcal{S}$, defined as

$$\tau = \inf\left\{t \geq 0, s(t) \notin \mathring{\mathcal{S}}\right\}. \tag{8}$$

For all admissible policies $\mathcal{P}$, we have

$$J_x(\mathcal{P}) = E_x \int_0^\tau e^{-\delta t} u(c(t))dt. \tag{9}$$

On the other hand, for any policy $\mathcal{P}$, we can construct an admissible policy which coincides with $\mathcal{P}$ until time τ (such that the process $s(t)$ jumps to the origin at time τ). The value function can then be rewritten as

$$V(x) = \sup_{\mathcal{P}\in\mathcal{U}} E_x \int_0^\tau e^{-\delta t} u(c(t))dt, \tag{10}$$

where $\mathcal{U}$ is the set of all policies.

We make the assumptions

$$\delta > \gamma\left(r + \frac{1}{2(1-\gamma)}\sum_{i=1}^{n}\left(\frac{\alpha_i - r}{\sigma_i}\right)^2\right), \tag{A.1}$$

$$0 \le \mu_i < 1, \quad \lambda_i \ge 0, \quad \lambda_i + \mu_i > 0 \quad \forall i = 1, \dots, n. \tag{A.2}$$

Remark 2.3. When the transaction costs are equal to zero (Merton's problem), the value function V is finite iff Assumption (A.1) is satisfied (see Davis and Norman [10] for $n = 1$, Karatzas et al. [17], and §7 below).

3. The variational inequality. We state the main theorem.

THEOREM 3.1. *Under Assumptions* (A.1) *and* (A.2),

(i) *the value function V defined in* (7) *or* (10) *is γ-Hölder continuous and concave in $\mathcal{S}$ and nondecreasing with respect to x_i for $i = 0, \dots, n$.*

(ii) *V is the unique viscosity solution of the variational inequality (VI):*

$$\max\left\{AV + u^*\left(\frac{\partial V}{\partial x_0}\right), \max_{1\le i\le n} L_i V, \max_{1\le i\le n} M_i V\right\} = 0 \quad \text{in } \mathring{\mathcal{S}}, \tag{11}$$

$$V = 0 \quad \text{on } \partial\mathcal{S}, \tag{12}$$

where

$$AV = \frac{1}{2}\sum_{i=1}^{n}\sigma_i^2 x_i^2 \frac{\partial^2 V}{\partial {x_i}^2} + \sum_{i=1}^{n}\alpha_i x_i \frac{\partial V}{\partial x_i} + r x_0 \frac{\partial V}{\partial x_0} - \delta V, \tag{13}$$

$$L_i V = -(1+\lambda_i)\frac{\partial V}{\partial x_0} + \frac{\partial V}{\partial x_i}, \tag{14}$$

$$M_i V = (1-\mu_i)\frac{\partial V}{\partial x_0} - \frac{\partial V}{\partial x_i}, \tag{15}$$

and u^* is the convex Legendre transform of u defined by

$$\begin{aligned} u^*(p) &= \max_{c\ge 0}(-cp + u(c)) \\ &= \left(\frac{1}{\gamma} - 1\right) p^{\frac{\gamma}{\gamma-1}}. \end{aligned} \tag{16}$$

The solvency region $\mathcal{S}$ is divided as follows:

$$B_i = \{x \in \mathcal{S},\ L_iV(x) = 0\}, \tag{17}$$

$$S_i = \{x \in \mathcal{S},\ M_iV(x) = 0\}, \tag{18}$$

$$NT_i = \mathcal{S} \setminus (B_i \cup S_i), \tag{19}$$

$$NT = \bigcap_{i=1}^{n} NT_i. \tag{20}$$

NT is the no-transaction region. Outside NT, an instantaneous transaction brings the position to the boundary of NT: buy stock i in B_i, sell stock i in S_i. After the initial transaction, the agent position remains in

$$\overline{NT} = \left\{ x \in \mathcal{S}, AV + u^*\left(\frac{\partial V}{\partial x_0}\right) = 0 \right\},$$

and further transactions occur only at the boundary (see [10]).

We shall first recall the definition of viscosity solutions and then prove points (i) and (ii) of Theorem 3.1 in §3.2 and §3.3, respectively.

3.1. Viscosity solutions of nonlinear elliptic equations. For simplicity, we restrict ourselves to equations with Dirichlet boundary conditions. Consider fully nonlinear elliptic equations of the form

$$F(D^2v, Dv, v, x) = 0 \quad \text{in } \mathcal{O}, \tag{21}$$

$$v = 0 \quad \text{on } \partial\mathcal{O}, \tag{22}$$

where F is a given continuous function in $S^N \times \mathbb{R}^N \times \mathbb{R} \times \mathcal{O}$, S^N is the space of symmetric $N \times N$ matrices, $\mathcal{O}$ is an open domain of $\mathbb{R}^N$, and the ellipticity of (21) is expressed by

$$F(A, p, v, x) \geq F(B, p, v, x) \quad \text{if } A \geq B,\ A, B \in S^N,\ p \in \mathbb{R}^N,\ v \in \mathbb{R},\ x \in \mathcal{O}. \tag{23}$$

A special case of (21) is given by

$$F(X, p, v, x) = \max_{\eta \in U} \left\{ \sum_{i,j=1}^{N} a_{ij}(x, \eta) X_{ij} + \sum_{i=1}^{N} b_i(x, \eta) p_i - \beta(x, \eta) v + u(x, \eta) \right\}, \tag{24}$$

where (23) is satisfied when the matrix $(a_{ij}(x, \eta))_{i,j}$ is symmetric nonnegative in $\mathcal{O} \times U$.

Bellman equations are clearly equations of this type, whereas variational inequalities like (11)–(12) can also be formulated in this form by using an additive discrete control which selects the equation which satisfies the maximum.

DEFINITION 3.2. *Let $v \in C(\overline{\mathcal{O}})$. Then v is a viscosity solution of* (21)–(22) *if the following relations hold, together with* (22):

$$F(X, p, v(x), x) \geq 0 \quad \forall (p, X) \in J^{2,+}v(x),\ \forall x \in \mathcal{O}, \tag{25}$$

$$F(X, p, v(x), x) \leq 0 \quad \forall (p, X) \in J^{2,-}v(x),\ \forall x \in \mathcal{O}, \tag{26}$$

where $J^{2,+}$ and $J^{2,-}$ are the second-order "superjets" defined by

$$J^{2,+}v(x) = \Big\{ (p, X) \in \mathbb{R}^N \times S^N,$$

$$\limsup_{\substack{y \to x \\ y \in \mathcal{O}}} \left[v(y) - v(x) - (p, y - x) - \frac{1}{2}(X(y - x), y - x) \right] |y - x|^{-2} \leq 0 \Big\}$$

and

$$J^{2,-}v(x) = \Big\{(p,X) \in \mathbb{R}^N \times S^N,$$
$$\liminf_{\substack{y \to x \\ y \in \mathcal{O}}} \Big[v(y) - v(x) - (p, y-x) - \frac{1}{2}(X(y-x), y-x)\Big] |y-x|^{-2} \geq 0\Big\}.$$

A viscosity subsolution (resp., supersolution) of (21) is similarly defined as an upper semicontinuous function satisfying (25) (resp., a lower semicontinuous function satisfying (26)) (see Crandall, Ishii, and Lions [9]).

3.2. Properties of the value function.

PROPOSITION 3.3. *The value function V is concave in $\mathcal{S}$.*

Proof. The dynamic (1) is linear, and the solvency region $\mathcal{S}$ is convex. Hence, for any θ in $[0,1]$, x and x' in $\mathcal{S}$, $\mathcal{P}$ in $\mathcal{U}(x)$, and $\mathcal{P}'$ in $\mathcal{U}(x')$, we have $\theta\mathcal{P}+(1-\theta)\mathcal{P}' \in \mathcal{U}(y)$ for $y = \theta x + (1-\theta)x'$ and

$$V(y) \geq J_y(\theta\mathcal{P} + (1-\theta)\mathcal{P}') = E\int_0^{+\infty} e^{-\delta t} u(\theta c(t) + (1-\theta)c'(t))dt.$$

Since u is concave we infer

$$V(y) \geq \theta J_x(\mathcal{P}) + (1-\theta) J_{x'}(\mathcal{P}').$$

Taking now the supremum over all $\mathcal{P}$ and $\mathcal{P}'$, we obtain that V is concave. □

As a consequence, V is locally Lipschitz continuous in $\overset{\circ}{\mathcal{S}}$. The continuity of V at the boundary is a consequence of the Proposition 3.5 below. First let us state the following lemma.

LEMMA 3.4. *Suppose* (A.1) *holds. Then there exists a positive constant a such that the functions*

$$(27)\quad \varphi_\nu(x) = a\left(x_0 + \sum_{i=1}^n (1-\nu_i)x_i\right)^\gamma \quad \text{with } \nu = (\nu_1, \ldots, \nu_n),\ \nu_i = -\lambda_i \text{ or } \mu_i,$$

are classical supersolutions of equation (11). *Consequently, the function*

$$(28)\qquad \varphi(x) = a\left(x_0 + \sum_{i=1}^n \min((1-\mu_i)x_i, (1+\lambda_i)x_i)\right)^\gamma$$

is a viscosity supersolution of equation (11) *such that $\varphi = 0$ on $\partial\mathcal{S}$.*

Proof. Denote

$$(29)\qquad \mathcal{W}_\nu(x) = x_0 + \sum_{i=1}^n (1-\nu_i)x_i.$$

Then

$$(30)\qquad \varphi_\nu(x) = a\mathcal{W}_\nu(x)^\gamma$$

and we have

$$(31)\qquad L_i\varphi_\nu(x) = -(\lambda_i + \nu_i)a\gamma\mathcal{W}_\nu(x)^{\gamma-1} \leq 0,$$

$$M_i\varphi_\nu(x) = (\mu_i - \nu_i)a\gamma\mathcal{W}_\nu(x)^{\gamma-1} \le 0, \tag{32}$$
$$A\varphi_\nu(x) = a\gamma\mathcal{W}_\nu(x)^{\gamma-2}G(x)$$

with

$$G(x) = \frac{1}{2}\sum_{i=1}^{n}\sigma_i^2 x_i^2(1-\nu_i)^2(\gamma-1) + \mathcal{W}_\nu(x)\left(\sum_{i=1}^{n}(\alpha_i - r)x_i(1-\nu_i)\right) + \mathcal{W}_\nu(x)^2\left(r - \frac{\delta}{\gamma}\right).$$

From Assumption (A.1), there exists $\eta > 0$ such that

$$\frac{\delta}{\gamma} - r - \eta \ge \frac{1}{2(1-\gamma)}\sum_{i=1}^{n}\left(\frac{\alpha_i - r}{\sigma_i}\right)^2.$$

This implies

$$G(x) \le -\eta\mathcal{W}_\nu(x)^2$$

and

$$A\varphi_\nu(x) \le -a\gamma\eta\mathcal{W}_\nu(x)^\gamma = -\gamma\eta\varphi_\nu(x). \tag{33}$$

Moreover

$$u^*\left(\frac{\partial\varphi_\nu}{\partial x_0}(x)\right) = u^*(a\gamma\mathcal{W}_\nu(x)^{\gamma-1}) = \left(\frac{1}{\gamma} - 1\right)(a\gamma)^{\frac{\gamma}{\gamma-1}}\mathcal{W}_\nu(x)^\gamma = (1-\gamma)(a\gamma)^{\frac{1}{\gamma-1}}\varphi_\nu(x).$$

The constant a can then be chosen such that

$$A\varphi_\nu + u^*\left(\frac{\partial\varphi_\nu}{\partial x_0}\right) \le 0 \quad \text{in } \mathring{\mathcal{S}}. \tag{34}$$

Since (31), (32), and (34) hold and $\varphi_\nu \ge 0$ on $\partial\mathcal{S}$, φ_ν is a classical supersolution of (11) (continuous in $\mathcal{S}$ and twice continuously differentiable in $\mathring{\mathcal{S}}$).

Now, φ can be rewritten as

$$\varphi = \min_\nu \varphi_\nu.$$

Consequently, φ is a viscosity supersolution of (11) as the minimum of continuous supersolutions and clearly vanishes on $\partial\mathcal{S}$. □

PROPOSITION 3.5. *Suppose* (A.1) *holds. The value function* V *satisfies*

$$0 \le V(x) \le \varphi(x) \qquad \forall x \in \mathcal{S}, \tag{35}$$

where φ *is the supersolution defined by* (28). *Consequently,* V *is continuous in* $\mathcal{S}$.

Proof. Consider $x \in \mathcal{S}$ and $\mathcal{P} \in \mathcal{U}$ and denote by τ the first exit time of $\mathring{\mathcal{S}}$ of the process $s(t)$ defined by (1) with $s(0^-) = x$. The function φ_ν defined in (27) has $\mathcal{C}^2$-regularity and is a classical supersolution of (11). Denote by A^c the operator

$A - c\frac{\partial}{\partial x_0}$ and by $\mathcal{L}_i^c(t)$ and $\mathcal{M}_i^c(t)$ the continuous parts of $\mathcal{L}_i(t)$ and $\mathcal{M}_i(t)$. We apply Ito's formula for càdlàg processes (see Meyer [25]) to $e^{-\delta t}\varphi_\nu(s(t))$. For any stopping time θ, the process

$$
\begin{aligned}
&e^{-\delta(\theta\wedge\tau)}\varphi_\nu(s(\theta\wedge\tau)) \\
&-\int_0^{\theta\wedge\tau} e^{-\delta t}\left\{A^{c(t)}\varphi_\nu(s(t))dt + \sum_{i=1}^n [L_i\varphi_\nu(s(t))d\mathcal{L}_i^c(t) + M_i\varphi_\nu(s(t))d\mathcal{M}_i^c(t)]\right\} \\
&-\sum_{0\le t\le\theta\wedge\tau} e^{-\delta t}[\varphi_\nu(s(t)) - \varphi_\nu(s(t^-))]
\end{aligned}
$$

is a martingale.

Since $s(t)$ has a jump only when $\mathcal{L}_i(t)$ or $\mathcal{M}_i(t)$ is discontinuous, we have

$$
\begin{aligned}
\mathcal{W}_\nu(s(t)) &= \mathcal{W}_\nu(s(t^-)) \\
&\quad -\sum_{i=1}^n [(\lambda_i+\nu_i)(\mathcal{L}_i(t) - \mathcal{L}_i(t^-)) + (\mu_i - \nu_i)(\mathcal{M}_i(t) - \mathcal{M}_i(t^-))] \\
&\le \mathcal{W}_\nu(s(t^-)).
\end{aligned}
$$

Hence,

$$\varphi_\nu(s(t)) \le \varphi_\nu(s(t^-)).$$

In addition, $\mathcal{L}_i^c$ and $\mathcal{M}_i^c$ are nondecreasing. Consequently

$$M_t^\nu = e^{-\delta(t\wedge\tau)}\varphi_\nu(s(t\wedge\tau)) + \int_0^{t\wedge\tau} e^{-\delta\theta}u(c(\theta))d\theta \tag{36}$$

is a supermartingale, as is the process $\min_\nu M_t^\nu$. Therefore

$$E\int_0^\tau e^{-\delta t}u(c(t))dt \le \varphi(x).$$

Taking the supremum over all policies $\mathcal{P}\in\mathcal{U}$, we get $V(x)\le\varphi(x)$. As $V(x)\ge 0$, we conclude that $V(x) = 0$ on ∂S and that V is continuous on ∂S. Since V is locally Lipschitz continuous in $\mathring{\mathcal{S}}$, V is continuous in $\mathcal{S}$. □

The regularity of V can be stated as follows.

PROPOSITION 3.6. *Suppose* (A.1) *holds. Then V is uniformly γ-Hölder continuous in $\mathcal{S}$, that is,*

$$\exists C>0,\quad |V(x)-V(x')| \le C\|x-x'\|^\gamma \qquad \forall x,x'\in\mathcal{S}. \tag{37}$$

Proof. Consider two initial positions x and x', and denote by τ (resp., τ') the first exit time of $\mathring{\mathcal{S}}$ of the process $s(t)$ (resp., $s'(t)$) defined by (1) and $s(0^-) = x$ (resp., $s'(0^-) = x'$). We have

$$
\begin{aligned}
V(x) - V(x') &= \sup_{\mathcal{P}\in\mathcal{U}} E\int_0^\tau e^{-\delta t}u(c(t))dt - \sup_{\mathcal{P}\in\mathcal{U}} E\int_0^{\tau'} e^{-\delta t}u(c(t))dt \\
&\le \sup_{\mathcal{P}\in\mathcal{U}} E\left(\int_0^\tau e^{-\delta t}u(c(t))dt - \int_0^{\tau'} e^{-\delta t}u(c(t))dt\right) \\
&\le \sup_{\mathcal{P}\in\mathcal{U}} E\int_{\tau\wedge\tau'}^\tau e^{-\delta t}u(c(t))dt.
\end{aligned}
$$

Using the supermartingale property of $\min_{\nu} M_t^{\nu}$ defined in (36), we get

$$E\int_{\tau\wedge\tau'}^{\tau} e^{-\delta t}u(c(t))dt \le E(e^{-\delta(\tau\wedge\tau')}\varphi(s(\tau\wedge\tau')) - e^{-\delta\tau}\varphi(s(\tau))),$$

and since φ vanishes on $\partial\mathcal{S}$, we have

$$V(x) - V(x') \le \sup_{\mathcal{P}\in\mathcal{U}} E(e^{-\delta\tau'}(\varphi(s(\tau')) - \varphi(s'(\tau')))1_{\tau'<\tau}),$$

where 1_A denotes the characteristic function of the set A.

Let us fix for instance $\|x\| = \sup_{i=0,\dots,n}|x_i|$. The function φ is γ-Hölder continuous, that is,

$$|\varphi(x) - \varphi(x')| \le C\|x - x'\|^{\gamma}$$

for some positive constant C. We thus get

$$V(x) - V(x') \le C\sup_{\mathcal{P}\in\mathcal{U}} E(e^{-\delta\tau'}\|(s(\tau') - s'(\tau')\|^{\gamma}1_{\tau'<\tau}). \tag{38}$$

The process $\Sigma(t) = s(t) - s'(t)$ is a diffusion process with generator $A + \delta I$ and initial value $\Sigma(0) = x - x'$.

If the function $\psi(x) = \|x\|^{\gamma}$ would satisfy $A\psi \le 0$, then $\psi(\Sigma(\tau\wedge t))e^{-\delta(\tau\wedge t)}$ would be a supermartingale which would readily lead to (37). Because ψ is not smooth, we consider the function $\psi_\beta(x) = \sum_{i=0}^{n}(x_i^2 + \beta)^{\gamma/2}$ with $\beta > 0$. We have

$$A\psi_\beta = \gamma(x_0^2+\beta)^{(\frac{\gamma}{2}-1)}\left(\left(r - \frac{\delta}{\gamma}\right)x_0^2 - \frac{\delta}{\gamma}\beta\right) + \sum_{i=1}^{n}\gamma(x_i^2+\beta)^{(\frac{\gamma}{2}-2)}f_i$$

with

$$f_i = x_i^4\left(\frac{1}{2}\sigma_i^2(\gamma-1) + \alpha_i - \frac{\delta}{\gamma}\right) + x_i^2\beta\left(\frac{1}{2}\sigma_i^2 + \alpha_i - \frac{2\delta}{\gamma}\right) - \frac{\delta}{\gamma}\beta^2.$$

Assumption (A.1) implies

$$r - \frac{\delta}{\gamma} < 0 \quad \text{and} \quad \frac{1}{2}\sigma_i^2(\gamma-1) + \alpha_i - \frac{\delta}{\gamma} < 0.$$

Consequently, there exists a positive constant C such that

$$A\psi_\beta \le C\beta^{\gamma/2}.$$

Applying Ito's formula to ψ_β, we obtain

$$E(e^{-\delta(\tau'\wedge\tau)}\psi_\beta(\Sigma(\tau'\wedge\tau))) \le \psi_\beta(x - x') + \frac{C}{\delta}\beta^{\gamma/2}. \tag{39}$$

Taking the limit of (39) when β goes to zero and using

$$\psi(x) \le \psi_0(x) \le (n+1)\psi(x)$$

we get

$$E(e^{-\delta(\tau'\wedge\tau)}\psi(\Sigma(\tau'\wedge\tau))) \le (n+1)\psi(x - x'),$$

which leads, together with (38), to the desired estimate (37). □

PROPOSITION 3.7. *V is nondecreasing with respect to x_i for $i = 0, \dots, n$.*

Proof. Let us denote explicitly by $s(t,x)$ the process $s(t)$ defined in (1) with initial value x and by τ_x the exit time of $\overset{\circ}{\mathcal{S}}$ of $s(t,x)$. Because

$$V(x) = \sup_{\mathcal{P}\in\mathcal{U}} E \int_0^{\tau_x} e^{-\delta t} u(c(t))dt$$

and u is positive, it is enough to prove the nondecreasing property of the stopping time τ_x for any control process $\mathcal{P}$.

Define $y(t,x)$ by

$$\begin{cases} y_0(t,x) &= e^{-rt} s_0(t,x), \\ y_i(t,x) &= e^{-(\alpha_i - \frac{1}{2}\sigma_i^2)t - \sigma_i W_i(t)} s_i(t,x), \quad i = 1, \dots, n. \end{cases}$$

The process $y(t,x)$ evolves according to

$$(40)\begin{cases} dy_0(t,x) &= e^{-rt}\left(-c(t)dt + \sum_{i=1}^{n}(-(1+\lambda_i)d\mathcal{L}_i(t) + (1-\mu_i)d\mathcal{M}_i(t))\right), \\ dy_i(t,x) &= e^{-(\alpha_i - \frac{1}{2}\sigma_i^2)t - \sigma_i W_i(t)}(d\mathcal{L}_i(t) - d\mathcal{M}_i(t)) \end{cases}$$

and satisfies $y(0,x) = x$. Hence, we can write

$$y(t,x) = x + Y(t,\mathcal{P}),$$

where $Y(t,\mathcal{P})$ depends only on $\mathcal{P}$. Consequently,

$$(41) \qquad s(t,x) = (e^{rt}x_0, (e^{(\alpha_i - \frac{1}{2}\sigma_i^2)t + \sigma_i W_i(t)} x_i)_{i=1,\dots,n}) + S(t,\mathcal{P}),$$

where $S(t,\mathcal{P})$ is a process which is independent of x.

Consider $\tilde{x} \geq x$ (i.e., $\tilde{x}_i \geq x_i \ \forall i = 0, \dots, n$) and fix $\mathcal{P}$ in $\mathcal{U}$. We have from (41)

$$s(t,x) \leq s(t,\tilde{x})$$

and

$$\mathcal{W}(s(t,x)) \leq \mathcal{W}(s(t,\tilde{x})),$$

where $\mathcal{W}$ is defined in (3).

Since

$$\tau_{\tilde{x}} = \inf\{t \geq 0,\ \mathcal{W}(s(t,\tilde{x})) \leq 0\}$$

for any $t > \tau_{\tilde{x}}$, there exists t' such that $\tau_{\tilde{x}} < t' < t$ and $\mathcal{W}(s(t',\tilde{x})) \leq 0$. This implies $\mathcal{W}(s(t',x)) \leq 0$ and $t > t' \geq \tau_x$. Consequently, $\tau_{\tilde{x}} \geq \tau_x$ and $V(\tilde{x}) \geq V(x)$. □

3.3. Existence and uniqueness results. First, we show that the value function V is a viscosity solution of the variational inequality (11). The problem is reduced to prove a weak dynamic programming principle (see Fleming and Soner [13]).

LEMMA 3.8. *There exists $C > 0$ such that*

$$(42) \qquad |J_x(\mathcal{P}) - J_{x'}(\mathcal{P})| \leq C\|x - x'\|^\gamma \quad \forall x, x' \in \mathcal{S}, \quad \forall \mathcal{P} \in \mathcal{U},$$

where $J_x(\mathcal{P})$ is given in (9).

Proof. Estimate (42) is readily obtained from the proof of Proposition 3.6. □

PROPOSITION 3.9. *The weak dynamic programming principle is satisfied for the value function V, that is,*

$$V(x) = \sup_{\mathcal{P}\in\mathcal{U}} E\left(\int_0^{\theta\wedge\tau} e^{-\delta t}u(c(t))dt + e^{-\delta(\theta\wedge\tau)}V(s((\theta\wedge\tau)^-))\right) \quad \forall x \in \mathcal{S} \tag{43}$$

for any stopping time θ.

Proof. By means of the Markov property, we have for all $\mathcal{P}$ in $\mathcal{U}$

$$E^{\mathcal{F}^{\theta\wedge\tau}}\int_0^{\tau} e^{-\delta t}u(c(t))dt = \int_0^{\theta\wedge\tau} e^{-\delta t}u(c(t))dt + e^{-\delta(\theta\wedge\tau)}J_{s((\theta\wedge\tau)^-)}(\mathcal{P}'),$$

with $\mathcal{P}'$ equal to $\mathcal{P}$ "shifted" by $\theta\wedge\tau$. Note that $\mathcal{P}'$ may not be admissible. The correct method would be to proceed with admissible systems composed with a filtration $(\Omega, \mathcal{F}_t, P)$, a Wiener process $W = (W_i)_{i=1,\dots,n}$ in $\mathbb{R}^n$, and an admissible control process $\mathcal{P}$ and consider V as the supremum of $J_x(\mathcal{P})$ over all admissible systems instead of the supremum over all admissible policies. We give here a formal proof. Rigorous proofs are given in Fleming and Soner [13], Nisio [26], El Karoui [11], and Lions [19]. Thus,

$$\begin{aligned} J_x(\mathcal{P}) &= E\left(\int_0^{\theta\wedge\tau} e^{-\delta t}u(c(t))dt + e^{-\delta(\theta\wedge\tau)}J_{s((\theta\wedge\tau)^-)}(\mathcal{P}')\right) \\ &\le E\left(\int_0^{\theta\wedge\tau} e^{-\delta t}u(c(t))dt + e^{-\delta(\theta\wedge\tau)}V(s((\theta\wedge\tau)^-))\right). \end{aligned}$$

By taking the supremum over all policies $\mathcal{P}$, we deduce one inequality side of (43). For the reverse inequality, we need to construct nearly optimal controls for each initial state x in a measurable way. To that purpose, consider $\varepsilon > 0$ and $\{\mathcal{S}^k\}_{k=1}^{\infty}$ a sequence of disjoint subsets of $\mathcal{S}$ such that

$$\bigcup_{k=1}^{\infty} \mathcal{S}^k = \mathcal{S}, \qquad \text{diameter}(\mathcal{S}^k) < \varepsilon.$$

For any k, take x^k in $\mathcal{S}^k$ and $\mathcal{P}^k = (c^k, (\mathcal{L}_i^k, \mathcal{M}_i^k)_{i=1,\dots,n})$ in $\mathcal{U}$ such that

$$V(x^k) - \varepsilon \le J_{x^k}(\mathcal{P}^k). \tag{44}$$

Now, for a given stopping time θ and an arbitrary policy $\mathcal{P}$ in $\mathcal{U}$, we define $\mathcal{P}^\theta = (c^\theta, (\mathcal{L}_i^\theta, \mathcal{M}_i^\theta)_{i=1,\dots,n})$ with

$$\begin{aligned} c^\theta(t) &= c(t)1_{t<\theta} + c^k(t-\theta)1_{t\ge\theta}, \\ \mathcal{L}_i^\theta(t) &= \mathcal{L}_i(t)1_{t<\theta} + (\mathcal{L}_i(\theta^-) + \mathcal{L}_i^k(t-\theta))1_{t\ge\theta}, \\ \mathcal{M}_i^\theta(t) &= \mathcal{M}_i(t)1_{t<\theta} + (\mathcal{M}_i(\theta^-) + \mathcal{M}_i^k(t-\theta))1_{t\ge\theta} \end{aligned}$$

for $s(\theta^-) \in \mathcal{S}^k$. Using (42) and (44) we have

$$\begin{aligned} J_{s(\theta^-)}(\mathcal{P}^k) &= (J_{s(\theta^-)}(\mathcal{P}^k) - J_{x^k}(\mathcal{P}^k)) + J_{x^k}(\mathcal{P}^k) \\ &\ge -C\varepsilon^\gamma - \varepsilon + V(x^k) \\ &\ge -2C\varepsilon^\gamma - \varepsilon + V(s(\theta^-)). \end{aligned}$$

Denote by I the right-hand side of (43). There exists a policy $\mathcal{P}$ such that

$$I - \varepsilon \leq E\left(\int_0^{\theta\wedge\tau} e^{-\delta t}u(c(t))dt + e^{-\delta(\theta\wedge\tau)}V(s((\theta\wedge\tau)^-))\right),$$

and using the Markov property, we get

$$I - \varepsilon \leq J_x(\mathcal{P}^{\theta\wedge\tau}) + (2C\varepsilon^\gamma + \varepsilon)$$

and

$$I - 2C\varepsilon^\gamma - 2\varepsilon \leq J_x(\mathcal{P}^{\theta\wedge\tau}) \leq V(x),$$

which leads to (43). □

COROLLARY 3.10. *The value function $V(x)$ defined by* (10) *is a viscosity solution of the variational inequality* (11)–(12).

In the case of pure diffusion processes, this is a standard result of the theory of viscosity solutions (see Lions [20]). For singular stochastic control problems, we refer to Fleming and Soner [13, Chap. 8, Thm. 5.1].

PROPOSITION 3.11. *Under Assumptions* (A.1) *and* (A.2), *the value function V is the unique viscosity solution of the variational inequality* (11)–(12) *in the class of continuous functions in $\mathcal{S}$ which satisfy*

$$|V(x)| \leq C(1 + \|x\|^\gamma) \qquad \forall x \in \mathcal{S}. \tag{45}$$

Proof. By Corollary 3.10 and equation (35), the value function V is a viscosity solution of (11)–(12) and satisfies (45). Uniqueness is a consequence of the following maximum principle.

LEMMA 3.12. *If v is a viscosity subsolution and v' is a viscosity supersolution of* (11) *which satisfy* (45) *and $v \leq v'$ on $\partial\mathcal{S}$, then $v \leq v'$ in $\mathcal{S}$.*

Indeed, a viscosity solution of (11)–(12) is both a subsolution and a supersolution with the boundary condition $v = 0$ on $\partial\mathcal{S}$. We prove Lemma 3.12 by using the Ishii technique; in particular we adapt the proofs of Theorems 3.3 and 5.1 of Crandall, Ishii, and Lions [9]. They are themselves based on the following corollary of Theorem 3.2 of [9].

LEMMA 3.13. *Let V be an upper semicontinuous function and V' be a lower semicontinuous function in an open domain $\mathcal{O}$ of $\mathbb{R}^N$. Consider $W(x,y) = V(x) - V'(y) - \frac{k}{2}|x-y|^2$ with $k > 0$ and suppose that $(\hat{x},\hat{y})$ is a local maximum of W. Then there exist two matrices X and Y in S^N such that*

$$(k(\hat{x}-\hat{y}), X) \in \overline{J}^{2,+}V(\hat{x}), \quad (k(\hat{x}-\hat{y}), Y) \in \overline{J}^{2,-}V'(\hat{y})$$

and

$$\begin{pmatrix} X & 0 \\ 0 & -Y \end{pmatrix} \leq 3k \begin{pmatrix} I & -I \\ -I & I \end{pmatrix}. \tag{46}$$

In this statement, $|.|$ denotes the euclidian norm and I the identity $N \times N$ matrix and $\overline{J}^{2,+}$ is defined as follows:

$$\overline{J}^{2,+}v(x) = \{(p,X) \in \mathbb{R}^N \times S^N,\ \exists (x_n, p_n, X_n) \in \mathcal{O} \times \mathbb{R}^N \times S^N,$$
$$(p_n, X_n) \in J^{2,+}v(x_n),\ \text{and}\ (x_n, v(x_n), p_n, X_n) \underset{n\to\infty}{\to} (x, v(x), p, X)\}.$$

$\overline{J}^{2,-}$ is similarly defined. If F is a continuous function in $S^N \times \mathbb{R}^N \times \mathbb{R} \times \mathcal{O}$ satisfying the elliptic condition (23), and v is a viscosity subsolution of (21), we have

$$F(X, p, v(x), x) \geq 0 \quad \forall (p, X) \in \overline{J}^{2,+} v(x), \quad \forall x \in \mathcal{O}. \tag{47}$$

Consider now v and v' as in Lemma 3.12 and argue by contradiction in order to prove $v \leq v'$ in $\mathcal{S}$. Suppose that there exists z in $\mathcal{S}$ such that $v(z) - v'(z) > 0$. For $k > 0$, define the function w_k in $\mathcal{S} \times \mathcal{S}$ as

$$w_k(x, y) = v(x) - v'(y) - \frac{k}{2}|x - y|^2 - \varepsilon(\mathcal{W}_\nu(x)^{\gamma'} + \mathcal{W}_\nu(y)^{\gamma'}),$$

where

$$\mathcal{W}_\nu(x) = x_0 + \sum_{i=1}^{n} (1 - \nu_i) x_i$$

and ν, ε, and γ' are parameters which will be chosen further. In addition, denote

$$m_k = \sup_{(x,y) \in \mathcal{S} \times \mathcal{S}} w_k(x, y).$$

In the following, C, C_1, and C_2 denote generic constants.

LEMMA 3.14. *For $\nu = (\nu_i)_{i=1,\ldots,n}$ with $-\lambda_i < \nu_i < \mu_i$, there exist C_1 and $C_2 > 0$ such that*

$$C_1|x| \leq \mathcal{W}_\nu(x) \leq C_2|x| \quad \forall x \in \mathcal{S}. \tag{48}$$

Proof. The second inequality of (48) is straightforward. To obtain the first inequality, we use the nonnegativity of $\mathcal{W}$ (defined in (3)) in $\mathcal{S}$:

$$\mathcal{W}_\nu(x) = \mathcal{W}(x) - \sum_{i=1}^{n} \min((\nu_i - \mu_i)x_i, (\nu_i + \lambda_i)x_i) \geq C \sum_{i=1}^{n} |x_i| \geq 0.$$

Moreover,

$$|x_0| = \left| \mathcal{W}_\nu(x) - \sum_{i=1}^{n} (1 - \nu_i) x_i \right| \leq \mathcal{W}_\nu(x) + C \sum_{i=1}^{n} |x_i| \leq C \mathcal{W}_\nu(x).$$

Consequently,

$$|x| \leq C \mathcal{W}_\nu(x). \qquad \square$$

Fix $\gamma' > \gamma$ such that Assumption (A.1) is still valid with γ' instead of γ, and ν as in Lemma 3.14. This guarantees $m_k < +\infty$ (see Lemma 3.15). On the other hand, we have

$$m_k \geq \overline{m} = \sup_{x \in \mathcal{S}} \{v(x) - v'(x) - 2\varepsilon \mathcal{W}_\nu(x)^{\gamma'}\} \geq v(z) - v'(z) - 2\varepsilon \mathcal{W}_\nu(z)^{\gamma'}.$$

As $v(z) > v'(z)$, there exists $\varepsilon > 0$ such that $m_k \geq \overline{m} > 0$ for any k; in the following, we consider such ε.

LEMMA 3.15. *Consider $\gamma' > \gamma$ and ν as in Lemma 3.14. There exist x_k, y_k in $\mathcal{S}$ such that*

$$m_k = w_k(x_k, y_k) < +\infty,$$

$$k|x_k - y_k|^2 \underset{k\to\infty}{\to} 0, \tag{49}$$

and

$$m_k \underset{k\to\infty}{\to} \overline{m} \equiv \sup_{x\in\mathcal{S}}\{v(x) - v'(x) - 2\varepsilon\mathcal{W}_\nu(x)^{\gamma'}\}. \tag{50}$$

Proof. Since v and v' satisfy (45), we have

$$m_k \le C + \sup_{x\in\mathcal{S}}(C_1|x|^\gamma - C_2|x|^{\gamma'}) < +\infty.$$

Let (x^n, y^n) be a maximizing sequence:

$$w_k(x^n, y^n) \ge m_k - \frac{1}{n} \ge \overline{m} - \frac{1}{n},$$

which implies that

$$C_2|x^n|^{\gamma'} - C_1|x^n|^\gamma \le C.$$

Hence, x^n is bounded, and similarly y^n is bounded. Consequently, there exists a converging subsequence of (x^n, y^n), and the limit $(x_k, y_k) \in \mathcal{S}\times\mathcal{S}$ realizes the maximum of w_k. As

$$v(x_k) - v'(y_k) - \varepsilon(\mathcal{W}_\nu(x_k)^{\gamma'} + \mathcal{W}_\nu(y_k)^{\gamma'}) = m_k + \frac{k}{2}|x_k - y_k|^2 \ge 0$$

for any k, we conclude that x_k, y_k, and $k|x_k - y_k|^2$ are bounded. Moreover, for any subsequence of (x_k, y_k) converging to $(\hat{x}, \hat{y})$ when k goes to infinity, we have $\hat{x} = \hat{y}$, and using $m_k \ge \overline{m}$, we get

$$\limsup_{k\to\infty} \frac{k}{2}|x_k - y_k|^2 \le v(\hat{x}) - v'(\hat{x}) - 2\varepsilon\mathcal{W}_\nu(\hat{x})^{\gamma'} - \overline{m} \le 0.$$

Consequently, (49) and (50) are satisfied. □

Now, since $\overline{m} > 0$ and $v \le v'$ on $\partial\mathcal{S}$, the limit $\hat{x}$ of x_k and y_k is in $\mathring{\mathcal{S}}$; then for any converging subsequence of (x_k, y_k), we have $(x_k, y_k) \in \mathring{\mathcal{S}}\times\mathring{\mathcal{S}}$ for large k. Applying Lemma 3.13 with $V = v - \varepsilon\mathcal{W}_\nu{}^{\gamma'}$ and $V' = v' + \varepsilon\mathcal{W}_\nu{}^{\gamma'}$ at the point (x_k, y_k) in $\mathring{\mathcal{S}}\times\mathring{\mathcal{S}}$, we obtain that there exist X, Y in S^{n+1} satisfying (46) such that

$$\begin{aligned}(p_k, X_k) &\equiv \Big(k(x_k - y_k) + \varepsilon\gamma'\mathcal{W}_\nu(x_k)^{\gamma'-1}\hat{p}, X + \varepsilon\gamma'(\gamma'-1)\mathcal{W}_\nu(x_k)^{\gamma'-2}A\Big)\\ &\in \overline{J}^{2,+}v(x_k),\end{aligned} \tag{51}$$

$$\begin{aligned}(p'_k, Y_k) &\equiv \Big(k(x_k - y_k) - \varepsilon\gamma'\mathcal{W}_\nu(y_k)^{\gamma'-1}\hat{p}, Y - \varepsilon\gamma'(\gamma'-1)\mathcal{W}_\nu(y_k)^{\gamma'-2}A\Big)\\ &\in \overline{J}^{2,-}v'(y_k)\end{aligned} \tag{52}$$

with $\hat{p} = (1, 1-\nu_1, \dots, 1-\nu_n)$ and $A = \hat{p}^t\hat{p}$.

Denote

$$F(X,p,v,x) = \max\left(F_0(X,p,v,x) + u^*(p_0), \max_{1\le i\le n} G_i(p), \max_{1\le i\le n} H_i(p)\right),$$

$$F_0(X,p,v,x) = \frac{1}{2}\sum_{i=1}^{n}\sigma_i^2 x_i^2 X_{ii} + \sum_{i=1}^{n}\alpha_i x_i p_i + r x_0 p_0 - \delta v,$$

$$G_i(p) = -(1+\lambda_i)p_0 + p_i,$$

$$H_i(p) = (1-\mu_i)p_0 - p_i,$$

where $X = (X_{ij})_{i,j=0,\dots,n}$, $p = (p_i)_{i=0,\dots,n}$.

Note that although F is continuous, F takes its values in $\mathbb{R}\cup\{+\infty\}$, since $F = +\infty$ when $p_0 \le 0$. This leads to a difficulty to obtain a uniform continuity property similar to [9, eq. (3.14)], and consequently straightforward application of the results of [9] cannot be used. Moreover, as the discount factor δ appears only in the F_0 component of F and not in G_i and F_i, property [9, eq. (3.13)], that is,

$$F(X,p,v,x) - F(X,p,v',x) \le -\lambda(v-v') \text{ for } v \ge v', \text{ with } \lambda > 0,$$

is not satisfied.

Using that v is a viscosity subsolution and v' is a viscosity supersolution of (11) (that is, of $F(D^2v, Dv, v, x) = 0$ in $\mathring{\mathcal{S}}$) and using (51) and (52), we get

$$F(X_k, p_k, v(x_k), x_k) \ge 0,$$
$$F(Y_k, p'_k, v'(y_k), y_k) \le 0.$$

This last inequality implies $G_i(p'_k) \le 0$ and $H_i(p'_k) \le 0$, and by linearity of G_i and H_i, we obtain

$$G_i(p_k) = G_i(p'_k) - \varepsilon\gamma'(\mathcal{W}_\nu(x_k)^{\gamma'-1} + \mathcal{W}_\nu(y_k)^{\gamma'-1})(\lambda_i+\nu_i) < 0$$

and

$$H_i(p_k) = H_i(p'_k) + \varepsilon\gamma'(\mathcal{W}_\nu(x_k)^{\gamma'-1} + \mathcal{W}_\nu(y_k)^{\gamma'-1})(\nu_i-\mu_i) < 0.$$

This leads to

$$F_0(X_k, p_k, v(x_k), x_k) + u^*((p_k)_0) \ge 0 \ge F_0(Y_k, p'_k, v'(y_k), y_k) + u^*((p'_k)_0).$$

Using now that u^* is nonincreasing and $(p'_k)_0 < (p_k)_0$, we obtain

$$F_0(X_k, p_k, v(x_k), x_k) - F_0(Y_k, p'_k, v'(y_k), y_k) \ge 0.$$

Hence,

$$\begin{aligned}(53)\qquad 0 \le{}& \frac{1}{2}\sum_{i=1}^{n}\sigma_i^2((x_k)_i^2 X_{ii} - (y_k)_i^2 Y_{ii})\\ &+k\left(\sum_{i=1}^{n}\alpha_i((x_k)_i - (y_k)_i)^2 + r((x_k)_0 - (y_k)_0)^2\right)\\ &-\delta(v(x_k) - v'(y_k) - \varepsilon(\mathcal{W}_\nu(x_k)^{\gamma'} + \mathcal{W}_\nu(y_k)^{\gamma'}))\\ &+\varepsilon(f(x_k) + f(y_k)),\end{aligned}$$

where

$$\begin{aligned} f(x) &= \frac{1}{2}\sum_{i=1}^{n}\sigma_i^2 x_i^2\gamma'(\gamma'-1)\mathcal{W}_\nu(x)^{\gamma'-2}A_{ii} \\ &\quad + \sum_{i=1}^{n}\alpha_i x_i\gamma'\mathcal{W}_\nu(x)^{\gamma'-1}\hat{p}_i + rx_0\gamma'\mathcal{W}_\nu(x)^{\gamma'-1}\hat{p}_0 \\ &\quad - \delta\mathcal{W}_\nu(x)^{\gamma'} \\ &= \gamma'\mathcal{W}_\nu(x)^{\gamma'}\left[r - \frac{\delta}{\gamma'} + \sum_{i=1}^{n}(\alpha_i - r)y_i + \frac{(\gamma'-1)}{2}\sum_{i=1}^{n}\sigma_i^2 y_i^2\right] \end{aligned}$$

with

$$y_i = \frac{(1-\nu_i)x_i}{\mathcal{W}_\nu(x)}.$$

We have

$$f(x) \le \gamma'\mathcal{W}_\nu(x)^{\gamma'}\left[r - \frac{\delta}{\gamma'} + \frac{1}{2(1-\gamma')}\sum_{i=1}^{n}\left(\frac{\alpha_i - r}{\sigma_i}\right)^2\right],$$

and since γ' is such that (A.1) is satisfied, $f(x) \le 0 \ \ \forall x \in \mathbb{R}^{n+1}$. Using (46), we see that the first term of the right-hand side of (53) is bounded by $Ck|x_k - y_k|^2$. Hence,

$$0 \le Ck|x_k - y_k|^2 - \delta m_k \underset{k\to\infty}{\to} -\delta\overline{m} < 0.$$

We thus get a contradiction, and Lemma 3.12 and Proposition 3.11 are proven. □

4. Change of variables.

4.1. Reduction of the state dimension. The value function V defined by (7) has the homothetic property (see [10])

$$\forall \rho > 0,\ V(\rho x) = \rho^\gamma V(x). \tag{54}$$

Consequently, the $(n+1)$-dimensional VI (11)–(12) satisfied by V can be reduced to a n-dimensional VI by using the following homogeneous model, that is, by considering the new state variables:

$$\begin{cases} \rho = x_0 + \sum_{i=1}^{n}(1-\mu_i)x_i & \text{(net wealth)}, \\ y_i = \dfrac{(1-\mu_i)x_i}{\rho}, \quad i = 1,\dots,n & \text{(fraction of net wealth invested in stock } i) \end{cases} \tag{55}$$

and the new control variable

$$C = \frac{c}{\rho} \quad \text{(fraction of net wealth dedicated to consumption)}. \tag{56}$$

The function $V(x)$ can be written as

$$\begin{aligned} V(x) &= V\left(\rho\left(1 - \sum_{i=1}^{n} y_i\right), \frac{\rho y_1}{(1-\mu_1)}, \dots, \frac{\rho y_n}{(1-\mu_n)}\right) \\ &= \rho^\gamma W(y), \end{aligned} \tag{57}$$

where the function

$$W(y) = V\left(1 - \sum_{i=1}^{n} y_i, \frac{y_1}{(1-\mu_1)}, \ldots, \frac{y_n}{(1-\mu_n)}\right) \tag{58}$$

is defined in

$$\tilde{\mathcal{S}} = \left\{ y = (y_1, \ldots, y_n) \in \mathbb{R}^n,\ 1 - \sum_{i=1}^{n} \frac{\lambda_i + \mu_i}{1-\mu_i} \{y_i\}^- \geq 0 \right\}$$

with $\{y\}^- = \max(0, -y)$.

Using inequality (35) we deduce that the function W is bounded in $\tilde{\mathcal{S}}$:

$$0 \leq W(y) \leq \varphi\left(1 - \sum_{i=1}^{n} y_i, \frac{y_1}{(1-\mu_1)}, \ldots, \frac{y_n}{(1-\mu_n)}\right) \leq a. \tag{59}$$

The function W is the unique viscosity solution of

$$\max\left(\tilde{A}W + u^*(BW), \max_{1\leq i\leq n} \tilde{L}_i W, \max_{1\leq i\leq n} \tilde{M}_i W\right) = 0 \quad \text{in } \overset{\circ}{\tilde{\mathcal{S}}}, \tag{60}$$

$$W = 0 \quad \text{on } \partial\tilde{\mathcal{S}}, \tag{61}$$

where

$$\tilde{A}W = \sum_{j,k=1}^{n} a_{jk} \frac{\partial^2 W}{\partial y_j \partial y_k} + \sum_{j=1}^{n} b_j \frac{\partial W}{\partial y_j} - \beta W, \tag{62}$$

$$BW = \gamma W - \sum_{j=1}^{n} y_j \frac{\partial W}{\partial y_j}, \tag{63}$$

$$\tilde{L}_i W = \frac{\partial W}{\partial y_i} - \left(\frac{\lambda_i + \mu_i}{1-\mu_i}\right) BW, \tag{64}$$

$$\tilde{M}_i W = -\frac{\partial W}{\partial y_i} \tag{65}$$

and

$$a_{jk} = \frac{y_j y_k}{2} \sum_{i=1}^{n} \sigma_i^2 (\delta_{ki} - y_i)(\delta_{ji} - y_i), \tag{66}$$

$$b_j = y_j \sum_{i=1}^{n} [(\gamma - 1)\sigma_i^2 y_i + \alpha_i - r](\delta_{ij} - y_i), \tag{67}$$

$$\beta = \delta - \gamma\left(r + \sum_{i=1}^{n} \left[(\alpha_i - r)y_i + \frac{\gamma - 1}{2}\sigma_i^2 y_i^2\right]\right). \tag{68}$$

The symbol δ_{ij} denotes the Kronecker index, which is equal to 0 when $i \neq j$ and equal to 1 when $i = j$.

Using the properties of V and (60), we deduce that W is concave, nonnegative, and nondecreasing with respect to each coordinate y_i.

Remark 4.1. Equations (60)–(61) depend only on $\nu = (\nu_i)_{i=1\ldots n}$ with $\nu_i = (\lambda_i + \mu_i)/(1 - \mu_i)$, and so does the function W. Denote by $V_{\lambda,\mu}$ the value function (7) in order to express explicitly the dependency of V on the transaction costs and by $W_{\lambda,\mu}$ the solution of (60)–(61). We have

$$\begin{aligned} W_{\lambda,\mu}(y) &= W_{\nu,0}(y) \\ &= V_{\nu,0}\left(1 - \sum_{i=1}^{n} y_i, y_1, \ldots, y_n\right). \end{aligned} \tag{69}$$

Using (54), we get

$$V_{\lambda,\mu}(x) = V_{\nu,0}(x_0, (1-\mu_1)x_1, \ldots, (1-\mu_n)x_n). \tag{70}$$

Consequently, it is sufficient to compute the value function V when the transaction costs on sale are equal to zero.

This remark could have been observed directly from the model. Indeed, the quantity $s_i(t)$ represents the amount of money in the ith risky asset at time t, that is, the quantity of the ith asset multiplied by the reference price $P_i(t)$. This reference price is useless in practice unless the transaction costs are time dependent. What matters for the investor is the buying price $(1 + \lambda_i)P_i$ and the selling price (or net price) $(1 - \mu_i)P_i$. The relevant quantity to consider is the net value of the ith asset, that is, $(1 - \mu_i)s_i$. Purchase of dL_i units of the ith asset increases the net value of this asset by $dL'_i = (1 - \mu_i)dL_i$ and requires a payment of $(1 + \nu_i)dL'_i$, whereas sale of dM_i units reduces the net value by $dM'_i = (1 - \mu_i)dM_i$ and realizes effectively dM'_i in cash. Consequently, by using a formulation of the problem based on the net values $(1 - \mu_i)s_i$ of the assets, the value function depends only on the coefficients ν_i, where ν_i represents the proportional transaction cost on purchase with respect to the net price of the ith asset.

4.2. Additional treatment for numerical purpose. Our purpose is now to solve equations (60)–(61).

In order to simplify the numerical computation, we restrict the admissible region $\mathcal{S}$ to

$$\mathcal{S}^+ = \left\{ x \in \mathbb{R}^{n+1},\ x_1, \ldots, x_n \geq 0,\ x_0 + \sum_{i=1}^{n} (1 - \mu_i)x_i \geq 0 \right\};$$

that is, we suppose that the amounts of money allocated in the risky assets are nonnegative, while the amount of money in the bank account can be negative as long as the net wealth remains nonnegative. This is not restrictive since, when $\alpha_i > r$, the no-transaction cone is inside $\mathcal{S}^+$ and a trajectory which starts in $\mathcal{S}^+$ remains in $\mathcal{S}^+$ (see [10] for $n = 1$).

This leads to the study of VI (60) in the domain $(\mathbb{R}^+)^n$:

$$\max\left(\tilde{A}W + u^*(BW), \max_{1\leq i\leq n} \tilde{L}_i W, \max_{1\leq i\leq n,\ y_i>0} \tilde{M}_i W\right) = 0 \quad \text{in } (\mathbb{R}^+)^n. \tag{71}$$

This VI degenerates at the boundary and is valid up to the boundary, but the controls which make the trajectory go out of the domain are not admissible. Note that the function W has bounded derivatives in $(\mathbb{R}^+)^n$.

We proceed with a technical change of variables which brings $(\mathbb{R}^+)^n$ to $[0,1]^n$, namely,

$$\text{(72)} \qquad \begin{cases} \psi(z) = \theta(z)W(y), \\ \theta(z) = \displaystyle\prod_{i=1}^{n}(1-z_i), \\ z_i = \dfrac{y_i}{1+y_i}, \quad i = 1,\dots,n. \end{cases}$$

The function ψ is bounded and concave with respect to z_i, $i = 1,\dots,n$, has bounded derivatives, and satisfies

$$\text{(73)} \qquad \begin{cases} \max\left(\overline{A}\psi + \sup\limits_{C\geq 0}\left(-C\overline{B}\psi + \theta(z)\dfrac{C^\gamma}{\gamma}\right), \max\limits_{1\leq i\leq n}\overline{L}_i\psi, \max\limits_{i,\, z_i>0}\overline{M}_i\psi\right) = 0 \quad \text{in } [0,1)^n, \\ \psi = 0 \quad \text{on } [0,1]^n \cap \{z_i = 1\} \quad \forall i = 1,\dots,n, \end{cases}$$

where

$$\begin{aligned} \overline{A}\psi &= \sum_{j,k=1}^{n} a'_{jk}\frac{\partial^2\psi}{\partial z_j \partial z_k} + \sum_{j=1}^{n} b'_j\frac{\partial\psi}{\partial z_j} - \beta'\psi, \\ \overline{B}\psi &= \left(\gamma - \sum_{j=1}^{n} z_j\right)\psi - \sum_{j=1}^{n} z_j(1-z_j)\frac{\partial\psi}{\partial z_j}, \\ \overline{L}_i\psi &= (1-z_i)\left(\psi + (1-z_i)\frac{\partial\psi}{\partial z_i}\right) - \lambda_i\overline{B}\psi, \\ \overline{M}_i\psi &= -(1-z_i)\left(\psi + (1-z_i)\frac{\partial\psi}{\partial z_i}\right), \end{aligned}$$

with

$$\begin{aligned} a'_{jk} &= \frac{1}{2}z_j(1-z_j)z_k(1-z_k)\overline{a}_{jk}, \\ b'_j &= z_j(1-z_j)\left(\overline{b}_j + \sum_{\substack{k=1 \\ k\neq j}}^{n} z_k\overline{a}_{jk}\right), \\ \beta' &= \beta - \sum_{j=1}^{n} z_j\overline{b}_j - \sum_{\substack{j,k=1 \\ j\neq k}}^{n} \overline{a}_{jk}\frac{z_j z_k}{2}, \\ \overline{a}_{jk} &= \sum_{i=1}^{n}\sigma_i^2\left(\delta_{ki} - \frac{z_i}{1-z_i}\right)\left(\delta_{ji} - \frac{z_i}{1-z_i}\right), \\ \overline{b}_j &= \sum_{i=1}^{n}\left((\gamma-1)\sigma_i^2\frac{z_i}{1-z_i} + \alpha_i - r\right)\left(\delta_{ij} - \frac{z_i}{1-z_i}\right) \end{aligned}$$

and β defined in (68).

The numerical study is organized as follows: equation (73) is solved by using the numerical methods explained in §5 below. Then a reverse change of variable is performed in order to display the numerical results for equation (71) (see §6).

5. Numerical methods. We consider equations of the form

$$\begin{cases} \max\limits_{P \in \mathcal{P}_{ad}} (A^P W + u(P)) = 0 & \text{in } \Omega = [0,1]^m \setminus \Gamma, \\ W = 0 & \text{on } \Gamma, \end{cases} \tag{74}$$

where A^P is a second-order degenerate elliptic operator

$$A^P W(x) = \sum_{i,j=1}^{m} a_{ij}(x,P) \frac{\partial^2 W}{\partial x_i \partial x_j}(x) + \sum_{i=1}^{m} b_i(x,P) \frac{\partial W}{\partial x_i}(x) - \beta(x,P) W(x)$$

with

$$\sum_{i,j=1}^{m} a_{ij}(x,P)\eta_i\eta_j \geq 0, \qquad \beta(x,P) \geq 0 \qquad \forall x \in \Omega,\ \eta \in \mathbb{R}^m,\ P \in \mathcal{P}_{ad}.$$

$\mathcal{P}_{ad}$ is a closed subset of $\mathbb{R}^k$ (which may depend on x) and Γ is a part of the boundary $\partial\Omega$, which consists of faces of the m-cube $[0,1]^m$. On $\partial\Omega \setminus \Gamma$, the operator A^P is degenerate.

In §3, we have proven that the value function (7), within a change of variables, is the unique viscosity solution of an equation of type (74). This solution can be approximate by the following numerical method: (i) Discretize (74) by using a consistent finite-difference approximation which satisfies the discrete maximum principle (DMP) (recalled below). (ii) Solve the discrete equation by means of the value iteration (successive approximation) algorithm or the Howard algorithm (policy iteration). This method does not require any stronger regularity condition on the viscosity solution (see Barles and Souganidis [3], Fleming and Soner [13]). The algorithms mentioned in (ii) may be replaced by the (full) multigrid–Howard algorithm (FMGH), introduced in Akian [1], [2] and based on the Howard algorithm and the multigrid method. This algorithm is more efficient, but proof of convergence has been obtained only when the DMP is satisfied, the feedbacks are regular, and the Bellman equation is strongly elliptic.

For the numerical solution of (74), we use a classical finite-difference discretization in a regular grid and the FMGH algorithm. Convergence arguments used in [1], [2] cannot be applied here since the DMP is not satisfied (because of the presence of mixed derivatives), the equation is degenerate, and the control is singular. Nevertheless, numerical experiments show that this numerical method converges.

This procedure and the computer implementation are treated by using the expert system *Pandore* (see Chancelier, et al. [6], Akian [2]), which has been developed to automate studies in stochastic control.

5.1. Discretization. Let $h = 1/N$ ($N \in \mathbb{N}^*$) denote the finite-difference step in each coordinate direction, e_i the unit vector in the ith coordinate direction, and $x = (x_1, \ldots, x_m)$ a point of the uniform grid $\Omega_h = \Omega \cap (h\mathbb{Z})^m$. Equation (74) is discretized by replacing the first- and second-order derivatives of W by the following approximation:

$$\frac{\partial W}{\partial x_i} \sim \frac{W(x + he_i) - W(x - he_i)}{2h} \tag{75}$$

or

$$\frac{\partial W}{\partial x_i} \sim \begin{cases} \dfrac{W(x+he_i)-W(x)}{h} & \text{when} \quad b_i(x,P) \geq 0, \\ \dfrac{W(x)-W(x-he_i)}{h} & \text{when} \quad b_i(x,P) < 0. \end{cases} \tag{76}$$

$$\frac{\partial^2 W}{\partial x_i^2}(x) \sim \frac{W(x+he_i)-2W(x)+W(x-he_i)}{h^2}, \tag{77}$$

$$\begin{aligned} \frac{\partial^2 W}{\partial x_i \partial x_j}(x) \sim{}& \frac{W(x+he_i+he_j)-W(x+he_i-he_j)}{4h^2} \\ &+ \frac{W(x-he_i-he_j)-W(x-he_i+he_j)}{4h^2} \qquad \text{for } i \neq j. \end{aligned} \tag{78}$$

Approximation (75) may be used when A is uniformly elliptic, whereas (76) has to be used when A is degenerate (see Kushner [18]). These differences are computed in the entire grid Ω_h by extending W to the "boundary" of Ω_h in $(h\mathbb{Z})^m$:

$$\begin{aligned} W(x) &= 0 \qquad \forall x \in \Gamma \cap (h\mathbb{Z})^m, \\ W(x-he_i) &= W(x) \quad \forall x \in \{x_i = 0\} \cap \Omega_h, \\ W(x+he_i) &= W(x) \quad \forall x \in \{x_i = 1\} \cap \Omega_h. \end{aligned}$$

We obtain a system of N_h nonlinear equations of N_h unknowns $\{W_h(x),\ x \in \Omega_h\}$:

$$\max_{P \in \mathcal{P}_{ad}} (A_h^P W_h + u(P))(x) = 0 \quad \forall x \in \Omega_h, \tag{79}$$

where $N_h = \sharp\Omega_h \sim 1/h^m$. Let $\mathcal{P}_h$ denote the set of control functions $P : \Omega_h \to \mathcal{P}_{ad}$ and $\mathcal{V}_h$ the set of functions from Ω_h into $\mathbb{R}$. Equation (79) can be rewritten

$$\max_{P \in \mathcal{P}_h} (A_h^P W_h + u(P)) = 0, \quad W_h \in \mathcal{V}_h.$$

Then, the operator A_h^P, depending on P in $\mathcal{P}_h$, maps $\mathcal{V}_h$ into itself (or is a $N_h \times N_h$ matrix).

Because of the degeneracy of the operator A^P at some points of the closed m-cube $\overline{\Omega}$ and the presence of mixed derivatives, A_h^P does not satisfy the usual DMP (i.e., $(A_h^P W_h(x) \leq 0 \ \ \forall x \in \Omega_h) \Rightarrow (W_h(x) \geq 0 \ \ \forall x \in \Omega_h)$). Consequently, equation (79) may not be stable, even for small step h. However, A_h^P can be written as the sum of a symmetric negative definite operator and an operator which satisfies the DMP; we thus infer the stability of A_h^P, which is confirmed by numerical experiments.

We describe below the available algorithms to solve equation (79).

5.2. The value iteration method. Suppose that the $N_h \times N_h$ matrix A_h^P satisfies

$$(A_h^P)_{ij} \geq 0 \ \ \forall i \neq j, \qquad \sum_{j=1}^{N_h} (A_h^P)_{ij} = -\lambda < 0 \ \ \forall i, \tag{80}$$

which implies that A_h^P satisfies the DMP. Equation (79) can be rewritten as

$$W_h = \frac{1}{1+\lambda k} \max_{P \in \mathcal{P}_h} (M^P W_h + k u(P)), \tag{81}$$

where $k > 0$ and $M^P = I + k(A_h^P + \lambda I)$ is a Markov matrix. (I is the $N_h \times N_h$ identity matrix.) Equation (79) can then be interpreted as the dynamic programming equation of a control problem of Markov chain with discount factor $1/(1+\lambda k)$, instantaneous cost $ku(P)$, and transition matrix M^P:

$$\max_{(P_n)} \sum_{n=0}^{\tau} \frac{k}{(1+\lambda k)^{n+1}} u(X_n, P_n).$$

The value iteration method (see Bellman [5]) consists in the contraction iteration

$$W^{n+1} = \frac{1}{1+\lambda k} \max_{P \in \mathcal{P}_h} (M^P W^n + ku(P)). \tag{82}$$

The contracting factor is $1/(1+\lambda k) = 1 - \mathcal{O}(h^2)$ and the complexity[1] of the method is

$$C_h = \mathcal{O}\left(\frac{-\log h}{h^2} N_h\right) = \mathcal{O}(-h^{-(2+m)} \log h) = \mathcal{O}(N_h^{1+2/m} \log N_h).$$

When the operator A_h^P does not satisfy the DMP, equation (79) cannot be interpreted as a discrete Bellman equation. Nevertheless, the iterative method (82) can still be used if we find λ and k such that the L^2 norm of M^P (which is no more a Markov matrix) is lower than 1 for all P. This condition may be obtained for instance when the discount factor $\beta(x, P)$ is large enough.

An example of the use of the value iteration algorithm is given in Sulem [30] for solving the one-dimensional investment-consumption problem.

5.3. The multigrid–Howard algorithm. Another classical algorithm is the Howard algorithm (see Howard [16], Bellman [4], [5]), also named policy iteration. It consists of an iteration algorithm on the control and value functions (starting from P^0 or W^0):

$$\text{for } n \geq 1, \quad P^n \in \operatorname*{Argmax}_{P \in \mathcal{P}_h} (A_h^P W^{n-1} + u(P)), \tag{83}$$

$$\text{for } n \geq 0, \quad W^n \text{ is the solution of } A_h^{P^n} W + u(P^n) = 0. \tag{84}$$

When A_h^P satisfies the DMP, the sequence W^n decreases and converges to the solution of (79) and the convergence is in general superlinear [4], [5], [1], [2].

The exact computation of step (84) is expensive in dimension $m \geq 2$. (The complexity of a direct method is $\mathcal{O}(N_h^{3-2/m})$.) We thus use the multigrid–Howard algorithm introduced in [1], [2]: in (84), W^n is computed by a multigrid method with initial value W^{n-1}. The advantage is that each multigrid iteration takes a computing time of $\mathcal{O}(N_h)$ and contracts the error by a factor independent of the discretization step h. For a detailed description of the multigrid algorithm, see, for example, McCormick [22], Hackbusch [14], and Hackbusch and Trottenberg [15].

Let $\mathcal{M}^P$ denote the operator of an iteration of the multigrid method associated with the equation $A_h^P W + u(P) = 0$. Starting from W^0, we proceed with the following iteration:

$$\left\{ \begin{array}{l} (83), \\ \text{for } n \geq 1 \left\{ \begin{array}{l} W^{n,0} = W^{n-1}, \\ \text{for } i = 1 \text{ to } m_n, \ W^{n,i} = \mathcal{M}^{P^n}(W^{n,i-1}), \\ W^n = W^{n,m_n}. \end{array} \right. \end{array} \right. \tag{85}$$

[1] The number of elementary operations for computing an approximation of the solution of (79) with an error in the order of the discretization error.

This algorithm converges to the solution W_h^* of (79) if W^0 is sufficiently close to W_h^* and m_n is large enough (independently of the step h) [1], [2].

We introduce now the FMGH algorithm, which solves equation (79) from any initial value W^0.

5.4. The FMGH algorithm. This algorithm [1], [2] fully uses the idea of the full multigrid method (see, for example, Hackbusch and Trottenberg [15]).

Consider the sequence of grids $(\Omega_k)_{k\geq 1}$ of steps $h_k = 2^{-k}$ and denote by $\mathcal{I}_k^{k+1}$ the operator of the m-linear interpolation from $\mathcal{V}_{h_k}$ into $\mathcal{V}_{h_{k+1}}$.

If $W_k \in \mathcal{V}_{h_k}$, $W_{k+1} = \mathcal{I}_k^{k+1} W_k$ is defined by

$$\begin{cases} W_{k+1}(x) = W_k(x) & \forall x \in \Omega_k \subset \Omega_{k+1}, \\ W_{k+1}\left(\dfrac{x+y}{2}\right) = \dfrac{W_{k+1}(x) + W_{k+1}(y)}{2} & \forall x, y \in \Omega_{k+1} \text{ such that } \dfrac{x+y}{2} \in \Omega_{k+1} \\ & \text{and } x,\ y \text{ are in the same cell of } \Omega_k, \end{cases}$$

where a cell of Ω_h is a m-cube of width h included in $\overline{\Omega}$ and with vertices in $(h\mathbb{Z})^m$.

The FMGH algorithm is defined as

For $1 \leq k \leq \overline{k}$,

$W_k^{\overline{n}}$ is the $\overline{n}$th iteration of the sequence defined by (85) in the grid Ω_k of initial value W_k^0.

For $1 \leq k < \overline{k}$,

$W_{k+1}^0 = \mathcal{I}_k^{k+1} W_k^{\overline{n}}$.

Under appropriate assumptions (strong ellipticity, DMP, regularity of the feedback; see [1], [2]), the error between $W_k^{\overline{n}}$ and the solution W_k^* of (79) with $h = h_k$ is in the order of the discretization error for any k. This property is realized for any initial value W_1^0, if the numbers m_n and $\overline{n}$ are large enough (but independent of the level k). Consequently, this algorithm solves equation (79) (with an error in the order of the discretization error) with a computing time of $\mathcal{O}(N_h)$.

6. Numerical results. Equation (71) is solved in $(\mathbb{R}^+)^n$ by using the FMGH algorithm for $n = 1$ and $n = 2$ and various numerical values of the parameters.

Remark 6.1. The regions B_i and S_i defined in (17) and (18) are characterized by

$$B_i = \{x \in \mathcal{S},\ \tilde{L}_i W(y) = 0,\ y \text{ given by (55)}\},$$
$$S_i = \{x \in \mathcal{S},\ \tilde{M}_i W(y) = 0,\ y \text{ given by (55)}\},$$

where the operators $\tilde{L}_i$ and $\tilde{M}_i$ are defined in (64) and (65). By extension we use the notation

$$B_i = \{y \in (\mathbb{R}^+)^n,\ \tilde{L}_i W(y) = 0\},$$
$$S_i = \{y \in (\mathbb{R}^+)^n,\ \tilde{M}_i W(y) = 0\},$$
$$NT_i = (\mathbb{R}^+)^n \backslash (B_i \cup S_i), \tag{86}$$
$$NT = \bigcap_{i=1}^{n} NT_i.$$

6.1. One risky asset. Numerical tests are performed with $\gamma = 0.3$, $\delta = 10\%$, $r = 7\%$, $\alpha_1 = 11\%$, $\sigma_1 = 30\%$, $\nu = \nu_1 = (\lambda_1 + \mu_1)/(1 - \mu_1) = 0.1, 0.3, 0.5, 1, 2, 3$ or 4%. These values of ν are obtained for example when $\lambda_1 = \mu_1 \simeq \nu/2 = 0.05$, 0.15, 0.25, 0.5, 1, 1.5, 2%.

When $\nu > 0$, the regions B_1 and S_1 are of the form (see §7) : $B_1 = [0, \pi^-]$ and $S_1 = [\pi^+, +\infty)$ with $0 < \pi^- < \pi^+$. When $\nu = 0$ (no transaction costs), the optimal policy is to keep a constant proportion of risky asset equal to π_1^* (given by (90) below), that is $\pi^+ = \pi^- = \pi_1^*$. In our example, $\pi_1^* = 0.635$. The values of π^+ and π^- are given in Table 1 and displayed in Fig. 1 as functions of ν.

TABLE 1.

ν (%)	0.1	0.3	0.5	1	2	3	4
π^-	0.56	0.54	0.52	0.47	0.42	0.39	0.36
π^+	0.68	0.68	0.68	0.68	0.68	0.68	0.68

The graphs of π^+ and π^- are similar to those obtained by Davis and Norman [10] who already observed that the "sell-barrier" is very insensitive to the transaction cost, while the "buy-barrier" decreases rapidly as ν increases. Indeed, even if the selling cost is high, the risky asset must be sold before it can be realized for consumption. On the other hand, it may not be worthwhile to invest in the risky asset if the transaction costs are too high.

The value function W, solution of (71), and the optimal consumption C are displayed in Figs. 2 and 3.

From equations (71) and (86), we obtain $W(y) = c(1 + \nu y)^\gamma$ in B_1, where c is a constant depending on ν. In S_1, W is constant and seems insensitive to the transaction costs. This means that when the initial proportion in the risky asset is in S_1, the probability of a future purchase is small. On the other hand, if the initial proportion invested in stock is in B_1, loss of profit (when ν increases) is due mainly to the first transaction.

The values of C are not relevant in B_1 and S_1 since the investor makes transactions and thus does not consume. As expected, C decreases in $[\pi^-, \pi^+]$, as does the fraction of wealth in cash.

6.2. Two risky assets. We set $\gamma = 0.3$, $\delta = 10\%$, and $r = 7\%$ and fix the parameters of the first risky asset to $\alpha_1 = 11\%$, $\sigma_1 = 30\%$, and $\nu_1 = (\lambda_1 + \mu_1)/(1 - \mu_1) = 1\%$.

Four tests are performed:

test 1: $\alpha_2 = 15\%$, $\sigma_2 = 35\%$, $\nu_2 = 2\%$,
test 2: $\alpha_2 = 15\%$, $\sigma_2 = 35\%$, $\nu_2 = 0.5\%$,
test 3: $\alpha_2 = 15\%$, $\sigma_2 = 35\%$, $\nu_2 = 1\%$,
test 4: $\alpha_2 = 20\%$, $\sigma_2 = 50\%$, $\nu_2 = 1\%$.

For test 1, the value function W, the optimal consumption C, and their contour lines are displayed in Figs. 4–7.

The partition of the domain is displayed for each test in Figs. 8–11. As expected, nine regions appear: buy (resp., sell) asset i when y_i is below (resp., above) a critical level π_i^- (resp., π_i^+) depending on y_j ($j \neq i$) and no transaction between π_i^- and π_i^+.

After the first transaction, the position of the investor evolves as a diffusion process with reflection on the boundary of NT. The direction of the reflection is given by the equation $L_i W = 0$ on the frontier with B_i and $M_i W = 0$ on the frontier with S_i.

Note that the no-transaction interval for the first asset $NT_1 \cap \{y_2 = \text{constant}\} \simeq [0.39, 0.78]$ is much larger than the no-transaction interval $[0.47, 0.68]$ obtained in dimension 1, when only one asset (with same parameters) is available. This is not surprising since the second asset has larger expected rate of return; it is thus more interesting to make transactions on the second asset.

We observe that the boundaries of the regions B_i and S_i seem at first to be straight lines (y_i = constant). This would mean that the investment policies are decoupled although the dynamics are correlated. In fact, when the cost for purchase ν_2 grows, the region NT_2 grows as expected but the boundaries of S_1 and B_1 are also perturbed. Moreover, a variation of α_2 and σ_2 affect both NT_2 and NT_1. A theoretical study of the boundaries is done below in order to confirm these remarks.

7. Theoretical analysis of the optimal strategy.

7.1. No transaction costs: The Merton problem. When the transaction costs are equal to zero, the optimal investment strategy is to keep a constant fraction of total wealth in each risky asset (see Merton [24], Sethi and Taksar [27], Karatzas, et al. [17], and Davis and Norman [10]). Indeed, set $\lambda = \mu = 0$ in equation (71). We obtain

$$(87)\quad \max\left(\tilde{A}W + u^*(BW),\ \max_{1\le i\le n} \frac{\partial W}{\partial y_i},\ \max_{1\le i\le n,\ y_i>0}\left(-\frac{\partial W}{\partial y_i}\right)\right) = 0 \quad \text{in } (\mathbb{R}^+)^n,$$

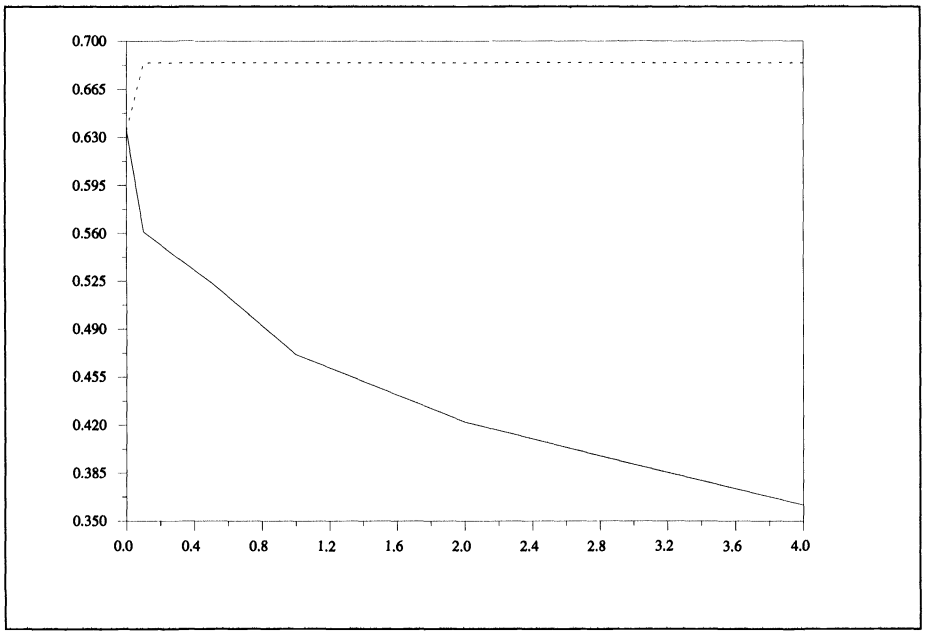


FIG. 1. *Graph of* π^+ *and* π^- *for* $n = 1$, $\gamma = 0.3$, $\delta = 10\%$, $r = 7\%$, $\alpha_1 = 11\%$, $\sigma_1 = 30\%$.

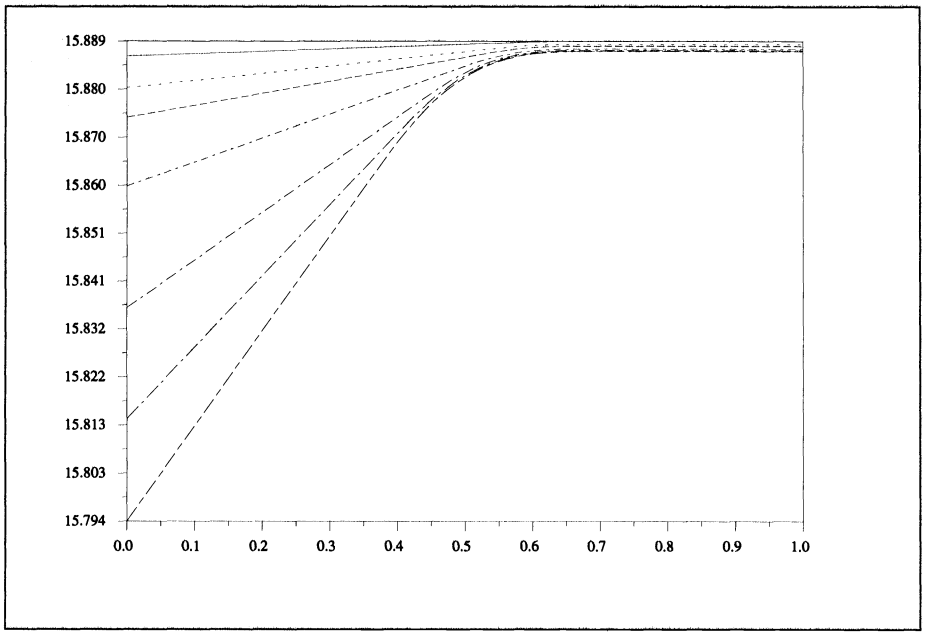


FIG. 2. *Value function W for $n = 1$, $\gamma = 0.3$, $\delta = 10\%$, $r = 7\%$, $\alpha_1 = 11\%$, $\sigma_1 = 30\%$.*

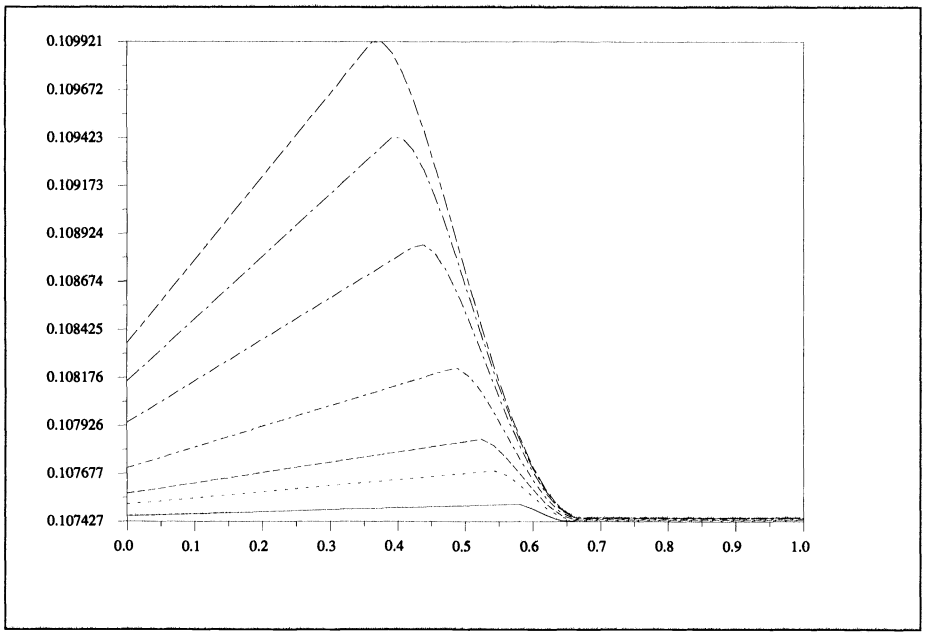


FIG. 3. *Optimal consumption C for $n = 1$, $\gamma = 0.3$, $\delta = 10\%$, $r = 7\%$, $\alpha_1 = 11\%$, $\sigma_1 = 30\%$.*

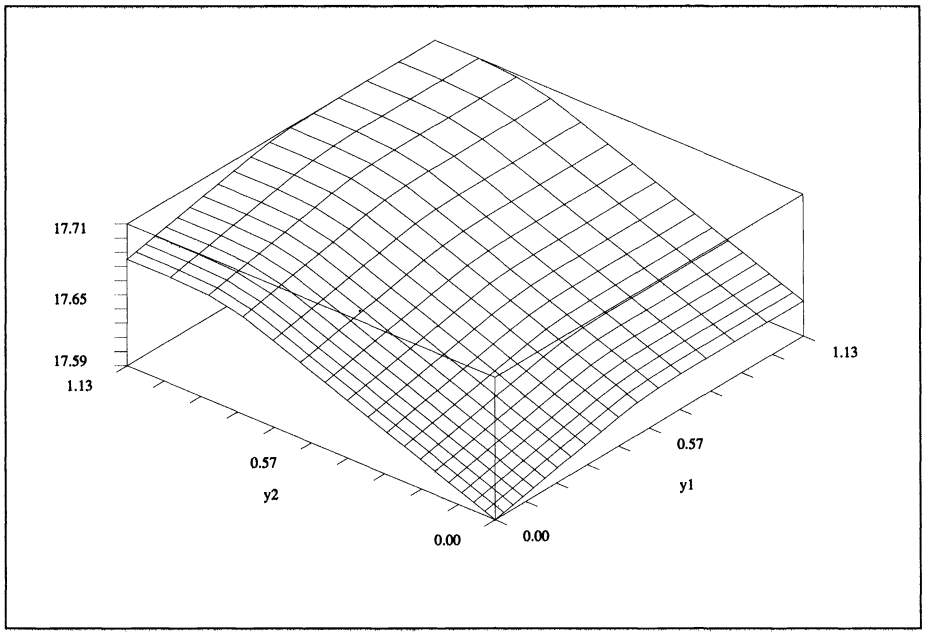


FIG. 4. *Value function W for* $\gamma = 0.3$, $\delta = 10\%$, $r = 7\%$, $\alpha = (11\%, 15\%)$, $\sigma = (30\%, 35\%)$, $\nu = (1\%, 2\%)$.

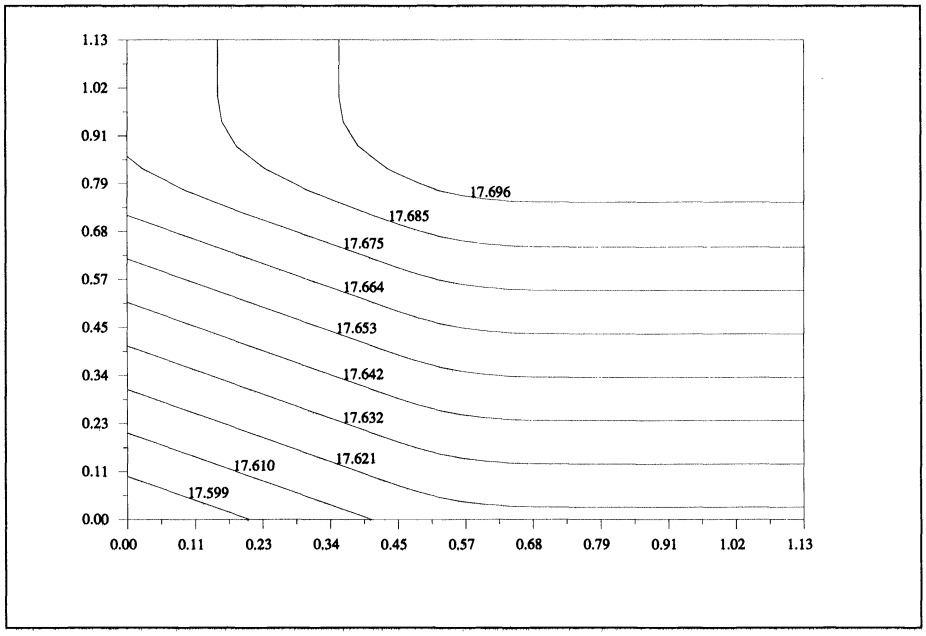


FIG. 5. *Value function W for* $\gamma = 0.3$, $\delta = 10\%$, $r = 7\%$, $\alpha = (11\%, 15\%)$, $\sigma = (30\%, 35\%)$, $\nu = (1\%, 2\%)$.

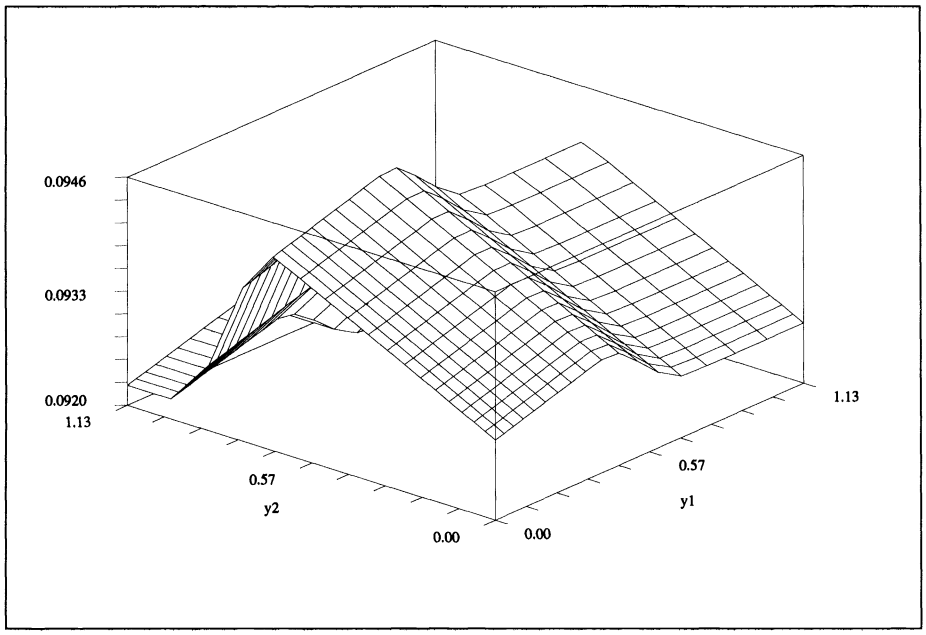


FIG. 6. *Optimal consumption C for* $\gamma = 0.3$, $\delta = 10\%$, $r = 7\%$, $\alpha = (11\%, 15\%)$, $\sigma = (30\%, 35\%)$, $\nu = (1\%, 2\%)$.

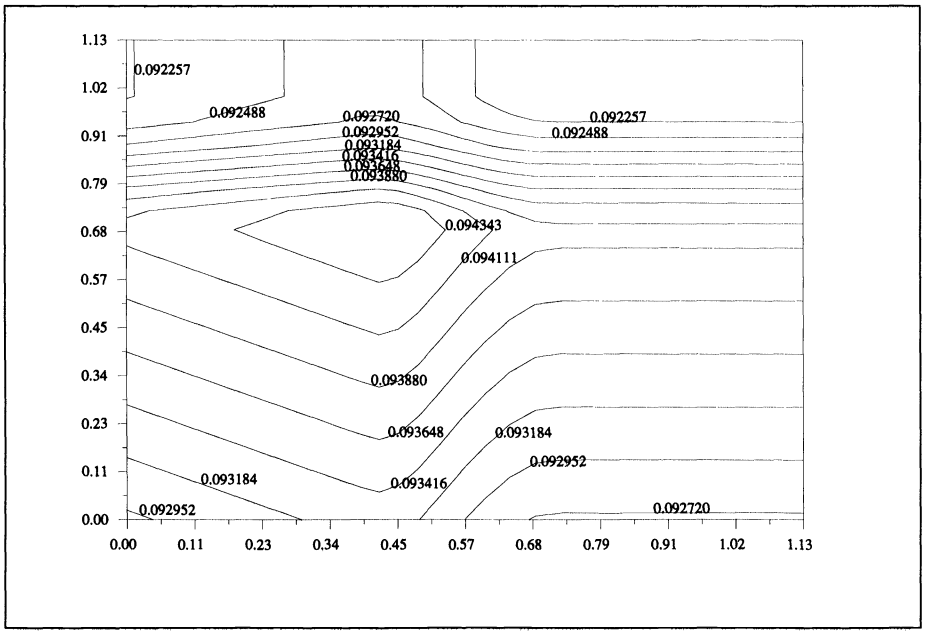


FIG. 7. *Optimal consumption C for* $\gamma = 0.3$, $\delta = 10\%$, $r = 7\%$, $\alpha = (11\%, 15\%)$, $\sigma = (30\%, 35\%)$, $\nu = (1\%, 2\%)$.

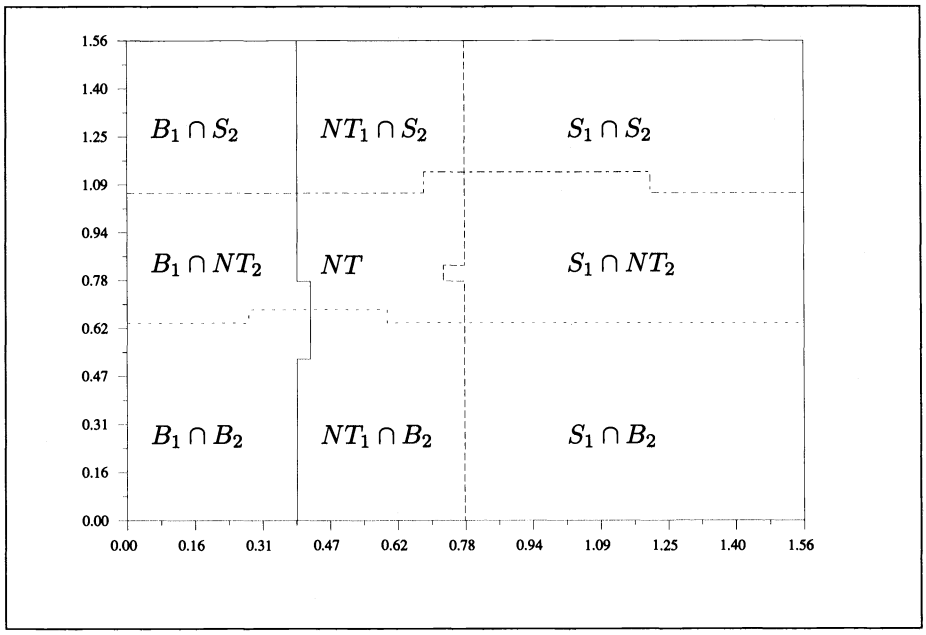


FIG. 8. *Boundaries of the regions* B_i, S_i, *and* NT_i *for* $\gamma = 0.3$, $\delta = 10\%$, $r = 7\%$, $\alpha = (11\%, 15\%)$, $\sigma = (30\%, 35\%)$, $\nu = (1\%, 2\%)$.

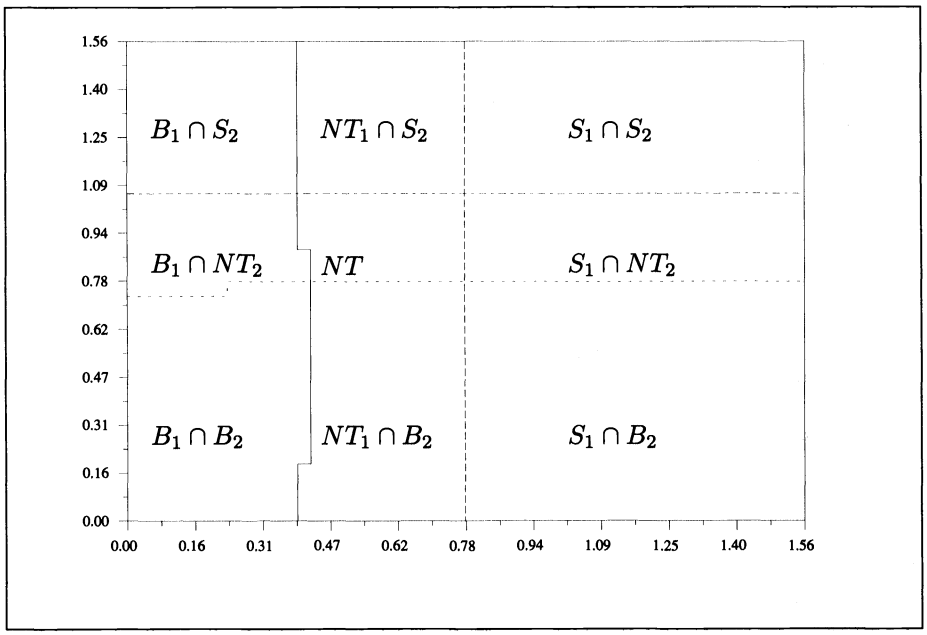


FIG. 9. *Boundaries of the regions* B_i, S_i, *and* NT_i *for* $\gamma = 0.3$, $\delta = 10\%$, $r = 7\%$, $\alpha = (11\%, 15\%)$, $\sigma = (30\%, 35\%)$, $\nu = (1\%, 0.5\%)$.

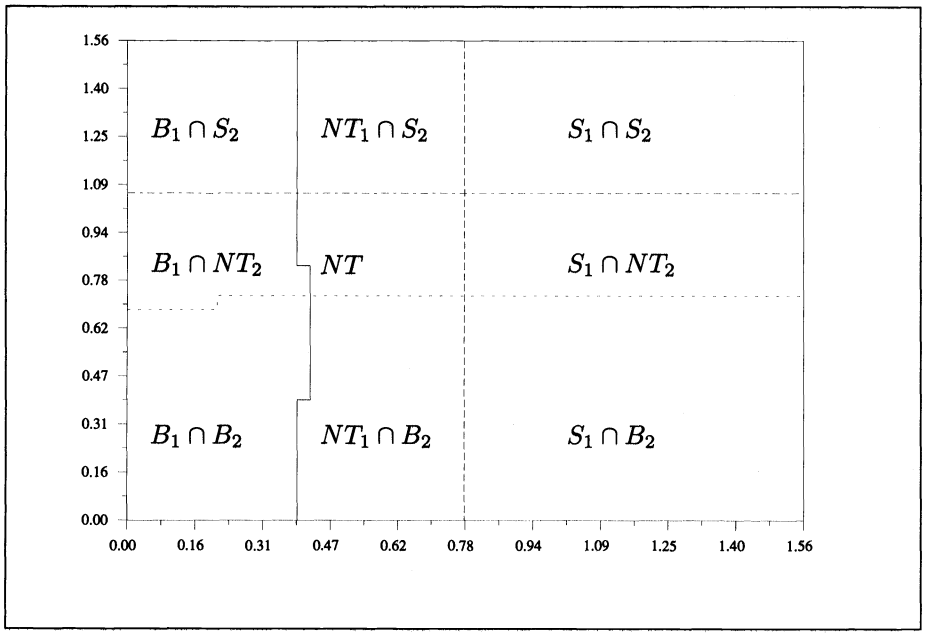


FIG. 10. *Boundaries of the regions B_i, S_i, and NT_i for $\gamma = 0.3$, $\delta = 10\%$, $r = 7\%$, $\alpha = (11\%, 15\%)$, $\sigma = (30\%, 35\%)$, $\nu = (1\%, 1\%)$.*

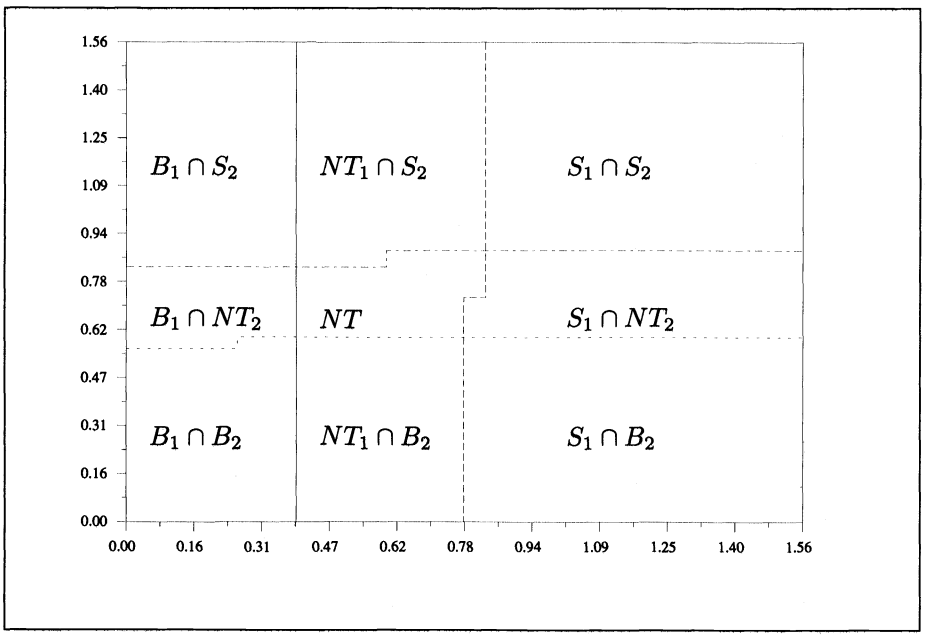


FIG. 11. *Boundaries of the regions B_i, S_i, and NT_i for $\gamma = 0.3$, $\delta = 10\%$, $r = 7\%$, $\alpha = (11\%, 20\%)$, $\sigma = (30\%, 50\%)$, $\nu = (1\%, 1\%)$.*

which is equivalent to

$$\begin{cases} W = \text{ constant}, \\ -\beta(y)W + u^*(\gamma W) \leq 0 \quad \forall y \in (\mathbb{R}^+)^n, \end{cases} \tag{88}$$

with $\beta(y)$ defined in (68). Uniqueness of the solution of (88) is not guaranteed since Assumption (A.2) is not satisfied, but the function W defined in (58) is the minimal solution of VI (87). Hence, we have

$$\max_{y \in (\mathbb{R}^+)^n} \{-\beta(y)W + u^*(\gamma W)\} = 0. \tag{89}$$

Equation (89) coincides with the Bellman equation of the problem where the proportion y_i is considered as a control variable (see [10]). Under Assumption (A.1), the optimal proportion denoted by π_i^* and called the Merton proportion is given by

$$\pi_i^* = \frac{\alpha_i - r}{\sigma_i^2(1-\gamma)}. \tag{90}$$

The optimal fraction of wealth dedicated to consumption is

$$C^* = \frac{1}{1-\gamma}\left(\delta - \gamma\left(r + \frac{1}{2(1-\gamma)}\sum_{i=1}^{n}\left(\frac{\alpha_i - r}{\sigma_i}\right)^2\right)\right),$$

and the value function W is equal to

$$W = \frac{C^{*(\gamma-1)}}{\gamma}.$$

The regions "sell i" and "buy i" are characterized by

$$B_i = \{y \in (\mathbb{R}^+)^n,\ y_i \leq \pi_i^*\},$$
$$S_i = \{y \in (\mathbb{R}^+)^n,\ y_i \geq \pi_i^*\}.$$

Note that these regions are not obtained by merely setting $\lambda = \mu = 0$ in (86) but by taking the limit of these expressions when λ and μ tend to 0.

7.2. A general shape of the transaction regions. In this section, we derive formally from VI (71), without numerical computation, the general shape of the transaction regions, given in Fig. 12. To that purpose, we assume the function W to be $\mathcal{C}^2$ in the interior of $(\mathbb{R}^+)^n$. Although this is not true in general, what is done below can be adapted by using the theory of viscosity solutions. This approach is used for example in Fleming and Soner [13] to obtain regularity results for the value function V and general properties of the transaction regions for $n = 1$.

From (71), we have $\tilde{M}_i W \leq 0$; in addition, the concavity of W implies that $\tilde{M}_i W = -\frac{\partial W}{\partial y_i}$ is nondecreasing with respect to y_i. Consequently, the region S_i defined in (86) can be written as

$$S_i = \{y \in (\mathbb{R}^+)^n,\ y_i \geq \pi_i^+(\hat{y})\},$$

where π_i^+ is some mapping of

$$\hat{y} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n).$$

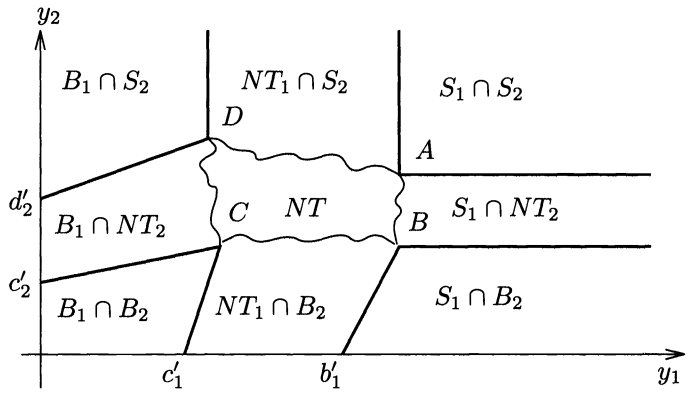


FIG. 12. *General shape of the transaction regions.*

To obtain a similar characterization for B_i, we consider another change of variables (ρ', y') obtained by substituting $-\lambda_i$ for μ_i in (55) for some fixed $i \in \{1, \dots, n\}$. Proceeding as above, and using Remark 6.1, we obtain

$$B_i = \{y \in (\mathbb{R}^+)^n,\ y_i' \leq \pi_i'^{-}(\hat{y}')\} \tag{91}$$

with

$$y_i' = \frac{(1+\nu_i)y_i}{1+\nu_i y_i} \quad \text{and} \quad \hat{y}' = \frac{1}{1+\nu_i y_i}\hat{y}, \tag{92}$$

where ν_i is defined in Remark 4.1. Since y_i' is non decreasing with respect to y_i, we get

$$B_i = \left\{ y \in (\mathbb{R}^+)^n,\ y_i \leq \pi_i^- \left(\frac{1}{1+\nu_i y_i}\hat{y} \right) \right\}.$$

Suppose $\pi_i^+ < +\infty$ and $\pi_i^- > 0$. This implies that π_i^+ and π_i^- are continuous functions and that the regions S_i and B_i are connected.

We restrict ourselves to the case $n = 2$, but what is done below can easily be generalized to $n > 2$.

In S_1, $\tilde{M}_1 W = -\frac{\partial W}{\partial y_1} = 0$. The function W is thus constant with respect to y_1 in S_1. Consequently the parts of the boundaries ∂B_2 and ∂S_2 included in S_1 are straight lines of equation $y_2 =$ constant. Similarly, using the change of variables (92) with $i = 1$, we infer that the parts of the boundaries ∂B_2 and ∂S_2 included in B_1 are straight lines of equation

$$y_2' = \frac{y_2}{1+\nu_1 y_1} = \text{constant}.$$

By symmetry, we get similar properties for the boundaries ∂B_1 and ∂S_1 as displayed in Fig. 12. No other property has been obtained for the boundary of NT.

A question which arises now is how is located the "Merton proportion" π^*. In general, π^* is not necessarily in the region NT. Nevertheless, we have the following proposition.

PROPOSITION 7.1. *We use the notation of Fig.* 12:

$$A = (a_1, a_2) = \partial S_1 \cap \partial S_2, \qquad B = (b_1, b_2) = \partial S_1 \cap \partial B_2,$$

$$C = (c_1, c_2) = \partial B_1 \cap \partial B_2, \qquad D = (d_1, d_2) = \partial B_1 \cap \partial S_2,$$

$$c_1' = \frac{c_1}{1+\nu_2 c_2}, \quad b_1' = \frac{b_1}{1+\nu_2 b_2}, \quad c_2' = \frac{c_2}{1+\nu_1 c_1}, \quad d_2' = \frac{d_2}{1+\nu_1 d_1}$$

and set

$$\tilde{\pi}_i^* = \begin{cases} \dfrac{\pi_i^*}{1+\nu_i - \nu_i \pi_i^*} & \text{if} \quad \pi_i^* < 1 + \dfrac{1}{\nu_i}, \\ +\infty & \text{otherwise.} \end{cases}$$

Then

$$\pi_1^* \le a_1, b_1', \quad \pi_2^* \le a_2, d_2'$$

and

$$d_1, c_1' \le \tilde{\pi}_1^*, \quad b_2, c_2' \le \tilde{\pi}_2^*.$$

Proof. We prove $\pi_i^* \le a_i$, $i = 1, 2$. The other inequalities are obtained similarly by using the change of variables (92). In $S_1 \cap S_2$, the function W is equal to a constant W_0 and satisfies (71), which reduces to

$$-\beta(y) W_0 + u^*(\gamma W_0) \le 0$$

with $\beta(y)$ given in (68). Hence,

$$-\beta(y) + (1-\gamma)\gamma^{\frac{1}{\gamma-1}} W_0^{\frac{1}{\gamma-1}} \le 0 \quad \forall y \in S_1 \cap S_2.$$

On the other hand, the point A is in $S_1 \cap S_2 \cap \overline{NT}$. Assuming that W is C^2 at point A, we obtain

$$\tilde{A} W + u^*(BW) = 0 \tag{93}$$

and

$$-\beta(A) + (1-\gamma)\gamma^{1/\gamma - 1} W_0^{\frac{1}{\gamma-1}} = 0.$$

Consequently

$$\beta(y) \ge \beta(A) \quad \forall y \in S_1 \cap S_2 = [a_1, +\infty) \times [a_2, +\infty).$$

As the function $\beta(y)$ is of the form $\beta_1(y_1) + \beta_2(y_2)$ with quadratic functions β_i, we get

$$\beta_i(y_i) \ge \beta_i(a_i) \quad \forall y_i \ge a_i.$$

Consequently, $a_i \ge \operatorname{Argmin} \beta_i = \pi_i^*$. □

7.3. Special case of no transaction cost for one of the risky assets. We suppose here $n = 2$, $\nu_1 = 0$, $\nu_2 > 0$. The VI (71) then reduces to

$$\max\left(\tilde{A}W + u^*(BW), \frac{\partial W}{\partial y_1}, \frac{\partial W}{\partial y_2} - \nu_2 BW, \max_{i=1,2,\, y_i>0}\left(-\frac{\partial W}{\partial y_i}\right)\right) = 0, \tag{94}$$

which implies that the function W is independent of y_1. Consequently the boundaries of B_2 and S_2 are horizontal straight lines of equation $y_2 = \pi_2^-$ and $y_2 = \pi_2^+$, respectively. Since equation (94) holds for all $y_1 \geq 0$ and W is the minimal solution of (94), we have

$$\max\left(\max_{y_1\geq 0}(\tilde{A}W + u^*(BW)), \frac{\partial W}{\partial y_2} - \nu_2 BW, -\frac{\partial W}{\partial y_2}\right) = 0 \quad \text{for } y_2 > 0.$$

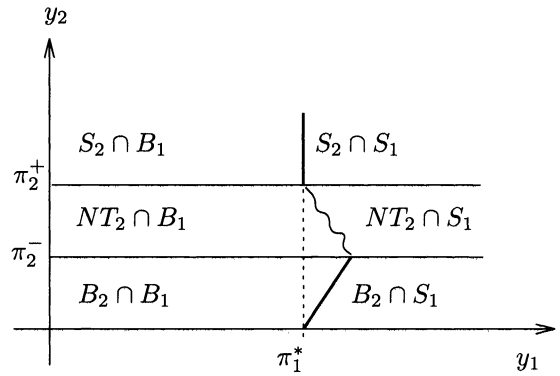


FIG. 13. *Boundaries of the transaction regions in the case of no transaction cost for the first risky asset.*

The regions B_1 and S_1 are delimited by the curve of equation $y_1 = \pi_1(y_2)$, where

$$\pi_1(y_2) = \operatorname*{Argmax}_{y_1\geq 0}(\tilde{A}W + u^*(BW))$$

is the solution of

$$y_1\sigma_1^2 y_2^2 \frac{\partial^2 W}{\partial y_2^2} - (2(\gamma-1)\sigma_1^2 y_1 + (\alpha_1 - r))y_2 \frac{\partial W}{\partial y_2} + \gamma((\alpha_1 - r) + (\gamma-1)\sigma_1^2 y_1)W = 0.$$

Consequently

$$\pi_1(y_2) = \frac{\pi_1^* BW}{BW + \dfrac{y_2}{1-\gamma}\dfrac{\partial BW}{\partial y_2}}.$$

In particular $\pi_1(0) = \pi_1^*$ and $\pi_1(y_2) = (1 + \nu_2 y_2)\pi_1^*$ in B_2. In S_2, W is constant and $\pi_1(y_2) = \pi_1^*$ (see Fig. 13). Moreover, by using the concavity of W, we obtain the estimate

$$0 < \pi_1(y_2) \leq \frac{\pi_1^*}{1 - \nu_2 y_2}.$$

REFERENCES

[1] M. AKIAN, *Analyse de l'algorithme multigrille FMGH de résolution d'équations d'Hamilton-Jacobi-Bellman*, in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Control and Information Sciences, 144, Springer-Verlag, New York, 1990, pp. 113–122.

[2] ———, *Méthodes multigrilles en contrôle stochastique*, Thèse de l'Université Paris IX-Dauphine, Paris, France, 1990.

[3] G. BARLES AND P. E. SOUGANIDIS, *Convergence of approximation schemes for fully nonlinear second-order equations*, Asymptotic Anal., 4 (1991), pp. 271–283.

[4] R. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

[5] ———, *Introduction to the Mathematical Theory of Control Processes*, Academic Press, New York, 1971.

[6] J. PH. CHANCELIER, C. GOMEZ, J. P. QUADRAT, AND A. SULEM, *Automatic study in stochastic control*, in Stochastic Differential Systems, Stochastic Control Theory and Applications, W. Fleming and P. L. Lions, eds., IMA Vol. Math. Appl., 10, Springer-Verlag, New York, 1987, pp. 79–86.

[7] P. L. CHOW, J. L. MENALDI, AND M. ROBIN, *Additive control of stochastic linear systems with finite horizon*, SIAM J. Control Optim., 23 (1985), pp. 858–899.

[8] G. M. CONSTANTINIDES, *Capital market equilibrium with transaction costs*, J. of Political Economy, 94 (1986), pp. 842–862.

[9] M. G. CRANDALL, H. ISHII, AND P. L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[10] M. DAVIS AND A. NORMAN, *Portfolio selection with transaction costs*, Math. Oper. Res., 15 (1990), pp. 676–713.

[11] N. EL KAROUI, *Les aspects probabilistes du contrôle stochastique*, Lectures Notes in Mathematics, 876, Springer-Verlag, New York, (1981), pp. 513–537.

[12] B. FITZPATRICK AND W. H. FLEMING, *Numerical methods for an optimal investment-consumption model*, Math. Oper. Res., 16 (1991), pp. 823–841.

[13] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[14] W. HACKBUSCH, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, Heidelberg, 1985.

[15] W. HACKBUSCH AND U. TROTTENBERG, EDS., *Multigrid Methods*, Lecture Notes in Mathematics, 960, Springer-Verlag, New York, 1981.

[16] R. A. HOWARD, *Dynamic Programming and Markov Process*, MIT Press, Cambridge, MA, 1960.

[17] I. KARATZAS, J. LEHOCZKY, S. SETHI, AND S. SHREVE, *Explicit solution of a general consumption/investment problem*, Math. Oper. Res., 11 (1986), pp. 261–294.

[18] H. J. KUSHNER, *Probability Methods in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.

[19] P. L. LIONS *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations, Part* 1: *The dynamic programming principle and applications*, Comm. Partial Differential Equations, 8 (1983), pp. 1101–1174.

[20] ———, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations, Part* 2: *Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1229–1276.

[21] M. J. P. MAGILL AND G. M. CONSTANTINIDES, *Portfolio selection with transaction costs*, J. Econ. Theory, 13 (1976), pp. 245–263.

[22] S. F. MCCORMICK, ED., *Multigrid Methods*, Frontiers in Applied Mathematics, 5, Society for Industrial and Applied Mathematics, Philadelphia, 1987.

[23] J. L. MENALDI AND M. ROBIN, *On some cheap control problems for diffusion processes*, Trans. Am. Math. Soc., 278 (1983), pp. 771–802.

[24] R. C. MERTON, *Optimum consumption and portfolio rules in a continuous time model*, J. Economic Theory, 3 (1971), pp. 373–413.

[25] P. A. MEYER, *Un cours sur les intégrales stochastiques*, Séminaire de Probabilités. Lectures Notes in Mathematics, 511, Springer-Verlag, Berlin, 1976, pp. 245–400.

[26] M. NISIO, *On non linear semigroup attached to stochastic optimal control*, Publ. Res. Ins. Math. Sci., 12 (1976), pp. 513–537.

[27] S. SETHI AND M. TAKSAR, *A note on Merton's "Optimum consumption and portfolio rules in continuous-time model,"* J. Econ. Theory, 46 (1988), pp. 395–401.

[28] S. E. SHREVE AND H. M. SONER, *Optimal investment and consumption with transaction costs*,

Ann. Appl. Probab., 4 (1994), pp. 909–962.

[29] S. E. SHREVE, H. M. SONER, AND V. XU, *Optimal investment and consumption with two bonds and transaction costs*, Math. Finance, 1 (1991), pp. 53–84.

[30] A. SULEM, *Application of stochastic control to portfolio selection with transaction costs*, Rapport de recherche INRIA, 1062 (1989).

[31] M. TAKSAR, M. J. KLASS AND D. ASSAF, *A diffusion model for optimal portfolio selection in the presence of brokerage fees*, Math. Oper. Res., 13 (1988), pp. 277–294.

[32] T. ZARIPHOPOULOU, *Investment-consumption model with transaction fees and Markov chain parameters*, SIAM J. Control Optim., 30 (1992), pp. 613–636.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 365–388, January 1996

016

ADAPTIVE CONTROL VIA A SIMPLE SWITCHING ALGORITHM*

JI FENG ZHANG† AND PETER E. CAINES‡

Abstract. In this paper we present an adaptive stabilization control for systems with unknown constant parameters and stochastic disturbances, which may be neither open-loop stable nor minimum phase. The ideas come from previous works [J. F. Zhang and H. F. Chen, *Adaptive stabilization under the weakest condition*, Proc. 31st Control and Design Conference, December 14–18, 1992, pp. 3620–3621, and H. F. Chen, *Continuous-Time Stochastic Adaptive Control Stabilizing the System and Minimizing the Quadratic Loss Function*, Tech. Report, Institute of Systems Science, Academia Sinica, Beijing, 1992], but here we not only simplify the construction procedure of an adaptive control but also reduce the computational load significantly, so that the adaptive control in this paper is more practical. Furthermore, parameter estimation is carried out in only a finite time period and, unlike previous work, the parameter estimates are generated by ordinary differential equations rather than stochastic differential equations.

Key words. adaptive control, parameter estimation, switching algorithm, continuous time, stochastic system

AMS subject classifications. 93C40, 93E15, 93E35

1. Introduction.

The switching control strategies of Zhang and Chen [1991] and Chen [1992] show that an alternation of excitation and control regimes can yield stabilizing controls. The idea is that, if a certain prediction error test fails at a specified instant, then a signal which is (in the limit) persistently exciting is applied. On the other hand, if the test is passed, then a particular certainty equivalence control law using the current estimate is applied. This strategy has common-sense appeal, despite the fact that the laws are somewhat complex in their present form. It is shown in the analysis of these laws that eventually the prediction error tests must always be passed, and hence it is shown that the system "locks on" to an acceptable control law. In summary, the adaptive control algorithms used in Zhang and Chen [1992] and Chen [1992] are as follows:

Step A) Introduce an appropriate criterion to judge whether or not the parameter estimate is satisfactory (for instance, a prediction error criterion).

Step B) Apply an excitation signal to the system, and estimate the unknown parameters via a least-squares (or related) algorithm until a "satisfactory" estimate is obtained according to the criterion; and after this,

Step C) construct a control law via the previously obtained "satisfactory" parameter estimates and use this law to control the system until some "unsatisfactory" property appears according to the criterion; and then

Step D) repeat this procedure through Steps B) and C).

If no "unsatisfactory" property appears at some stage in Step C), then the designed adaptive control law is used forever.

It is worth noticing that in some previous works (i) one or both of the derivatives dx_t and dy_t of the system state x and observation process y are required to be measurable in the parameter estimation procedure (see, e.g., Caines [1992]; Chen [1992];

* Received by the editors November 25, 1992; accepted for publication (in revised form) October 7, 1994. This research was supported by Canadian NSERC grant A 1329.

† Canadian Institute for Advanced Research, and Institute of Systems Science, Academia Sinica, Beijing 100080, People's Republic of China.

‡ Department of Electrical Engineering, McGill University, 3480 University Street, Montreal H3A 2A7, Canada.

Chen and Guo [1990]; Chen and Moore [1987]; Duncan and Pasik-Duncan [1990], [1991]; Gevers, Goodwin, and Wertz [1991]; Goodwin, et al. [1991]; Moore [1988]; Christopeit [1986]); (ii) the criteria used in steps A through C have to be verified at all time instants, which is an uncountable procedure because of the nature of the continuous time model (see, e.g., Chen [1992] and Zhang and Chen [1991]); (iii): the unknown parameters are always estimated no matter whether they are needed or not (see, e.g., Chen [1992]; Chen and Zhang [1992]; and Zhang and Chen [1991]); (iv) some external stochastic excitation signals are invoked (see, e.g., Chen and Zhang [1992] and Zhang and Chen [1992]).

In this paper, we formulate an adaptive control algorithm which (i) avoids use of dx_t or dy_t and the introduction of external stochastic signals in the procedures of parameter estimation and adaptive control, (ii) simplifies the criteria in steps A through C so that they are required to be verified at discrete time instants only, (iii) stop the parameter estimation procedure when it is not needed in order to make the adaptive control law more practical, and finally, (iv) does not use an external stochastic excitation signal. (In effect the Brownian motion w driving the system is exploited for this purpose.)

It may be conjectured that such an alternation of identification and control regimes will work in certain time-varying cases.

2. Full observation systems. In this section we consider the LQ adaptive control problem for the following system model:

$$dx_t = Ax_t dt + Bu_t dt + Cdw_t, \quad t \geq 0, \tag{2.1}$$

where $x_t \in \mathbb{R}^n$ and $u_t \in \mathbb{R}$ are the state and input of the system and $\{w_t, \mathcal{F}_t\}$ is a standard Wiener process in $\mathbb{R}^m$.

Using controls which at any instant t are based only on information available up to time t, we wish to stabilize the system (2.1). In this paper, this is achieved by the use of controls which are certain time-interleaved versions of an excitation signal and a signal designed via the certainty equivalence principle for the following quadratic loss function:

$$\min_{u \in \mathcal{U}} \limsup_{t \to \infty} J_t(u), \tag{2.2}$$

where

$$\mathcal{U} = \left\{ \{u_t : \ u_t \in \mathcal{F}_t \triangleq \sigma\{x_\tau, \ u_s, \ 0 \leq \tau \leq t, \ 0 \leq s < t\}, \ t \geq 0 \right\}, \tag{2.3}$$

$$J_t(u) = \frac{1}{t} \int_0^t (x_s^T Q_1 x_s + Q_2 u_s^2) ds, \quad Q_1 \geq 0, \quad Q_2 > 0. \tag{2.4}$$

This problem has been investigated in previous work; see, for instance, Zhang and Chen [1991], Chen [1992], where these authors presented the first rigorous stability analysis for such adaptive stabilization of a system, which might be neither open-loop stable nor minimum phase and might be subject to disturbances with an unknown bound.

Specification of the adaptive control law. First, we define a causal system which shall generate a disturbance input u', which shall be employed over an at most countable set of intervals; second, we define a linear state feedback control law, which shall be used over the intervals which interleave those during which the disturbance

is used; then, third, we give a rule for determining the switching times which depend solely upon the history of the system inputs and outputs.

Following the procedure described from step A to step D, we now find an adaptive stabilization control for model (2.1).

Assume that the control input u is defined on the interval $[0,t)$; then the input u' and a countable sequence of stopping times with no finite accumulation point are defined as follows:

Let $T>1$ and α be positive constants chosen arbitrarily. Define for $i=1, 2, \ldots, n+1$

$$\beta_i=(-1)^{i+1}\alpha^i\frac{(n+1)!}{i!\times(n+1-i)!} \quad \text{with} \quad 0!\triangleq 1,\ i!\triangleq 1\times 2\times\cdots\times i \tag{2.5}$$

and for some integer $k\geq 0$

$$u'_t=L_{T^k}+\beta_1 S_k u'_t+\cdots+\beta_{n+1}S_k^{n+1}u'_t,\quad t\in[T^k,\ T^{k+1}), \tag{2.6}$$

where $S_k u_t=\int_{T^k}^t u_s ds$ and $L_t=1+\int_0^t(\|\psi_s\|^2+\|z_s\|^2+u_s^2)ds$ with $\psi_t=[Sx_t^\tau,\ Su_t]^\tau$ and $z_t=[S\overline{x}_t^\tau,\ S\overline{u}_t]^\tau$. Here, S is the integral operator $Sx_t=\int_0^t x_s ds$; x_t and u_t are, respectively, the system state and system input, which is recursively given by (2.5)–(2.12); $\overline{x}_t$ and $\overline{u}_t$ are the solutions of the following equations, respectively:

$$(S+1)\overline{x}_t=x_t,\quad (S+1)\overline{u}_t=u_t,$$

i.e.,

$$\overline{x}_t=x_t-\int_0^t e^{-(t-\lambda)}x_\lambda d\lambda,\quad \overline{u}_t=u_t-\int_0^t e^{-(t-\lambda)}u_\lambda d\lambda.$$

It will be seen below that the function u which appears in the definitions above is equal to u' during the time intervals when the excitation input to the system is in use and is given as a linear function of the state x during the periods when u' is not being used as a system input.

Set $\theta=[A,\ B]^\tau$. Choosing an arbitrary θ_0, the unknown parameter θ is estimated via the least-squares method, which is modified so as to be active only over a sequence of intervals $[T^{\tau_{i-1}},T^{\sigma_i})$. Specifically, the estimate $\theta_t=[A_t,\ B_t]^\tau$ is given by

$$\dot{\theta}_t=P_t\psi'_t(\overline{x}_t-\theta_t^\tau\psi'_t)\quad \text{with}\quad P_t=\left(I+\int_0^t\psi'_s\psi_s'^\tau ds\right)^{-1} \tag{2.7}$$

and

$$\psi'_t=\begin{cases} z_t & \text{if } t\in[0,T^{\tau_0}) \text{ or } t\in[T^{\tau_{i-1}},T^{\sigma_i}) \\ 0 & \text{if } t\in[T^{\sigma_i},T^{\tau_i})\end{cases}\quad \forall i\geq 1, \tag{2.8}$$

where $\{\tau_i\}$ and $\{\sigma_i\}$ are two stopping time sequences defined as follows: $0=\tau_0<\sigma_1<\tau_1<\sigma_2<\tau_2<\cdots$,

$$\begin{aligned}\sigma_i=\inf\{k>\tau_{i-1}:\ & k\in\mathcal{N},\quad (A_{T^k},B_{T^k},D) \text{ is controllable and observable,}\\ & \text{where here, and hereafter, } \mathcal{N} \text{ denotes the set of all positive}\\ & \text{integers, and } D \text{ is any square matrix such that } D^\tau D=Q_1\},\end{aligned} \tag{2.9}$$

$$\tau_i = \inf\left\{k > \sigma_i : k \in \mathcal{N}, \quad \int_0^{T^k} \|x_s\|^2 ds > T^{\sigma_i} \int_0^{T^{\sigma_i}} (\|x_s\|^2 + \|u_s''\|^2) ds \right.$$

$$\left. + T^{\sigma_i} T^k + T^{\sigma_i} \right\} \tag{2.10}$$

with the excitation input u'' given by

$$u_t'' = \begin{cases} u_t' & \text{if } t \in [0, T^{\tau_0}) \text{ or } t \in [T^{\tau_{i-1}}, T^{\sigma_i}) \\ 0 & \text{if } t \in [T^{\sigma_i}, T^{\tau_i}) \end{cases} \quad \forall i \geq 1. \tag{2.11}$$

Here Q_1 is given in (2.4) and u_t' is defined by (2.6).

The adaptive control u is generated by interleaving the excitation input u' and a linear feedback input as follows:

$$u_t = \begin{cases} u_t' & \text{if } t \in [0, T^{\tau_0}) \text{ or } t \in [T^{\tau_{i-1}}, T^{\sigma_i}) \text{ for some } i \geq 1, \\ -Q_2^{-1} B_{T^{\sigma_i}}^\tau R_{T^{\sigma_i}} x_t & \text{if } t \in [T^{\sigma_i}, T^{\tau_i}) \text{ for some } i \geq 1, \end{cases} \tag{2.12}$$

where Q_2 is the positive constant in (2.4), $B_{T^{\sigma_i}}$ is the estimate for B at time instant T^{σ_i} given by (2.7) and (2.8), and $R_{T^{\sigma_i}}$ is a solution of the following algebraic Riccati equation:

$$A_{T^{\sigma_i}}^\tau R_{T^{\sigma_i}} + R_{T^{\sigma_i}} A_{T^{\sigma_i}} - R_{T^{\sigma_i}} B_{T^{\sigma_i}} Q_2^{-1} B_{T^{\sigma_i}}^\tau R_{T^{\sigma_i}} + D^\tau D = 0.$$

Here $A_{T^{\sigma_i}}$ is the estimate for A at time instant T^{σ_i} given by (2.7) and (2.8).

Remark 2.1. From the definition (2.12) of u_t, it is easy to see that τ_i and σ_i are Markov times, i.e., $\sigma\{T^{\tau_i} \leq t\} \in \mathcal{F}_t$ and $\sigma\{T^{\sigma_i} \leq t\} \in \mathcal{F}_t$. Thus, $u \in \mathcal{U}$ and $\psi_t' \in \mathcal{F}_t$.

Remark 2.2. The excitation signal in (2.12) is generated by (2.6), in which a designer needs only determine the deterministic coefficients β_i. No additional stochastic signal is introduced except L_{T^k}, so we call this a deterministic-like excitation signal.

Remark 2.3. In (2.9), to get σ_i, the only thing one should do is to check the controllability and observability of (A_{T^k}, B_{T^k}, D) for every integer k at time instant T^k. While in (2.10), one need only check whether or not

$$\int_0^{T^k} \|x_s\|^2 ds > T^{\sigma_i} \int_0^{T^{\sigma_i}} (\|x_s\|^2 + \|u_s''\|^2) ds + T^{\sigma_i} T^k + T^{\sigma_i}$$

for every integer k at time instant T^k and such a set of time instants is evidently countable.

Remark 2.4. By (2.7) and (2.8) it is easy to see that $\theta_t = \theta_{T^{\sigma_i}}$ for all $t \in [T^{\sigma_i}, T^{\tau_i})$. In other words, unlike in Zhang and Chen [1991] or in Chen [1992], the LS parameter estimation is not carried out in the time interval $[T^{\sigma_i}, T^{\tau_i})$. Thus, if adaptive control (2.12) results in an integer i such that $\sigma_i < \infty$ and $\tau_i = \infty$, then the unknown parameter estimates will be locked on an acceptable value $\theta_{T^{\sigma_i}}$ forever.

LEMMA 2.1. *Let* $\lambda_{\min}^{(t)}$ *denote the smallest eigenvalue of matrix* P_t^{-1}. *Then the parameter estimate* θ_t *given by* (2.7)–(2.12) *has the following property:*

$$\|\theta_t - \theta\|^2 \leq \frac{c(t+1)}{\lambda_{\min}^{(t)}} \quad \forall t \geq 0,$$

where here and hereafter $c \geq 0$ *is a possibly random quantity which is independent of* t.

Proof. Let $\tilde{\theta}_t = \theta_t - \theta$. Then from (2.1) it follows that

$$x_t = x_0 + \theta^\tau \psi_t + Cw_t.$$

Substituting this into the first equation of (2.7) and noting (2.8), $(S+1)\overline{x}_t = x_t$, and $(S+1)z_t = \psi_t$ we get

$$\dot{\tilde{\theta}}_t = -P_t\psi_t'\psi_t'^\tau\tilde{\theta}_t + P_t\psi_t'[(S+1)^{-1}(Cw_t) + \varepsilon_t],$$

where here and hereafter ε_t denotes a time function which exponentially converges to zero.

From this and the second equation of (2.7) we obtain

$$\frac{d(\tilde{\theta}_t^\tau P_t^{-1}\tilde{\theta}_t)}{dt} = -(\tilde{\theta}_t^\tau\psi_t')^2 + 2\tilde{\theta}_t^\tau\psi_t'[(S+1)^{-1}(Cw_t) + \varepsilon_t],$$

which implies that

$$\begin{aligned} 0 &\le \tilde{\theta}_t^\tau P_t^{-1}\tilde{\theta}_t \\ &= \tilde{\theta}_0^\tau P_0^{-1}\tilde{\theta}_0 - \int_0^t (\tilde{\theta}_s^\tau\psi_s')^2 ds + 2\int_0^t \tilde{\theta}_s^\tau\psi_s'[(S+1)^{-1}(Cw_s) + \varepsilon_s]ds \\ &\le \tilde{\theta}_0^\tau P_0^{-1}\tilde{\theta}_0 + \int_0^t [(S+1)^{-1}(Cw_s) + \varepsilon_s]^2 ds = O(t), \end{aligned}$$

where for the last inequality we have invoked $\int_0^t[(S+1)^{-1}(Cw_s)]^2 ds = O(t+1)$ a.s. (e.g., Chen and Guo [1990]).

Therefore, Lemma 2.1 is true. □

LEMMA 2.2. *In system* (2.1), *if* (A, B) *is controllable and* $u_t = u_t'$ *for some* k *and all* $t \in (T^k, T^{k+1}]$, *then there exist* $c > 0$, $a > 1$, *and* $k_0 > 0$ *such that*

$$\lambda_{\min}^{(T^{k+1})} \ge ca^{T^{k+1}} L_{T^k} \quad \forall k \ge k_0. \tag{2.13}$$

Proof. See Appendix A.

THEOREM 2.1. *If* $\mathrm{Span}(B) \subset \mathrm{Span}(C)$ *and* (A, B, D) *is controllable and observable with* $D^\tau D = Q_1$, *then under the adaptive control* (2.12) *it is the case that*

(a) *there is an integer* i *such that* $\sigma_i < \infty$, $\tau_i = \infty$, *and* $\theta_t = \theta_{T^{\sigma_i}}$ $\forall t \ge T^{\sigma_i}$;

(b) $\lim_{t\to\infty} \frac{1}{t}\int_0^t x_s x_s^\tau ds$ *and* $\lim_{t\to\infty} \frac{1}{t}\int_0^t u_s^2 ds$ *exist and are finite a.s.*

Here and hereafter by $\mathrm{Span}(X)$ *we mean the linear space spanned by the column vectors of* X.

Proof. Let R be the solution of the following algebraic Riccati equation:

$$A^\tau R + RA - RBQ_2^{-1}B^\tau R + D^\tau D = 0.$$

Then it is well known that $\Phi \triangleq A - BQ_2^{-1}B^\tau R$ is stable, and hence there is a positive matrix P such that

$$P\Phi + \Phi^\tau P = -I.$$

From this, it is easy to see that there exists a small enough positive constant ε such that

$$P\Phi_t + \Phi_t^\tau P \le -\frac{1}{2}I \quad \forall \Phi_t \in \{\Phi_t : \|\Phi_t - \Phi\| < \varepsilon\}. \tag{2.14}$$

We now define

$$\Phi_t = \begin{cases} A & \text{if } t \in [0, T^{\tau_0}) \text{ or } t \in [T^{\tau_{i-1}}, T^{\sigma_i}), \\ A - BQ_2^{-1}B_{T^{\sigma_i}}^{\tau} R_{T^{\sigma_i}} & \text{if } t \in [T^{\sigma_i}, T^{\tau_i}). \end{cases}$$

Then the adaptive control system defined above is expressed as

$$dx_t = \Phi_t x_t dt + Bu_t'' dt + C dw_t.$$

From the definitions (2.9) and (2.10) of sequences $\{\sigma_i\}$ and $\{\tau_i\}$ we see that only three cases can possibly hold. The first case is that there exists an integer i such that $\tau_{i-1} < \infty$ and $\sigma_i = \infty$; the second is that $\sigma_i < \tau_i < \infty$ for every integer $i \geq 1$, and the third is there exists an integer i such that $\sigma_i < \infty$ and $\tau_i = \infty$. We shall now show the first two cases are impossible.

Case 1. It is impossible that an integer i such that $\tau_{i-1} < \infty$ and $\sigma_i = \infty$ exists.

From Lemmas 2.1 and 2.2, it is easy to see that if there were an integer i such that $\tau_{i-1} < \infty$ and $\sigma_i = \infty$, then there would be

$$A_{T^k} \xrightarrow[k \to \infty]{} A \text{ and } B_{T^k} \xrightarrow[k \to \infty]{} B.$$

Thus, by the assumption that (A, B, D) is controllable and observable, we see that there would exist a k such that $T^k \geq T^{\tau_{i-1}}$ and (A_{T^k}, B_{T^k}, D) is controllable and observable. This contradicts $\sigma_i = \infty$.

Case 2. It is impossible that $\sigma_i < \tau_i < \infty$ for every integer $i \geq 1$.

If for every integer $i \geq 1$, $\sigma_i < \tau_i < \infty$, then by Lemmas 2.1 and 2.2 it is easy to see that $\Phi_{T^{\sigma_i}} \xrightarrow[i \to \infty]{} \Phi$. Therefore, there exists i_0 such that for all $i \geq i_0$,

$$\|\Phi_{T^{\sigma_i}} - \Phi\| < \varepsilon,$$

which together with (2.14) implies that

$$\|\Phi_t\| \leq c \quad \text{and} \quad P\Phi_{T^{\sigma_i}} + \Phi_{T^{\sigma_i}}^{\tau} P \leq -\frac{1}{2} I. \tag{2.15}$$

Using Ito's formula (cf. Schwartz [1984]) we find that for $k \in [\sigma_i, \tau_i) \cap \mathcal{N}$

$$\begin{aligned} x_{T^k}^{\tau} P x_{T^k} \leq x_0^{\tau} P x_0 + \int_0^{T^{\sigma_i}} x_s^{\tau}(P\Phi_s + \Phi_s^{\tau} P) x_s ds - \frac{1}{2}\int_{T^{\sigma_i}}^{T^k} \|x_s\|^2 ds \\ +2\int_0^{T^{\sigma_i}} x_s^{\tau} P B u_s'' ds + 2\int_0^{T^k} x_s^{\tau} P C dw_t + \operatorname{tr}(C^{\tau} P C) T^k, \end{aligned} \tag{2.16}$$

where here and hereafter $\operatorname{tr}(X)$ denotes the trace of X.

Note that by Lemma 4 of Christopeit [1986], there exist random numbers c', c'', independent of t, such that for all k sufficiently large, say, for all $k \geq \sigma_m$,

$$\int_0^{T^k} x_s^{\tau} P C dw_s \leq c' \left(\int_0^{T^k} \|x_s^{\tau}\|^2 ds\right)^{\eta + \frac{1}{2}} + c' \leq c'' + \frac{1}{8}\int_0^{T^k} \|x_s^{\tau}\|^2 ds, \quad \eta \in \left(0, \frac{1}{2}\right).$$

Then from (2.16), for some random number c''' independent of time, we have

$$\begin{aligned} x_{T^k}^{\tau} P x_{T^k} \leq c''' + c''' \int_0^{T^{\sigma_m}} \|x_s\|^2 ds - \frac{3}{8}\int_{T^{\sigma_m}}^{T^k} \|x_s\|^2 ds \\ +c\int_0^{T^{\sigma_m}} \|u_s''\|^2 ds + \operatorname{tr}(C^{\tau} P C) T^k, \end{aligned}$$

and hence, for some random number c independent of time,

$$\int_0^{T^k} \|x_s\|^2 ds \le c \int_0^{T^{\sigma_m}} (\|x_s\|^2 + \|u_s''\|^2) ds + cT^k + c \quad \forall k \ge \sigma_m. \tag{2.17}$$

Now there exists i sufficiently large that $T^{\sigma_i} \ge c$, and so by (2.17) this gives

$$\int_0^{T^k} \|x_s\|^2 ds \le T^{\sigma_i} \int_0^{T^{\sigma_i}} (\|x_s\|^2 + \|u_s''\|^2) ds + T^{\sigma_i} T^k + T^{\sigma_i} \quad \forall k \ge \sigma_i,$$

which contradicts $\tau_i < \infty$.

So, there must exist an integer i such that $\sigma_i < \infty$ and $\tau_i = \infty$.

Thus by (2.7) and (2.8) we get Assertion (a) of Theorem 2.1. Furthermore, (2.10) together with $\tau_i = \infty$ implies that

$$\limsup_{k\to\infty} \frac{1}{T^k} \int_0^{T^k} \|x_s\|^2 ds < \infty \quad \text{a.s.} \tag{2.18}$$

Notice that $\text{Span}(B) \subset \text{Span}(C)$ and (A, B) is controllable implies that $(\Phi_{T^{\sigma_i}}, C)$ is controllable. Then by (2.18) and Lemma B.1 in Appendix B we see that $\Phi_{T^{\sigma_i}}$ is stable. Therefore, from Lemma 3 of Chen and Guo [1990] it follows that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t x_s x_s^\tau ds \quad \text{exists and is finite a.s.},$$

which together with (2.12) implies that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t u_s^2 ds = Q_2^{-2} B_{T^{\sigma_i}}^\tau R_{T^{\sigma_i}} \left(\lim_{t\to\infty} \frac{1}{t} \int_0^t x_s x_s^\tau ds \right) (B_{T^{\sigma_i}}^\tau R_{T^{\sigma_i}})^\tau \text{ exists and is finite a.s.}$$

This proves Assertion (b) of Theorem 2.1. □

3. Partially observed system.

3.1. Problem statement. In this section we consider the single-input single-output continuous-time system described by

$$A(S)y_t = y_0 + SB(S)u_t + C(S)w_t + S\eta_t \quad \forall t \ge 0, \tag{3.1}$$

where $A(S)$, $B(S)$, and $C(S)$ are polynomials in S with unknown coefficients:

$$A(S) = 1 + \sum_{i=1}^{p} a_i S^i, \quad B(S) = \sum_{i=1}^{q} b_i S^{i-1}, \quad C(S) = \sum_{i=0}^{l} c_i S^i; \tag{3.2}$$

$\{w_t, \mathcal{F}_t\}$ is a standard Wiener process with respect to a nondecreasing σ-algebras $\{\mathcal{F}_t\}$ defined on a probability space; y_t and u_t are the system output and input, respectively, and measurable with respect to $\mathcal{F}_t$; and η_t is unknown disturbance or unmodeled dynamics which is measurable with respect to $\mathcal{F}_t$.

As Zhang and Chen (1991) show, model (3.1) subject to (3.2) is very general and includes some widely used models. For instance, in the case where $l = p$ and $c_p = ga_p$, the input-output properties of (3.1) are equivalent to those of the following well-known

state-space representation (e.g., Gevers, Goodwin, and Wertz [1991]; Goodwin, et al. [1991]; Caines [1992]):

$$dx_t = Ax_t dt + Bu_t dt + Cdw_t + D\eta_t dt,$$
$$dy_t = D^\tau x_t dt + gdw_t,$$

with

$$A = \begin{bmatrix} -a_1 & 1 & & \\ -a_2 & & \ddots & \\ \vdots & & & 1 \\ -a_m & & & 0 \end{bmatrix}, \quad B = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}, \quad C = \begin{bmatrix} c_0 - ga_0 \\ \vdots \\ c_{m-1} - ga_{m-1} \end{bmatrix}, \quad D = \left.\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right\} m,$$

where $m = \max\{p, q\}$ and here and hereafter we set $a_0 = 1$, $a_i = 0$ for $i > p$, $b_j = 0$ for $j > q$, and $c_k = 0$ for $k > l$.

Let us denote the collection of unknown coefficients of $A(S)$ and $B(S)$ by θ:

$$\theta = [-a_1, \ldots, -a_p, \ b_1, \ldots, b_q]^\tau. \tag{3.3}$$

Let

$$F(S) = 1 + f_1 S + \cdots + f_{l+1} S^{l+1} \quad \text{with } f_{l+1} \neq 0$$

be an arbitrarily given stable polynomial of S; i.e., every S that satisfies $F(S) = 0$ has negative real part.

Denote by y_t^f and u_t^f the filtered value, respectively:

$$F(S)y_t^f = y_t, \quad F(S)u_t^f = u_t \tag{3.4}$$

and

$$\varphi_t^f = [Sy_t^f, \ldots, S^p y_t^f, \quad Su_t^f, \ldots, S^q u_t^f]^\tau. \tag{3.5}$$

Define

$$\varphi_t' = \begin{cases} \varphi_t^f & \text{if } t \in [0, T^{\tau_0}) \text{ or } t \in [T^{\tau_{i-1}}, T^{\sigma_i}) \text{ for some } i \geq 1, \\ 0 & \text{if } t \in [T^{\sigma_i}, T^{\tau_i}) \text{ for some } i \geq 1, \end{cases} \tag{3.6}$$

where $\{\tau_i\}$ and $\{\sigma_i\}$ are two stopping time sequences such that $\varphi_t^f \in \mathcal{F}_t$.

Then the unknown parameter θ is estimated as follows:

$$\dot{\theta}_t = P_t \varphi_t' (y_t^f - \theta_t^\tau \varphi_t') \quad \text{with} \quad P_t = \left(I + \int_0^t \varphi_s' \varphi_s'^\tau ds \right)^{-1}, \tag{3.7}$$

where θ_0 is a constant chosen arbitrarily.

The purpose of this paper is to design a θ_t-based adaptive control so that the closed-loop system is stabilized in the sense that

$$\sup_{t \geq 0} \frac{1}{t+1} \int_0^t (y_s^2 + u_s^2) ds < \infty \quad \text{a.s.} \tag{3.8}$$

under the following assumptions:

A.1. $A(S)$ and $SB(S)$ are coprime, $b_q \neq 0$, $l \leq \min\{p, q-1\}$, and p and q are known.

A.2. $\sup_{t\geq 0} \frac{1}{t+1} \int_0^t \eta_s^2 ds < \infty$.

From Zhang and Chen [1991] it follows that Assumptions A.1 and A.2 are as weak as the following necessary and sufficient ones even when θ is known:

A.1′. The greatest common factor of $A(S)$ and $SB(S)$ is 1 or a stable polynomial, $b_q \neq 0$, the order of the greatest common factor, $l \leq \min\{p, q-1\}$, and p and q are known.

A.2′. $\sup_{t\geq 0} \frac{1}{t+1} \int_0^t \eta_s^2 ds < \infty$ a.s.

However, for simplicity of notation, in this paper we use Assumptions A.1 and A.2.

Remark 3.1. We now look at how to calculate the filtered values y_t^f and u_t^f of y_t and u_t, respectively, with respect to filter $F(S)$ in (3.4).

Let

$$D_F = \begin{bmatrix} -f_1 & -f_2 & \cdots & -f_{l+1} \\ 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & 1 & 0 \end{bmatrix}, \quad Y_t = \begin{bmatrix} y_t^f \\ Sy_t^f \\ \vdots \\ S^l y_t^f \end{bmatrix}, \quad H_l = \left.\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right\} l+1.$$

Then from (3.4) we see that

$$Y_t = D_F S Y_t + H_l y_t,$$

which is equivalent to

$$Y_t = D_F \int_0^t e^{D_F(t-\lambda)} H_l y_\lambda d\lambda + H_l y_t.$$

Thus we have

$$y_t^f = H_l^\tau Y_t = y_t + H_l^\tau D_F \int_0^t e^{D_F(t-\lambda)} H_l y_\lambda d\lambda.$$

Similarly, we can get

$$u_t^f = u_t + H_l^\tau D_F \int_0^t e^{D_F(t-\lambda)} H_l u_\lambda d\lambda.$$

3.2. Adaptive control. We first look at what the stabilization control is in the case where θ is known. To see this, we introduce the following lemma (e.g., Chen and Guo [1990]).

LEMMA 3.1. *Let $k \geq 0$ be an integer and $E(S) = 1 + e_1 S + \cdots + e_k S^k$ be a stable polynomial with $e_k \neq 0$. Then there is a (nonrandom) constant $C_e \geq 1$ (depending on $E(S)$ only) such that*

$$\sum_{i=0}^{k} \int_0^t \left(\frac{S^i}{E(S)} z_\lambda \right)^2 d\lambda \leq C_e \int_0^t z_\lambda^2 d\lambda$$

for any square-integrable process $\{z_t\}$.

If $A(S)$, $SB(S)$ are coprime and $b_q \neq 0$, then for any polynomial

$$E(S) = 1 + e_1 S + \cdots + e_{p+q} S^{p+q} \quad \text{with} \quad e_{p+q} \neq 0, \tag{3.9}$$

there exists a unique polynomial pair $(G(S), H(S))$ such that

$$A(S)G(S) - SB(S)H(S) = E(S) \quad \text{with } \partial(G(S)) \leq q-1 \text{ and } \partial(H(S)) = p, \tag{3.10}$$

where here and hereafter $\partial(X(S))$ denotes the degree of the polynomial $X(S)$ in S.

From (3.10) and (3.1) it is clear that

$$\begin{aligned} E(S)y_t &= A(S)G(S)y_t - SB(S)H(S)y_t \\ &= G(S)[A(S)y_t - SB(S)u_t] + SB(S)[G(S)u_t - H(S)y_t] \\ &= G(S)[y_0 + C(S)w_t + S\eta_t] + SB(S)[G(S)u_t - H(S)y_t] \end{aligned}$$

and

$$\begin{aligned} E(S)u_t &= A(S)G(S)u_t - SB(S)H(S)u_t \\ &= H(S)[A(S)y_t - SB(S)u_t] + A(S)[G(S)u_t - H(S)y_t] \\ &= H(S)[y_0 + C(S)w_t + S\eta_t] + A(S)[G(S)u_t - H(S)y_t]. \end{aligned}$$

Therefore, in the case where θ is known, $l \leq \min\{p, q-1\}$ and Assumptions A.1 and A.2 hold, for any given stable $E(S)$ subject to (3.9), if the control is defined as follows:

$$G(S)u_t - H(S)y_t = 0, \quad t \geq 0,$$

then the system is stabilized in the average sense (3.8).

Similar to §2, we now introduce a deterministic-like excitation signal u'_t. Let $T > 1$ and α be positive constants chosen arbitrarily. Define for $i = 1, 2, \ldots, p+q$,

$$\beta_i = (-1)^{i+1}\alpha^i \frac{(p+q)!}{i! \times (p+q-i)!} \quad \text{with} \quad 0! \triangleq 1, \ i! \triangleq 1 \times 2 \times \cdots \times i,$$

and for some integer $k \geq 0$,

$$u'_t = L_{T^k} + \beta_1 S_k u'_t + \cdots + \beta_{p+q} S_k^{p+q} u'_t, \quad t \in [T^k, T^{k+1}),$$

where $S_k u_t = \int_{T^k}^t u_s ds$ and $L_t = 1 + \int_0^t (\|\varphi_s^f\|^2 + \|\varphi_s\|^2 + u_s^2) ds$ with φ_t^f defined by (3.5) and φ_t defined by

$$\varphi_t = [Sy_t, \ldots, S^p y_t, \quad Su_t, \ldots, S^q u_t]^\tau. \tag{3.11}$$

For any $k \geq 1$, write θ_{T^k} in the component form

$$\theta_{T^k} = [-a_{1T^k}, \ldots, -a_{pT^k}, \quad b_{1T^k}, \ldots, b_{qT^k}]^\tau$$

and set

$$A_k(S) = 1 + \sum_{i=1}^p a_{iT^k} S^i, \quad B_k(S) = \sum_{i=1}^q b_{iT^k} S^{i-1}. \tag{3.12}$$

Let $E(S)$ be a stable polynomial subject to (3.9). Then by Lemma 3.1 there is a constant C_e (depending on $E(S)$ only) such that

$$\sum_{i=0}^{p+q} \int^t \left(\frac{S^i}{E(S)} z_\lambda \right)^2 d\lambda \leq C_e \int_0^t z_\lambda^2 d\lambda \tag{3.13}$$

for any square-integrable process $\{z_\lambda\}$.

Let $R(S)$ be a stable polynomial of S with $\partial(R(S)) = \min\{p+1, q\}$ and let ξ_t^u, ξ_t^y denote the filtered values of u_t, y_t with respect to $R(S)$, respectively:

$$R(S)\xi_t^u = u_t, \quad R(S)\xi_t^y = y_t \quad \forall t \geq 0. \tag{3.14}$$

Actually, the filtered values ξ_t^u and ξ_t^y can be calculated as in Remark 3.1.

Set $\zeta_t = [S\xi_t^y, \dots, S^p\xi_t^y, S\xi_t^u, \dots, S^q\xi_t^u]^\tau$. Then by (3.11) and (3.14) we get

$$R(S)\zeta_t = \varphi_t. \tag{3.15}$$

In the following discussion, for a given polynomial $Z(S) = z_0 + z_1 S + \cdots + z_r S^r$, its norm is defined as

$$\|Z(S)\| = \left(\sum_{i=0}^{r} z_i^2\right)^{1/2}.$$

Define switching times $1 = \tau_0 < \sigma_1 < \tau_1 < \sigma_2 < \tau_2 \cdots$ as follows:

$$\begin{aligned}\sigma_i = \inf\Big\{ k > \tau_{i-1} :\ & A_k(S)G_k(S) - SB_k(S)H_k(S) = E(S) \text{ is solvable} \\ & \text{with respect to } G_k(S) \text{ and } H_k(S) \text{ subject to} \\ & \partial(G_k(S)) \leq q-1 \quad \text{and} \quad \partial(H_k(S)) = p; \quad \text{and} \\ & \|G_k(S)\|^2 + \|H_k(S)\|^2 \leq \frac{k}{2C_e(p+q+1)}; \\ & \int_0^{T^k} [\xi_s^y - \theta_{T^k}^\tau \zeta_s]^2\, ds \leq k^{-2} f(k, T^k) \Big\},\end{aligned} \tag{3.16}$$

$$\begin{aligned}\tau_i = \min\Big\{ k > \sigma_i :\ & \text{there exists } t \in (T^{\sigma_i}, T^k] \text{ such that} \\ & \int_0^t [\xi_s^y - \theta_{T^{\sigma_i}}^\tau \zeta_s]^2\, ds > \sigma_i^{-2} f(\sigma_i, t) \Big\},\end{aligned} \tag{3.17}$$

where C_e is given in (3.13), and

$$f(x,t) = (t+1) \sup_{0 \leq \lambda \leq t} \left\{ x^3 + \frac{1}{\lambda+1} \int_0^\lambda \left[\sum_{j=0}^{p+1} (S^j \xi_s^y)^2 + \sum_{j=0}^{q} (S^j \xi_s^u)^2 \right] ds \right\}. \tag{3.18}$$

Similar to (2.12) we define the adaptive control u_t as follows:

$$u_t = \begin{cases} u_t' & \text{if } t \in [0, T^{\tau_0}) \text{ or } t \in [T^{\tau_{i-1}}, T^{\sigma_i}) \text{ for some } i \geq 1, \\ H_{\sigma_i}(S)y_t - (G_{\sigma_i}(S) - 1)u_t & \text{if } t \in [T^{\sigma_i}, T^{\tau_i}) \text{ for some } i \geq 1, \end{cases} \tag{3.19}$$

where $H_{\sigma_i}(S)$ and $G_{\sigma_i}(S)$ are generated by (3.16), (3.12), and (3.4)–(3.7).

In this case, similar to Lemmas 2.1 and 2.2 we may obtain the following results.

LEMMA 3.2. *Let $\lambda_{\min}^{(t)}$ denote the smallest eigenvalue of matrix R_t^{-1}. Then the parameter estimate θ_t given by* (3.4)–(3.7) *has the following property:*

$$\|\theta_t - \theta\|^2 \leq \frac{c(t+1)}{\lambda_{\min}^{(t)}} \quad \forall t \geq 0$$

for some time-independent random variable $c \geq 0$.

Proof. Let $\widetilde{\theta}_t = \theta_t - \theta$. Then from (3.1)–(3.3) and (3.5) it follows that

$$y_t^f = \theta^\tau \varphi_t^f + F^{-1}(S)[C(S)w_t + S\eta_t] + \varepsilon_t,$$

where we recall that ε_t denotes a function of time decaying exponentially to zero. Substituting this into the first equation of (3.7) and noting (3.6), we get

$$\dot{\widetilde{\theta}}_t = -P_t \varphi_t' \varphi_t'^\tau \widetilde{\theta}_t + P_t \varphi_t' \left[F^{-1}(S)(C(S)w_t + S\eta_t) + \varepsilon_t\right].$$

Then combining this with the second equation of (3.7) gives

$$\frac{d(\widetilde{\theta}_t^\tau P_t^{-1} \widetilde{\theta}_t)}{dt} = -(\widetilde{\theta}_t^\tau \varphi_t')^2 + 2\widetilde{\theta}_t^\tau \varphi_t' \left[F^{-1}(S)(C(S)w_t + S\eta_t) + \varepsilon_t\right],$$

which together with $\int_0^t [F^{-1}(S)(C(S)w_s]^2 ds = O(t)$ (cf. Lemma 3 of Chen and Guo [1990]) implies that

$$\begin{aligned} 0 &\leq \widetilde{\theta}_t^\tau P_t^{-1} \widetilde{\theta}_t \\ &\leq \widetilde{\theta}_0^\tau P_0^{-1} \widetilde{\theta}_0 - \int_0^t (\widetilde{\theta}_s^\tau \varphi_s')^2 ds + 2 \int_0^t \widetilde{\theta}_s^\tau \varphi_s' \left[F^{-1}(S)(C(S)w_s + S\eta_s) + \varepsilon_t\right] ds \\ &\leq \widetilde{\theta}_0^\tau P_0^{-1} \widetilde{\theta}_0 + \int_0^t \left[F^{-1}(S)(C(S)w_s + S\eta_s) + \varepsilon_t\right]^2 ds = O(t+1), \end{aligned}$$

where for the final bound we have used Lemma 3.1 and Assumption A.2. □

LEMMA 3.3. *Under Assumptions* A.1 *and* A.2, *if* $u_t = u_t'$ *for some* k *and any* $t \in (T^k, T^{k+1}]$, *then there exist* $c > 0$, $a > 1$, *and* $k_0 > 0$ *such that*

$$\lambda_{\min}^{(T^{k+1})} \geq c a^{T^{k+1}} L_{T^k} \quad \forall k \geq k_0.$$

Proof. The proof resembles that of Lemma 2.2 and is given in Appendix C.

THEOREM 3.1. *Under Assumptions* A.1 *and* A.2 *and the adaptive control* (3.4)–(3.7), (3.16)–(3.19), *we get that*

(a) *there is an integer* i *such that* $\sigma_i < \infty$, $\tau_i = \infty$, *and* $\theta_t = \theta_{T^{\sigma_i}}$ $\forall t \geq T^{\sigma_i}$;

(b) $\sup_{t \geq 0} \frac{1}{t+1} \int_0^t (y_s^2 + u_s^2) ds < \infty$ *a.s.*

Proof. We first show that it is impossible that $\tau_i < \infty$ and $\sigma_{i+1} = \infty$ on a sample set $\mathcal{D}$ with positive probability for an integer-valued random variable $i \geq 0$.

In fact, if there were a sample set $\mathcal{D}$ of positive probability, i.e., for which $P(\mathcal{D}) > 0$, which was such that for every sample $\omega \in \mathcal{D}$, there were an $i(\omega) \geq 0$ (for simplicity, we drop ω below) such that $\tau_i < \infty$ and $\sigma_{i+1} = \infty$, then $u_t = u_t'$ for all $t \geq \tau_i$. Thus, by Lemmas 3.2 and 3.3 we would have that for some constant $a > 1$

$$\|\theta_{T^k} - \theta\|^2 = O\left(\frac{T^k}{a^{T^k} L_{T^k}}\right) = O\left(\frac{1}{k^3}\right) \quad \text{a.s. on } \mathcal{D} \quad \forall k > \tau_i, \tag{3.20}$$

which together with Lemma D.1 in Appendix D implies that there exists an integer $k_1 \geq 0$ such that for any $k \geq k_1$, $A_k(S)G_k(S) - SB_k(S)H_k(S) = E(S)$ is solvable with respect to $G_k(S)$ and $H_k(S)$ subject to $\partial(G_k(S)) \leq q - 1$ and $\partial(H_k(S)) = p$, and $\|G_k(S)\|^2 + \|H_k(S)\|^2 \leq k/(2C_e(p+q+1))$.

From Lemma 3 of Chen and Guo [1990] and the fact that $\partial(R(S)) \geq \partial(C(S)) + 1$ it follows that

$$\int_0^t \left(\frac{C(S)}{R(S)} w_s\right)^2 ds = O(t) \quad \text{a.s.},$$

while from (3.1), (3.3), (3.14), and (3.15) it follows that

$$\xi_s^y - \theta_t^\tau \zeta_s = (\theta - \theta_t)^\tau \zeta_s + \frac{C(S)}{R(S)} w_s + \frac{S}{R(S)} \eta_s + \varepsilon_t \quad \forall t,\ s \geq 0.$$

Therefore, by Assumption A.2, Lemma 3.1, and (3.18) we find that

$$\begin{aligned}
&\frac{1}{f(k, T^k)} \int_0^{T^k} [\xi_s^y - \theta_{T^k}^\tau \zeta_s]^2 \, ds \\
&\leq \frac{4}{f(k, T^k)} \left[\int_0^{T^k} ((\theta - \theta_{T^k})^\tau \zeta_s)^2 \, ds + \int_0^{T^k} \left(\frac{C(S)}{R(S)} w_s\right)^2 ds \right. \\
&\quad \left. + \int_0^{T^k} \left(\frac{S}{R(S)} \eta_s\right)^2 ds + \int_0^{T^k} (\varepsilon_s)^2 \, ds \right] \\
&= O\left(\|\theta_{T^k} - \theta\|^2 + \frac{1}{k^3} \right) = O\left(\frac{1}{k^3}\right) \quad \text{a.s. on } \mathcal{D}, \qquad (3.21)
\end{aligned}$$

where (3.20) is invoked for the last equality.

From (3.21) we conclude that there exists an integer $k_2 \geq k_1$ such that for any $k \geq k_2$,

$$\frac{1}{f(k, T^k)} \int_0^{T^k} [\xi_s^y - \theta_{T^k}^\tau \zeta_s]^2 \, ds \leq k^{-2} \quad \text{a.s. on } \mathcal{D}. \qquad (3.22)$$

Thus, $\sigma_{i+1} < \infty$ a.s. on $\mathcal{D}$. This contradicts $\sigma_{i+1} = \infty$ on $\mathcal{D}$ and $P(\mathcal{D}) > 0$.

We now prove that $\tau_i = \infty$ a.s. for some integer-valued random variable $i \geq 1$.

In fact, from Lemmas 3.2 and 3.3 it follows that for some $a > 1$,

$$\|\theta_{T^{\sigma_i}} - \theta\|^2 = O\left(\frac{T^{\sigma_i}}{a^{T^{\sigma_i}} L_{T^{\sigma_i}}}\right) = O\left(\frac{1}{\sigma_i^3}\right). \qquad (3.23)$$

As in (3.21) we would have

$$\frac{1}{f(\sigma_i, t)} \int_0^t [\xi_s^y - \theta_{T^{\sigma_i}}^\tau \zeta_s]^2 \, ds = O\left(\|\theta_{T^{\sigma_i}} - \theta\|^2 + \frac{1}{\sigma_i^3} \right) = O\left(\frac{1}{\sigma_i^3}\right) \leq \sigma_i^{-2},$$

where the last inequality is valid for some large enough i and $t \geq T^{\sigma_i}$ because of (3.23). Hence there must be $\tau_i = \infty$ for some i; i.e., assertion (a) is true. We now prove assertion (b). From assertion (a) and (3.19) it follows that for some $i \geq 1$,

$$H_{\sigma_i}(S) y_t - G_{\sigma_i}(S) u_t = 0, \quad t \geq T^{\sigma_i}. \qquad (3.24)$$

Henceforth, for simplicity of notation, we shall write $\theta_{\sigma_i}^\tau$ for $\theta_{T^k}^\tau$.

In view of (3.16) we get

$$\begin{aligned} E(S)S^k y_t &= S^k A_{\sigma_i}(S)G_{\sigma_i}(S)y_t - S^{k+1}B_{\sigma_i}(S)H_{\sigma_i}(S)y_t \\ &= S^k G_{\sigma_i}(S)[A_{\sigma_i}(S)y_t - SB_{\sigma_i}(S)u_t] \\ &\quad +S^{k+1}B_{\sigma_i}(S)[G_{\sigma_i}(S)u_t - H_{\sigma_i}(S)y_t], \quad k=0,\ 1,\dots,p+1; \end{aligned} \tag{3.25}$$

$$\begin{aligned} E(S)S^k u_t &= S^k H_{\sigma_i}(S)[A_{\sigma_i}(S)y_t - SB_{\sigma_i}(S)u_t] \\ &\quad +S^k A_{\sigma_i}(S)[G_{\sigma_i}(S)u_t - H_{\sigma_i}(S)y_t], \quad k=0,\ 1,\dots,q. \end{aligned} \tag{3.26}$$

Thus, noting that $A_{\sigma_i}(S)y_t - SB_{\sigma_i}(S)u_t = y_t - \theta_{\sigma_i}^\tau S\varphi_t$, by (3.14), (3.25), and inequality $(a+b)^2 \le 2a^2 + 2b^2$ we get

$$\begin{aligned} S^j\xi_t^y &= R(S)^{-1}S^j y_t \\ &= E(S)^{-1}S^j G_{\sigma_i}(S)R(S)^{-1}(y_t - \theta_{\sigma_i}^\tau\varphi_t) \\ &\quad +E(S)^{-1}R(S)^{-1}S^{j+1}B_{\sigma_i}(S)[G_{\sigma_i}(S)u_t - H_{\sigma_i}(S)y_t] \\ &= E(S)^{-1}S^j G_{\sigma_i}(S)(\xi_t^y - \theta_{\sigma_i}^\tau\zeta_t) \\ &\quad +E(S)^{-1}R(S)^{-1}S^{j+1}B_{\sigma_i}(S)[G_{\sigma_i}(S)u_t - H_{\sigma_i}(S)y_t] \\ &\quad j=0,\ 1,\dots,p+1, \end{aligned}$$

and, furthermore, we have

$$\begin{aligned} \sum_{j=0}^{p+1}(S^j\xi_s^y)^2 &\le 2\|G_{\sigma_i}(S)\|^2\sum_{j=0}^{p+1}\sum_{k=0}^{q-1}[S^{j+k}E^{-1}(S)(\xi_s^y - \theta_{\sigma_i}^\tau\zeta_s)]^2 \\ &\quad +2\sum_{j=0}^{p+1}\left(E^{-1}(S)R^{-1}(S)S^{j+1}B_{\sigma_i}(S)[G_{\sigma_i}(S)u_s - H_{\sigma_i}(S)y_s]\right)^2 \\ &\le 2\|G_{\sigma_i}(S)\|^2(p+q+1)\sum_{j=0}^{p+q}[S^jE^{-1}(S)(\xi_s^y - \theta_{\sigma_i}^\tau\zeta_s)]^2 \\ &\quad +2\sum_{j=0}^{p+1}\left(E^{-1}(S)R^{-1}(S)S^{j+1}B_{\sigma_i}(S)[G_{\sigma_i}(S)u_s - H_{\sigma_i}(S)y_s]\right)^2, \end{aligned} \tag{3.27}$$

and similarly, by (3.26) we get

$$\begin{aligned} \sum_{j=0}^{q}(S^j\xi_s^u)^2 &\le 2\|H_{\sigma_i}(S)\|^2\sum_{j=0}^{q}\sum_{k=0}^{p}[S^{j+k}E^{-1}(S)(\xi_s^y - \theta_{\sigma_i}^\tau\zeta_s)]^2 \\ &\quad +2\sum_{j=0}^{q}\left(E^{-1}(S)R^{-1}(S)S^jA_{\sigma_i}(S)[G_{\sigma_i}(S)u_s - H_{\sigma_i}(S)y_s]\right)^2 \\ &\le 2\|H_{\sigma_i}(S)\|^2(p+q+1)\sum_{j=0}^{p+q}[S^jE^{-1}(S)(\xi_s^y - \theta_{\sigma_i}^\tau\zeta_s)]^2 \\ &\quad +2\sum_{j=0}^{q}\left(E^{-1}(S)R^{-1}(S)S^jA_{\sigma_i}(S)[G_{\sigma_i}(S)u_s - H_{\sigma_i}(S)y_s]\right)^2. \end{aligned} \tag{3.28}$$

By Lemma 3.1 and (3.24) we see that

$$\limsup_{t\to\infty}\frac{1}{t+1}\sum_{j=0}^{p+1}\int_0^t\left(E^{-1}(S)R^{-1}(S)S^{j+1}B_{\sigma_i}(S)[G_{\sigma_i}(S)u_s - H_{\sigma_i}(S)y_s]\right)^2 ds < \infty \quad \text{a.s.}$$

and that

$$\limsup_{t\to\infty}\frac{1}{t+1}\sum_{j=0}^{q}\int_0^t \left(E^{-1}(S)R^{-1}(S)S^jA_{\sigma_i}(S)[G_{\sigma_i}(S)u_s - H_{\sigma_i}(S)y_s]\right)^2 ds < \infty \quad \text{a.s.}$$

Therefore, by (3.27), (3.28), and Lemma 3.1 we conclude that for some $\nu_1 < \infty$ which is independent of t

(3.29)

$$\begin{aligned}
&\frac{1}{t+1}\int_0^t \left[\sum_{k=0}^{p+1}(S^k\xi_s^y)^2 + \sum_{k=0}^{q}(S^k\xi_s^u)^2\right] ds \\
&\le 2(p+q+1)[\|G_{\sigma_i}(S)\|^2 + \|H_{\sigma_i}(S)\|^2]\frac{1}{t+1}\sum_{j=0}^{p+q}\int_0^t \left[\frac{S^j}{E(S)}(\xi_s^y - \theta_{\sigma_i}^\tau\zeta_s)\right]^2 ds + \nu_1 \\
&\le \frac{2(p+q+1)\sigma_i}{2C_e(p+q+1)}\cdot\frac{C_e}{t+1}\int_0^t(\xi_s^y - \theta_{\sigma_i}^\tau\zeta_s)^2 ds + \nu_1 \\
&\le \sigma_i\cdot\sigma_i^{-2}\frac{1}{t+1}f(\sigma_i,t) + \nu_1 \quad \text{a.s.}, \quad t \ge T^{\sigma_i},
\end{aligned}$$

where (3.17), $\sigma_i < \infty$, and $\tau_i = \infty$ a.s. have been used for the last inequality.

Set

$$\nu_2 = \nu_1 + \sup_{0\le\lambda\le T^{\sigma_i}}\left\{\frac{1}{\lambda+1}\int_0^\lambda\left[\sum_{k=0}^{p+1}(S^k\xi_s^y)^2 + \sum_{k=0}^{q}(S^k\xi_s^u)^2\right] ds\right\}.$$

Then from (3.29) and (3.18) it follows that

$$\begin{aligned}
&\sup_{0\le\lambda\le t}\left\{\frac{1}{\lambda+1}\int_0^\lambda\left[\sum_{k=0}^{p+1}(S^k\xi_s^y)^2 + \sum_{k=0}^{q}(S^k\xi_s^u)^2\right] ds\right\} \le \sigma_i^{-1}\frac{1}{t+1}f(\sigma_i,t) + \nu_2 \\
&\le \sigma_i^2 + \nu_2 + \sigma_i^{-1}\sup_{0\le\lambda\le t}\left\{\frac{1}{\lambda+1}\int_0^\lambda\left[\sum_{k=0}^{p+1}(S^k\xi_s^y)^2 + \sum_{k=0}^{q}(S^k\xi_s^u)^2\right] ds\right\} \quad \text{a.s.},
\end{aligned}$$

i.e.,

$$\begin{aligned}
&\limsup_{t\to\infty}\sup_{0\le\lambda\le t}\left\{\frac{1}{\lambda+1}\int_0^\lambda\left[\sum_{k=0}^{p+1}(S^k\xi_s^y)^2 + \sum_{k=0}^{q}(S^k\xi_s^u)^2\right] ds\right\} \\
&\le (1-\sigma_i^{-1})^{-1}\left[\nu_2 + \sigma_i^2\right] < \infty \quad \text{a.s.}
\end{aligned}$$

From this and (3.14), assertion (b) follows. □

Appendix A. Proof of Lemma 2.2. Let

$$N = \begin{bmatrix} \beta_1 & \beta_2 & \dots & \beta_{n+1} \\ 1 & 0 & \dots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & 1 & 0 \end{bmatrix}, \quad H = [1, \underbrace{0,\dots,0}_{n}]^\tau, \tag{A.1}$$

and, for any $t \in [T^k,\ T^{k+1})$,

$$U_t(k) = [u_t,\ S_k u_t, \dots, S_k^n u_t]^\tau.$$

Then from the definition of u_t it follows that for any $t \in (T^k,\ T^{k+1})$,

$$\frac{dU_t(k)}{dt} = NU_t(k) \quad \text{with } U_{T^k}(k) = HL_{T^k}, \tag{A.2}$$

i.e.,

$$S_kU_t(k) = N^{-1}U_t(k) - N^{-1}HL_{T^k} \quad \forall t \in [T^k,\ T^{k+1}). \tag{A.3}$$

We first to show that there exist constants $c > 0$, $\gamma > 1$ and k_0 such that

$$\lambda_{\min}\left(\int_{T^k}^{T^{k+1}} (S_kU_s(k))(S_kU_s(k))^\tau ds\right) \geq c\gamma^{T^{k+1}}L_{T^k}^2 \quad \forall k \geq k_0, \tag{A.4}$$

where here and hereafter $\lambda_{\min}(X)$ denotes the minimal eigenvalue of matrix X.

From (2.5) and (A.1) we see that the characteristic polynomial $\det(xI-N) = (x-\alpha)^{n+1}$ of N coincides with the minimal polynomial of N. Thus there is a nonsingular $(n+1)\times(n+1)$ matrix P such that

$$\overline{\Lambda} \triangleq P^{-1}NP = \begin{bmatrix} \alpha & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & \alpha \end{bmatrix}_{(n+1)\times(n+1)}. \tag{A.5}$$

Let $\overline{U}_t(k) = P^{-1}U_t(k)$ and $\overline{H} = P^{-1}H$. Then (A.2) is equivalent to

$$\frac{d\overline{U}_t(k)}{dt} = \overline{\Lambda}\cdot\overline{U}_t(k) \quad \text{with } \overline{U}_{T^k}(k) = \overline{H}L_{T^k} \quad \forall t \in (T^k,\ T^{k+1}). \tag{A.6}$$

Noting that for a given positive semidefinite matrix U,

$$\lambda_{\min}(PUP^\tau) \geq \lambda_{\min}(PP^\tau)\lambda_{\min}(U)$$

and $\lambda_{\max}(PP^\tau) \leq \|P\|^2$, by (A.3) and inequality $a^2 \geq \frac{1}{2}b^2 - (b-a)^2$ we have

$$\begin{aligned}
&\lambda_{\min}\left(\int_{T^k}^{T^{k+1}} (S_kU_s(k))(S_kU_s(k))^\tau ds\right) \\
&\geq \frac{1}{2}\lambda_{\min}(N^{-1}N^{-\tau})\lambda_{\min}\left(\int_{T^k}^{T^{k+1}} U_s(k)U_s^\tau(k)ds\right) - \|N^{-1}\|^2T^{k+1}L_{T^k}^2 \\
&\leq \frac{1}{2}\lambda_{\min}(N^{-1}N^{-\tau})\lambda_{\min}(PP^\tau)\,\lambda_{\min}\left(\int_{T^k}^{T^{k+1}} \overline{U}_s(k)\overline{U}_s^\tau(k)ds\right) - \|N^{-1}\|^2T^{k+1}L_{T^k}^2.
\end{aligned}$$

Thus, in order to show (A.4), it suffices to prove that there exist constants $c > 0$, $\gamma > 1$, and k_0 such that

$$\lambda_{\min}\left(\int_{T^k}^{T^{k+1}} \overline{U}_s(k)\overline{U}_s^\tau(k)ds\right) \geq c\gamma^{T^{k+1}}L_{T^k}^2 \quad \forall k \geq k_0. \tag{A.7}$$

From (A.6) it follows that

$$\overline{U}_t(k) = e^{\overline{\Lambda}t}\overline{H}L_{T^k} \quad \forall t \in [T^k, T^{k+1}),$$

which implies that

$$\begin{aligned}\int_{T^k}^{T^{k+1}} \overline{U}_s(k)\overline{U}_s^\tau(k)ds &= L_{T^k}^2 \int_{T^k}^{T^{k+1}} e^{\overline{\Lambda}s}\overline{H}\cdot\overline{H}^\tau e^{\overline{\Lambda}^\tau s}ds \\ &\geq L_{T^k}^2 \int_{T^{k+1}-1}^{T^{k+1}} e^{\overline{\Lambda}s}\overline{H}\cdot\overline{H}^\tau e^{\overline{\Lambda}^\tau s}ds \\ &= L_{T^k}^2 e^{\overline{\Lambda}(T^{k+1}-1)}\left(\int_0^1 e^{\overline{\Lambda}s}\overline{H}\cdot\overline{H}^\tau e^{\overline{\Lambda}^\tau s}ds\right)e^{\overline{\Lambda}^\tau(T^{k+1}-1)} \\ &\geq L_{T^k}^2\lambda_{\min}\left(\int_0^1 e^{\overline{\Lambda}s}\overline{H}\cdot\overline{H}^\tau e^{\overline{\Lambda}^\tau s}ds\right)e^{\overline{\Lambda}(T^{k+1}-1)}e^{\overline{\Lambda}^\tau(T^{k+1}-1)}. \qquad \text{(A.8)}\end{aligned}$$

Note that (N, H) is controllable and hence $(\overline{\Lambda}, \overline{H})$ is controllable. Therefore,

$$\lambda_{\min}\left(\int_0^1 e^{\overline{\Lambda}s}\overline{H}\cdot\overline{H}^\tau e^{\overline{\Lambda}^\tau s}ds\right) > 0. \qquad \text{(A.9)}$$

Set

$$\Sigma_t = \begin{bmatrix} 1 & & & \\ (t-1) & \ddots & & \\ \vdots & \ddots & \ddots & \\ \dfrac{(t-1)^n}{n!} & \cdots & (t-1) & 1 \end{bmatrix}.$$

Then from (A.5) we see that

$$e^{\overline{\Lambda}(t-1)} = e^{\alpha(t-1)}\Sigma_t. \qquad \text{(A.10)}$$

It is evident that

$$\det(\Sigma_t\Sigma_t^\tau) = 1 \quad \text{and} \quad \lambda_{\max}(\Sigma_t\Sigma_t^\tau) \leq (n+1)\sum_{i=0}^{n}(t-1)^{2i},$$

where $\lambda_{\max}(X)$ denotes the maximum eigenvalues of X.

Thus, from the fact that $\det(X) = \prod_{i=1}^{n+1}\lambda_i(X)$ for any $(n+1)\times(n+1)$ matrix with eigenvalues $\lambda_i(X)$ $(i = 1, \ldots, n+1)$ it follows that

$$\lambda_{\min}(\Sigma_t\Sigma_t^\tau) \geq [\lambda_{\max}(\Sigma_t\Sigma_t^\tau)]^{-n} \geq \left[(n+1)\sum_{i=0}^{n}(t-1)^{2i}\right]^{-n}$$

$$\geq (n+1)^{-2n}(t-1)^{-2n^2} \quad \forall t \geq 2.$$

From this and (A.10) we get that

$$\lambda_{\min}\left(e^{\overline{\Lambda}(t-1)}e^{\overline{\Lambda}^\tau(t-1)}\right) \geq e^{2\alpha(t-1)}\cdot(n+1)^{-2n}(t-1)^{-2n^2} \quad \forall t \geq 2,$$

which together with $\alpha > 0$, (A.9), and (A.8) implies the desired result (A.7). Therefore, (A.4) is true.

We are now in a position to prove (2.13).

Write

$$\operatorname{Adj}(I - AS) = I + A_1 S + \cdots + A_{n-1} S^{n-1},$$
$$A(S) \triangleq \det(I - AS) = a_0 + a_1 S + \cdots + a_n S^n$$

and set

$$M = \begin{bmatrix} 0 & B & A_1 B & \dots & A_{n-1} B \\ a_0 & a_1 & a_2 & \dots & a_n \end{bmatrix},$$

where I denotes $n \times n$ identity matrix.

Clearly, we have

$$A(S)\psi_\lambda = \begin{bmatrix} \operatorname{Adj}(I - AS) S^2 B u_\lambda + \operatorname{Adj}(I - AS)(x_0 \lambda + CSw_\lambda) \\ A(S) S u_\lambda \end{bmatrix}$$
$$= MSU_t + \begin{bmatrix} \operatorname{Adj}(I - AS)(x_0 \lambda + CSw_\lambda) \\ 0 \end{bmatrix},$$

where

$$U_t = [u_t,\ Su_t, \dots, S^n u_t]^\tau.$$

Therefore, we get

$$\lambda_{\min}\left(\int_{T^k}^t [A(S)\psi_s][A(S)\psi_s]^\tau ds\right)$$
$$\geq \frac{1}{2}\lambda_{\min}\left(M \int_{T^k}^t (SU_s)(SU_s)^\tau ds M^\tau\right) - c\sum_{i=0}^n \int_0^t \left[s^{2i} + \|S^i w_s\|^2\right] ds. \tag{A.11}$$

Using the argument in, e.g., Zhang and Chen [1991] we can obtain

$$\int_0^t \|S^i w_s\|^2 ds \leq c t^{2i+3}, \quad i = 0, 1, 2, \dots, n, \tag{A.12}$$

and

$$\lambda_{\min}\left(\int_{T^k}^t [A(S)\psi_s][A(S)\psi_s]^\tau ds\right) = \min_{\|x\|=1} \int_{T^k}^t \left|\sum_{i=0}^n a_i S^i x^\tau \psi_s\right|^2 ds$$
$$\leq c\sum_{i=0}^n t^{2i+1} \lambda_{\min}\left(\int_{T^k}^t \psi_s \psi_s^\tau ds\right) + c\left(\sum_{i=0}^n t^{2i+1}\right) \int_0^{T^k} \|\psi_s\|^2 ds. \tag{A.13}$$

From (A.11)–(A.13) we have

$$\lambda_{\min}\left(\int_{T^k}^t \psi_s \psi_s^\tau ds\right) \geq \frac{1}{c}\lambda_{\min}(MM^\tau)\left(\sum_{i=0}^n t^{2i+1}\right)^{-1} \lambda_{\min}\left(\int_{T^k}^t (SU_s)(SU_s)^\tau ds\right)$$
$$-c\left(\sum_{i=0}^n t^{2i+1}\right)^{-1}\left(\sum_{i=0}^{n+1} t^{2i+1}\right) - \int_0^{T^k} \|\psi_s\|^2 ds. \tag{A.14}$$

By induction we can show that

$$|S^i u_t - S_k^i u_t| \le t^{i-1} \int_0^{T^k} |u_s| ds \quad \forall i = 1, 2, \dots, \quad \forall t \in [T^k, T^{k+1}).$$

From this we see that for any $x \in I\!R^{n+1}$ with $\|x\| = 1$ and $t \in [T^k, T^{k+1})$,

$$\begin{aligned}
\int_{T^k}^t (x^\tau S U_s)^2 ds &\ge \frac{1}{2} \int_{T^k}^t [x^\tau S_k U_s(k)]^2 ds - \int_{T^k}^t \|S_k U_s(k) - S U_s\|^2 ds \\
&\ge \frac{1}{2} \int_{T^k}^t [x^\tau S_k U_s(k)]^2 ds - \int_{T^k}^t \sum_{i=1}^{n+1} |S^i u_s - S_k^i u_s|^2 ds \\
&\ge \frac{1}{2} \int_{T^k}^t [x^\tau S_k U_s(k)]^2 ds - \sum_{i=1}^{n+1} \left(\int_0^{T^k} |u_s| ds \right)^2 \int_{T^k}^t s^{2(i-1)} ds \\
&\ge \frac{1}{2} \int_{T^k}^t [x^\tau S U_s(k)]^2 ds - \sum_{i=1}^{n+1} \left(\int_0^{T^k} u_s^2 ds \right) T^{2i(k+1)} \\
&\ge \frac{1}{2} \int_{T^k}^t [x^\tau S U_s(k)]^2 ds - (n+1) T^{2(n+1)(k+1)} \int_0^{T^k} u_s^2 ds,
\end{aligned}$$

which implies that

$$\begin{aligned}
\lambda_{\min} \left(\int_{T^k}^{T^{k+1}} (S U_s)(S U_s)^\tau ds \right) \ge \frac{1}{2} \lambda_{\min} &\left(\int_{T^k}^{T^{k+1}} (S_k U_s(k))(S_k U_s(k))^\tau ds \right) \\
&- (n+1) T^{2(n+1)(k+1)} L_{T^k}.
\end{aligned} \tag{A.15}$$

This together with (A.4) and (A.14) leads to

$$\lambda_{\min} \left(\int_{T^k}^{T^{k+1}} \psi_s \psi_s^\tau ds \right) \ge c^{-1} T^{-(2n+1)(k+1)} \gamma^{T^{k+1}} L_{T^k}^2 - c T^{(2n+3)(k+1)} L_{T^k}. \tag{A.16}$$

With $(S+1) z_t = \psi_t$ in mind, we get

$$\begin{aligned}
\lambda_{\min} \left(\int_{T^k}^{T^{k+1}} \psi_s \psi_s^\tau ds \right) &= \lambda_{\min} \left(\int_{T^k}^{T^{k+1}} [(S+1) z_s][(S+1) z_s]^\tau ds \right) \\
&= \min_{\|x\|=1} \int_{T^k}^{T^{k+1}} |x^\tau z_s + S x^\tau z_s|^2 ds \\
&\le 4 T^{2(k+1)} \lambda_{\min} \left(\int_{T^k}^{T^{k+1}} z_s z_s^\tau ds \right) + 2 T^{2(k+1)} \int_0^{T^k} \|z_s\|^2 ds,
\end{aligned}$$

i.e.,

$$\begin{aligned}
\lambda_{\min} \left(\int_{T^k}^{T^{k+1}} z_s z_s^\tau ds \right) \ge 4^{-1} T^{-2(k+1)} &\lambda_{\min} \left(\int_{T^k}^{T^{k+1}} \psi_s \psi_s^\tau ds \right) \\
&- 2^{-1} \int_0^{T^k} \|z_s\|^2 ds.
\end{aligned}$$

From this, (A.16), and the definition of L_{T^k} it follows that

$$\lambda_{\min}\left(\int_{T^k}^{T^{k+1}} z_s z_s^\tau ds\right) \geq c^{-1}T^{-(2n+3)(k+1)}\gamma^{T^{k+1}}L_{T^k}^2 - cT^{(2n+1)(k+1)}L_{T^k},$$

which implies the desired result (2.13). □

Appendix B. The following lemma is based on Chen [1992].

LEMMA B.1. *If $T > 1$ is a constant, (F, C) is controllable, $F \in \mathcal{F}_\sigma$ with $\sigma < \infty$ a.s. being a stopping time, and if*

$$\limsup_{k\to\infty} \frac{1}{T^k}\int_0^{T^k} \|x_s\|^2 ds < \infty \quad a.s. \tag{B.1}$$

for the system

$$dx_t = Fx_t dt + C dw_t, \quad t \geq \sigma, \tag{B.2}$$

then F must be stable a.s.

Proof. Assume that F^τ has an eigenvalue λ with $\Re(\lambda) \geq 0$, where $\Re(x)$ denotes the real part of a complex number x. Let y be the corresponding eigenvector, i.e., $F^\tau y = \lambda y$. Then by (B.2) we get

$$d[\Re(y^\tau x_t) + i\Im(y^\tau x_t)] = [\Re(\lambda) + i\Im(\lambda)][\Re(y^\tau x_t) + i\Im(y^\tau x_t)]dt + y^\tau C dw_t \quad \forall t \geq \sigma,$$

i.e.

$$dz_t = \begin{bmatrix} \Re(\lambda) & -\Im(\lambda) \\ \Im(\lambda) & \Re(\lambda) \end{bmatrix} z_t dt + \begin{bmatrix} \Re(y^\tau C) \\ \Im(y^\tau C) \end{bmatrix} dw_t, \tag{B.3}$$

where $\Im(x)$ denotes the imaginary part of a complex number x and

$$z_t = \begin{bmatrix} \Re(y^\tau x_t) \\ \Im(y^\tau x_t) \end{bmatrix}.$$

Using Ito's formula, by (B.3) we obtain

$$\begin{aligned} dz_t^\tau z_t &= 2z_t^\tau \begin{bmatrix} \Re(\lambda) & -\Im(\lambda) \\ \Im(\lambda) & \Re(\lambda) \end{bmatrix} z_t dt + 2z_t^\tau \begin{bmatrix} \Re(y^\tau C) \\ \Im(y^\tau C) \end{bmatrix} dw_t + \|y^\tau C\|^2 dt \\ &= 2\Re(\lambda)\|z_t\|^2 dt + 2z_t^\tau \begin{bmatrix} \Re(y^\tau C) \\ \Im(y^\tau C) \end{bmatrix} dw_t + \|y^\tau C\|^2 dt, \end{aligned}$$

which implies (Christopeit [1986]) that for any $\eta \in (0, 1/2)$,

$$\|z_t\|^2 = \|z_\sigma\|^2 + 2\Re(\lambda)\int_\sigma^t \|z_s\|^2 ds + O\left(\left(\int_\sigma^t \|z_s\|^2 ds\right)^{\frac{1}{2}+\eta}\right) + \|y^\tau C\|^2(t-\sigma). \tag{B.4}$$

Noting that (B.1) implies that

$$\limsup_{k\to\infty} \frac{1}{T^k}\int_\sigma^{T^k} \|z_s\|^2 ds < \infty, \tag{B.5}$$

by (B.4) we have

$$\text{(B.6)}\quad \|z_t\|^2 = \|z_\sigma\|^2 + 2\Re(\lambda)\int_\sigma^t \|z_s\|^2 ds + O\left(t^{\frac{1}{2}+\eta}\right) + \|y^\tau C\|^2(t-\sigma) \quad \forall \eta \in (0, 1/2).$$

From controllability of (F, C) it is easy to see that $\|y^\tau C\| \neq 0$, and hence from (B.6) and $\Re(\lambda) \geq 0$ it follows that for some $t_0 \geq \sigma$ and $c > 0$,

$$\|z_t\|^2 \geq ct \quad \forall t \geq t_0, \quad t_0 \text{ random},$$

which contradicts (B.5). □

Appendix C. Proof of Lemma 3.3. As in Appendix A, we can show that there exist constants $c > 0$, $\gamma > 1$, and k_0 such that

$$\text{(C.1)}\qquad \lambda_{\min}\left(\int_{T^k}^{T^{k+1}} (S_k U_s(k))(S_k U_s(k))^\tau ds\right) \geq c\gamma^{T^{k+1}} L_{T^k}^2 \quad \forall k \geq k_0,$$

where here and hereafter

$$U_t(k) = [u_t,\ S_k u_t, \ldots, S_k^{p+q-1} u_t]^\tau.$$

Set $W_t = y_0 t + SC(S)w_t + S^2\eta_t$ and $M = [M_1,\ M_2]^\tau$ with

$$M_1^\tau \triangleq \left(\begin{array}{cccccccccc} \overbrace{0 & b_1 & \ldots & \ldots & \ldots & \ldots & b_q & 0 & \ldots & 0}^{p+q} \\ 0 & 0 & \ddots & & & & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & & \ddots & 0 \\ 0 & \ldots & 0 & 0 & b_1 & \ldots & \ldots & \ldots & \ldots & b_q \end{array}\right)\Bigg\}\, p$$

and

$$M_2^\tau \triangleq \left(\begin{array}{cccccccccc} \overbrace{1 & a_1 & \ldots & \ldots & \ldots & \ldots & a_q & 0 & \ldots & 0}^{p+q} \\ 0 & 1 & \ddots & & & & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & & \ddots & 0 \\ 0 & \ldots & 0 & 1 & a_1 & \ldots & \ldots & \ldots & \ldots & a_q \end{array}\right)\Bigg\}\, q.$$

Then from (3.1) and (3.11) it follows that

$$\text{(C.2)}\qquad A(S)\varphi_t = MSU_t + [W_t,\ SW_t, \ldots, S^{p-1}W_t,\ \underbrace{0, \ldots, 0}_{q}]^\tau,$$

where here and hereafter,

$$\text{(C.3)}\qquad U_t = [u_t,\ Su_t, \ldots, S^{p+q-1}u_t]^\tau.$$

Notice that by Assumption A.2, for $i = 1, 2, \ldots, p$,

$$\int_0^t (S^i\eta_s)^2 ds \leq \int_0^t \frac{t^{2i} - s^{2i}}{2i(2i-1)[(i-1)!]^2}\eta_s^2 ds \leq ct^{2i+1}.$$

Then similar to (A.11)–(A.14), by (C.2) we obtain

$$\lambda_{\min}\left(\int_{T^k}^{t}\varphi_s\varphi_s^\tau ds\right) \geq \frac{1}{c}\left(\sum_{i=0}^{p}t^{2i+1}\right)^{-1}\lambda_{\min}\left(\int_{T^k}^{t}(SU_s)(SU_s)^\tau ds\right)$$

$$-c\left(\sum_{i=0}^{p}t^{2i+1}\right)^{-1}\sum_{i=0}^{p+l}(t+1)^{2i+1}-\int_0^{T^k}\|\varphi_s\|^2 ds. \tag{C.4}$$

By (C.3) and along the argument of (A.15) we get

$$\lambda_{\min}\left(\int_{T^k}^{T^{k+1}}(SU_s)(SU_s)^\tau ds\right) \geq \frac{1}{2}\lambda_{\min}\left(\int_{T^k}^{T^{k+1}}(S_kU_s(k))(S_kU_s(k))^\tau ds\right)$$

$$-(p+q)T^{2(p+q)(k+1)}L_{T^k}.$$

This together with (C.1) and (C.4) leads to

$$\lambda_{\min}\left(\int_{T^k}^{T^{k+1}}\varphi_s\varphi_s^\tau ds\right) \geq c^{-1}T^{-(2p+1)(k+1)}\gamma^{T^{k+1}}L_{T^k}^2 - cT^{2(p+q+l+2)(k+1)}L_{T^k}. \tag{C.5}$$

With $F(S)\varphi_t^f = \varphi_t$ in mind, as in (A.13) we get

$$\lambda_{\min}\left(\int_{T^k}^{T^{k+1}}\varphi_s\varphi_s^\tau ds\right) = \lambda_{\min}\left(\int_{T^k}^{T^{k+1}}(F(S)\varphi_s^f)(F(S)\varphi_s^f)^\tau ds\right)$$

$$= \min_{\|x\|=1}\int_{T^k}^{T^{k+1}}\left|\sum_{i=0}^{l+1}f_iS^ix^\tau\varphi_s^f\right|^2 ds$$

$$\leq cT^{(2l+3)(k+1)}\lambda_{\min}\left(\int_{T^k}^{T^{k+1}}\varphi_s^f(\varphi_s^f)^\tau ds\right) + cT^{(2l+3)(k+1)}\int_0^{T^k}\|\varphi_s^f\|^2 ds,$$

i.e.,

$$\lambda_{\min}\left(\int_{T^k}^{T^{k+1}}\varphi_s^f(\varphi_s^f)^\tau ds\right) \geq cT^{-(2l+3)(k+1)}\lambda_{\min}\left(\int_{T^k}^{T^{k+1}}\varphi_s\varphi_s^\tau ds\right)$$

$$-\int_0^{T^k}\|\varphi_s^f\|^2 ds.$$

From this, (C.5) and the definition of L_{T^k} it follows that

$$\lambda_{\min}\left(\int_{T^k}^{T^{k+1}}\varphi_s^f(\varphi_s^f)^\tau ds\right) \geq c^{-1}T^{-2(p+l+2)(k+1)}\gamma^{T^{k+1}}L_{T^k}^2 - cT^{2(p+q+l+2)(k+1)}L_{T^k},$$

which implies the desired result, Lemma 3.3. □

Appendix D.

LEMMA D.1. *If $A(S)$ and $SB(S)$ are coprime, $b_q \neq 0$ and*

$$\theta_{T^k} \xrightarrow[k\to\infty]{} \theta \ \text{a.s.},$$

then there is an integer-valued K, possibly depending on a sample path such that for all $k \geq K$, $A_k(S)G_k(S) - SB_k(S)H_k(S) = E(S)$ is solvable with respect to $G_k(S)$ and $H_k(S)$ subject to $\partial(G_k(S)) \leq q-1$ and $\partial(H_k(S)) = p$, and such that $\|G_k(S)\|^2 + \|H_k(S)\|^2 \leq k/(2C_e(p+q+1))$.

Proof. Let

$$\text{(D.1)}\quad M_3^\tau = \overbrace{\begin{pmatrix} 1 & a_1 & \dots & \dots & \dots & \dots & a_p & 0 & \dots & 0 & 0 \\ 0 & 1 & \ddots & & & & & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & & \ddots & 0 & 0 \\ 0 & \dots & 0 & 1 & a_1 & \dots & \dots & \dots & \dots & a_p & 0 \end{pmatrix}}^{p+q+1} \Bigg\} q,$$

$$\text{(D.2)}\quad M_4^\tau = \overbrace{\begin{pmatrix} 0 & -b_1 & \dots & \dots & \dots & \dots & -b_q & 0 & \dots & 0 \\ 0 & 0 & \ddots & & & & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & & \ddots & 0 \\ 0 & \dots & 0 & 0 & -b_1 & \dots & \dots & \dots & \dots & -b_q \end{pmatrix}}^{p+q+1} \Bigg\} p+1,$$

$$\text{(D.3)}\quad M' = [M_3,\ M_4],$$

$$H_e = [1,\ e_1,\ \dots,\ e_{p+q}]^\tau.$$

Replacing a_i and b_j by their estimates a_{iT^k} and b_{jT^k}, respectively, in (D.1)–(D.3) for $i = 1, \dots, p$ and $j = 1, \dots, q$, we correspondingly denote M_3, M_4, and M' by $M_{3,k}$, $M_{4,k}$, and M_k'. Furthermore, if M_k' is nonsingular, we set $\Psi_k = (M_k')^{-1}H_e$.

Since $A(S)$ and $SB(S)$ are coprime and $b_q \neq 0$, we see that M' given by (D.1)–(D.3) is nonsingular. Let $\Psi = (M')^{-1}H_e$ and $G(S) = \sum_{i=0}^{q-1} g_i S^i$, $H(S) = \sum_{i=0}^{p} h_i S^i$ with

$$[g_0,\ g_1, \dots, g_{q-1},\quad h_0,\ h_1, \dots, h_p] \triangleq \Psi.$$

Then recalling that

$$\theta_{T^k} \xrightarrow[k \to \infty]{} \theta,$$

we see that there is an integer $K' \geq 0$ such that for all $k \geq K'$, M_k' is nonsingular,

$$M_k' \xrightarrow[k \to \infty]{} M', \text{ and } \Psi_k \xrightarrow[k \to \infty]{} \Psi.$$

Furthermore for all $k \geq K'$, if we set $G_k(S) = \sum_{i=0}^{q-1} g_{i,k} S^i$, $H_k(S) = \sum_{i=0}^{p} h_{i,k} S^i$ with

$$[g_{0,k},\ g_{1,k}, \dots, g_{q-1,k},\quad h_{0,k},\ h_{1,k}, \dots, h_{p,k}] \triangleq \Psi_k,$$

then we have $\partial(G_k(S)) \leq q-1$, $\partial(H_k(S)) = p$ and

$$A_k(S)G_k(S) - SB_k(S)H_k(S) = E(S).$$

Noting that

$$\Psi_k \xrightarrow[k \to \infty]{} \Psi$$

we see that there exists an integer $K \geq K'$ such that for all $k \geq K$, $\|G_k(S)\|^2 + \|H_k(S)\|^2 = \|\Psi_k\|^2 \leq k/(2C_e(p+q+1))$. □

REFERENCES

P. E. CAINES (1992), *Continuous time stochastic adaptive control: Non-explosion, ε-consistency and stability*, Systems Control Lett., 19, pp. 169–176.

H. F. CHEN (1990), *Continuous-time stochastic adaptive control stabilizing the system and minimizing the quadratic loss function*, Tech. report, Institute of Systems Science, Academia Sinica, Beijing.

H. F. CHEN AND L. GUO (1990), *Continuous-time stochastic adaptive control—robustness and asymptotic properties*, SIAM J. Control Optim., 28, pp. 513–527.

H. F. CHEN AND J. B. MOORE (1987), *Convergence rate of continuous-time stochastic ELS parameter estimation*, IEEE Trans. Automatic Control, 32, pp. 267–269.

H. F. CHEN AND J. F. ZHANG (1992), *Adaptive stabilization of unstable and nonminimum-phase stochastic systems*, Systems Control Lett., 19, pp. 27–38.

N. CHRISTOPEIT (1986), *Quasi-least-squares estimation in semimartingale regression models*, Stochastics, 16, pp. 255–278.

T. E. DUNCAN AND E. PASIK-DUNCAN (1990), *Adaptive control of continuous-time linear stochastic systems*, Math. Control Signals Systems, 3, pp. 45–60.

——— (1991), *Some methods for the adaptive control continuous time linear stochastic systems*, in Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control, L. Gerencsér and P. E. Caines, eds., Springer-Verlag, Berlin, Heidelberg.

M. GEVERS, G. C. GOODWIN, AND V. WERTZ (1991), *Continuous-time stochastic adaptive control*, SIAM J. Control Optim., 29, pp. 264–282.

G. C. GOODWIN, M. GEVERS, D. Q. MAYNE, AND V. WERTZ (1991), *Stochastic adaptive control: results and perspective*, in Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control, L. Gerencsér and P. E. Caines, eds., Springer-Verlag, Berlin, Heidelberg.

L. GUO (1994), *Existence and convergence of continuous-time AML*, Systems Control Lett., 22, pp. 111–121.

J. B. MOORE (1988), *Convergence of continuous time stochastic ELS parameter estimation*, Stochastic Process. Appl., 27, pp. 195–215.

L. SCHWARTZ (1984), *Semimartingales and Their Stochastic Calculus on Manifolds*, Presses de l'Université de Montréal, Montréal, pp. 99–104.

J. F. ZHANG AND H. F. CHEN (1992), *Adaptive stabilization under the weakest condition*, Proc. 31st CDC, Dec. 14–18, Tucson, AZ, pp. 3620–3621.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 1, pp. 389–406, January 1996

017

MULTIPLICATIVE INTERIOR GRADIENT METHODS FOR MINIMIZATION OVER THE NONNEGATIVE ORTHANT*

ALFREDO N. IUSEM†, B. F. SVAITER†, AND MARC TEBOULLE‡

Abstract. We introduce a new class of multiplicative iterative methods for solving minimization problems over the nonnegative orthant. The algorithm is akin to a natural extension of gradient methods for unconstrained minimization problems to the case of nonnegativity constraints, with the special feature that it generates a sequence of iterates which remain in the interior of the nonnegative orthant. We prove that the algorithm combined with an appropriate line search is weakly convergent to a saddle point of the minimization problem, when the minimand is a differentiable function with bounded level sets. If the function is convex, then weak convergence to an optimal solution is obtained. Moreover, by using an appropriate regularized line search, we prove that the level set boundeness hypothesis can be removed, and full convergence of the iterates to an optimal solution is established in the convex case.

Key words. multiplicative iterative algorithms, gradient methods, proximal methods, φ-divergences

AMS subject classifications. 90C25, 90C30

1. Introduction.

Consider the problem of minimizing a continuously differentiable function $f : \mathbf{R}^n \to \mathbf{R}$ over the nonnegative orthant of $\mathbf{R}^n$,

$$P: \quad \min\{f(x) : x \in \mathbf{R}^n_+\},$$

where $\mathbf{R}^n_+ = \{x \in \mathbf{R}^n : x_j \geq 0 \ \ (1 \leq j \leq n)\}$. From now on, we assume that problem P has solutions.

In this paper, we are interested in multiplicative iterative methods, which starting with an initial point in the interior $\mathbf{R}^n_{++}$ of the nonnegative orthant, will generate a sequence of feasible interior points $\{x^k\} \in \mathbf{R}^n_{++}$ through an iteration of the form

$$x^{k+1}_j = x^k_j M(\nabla f(x^k)_j, \lambda_k), \quad j = 1, \ldots, n \tag{1}$$

where M is an appropriate mapping, $\nabla f(x^k)_j$ denotes the jth component of the gradient of f, and λ_k is a given positive parameter. As shown below, this class of methods can be seen as a natural extension of gradient methods for unconstrained minimization problems, with the special feature that they generate iterates in the interior of the nonnegative cone. Accordingly, such methods will be called *multiplicative interior gradient* algorithms.

The fact that all the iterates of these algorithms remain in the interior of the nonnegative orthant connects them with interior points methods for linear programming, which became popular after Karmarkar's method [9] proved to be computationally efficient. These methods apply a transformation to the current iterate x^k so as to

* Received by the editors February 9, 1994; accepted for publication (in revised form) October 11, 1994.

† Instituto de Matemática Pura e Aplicada,Estrada Dona Castorina, 110, 22460 Rio de Janeiro, RJ, Brazil. The research of the first author was partially supported by CNPq grant 301280/86.

‡ School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv, Tel-Aviv 69978, Israel. Currently on leave from Department of Mathematics and Statistics, University of Maryland, Baltimore County Campus, Baltimore, MD 21228. The research of this author was partially supported by NSF grants DMS-9201297 and DMS-94011871.

move it away from the boundary of the orthant, and then use minus the gradient of the objective at x^k (projected onto the corresponding affine manifold, if there are linear equality constraints besides positivity) as the moving direction. Then a stepsize is chosen so as to preserve positivity, a new point is obtained, and finally the transformation is reversed. Typical transformations include projective transformations, as in [9], and affine scalings, as in [3], possibly the first among methods of this kind. A survey on this type of algorithms can be found in [6]. In this paper we consider only positivity constraints (so that no projection onto a linear manifold is needed), and we use a distancelike function, called a φ-divergence, which has a penalization effect. We use $-\nabla f(x^k)$ and the φ-divergence in order to generate the direction, and it is automatically ensured that a unit stepsize in this direction will produce a fully positive point. No transformation or change of variables is explicitly performed.

The motivation for considering multiplicative interior gradient algorithms stems from the recent work of Eggermont [4], who studied algorithms of the form (1) with particular choices of the mapping M. As demonstrated in [4], multiplicative iterative algorithms are useful in various interesting applications such as image reconstruction and inverse problems. For further details, we refer the reader to [4] and references therein.

The construction of the family of multiplicative interior gradient algorithms introduced here for solving problem P is derived by imitating the classical gradient method, which can be interpreted as solving

$$x^{k+1} = \operatorname{argmin}\{x^t \nabla f(x^k) + (2\lambda_k)^{-1} \|x - x^k\|^2\}, \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbf{R}^n$. This algorithm can alternatively be also viewed as an *explicit* realization of the proximal regularization method for solving $\inf\{f(x)\colon x \in \mathbf{R}^n\}$; for the proximal minimization algorithm, see, e.g., Rockafellar [12] and Lemaire [10]. To incorporate the nonnegativity constraints of problem P, we replace here the quadratic kernel by a distancelike function, based on the φ-divergence of Csiszár (see, e.g., [2]).

Let $\varphi\colon (0,\infty) \to [0,\infty)$ be a strictly convex and thrice continuously differentiable function which satisfies

$$\varphi(1) = \varphi'(1) = 0, \quad \varphi''(1) > 0, \quad \lim_{t\to 0} \varphi'(t) = -\infty.$$

The class of such functions will be denoted by Φ_1. Define $d_\varphi\colon \mathbf{R}^n_{++} \times \mathbf{R}^n_{++} \to \mathbf{R}_+$ as

$$d_\varphi(x,y) = \sum_{j=1}^{n} y_j \varphi\left(\frac{x_j}{y_j}\right). \tag{3}$$

When $\varphi \in \Phi_1$, d_φ is said to be a φ-divergence. In the context of proximal methods, φ-divergences are studied in [13], where several of their properties are presented. It is easily seen that $d_\varphi(x,y) \geq 0$, with equality if and only if $x = y$, and thus d_φ can be interpreted as a kind of (nonsymmetric) distance between two points in the positive orthant of $\mathbf{R}^n$. In analogy with (2), we define the basic multiplicative interior gradient algorithm (BMIG) as a method generating a sequence $\{x^k\} \subset \mathbf{R}^n_{++}$ according to

$$x^0 > 0, \tag{4}$$

$$x^{k+1} = \operatorname*{argmin}_{x\in\mathbf{R}^n_+}\{x^t \nabla f(x^k) + \lambda_k d_\varphi(x, x^k)\} \tag{5}$$

with $\varphi \in \Phi_1$.

Alternatively, BMIG can also be seen as an explicit version of the φ-divergence proximal algorithm we recently introduced in [8]:

$$x^0 > 0, \tag{6}$$

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathbf{R}^n_+} \{f(x) + \lambda_k d_\varphi(x, x^k)\}, \tag{7}$$

where the function f in (7) is replaced by its linear approximation around x^k in (5). For a convergence analysis and applications of (6)–(7) see [8]. Noting that (5) basically reduces to solving, for x^{k+1},

$$\varphi'\left(\frac{x_j}{x_j^k}\right) = -\lambda_k^{-1} \nabla f(x^k)_j, \quad j = 1, \ldots, n, \tag{8}$$

and assuming that φ' can be easily inverted (as in the case of most relevant examples), we get the explicit formula which is the basis of our algorithm:

$$x_j^{k+1} = x_j^k (\varphi')^{-1}(-\lambda_k^{-1} \nabla f(x^k)_j). \tag{9}$$

Before discussing in the next section the advantages and drawbacks of this approach, we present some of the multiplicative algorithms resulting from iteration (5) for several relevant choices of φ, illustrating the unified framework emerging from our approach to generate multiplicative algorithms.

Example 1. Let $\varphi_1(t) = t - \log t - 1$; then (9) gives

$$x_j^{k+1} = x_j^k (1 + \lambda_k^{-1} \nabla f(x_k)_j)^{-1}, \qquad j = 1, \ldots, n.$$

Example 2. $\varphi_2(t) = t \log t - t + 1$,

$$x_j^{k+1} = x_j^k \exp(-\lambda_k^{-1} \nabla f(x_k)_j)^{-1}, \qquad j = 1, \ldots, n.$$

Example 3. $\varphi_3(t) = (\sqrt{t} - 1)^2$,

$$x_j^{k+1} = x_j^k (1 + \lambda_k^{-1} \nabla f(x_k)_j)^{-2}, \qquad j = 1, \ldots, n.$$

Example 4. $\varphi_4(t) = \frac{1}{2}(t-1)^2$,

$$x_j^{k+1} = x_j^k (1 - \lambda_k^{-1} \nabla f(x_k)_j), \qquad j = 1, \ldots, n.$$

The algorithm in Example 1 was introduced and studied in Eggermont [4], which under appropriate assumptions proved its convergence in the convex case. The algorithms in Examples 2 and 3 appear to be new in the literature and can also be seen as explicit realizations of the φ-divergence proximal methods (see also §7 in [8]). Note that for the first three examples $\varphi \in \Phi_1$. The algorithm which emerges from Example 4 was also pointed out by Eggermont as another possible iterative method for solving P in the convex case; however, no convergence results were established. Note that since φ_4 is not in the class Φ_1, positivity-preserving safeguards are needed; see [7] for the analysis of such an algorithm and its application to maximum-likelihood estimation problems. We also observe the interesting similarity between the algorithms of Examples 2 and 4, the latter being just a "linearization" of the algorithm of Example 2, although it is not clear that this can be used in the analysis of the algorithm given in Example 4. As discussed in the following section, in order to obtain satisfactory convergence results, a linear search must be added to algorithm BMIG, producing the complete form of the multiplicative interior gradient (MIG), which we present in §3.

The iterative formulae for MIG with φ as in Examples 1–3 can be found in (50)–(52), where α_k is the optimal stepsize found in the linear search.

2. Preliminary discussion. The reason for considering iteration (5) as an algorithm for problem P follows from the fact that its fixed points are easily seen to be closely connected to the solutions of problem P. If we take $\lambda_k = \lambda$ a fixed positive constant, and x^* a fixed point of (9) we have

$$x_j^* = x_j^*(\varphi')^{-1}(-\lambda^{-1}\nabla f(x^*)_j), \tag{10}$$

and therefore either $x_j^* = 0$ or $\lambda^{-1}\nabla f(x^*)_j = \varphi'(1) = 0$, implying $\nabla f(x^*)_j = 0$. It follows that if the sequence $\{x^k\}$ generated by (5) is convergent, then its limit x^* satisfies $x^* \geq 0$, $\nabla f(x^*)^t x^* = 0$, and it can be proved that when f is convex, we also have $\nabla f(x^*) \geq 0$, in which case x^* is a solution of P (see §4).

Iteration (5) seems therefore, in view of the explicit formula (9), also an appealing option for implementing an approximate version of the φ-divergence proximal method. However, in its bare formulation given in (5) and (9), the method exhibits some serious problems. The first one refers to the existence of solutions of (5). While in the case of the φ-divergence proximal method, existence of solutions for (7) are guaranteed by the existence of solutions of problem P together with positivity and strict convexity of d_φ, this is not true anymore for (5), since the objective $\nabla f(x^k)^t x$ may fail to attain its minimum on $\mathbf{R}^n_+$ (in fact this happens whenever $\nabla f(x^k)_j < 0$ for some j). In terms of (8), lack of solution of the subproblem in (5) translates as $-\lambda_k^{-1}\nabla f(x^k)_j^t x$ not belonging to $\mathrm{Im}\varphi'$, the range of φ'. Since $\varphi \in \Phi_1$, it holds that $\lim_{t\to 0}\varphi'(t) = -\infty, \varphi'(1) = 0$, and φ' is increasing so that $\mathrm{Im}\varphi' = (-\infty, \eta)$ with $\eta = \lim_{t\to\infty}\varphi'(t) > 0$. If $-\lambda_k^{-1}\nabla f(x^k)_j > \eta$, then the minimization problem in (5) has no solutions. This cannot happen if $\eta = \infty$, as for φ_2 of Example 2, but it can occur with φ_1 of Example 1, for which $\eta = 1$. In this case the right-hand side in that example could be negative so that the iterate x^{k+1} would not be a solution of the subproblem given by (5). There are two ways to overcome this obstacle. One is to demand that φ satisfies $\lim_{t\to\infty}\varphi'(t) = \infty$, but then we must exclude several interesting choices of φ's from the analysis, such as φ_1 and φ_3 of Examples 1 and 3 (for both of which $\eta = 1$). Another option, which will be followed in this paper, is to adjust the regularization coefficients λ_k so that $-\nabla f(x^k)_j\lambda_k^{-1} < \eta$ for all k.

The second difficulty with iteration (5) is far more serious and stems from the fact that, while the fixed points of (5) are, in general, solutions of P, they are not necessarily attractors for the iteration. Another way to see this obstacle is the following: it is easy to check that iteration (7) implies $f(x^{k+1}) \leq f(x^k)$, and this descent feature of algorithm (7) is a key in the proof of its convergence [8]. This desirable property is not shared by BMIG, as the following examples show.

Example 5. Take $n = 1, \lambda_k = 1$ for all k, and $\varphi = \varphi_2$ as in Example 2. Here, $\mathrm{Im}\varphi_2' = (-\infty, \infty)$ so that x^{k+1} is always well defined by (5). Consider first the objective function $f(x) = ax^2 - bx$, with $a = 2/e - 1, b = (e+1)/(e-1)$. Note that f is a strongly convex quadratic function and the solution of problem P is $x^* = (e+1)/2$. If we start with $x^0 = 1$, then $x^1 = e, x^2 = 1$, and the sequence oscillates between 1 and e, forever missing x^*, which is the midpoint between these two limit points. For this choice of f the sequence $\{x^k\}$ is at least bounded, but this is not always the case. Consider now for instance $f(x) = 1/x + x^2$. This f is also strictly convex and the solution of P is $x^* = 2^{-1/3}$. It can be verified that if $x^0 > x^*$, then $\lim_{k\to\infty} x^{2k} = \infty$, $\lim_{k\to\infty} x^{2k+1} = 0$, and the values of such limits reverse when $x^0 < x^*$. Of course for

$x^0 = x^*$ we get $x^k = x^*$ for all k, but x^* is a repeller, rather than an attractor, for iteration (5), so that even a local convergence result is not attainable.

This obstacle can also be removed by choosing a sufficiently large value for λ_k, but the trouble is that such value cannot be determined with the information available at iteration k. Another way to overcome this difficulty is suggested by the fact that the direction $(x^{k+1} - x^k)$, with x^{k+1} given by (5), is a descent one. More precisely, the directional derivative of f at x^k in this direction is given by

$$(x^{k+1} - x^k)^t \nabla f(x^k) = \lambda_k \sum_{j=1}^{n} (x_j^k - x_j^{k+1}) \varphi'(x_j^{k+1}/x_j^k) \leq 0. \tag{11}$$

It follows that adding a line search in the segment between x^k and x^{k+1} to iteration (5) will at least produce a sequence with decreasing functional values. This will be the approach taken in this paper. We remark that even with the line search, iteration (5) is far easier to implement than the φ-divergence proximal method (6)–(7), since minimization in a segment substitutes for minimization in an orthant of $\mathbf{R}^n$ (i.e., for the solution of the $n \times n$ nonlinear system (8)).

The issue arises of what function should be minimized in the segment between x^k and the point given by the left-hand side of (8) (which we call y^k, reserving x^{k+1} for the result of the line search). Two obvious candidates are $f(x)$ and $f(x) + \lambda_k d_\varphi(x, x^k)$; both guarantee a decreasing (and henceforth convergent) objective-value sequence $\{f(x^k)\}$. From the point of view of the convergence analysis neither of these is the best choice. The reason is the following. In order to establish convergence of $\{x^k\}$ we will follow the approach developed in [8], proving quasi-Fejér convergence of $\{x^k\}$ to the set of solutions of problem P, meaning that x^{k+1} cannot be much farther away than x^k from any solution z (see definition and properties in §5). Quasi-Fejér convergence implies boundedness of $\{x^k\}$ and ultimately convergence to a solution. Under the choices given above for the line search objective we are not able to prove quasi-Fejér convergence. We have developed, however, in §4 some convergence results which hold when we choose just $f(x)$ as the minimand of the line search. We enforce boundedness of $\{x^k\}$ by assuming that f has one bounded level set which contains x^0 and get a so-called *weak convergence* result in the following sense:

DEFINITION 1. *A sequence $\{x^k\} \subset \mathbf{R}^n$ is said to be weakly convergent to a set $S \subset \mathbf{R}^n$ if*

(i) *the sequence is bounded,*

(ii) $\lim_{k\to\infty}(x^{k+1} - x^k) = 0$,

(iii) *all limit points of $\{x^k\}$ belong to S.*

In §4 we prove that for $\varphi \in \Phi_1$, arbitrary f, and λ_k in an appropriate interval, $\{x^k\}$ is weakly convergent to the set of points u which satisfy

$$u \geq 0, \quad u_j \nabla f(u)_j = 0, \quad j = 1, \ldots, n,$$

i.e., two of the three Karush–Kuhn–Tucker optimality conditions for problem P. We do not get $\nabla f(u) \geq 0$, so a limit point of $\{x^k\}$ could fail to be a solution of P. However, for a convex f, we prove $\nabla f(u) \geq 0$ for every limit point u, so $\{x^k\}$ is weakly convergent to the set of solutions of P. It follows that for a strictly convex f the sequence $\{x^k\}$ converges to the unique solution of P (boundedness of a level set is redundant in this case).

In §5 we prove that, adding an appropriately chosen regularization term to $f(x)$ in the line search, we can eliminate the level set boundedness hypothesis and furthermore

get full convergence when f is convex and φ is in a subset of the class of functions in Φ_1 defined by

$$\Psi := \{\varphi \in \Phi_1 : \ \varphi'(t) \leq \varphi''(1)\log t, \ \ \forall t > 0\}.$$

Note that the functions φ_i in the Examples 1–3 satisfy the above inequality. The function φ_4 of Example 4 does not belong to Ψ.

We decided to keep the weaker results for the nonregularized line search of §4 for three reasons. First, they hold for a wider choice of f and φ; second, most of these results will be used in the proofs of §5, so there is no duplication; and finally, it can be argued that our algorithm is akin to a natural extension to the case of positivity constraints of the steepest descent method for unconstrained optimization in the sense that in both of them a descent direction is chosen and a line search is performed in this line. The descent direction of our algorithm is not the steepest one but one which automatically takes care of the positivity constraints, so it is interesting to compare convergence results when the same objective, namely $f(x)$, is used in the line search of both algorithms, and it turns out that the results of §4 are the same that hold for the steepest-descent method under similar hypotheses on f (see, e.g., Polyak [11]).

3. Formal definition of algorithm MIG. Take $\varphi \in \Phi_1$. Let $\eta = \lim_{t\to\infty} \varphi'(t)$. It follows from the definition of Φ_1 that $\eta > 0$ (possibly $\eta = \infty$). For any vector $x \in \mathbf{R}^n$ let $x^- \in \mathbf{R}^n$ be defined by $x_j^- = \max\{0, -x_j\}$, and $||x||_\infty = \max_{1\leq j\leq n} |x_j|$ denotes the l_∞-norm. In order to define the algorithm we need four exogenous real constants $\hat{\beta}, \tilde{\beta}, \hat{\lambda}, \tilde{\lambda}$ satisfying

$$1 < \hat{\beta} \leq \tilde{\beta}, \tag{12}$$

$$0 < \hat{\lambda} \leq \tilde{\lambda}. \tag{13}$$

$\hat{\lambda}, \tilde{\lambda}$ are a priori candidates for lower and upper bounds of the regularization parameters λ_k. $\hat{\beta}$ and $\tilde{\beta}$ will be used to generate surrogate bounds $\hat{\gamma}_k, \tilde{\gamma}_k$ for λ_k, required in the dynamical adjustment of the λ_k's which guarantees that the minimization problem in (5) has solutions, even when $\eta < \infty$ (both $\hat{\gamma}_k$ and $\tilde{\gamma}_k$ are zero when $\eta = \infty$). The constants in (13) guarantee that the sequence $\{\lambda_k\}$ is bounded.

The multiplicative interior gradient algorithm (MIG).

Initialization.

$$x^0 > 0. \tag{14}$$

Iterative step. Given $x^k > 0$, define

$$\hat{\gamma}_k = \frac{\hat{\beta}}{\eta} \left\|\nabla f(x^k)^-\right\|_\infty, \tag{15}$$

$$\tilde{\gamma}_k = \frac{\tilde{\beta}}{\eta} \left\|\nabla f(x^k)^-\right\|_\infty, \tag{16}$$

and choose any $\lambda_k \in \mathbf{R}$ such that

$$\max\{\hat{\gamma}_k, \hat{\lambda}\} \leq \lambda_k \leq \max\{\tilde{\gamma}_k, \tilde{\lambda}\}. \tag{17}$$

Compute

$$y^k = \operatorname*{argmin}_{x\geq 0}\{\nabla f(x^k)^t x + \lambda_k d_\varphi(x, x^k)\}, \tag{18}$$

$$\alpha_k = \operatorname*{argmin}_{\alpha\in[0,1]}\{f(x^k + \alpha(y^k - x^k))\}, \tag{19}$$

$$x^{k+1} = x^k + \alpha_k(y^k - x^k). \tag{20}$$

We remark that if $\alpha_\ell = 0$ for some ℓ, then it follows from (14)–(20) that $x^{\ell+1} = x^\ell$ and consequently $x^k = x^\ell$ for all $k \geq \ell$; i.e., in practice the computation stops at $k = \ell$. It is not difficult to show that in such a case $\nabla f(x^\ell) = 0$; i.e., x^ℓ is a stationary point for problem P, but there is no need to prove this fact, because by considering that the sequence $\{x^k\}$ is always infinite (even if it stays always at the same point), our convergence proof simultaneously covers the cases of finite and infinite termination.

4. Convergence analysis for the MIG algorithm. The next two propositions and corollary do not use the specific form of the minimand in the line search given by (19), and so they hold also for the algorithm considered in §5. We start with an elementary property of φ and $\hat{\varphi}(t) := (t-1)\varphi'(t)$.

PROPOSITION 1. *Take* $\varphi \in \Phi_1$, $\{y^k\}, \{z^k\} \subset \mathbf{R}^n_{++}$, $\{z^k\}$ *bounded. It holds that*

i) *if* $\lim_{k\to\infty} d_\varphi(y^k, z^k) = 0$, *then* $\lim_{k\to\infty}(y^k - z^k) = 0$,

ii) *if* $\lim_{k\to\infty} d_{\hat{\varphi}}(y^k, z^k) = 0$, *then* $\lim_{k\to\infty}(y^k - z^k) = 0$.

Proof. i). Assume that the result is false so that there exists subsequences $\{y^{\ell_k}\}, \{z^{\ell_k}\}$ of $\{y^k\}, \{z^k\}$, respectively, such that $|y_j^{\ell_k} - z_j^{\ell_k}| > \varepsilon$ for some j and some $\varepsilon > 0$. Take ξ such that $z_j^k \leq \xi$ for all k, j. If $|y_j^{\ell_k} - z_j^{\ell_k}| > \varepsilon$, then either $z_j^{\ell_k} \geq y_j^{\ell_k} + \varepsilon > \varepsilon$ or $y_j^{\ell_k} \geq z_j^{\ell_k} + \varepsilon > \varepsilon$. In the first case $y_j^{\ell_k}/z_j^{\ell_k} \leq 1 - \varepsilon/z_j^{\ell_k} \leq 1 - \varepsilon/\xi < 1$ so that $\varphi(y_j^{\ell_k}/z_j^{\ell_k}) > \varphi(1 - \varepsilon/\xi) > 0$ (recall that $\varphi(t)$ is decreasing for $0 < t < 1$) and therefore

$$d_\varphi(y^{\ell_k}, z^{\ell_k}) \geq z_j^{\ell_k}\varphi(y_j^{\ell_k}/z_j^{\ell_k}) > \varepsilon\varphi\left(1 - \frac{\varepsilon}{\xi}\right). \tag{21}$$

In the second case $y_j^{\ell_k}/z_j^{\ell_k} \geq 1 + \varepsilon/z_j^{\ell_k} > 1$ so that

$$\varphi(y_j^{\ell_k}/z_j^{\ell_k}) \geq \varphi(1+\varepsilon/z_j^{\ell_k}) \geq \varphi(1+\varepsilon/2z_j^{\ell_k}) + \varphi'(1+\varepsilon/2z_j^{\ell_k})\varepsilon/2z_j^{\ell_k} \geq \varphi'(1+\varepsilon/2\xi)\varepsilon/2z_j^{\ell_k} \tag{22}$$

using convexity of φ and $\varphi(t) > 0$, $\varphi'(t) > 0$ for $t > 1$. It follows from (22) that

$$d_\varphi(y^{\ell_k}, z^{\ell_k}) \geq z_j^{\ell_k}\varphi(y_j^{\ell_k}/z_j^{\ell_k}) \geq \frac{\varepsilon}{2}\varphi'\left(1 + \frac{\varepsilon}{2\xi}\right). \tag{23}$$

From (21), (23), $d_\varphi(y^{\ell_k}, z^{\ell_k}) \geq \varepsilon \min\{\varphi(1 - \frac{\varepsilon}{\xi}), \frac{1}{2}\varphi'(1 + \frac{\varepsilon}{2\xi})\} > 0$, in contradiction with $\lim_{k\to\infty} d_\varphi(y^k, z^k) = 0$.

ii) The proof follows from (i) using (11). □

The next proposition just checks that the MIG algorithm is well defined and preserves positivity. From now on $\{x^k\}$ refers to the sequence generated by (14)–(20).

PROPOSITION 2. x^k *is well defined and* $x^k > 0$ *for all* k.

Proof. The proof is by induction. The result holds for $k = 0$ by (14). Assume $x^k > 0$. We establish first the following facts:

a) There exists a unique $t_j^k > 0$ such that $\varphi'(t_j^k) = -\frac{\nabla f(x^k)_j}{\lambda_k}$.

b) If $\eta < \infty$, then there exists $\tilde{t} < \infty$ such that $t_j^k \leq \tilde{t}$ for all j, k.

Since $\varphi \in \Phi_1$, we know that φ' is increasing and $\lim_{t\to 0}\varphi'(t) = -\infty$. It follows that the equation $\varphi'(t) = s$, in the unknown t, has a unique solution for all $s \in \mathbf{R}$ if

$\eta = \infty$ and for all $s < \eta$ if $\eta < \infty$, so that (a) holds when $\eta = \infty$. We now prove (a) and (b) for $\eta < \infty$.

$$-\frac{\nabla f(x^k)_j}{\lambda_k} \leq \frac{\left\|\nabla f(x^k)^-\right\|_\infty}{\lambda_k} \leq \frac{\left\|\nabla f(x^k)^-\right\|_\infty}{\hat{\gamma}_k} = \frac{\eta}{\hat{\beta}} < \eta$$

using (17), (15), and (12). It follows that t_j^k exists (solving $\varphi'(t) = -\nabla f(x^k)_j/\lambda_k$) and, taking $\tilde{t}$ as the solution of $\varphi'(t) = \frac{\eta}{\tilde{\beta}}$, we conclude that $t_j^k \leq \tilde{t}$ from the fact that φ' is increasing. (a) and (b) have been established.

The gradient of the minimand in (18) has jth component equal to

$$\nabla f(x^k)_j + \lambda_k \varphi'\left(\frac{x_j}{x_j^k}\right). \tag{24}$$

Taking $x_j = t_j^k x_j^k$, it follows from (b) that (24) vanishes. Since the minimand in (18) is convex, we conclude that the vector y^k with components $y_j^k = t_j^k x_j^k$ is the solution of (18). Since $t_j^k > 0$ by (a) and $x_j^k > 0$ by inductive hypothesis, we get $y^k > 0$. A minimizer $\alpha_k \in [0, 1]$ of (19) exists by continuity of f. Hence $x^k > 0$, $y^k > 0$, and (20) imply $x^{k+1} > 0$. □

COROLLARY 1. *For all* $k \geq 0$

$$\varphi'\left(\frac{y_j^k}{x_j^k}\right) = -\frac{\nabla f(x^k)_j}{\lambda_k}.$$

Proof. The proof is immediate from Proposition 2. □

The following proposition establishes the monotonicity and boundedness properties of the algorithm under the hypothesis that x^0 belongs to a bounded level set of f.

For $\rho \in \mathbf{R}$, let $L(\rho) = \{x \in \mathbf{R}_+^n : f(x) \leq \rho\}$ and $f^* = \min\{f(x) : x \in \mathbf{R}_+^n\}$.

PROPOSITION 3. *If there exists* $\rho \in \mathbf{R}$ *such that* $L(\rho)$ *is bounded and* $x^0 \in L(\rho)$, *then*

i) $f^* \leq f(x^{k+1}) \leq f(x^k)$ *for all* k,

ii) $\{f(x^k)\}$ *converges,*

iii) $\{x^k\}$ *is bounded,*

iv) $\{\lambda_k\}$ *is bounded,*

v) $\{y^k\}$ *is bounded.*

Proof. i) The left inequality follows from Proposition 2, and the right one from (19), (20).

ii) The proof follows from (i).

iii) $f(x^k) \leq f(x^0)$ for all k by (i) so that $x^k \in L(\rho)$ for all k.

iv) Let $\mu = \max_{x \in L(\rho)}\{\|\nabla f(x)^-\|_\infty\} \cdot \mu < \infty$ by continuous differentiability of f and compactness of $L(\rho)$. From (16) and (iii) $\tilde{\gamma}_k \leq \frac{\tilde{\beta}}{\eta}\mu$, and from (17) $\lambda_k \leq \max\{\frac{\tilde{\beta}}{\eta}\mu, \tilde{\lambda}\}$.

v) $y_j^k = t_j^k x_j^k$ with t_j^k as in (a) of the proof of Proposition 2, which also gives $y^k > 0$. If $\eta < \infty$ we have $y_j^k < \tilde{t} x_j^k$ by (b) of the proof of Proposition 2, and the result follows from (iii). If $\eta = \infty$, then, using Corollary 1, (17), and μ as in (iv),

$$\varphi'(t_j^k) = \varphi'(\frac{y_j^k}{x_j^k}) = -\frac{\nabla f(x^k)_j}{\lambda_k} \leq \frac{\left\|\nabla f(x^k)^-\right\|_\infty}{\lambda_k} \leq \frac{\mu}{\tilde{\lambda}}$$

so that $t_j^k \leq (\varphi')^{-1}(\frac{\mu}{\lambda})$; i.e., t_j^k is bounded and the result follows from (iii). □

The next proposition establishes the second requirement in the definition of weak convergence.

PROPOSITION 4. *Under the hypothesis of Proposition* 3, $\lim_{k\to\infty}(x^{k+1} - x^k) = 0$.

Proof. Assume the result is false. Then by Proposition 3(iii) there exists a subsequence $\{x^{\ell_k}\}$ of $\{x^k\}$ such that $\lim_{k\to\infty} x^{\ell_k} = u \neq v = \lim_{k\to\infty} x^{\ell_k+1}$. Define h_k, $h\colon [0,1] \to \mathbf{R}$ as $h_k(t) = f((1-t)x^k + tx^{k+1})$, $h(t) = f((1-t)u + tv)$, so that $\lim_{k\to\infty} h_{\ell_k}(t) = h(t)$ for all $t \in [0,1]$. By Corollary 1, for all k

$$h_k'(0) = \nabla f(x^k)^t(x^{k+1} - x^k) = \alpha_k \nabla f(x^k)^t(y^k - x^k) = -\alpha_k \lambda_k \sum_{j=1}^n \varphi'\left(\frac{y_j^k}{x_j^k}\right)(y_j^k - x_j^k)$$

(25)

$$= -\alpha_k \lambda_k d_{\hat{\varphi}}(y^k, x^k) \leq 0.$$

From (25)

$$h'(0) = \lim_{k\to\infty} h_{\ell_k}'(0) \leq 0. \tag{26}$$

By (19), (20), $h_k(1) \leq h_k(t)$ for all $t \in [0,1]$, implying

$$h(1) \leq h(t) \qquad (t \in [0,1]). \tag{27}$$

By Proposition 3(ii) and the definitions of u and v, $f(u) = f(v)$, implying

$$h(0) = h(1). \tag{28}$$

From (27), (28), $h(0) \leq h(t)$ for all $t \in [0,1]$, implying $h'(0) \geq 0$, and therefore, using (26),

$$0 = h'(0) = \nabla f(u)^t(v - u). \tag{29}$$

We may assume, refining the subsequence $\{x^{\ell_k}\}$ if necessary, that $\lim_{k\to\infty} \alpha_{\ell_k} = \overline{\alpha}$ and, using Proposition 3(v), that $\lim_{k\to\infty} y^{\ell_k} = y$. From (20)

$$v - u = \overline{\alpha}(y - u). \tag{30}$$

$u \neq v$ implies $\overline{\alpha} \neq 0$. Replacing (30) in (29) and using Corollary 1

(31)

$$0 = \nabla f(u)^t(y - u) = \lim_{k\to\infty} \lambda_{\ell_k} \sum_{j=1}^n \varphi'\left(\frac{y_j^{\ell_k}}{x_j^{\ell_k}}\right)\left(y_j^{\ell_k} - x_j^{\ell_k}\right) = -\lim_{k\to\infty} \lambda_{\ell_k} d_{\hat{\varphi}}(y^{\ell_k}, x^{\ell_k}).$$

Since $\lambda_{\ell_k} \geq \hat{\lambda} > 0$, we conclude from (31) that $\lim_{k\to\infty} d_{\hat{\varphi}}(y^{\ell_k}, x^{\ell_k}) = 0$ and then, from Proposition 1(ii), that $y = u$, in which case we get, from (30), $u = v$, in contradiction with the assumption. □

The next proposition completes our weak convergence result for a general (i.e., not necessarily convex) f.

PROPOSITION 5. *Let u be a limit point of $\{x^k\}$. Under the hypothesis of Proposition* 3, *$u \geq 0$ and $u_j \nabla f(u)_j = 0$ for all j.*

Proof. $u \geq 0$ follows from Proposition 2. In order to prove $u_j \nabla f(u)_j = 0$, we define

$$\overline{h}_k(\alpha) = f((1-\alpha)x^k + \alpha y^k);$$

i.e., $\overline{h}_k(\alpha)$ is the minimand of (19), so that $\overline{h}_k'(\alpha_k) = \nabla f(x^{k+1})^t(y^k - x^k)$, and we have $\overline{h}_k'(\alpha_k) \geq 0$ if $\alpha_k = 0$, $\overline{h}_k'(\alpha_k) = 0$ if $\alpha_k \in (0,1)$, $\overline{h}_k'(\alpha_k) \leq 0$ if $\alpha_k = 1$. Let $\{x^{\ell_k}\}$ be a subsequence of $\{x^k\}$ such that $\lim_{k\to\infty} x^{\ell_k} = u$. Without loss of generality, i.e., refining the subsequence if necessary, we may assume, in view of Proposition 3(iv), 3(v), and (17), that $\lim_{k\to\infty} y^{\ell_k} = y$, $\lim_{k\to\infty} \lambda_{\ell_k} = \lambda > 0$. We claim that $y = u$. We consider two cases:

i) there exists a subsequence $\{x^{i_k}\}$ of $\{x^{\ell_k}\}$ such that $\alpha_{i_k} = 1$ for all k. In this case, by (20), $x^{i_k+1} = y^{i_k}$ and, using Proposition 4, $u = \lim_{k\to\infty} x^{i_k} = \lim_{k\to\infty} x^{i_k+1} = \lim_{k\to\infty} y^{i_k} = y$, and the claim holds.

ii) If (i) does not occur, then $\alpha_{\ell_k} \in [0,1)$ for large enough k, implying

$$0 \leq \overline{h}_k'(\alpha_{\ell_k}) = \nabla f(x^{\ell_k+1})^t(y^{\ell_k} - x^{\ell_k}). \tag{32}$$

From (32), using Corollary 1 and (17)

$$\begin{aligned}(\nabla f(x^{\ell_k+1}) - \nabla f(x^{\ell_k}))^t(y^{\ell_k} - x^{\ell_k}) &\geq -\nabla f(x^{\ell_k})^t(y^{\ell_k} - x^{\ell_k}) \\ &= \lambda_{\ell_k} \sum_{j=1}^{n} \varphi'(y_j^{\ell_k}/x_j^{\ell_k})(y_j^{\ell_k} - x_j^{\ell_k}) \\ &= \lambda_{\ell_k} d_{\hat{\varphi}}(y^{\ell_k}, x^{\ell_k}) \\ &\geq \hat{\lambda} d_{\hat{\varphi}}(y^{\ell_k}, x^{\ell_k}) \geq 0. \qquad (33)\end{aligned}$$

Taking limits in (33) as k goes to ∞, and using Propositions 3(iii), 3(v), and 4, $0 = \lim_{k\to\infty} d_{\hat{\varphi}}(y^{\ell_k}, x^{\ell_k})$. By Proposition 1(ii) $y - u = \lim_{k\to\infty}(y^{\ell_k} - x^{\ell_k}) = 0$ and the claim holds.

Take j such that $u_j \neq 0$. By Corollary 1, $-\nabla f(u)_j = -\lim_{k\to\infty} \nabla f(x^{\ell_k})_j = \lim_{k\to\infty} \lambda_{\ell_k} \varphi'(y_j^{\ell_k}/x_j^{\ell_k}) = \lambda\varphi'(y_j/u_j) = \lambda\varphi'(1) = 0$. We have shown that $\nabla f(u)_j = 0$ whenever $u_j \neq 0$, and the result is established. □

We summarize the results obtained up to now in the following theorem.

THEOREM 1. *If there exists ρ such that $L(\rho)$ is bounded and $x^0 \in L(\rho)$, then the sequence $\{x^k\}$ generated by MIG satisfies the following:*

i) *$\{x^k\}$ is bounded.*

ii) $\lim_{k\to\infty}(x^{k+1} - x^k) = 0$.

iii) *If u is a limit point of $\{x^k\}$, then $u \geq 0$ and $u_j \nabla f(u)_j = 0$ for all j.*

Proof. The proof follows from Propositions 3(iii), 4, and 5. □

For convex f, the result of Theorem 1 can be improved and weak convergence to the solution set of problem P can be established. We need first a preliminary lemma.

LEMMA 1. *Take $g: \mathbf{R}^n \to \mathbf{R}$ convex and continuously differentiable and $\{z^k\} \subset \mathbf{R}^n_{++}$. Let $S = \{u \in \mathbf{R}^n : u \geq 0, u_j \nabla g(u)_j = 0 \ (1 \leq j \leq n)\}$ and T equal the set of solutions of $\min g(x)$ such that $x \geq 0$, so that $T \subset S$. If*

i) *$\{z^k\}$ is weakly convergent to S,*

ii) *there exists g^* such that $g(u) = g^*$ for any limit point u of $\{z^k\}$,*

iii) *$z_j^{k+1} > z_j^k$ if and only if $\nabla g(z^k)_j < 0$,*

then $\{z^k\}$ is weakly convergent to T.

Proof. By convexity of g, $T = \{u \in \mathbf{R}^n : u \geq 0, \nabla g(u) \geq 0, u^t \nabla g(u) = 0\} = \{u \in S : \nabla g(u) \geq 0\}$, so it suffices to prove that $\nabla g(u) \geq 0$ for any limit point u of $\{z^k\}$. Take a limit point u of $\{z^k\}$ and assume that there exists j such that $\nabla g(u)_j < 0$, i.e., that $\emptyset \neq J = \{j : \nabla g(u)_j < 0\}$. Let $K = \{1, \ldots, n\} \backslash J$ and $V = \{x \in \mathbf{R}^n_+ : x_j = 0$ for $j \in J\}$. It follows from (i) that $u \in S$ and, from the definition of S, that $u \in V$.

Therefore, for all $j \in K$ it holds that $u_j \geq 0$, $\nabla g(u)_j \geq 0$, $u_j \nabla g(u)_j = 0$ which are precisely the Karush–Kuhn–Tucker conditions of $\min g(x)$ such that $x \in V$. By convexity of g, u minimizes g on V. Let $B = \{x \in \mathbf{R}^n_+ : \nabla f(x)_j < 0 \text{ for } j \in J\}$. By definition, $u \in B$ which is open as a subset of $\mathbf{R}^n_+$. Take a subsequence $\{z^{\ell_k}\}$ of $\{z^k\}$ such that $\lim_{k\to\infty} z^{\ell_k} = u$ and $\{z^{\ell_k}\} \subset B$.

Let $p_k = \max\{q < \ell_k : z^q \notin B\}$ (set $p_k = 0$ if $z^q \in B$ for all $q < \ell_k$). Then $z^q \in B$ for $p_k + 1 \leq q \leq \ell_k$ and, using (iii) iteratively,

$$z_j^{\ell_k} \geq z_j^{p_k+1} \quad \text{for all } j \in J. \tag{34}$$

If there exists p such that $p_k = p$ for large enough k (i.e., if the sequence $\{z^k\}$ stays in B for large enough k), then we get $0 = u_j = \lim_{k\to\infty} x_j^{\ell_k} \geq x_j^{p+1} > 0$ for all $j \in J$ (where $u_j = 0$ follows from $u \in S$ and the definition of J), which is a contradiction. It follows that $\lim_{k\to\infty} p_k = \infty$, so that $\{z^{p_k}\}$ is a subsequence of $\{z^k\}$. By (34), for $j \in J$, $\lim_{k\to\infty} z_j^{p_k+1} \leq \lim_{k\to\infty} z_j^{\ell_k} = u_j = 0$, implying $\lim_{k\to\infty} z_j^{p_k+1} = 0$, and by (i) and condition (ii) of Definition 1, we get $\lim_{k\to\infty} z_j^{p_k} = 0$ for all $j \in J$. By definition of p_k, we have $z^{p_k} \notin B$ whenever $p_k > 1$. Take a subsequence $\{z^{i_k}\}$ of $\{z^k\}$, which converges, say, to v. (Existence of $\{z^{i_k}\}$ follows from (i) and condition (i) in Definition 1.) Since B is open in $\mathbf{R}^n_+$, $v \notin B$ and therefore there exists $m \in J$ such that

$$\nabla g(v)_m \geq 0. \tag{35}$$

Let $W = \{x \in \mathbf{R}^n_+ : x_j = 0 \text{ for } j \in J \backslash \{m\}\}$. The Karush–Kuhn–Tucker conditions for $\min g(x)$ such that $x \in W$ are

$$x_j \geq 0, \quad \nabla g(x)_j \geq 0, \quad x_j \nabla g(x)_j \geq 0 \quad \text{for } j \in K \cup \{m\}. \tag{36}$$

We claim that v satisfies (36). Consider first $j \in K$. Note that for $j \in J$, $v_j = \lim_{k\to\infty} z_j^{i_k} = \lim_{k\to\infty} z_j^{p_k} = 0$, so that $v \in V$. By (ii), $g(v) = g(u) = g^*$. Since u minimizes g on V, we conclude that v also minimizes g on V, and therefore it satisfies the Karush–Kuhn–Tucker conditions for $\min g(x)$ such that $x \in V$, which are precisely (36) with $j \in K$. Consider now $j = m$. The first and third conditions, in (36) hold by (i), and the second by (35). So v satisfies (36) and, by convexity of g, v minimizes g on W. Since $u \in V \subset W$ and $g(u) = g(v)$ by (ii), it follows that u also minimizes g on W, and therefore, it satisfies (36). In particular $\nabla g(u)_m \geq 0$. Since $m \in J$ this contradicts the definition of J. It follows that $J = \emptyset$ and therefore $\nabla g(u) \geq 0$. □

Next we give the weak convergence result for convex f.

THEOREM 2. *If there exists ρ such that $L(\rho)$ is bounded, and $x^0 \in L(\rho)$ and f is convex, then the sequence $\{x^k\}$ generated by Algorithm MIG satisfies the following:*

i) *$\{x^k\}$ is bounded.*

ii) $\lim_{k\to\infty}(x^{k+1} - x^k) = 0$.

iii) *Every limit point u of $\{x^k\}$ solves problem P.*

Proof. The statement of the theorem just says that $\{x^k\}$ is weakly convergent to the set of solutions of problem P. We use Lemma 1 with $\{z^k\} = \{x^k\}$, $g = f$. Hypotheses (i) and (ii) of Lemma 1 hold by Theorem 1 and Proposition 3(i), respectively. For hypothesis (iii) of Lemma 1, since $x_j^{k+1} - x_j^k = \alpha_k(y_j^k - x_j^k)$ with $\alpha_k \geq 0$ by (20), we have $x_j^{k+1} - x_j^k > 0$ iff $y_j^k > x_j^k$ iff $\varphi'(y_j^k/x_j^k) > 0$ iff $\nabla f(x^k)_j < 0$, using Corollary 1 and $\lambda_k > 0$. □

COROLLARY 2. *If f is strictly convex, then the sequence $\{x^k\}$ generated by algorithm MIG converges to the solution x^* of problem P.*

Proof. Note first that existence of solutions of problem P and strict convexity of f imply that all level sets of f are bounded, and so the hypothesis that $x^0 \in L(\rho)$ in Theorem 2 is redundant. By Theorem 2(iii) any limit point of $\{x^k\}$ is a solution of problem P. For strictly convex f problem P has a unique solution, so there is a unique limit point and the result follows. □

5. The algorithm with regularized line searches. In this section we present a regularization of the line search in (19) which allows us to establish full convergence of $\{x^k\}$ to a solution of problem P when f is convex and $\varphi \in \Psi$, without any hypothesis on boundedness of the level sets of f. Before presenting this modification of algorithm MIG, which will be called algorithm RMIG, we need some additional notation.

Given $\varphi \in \Phi_1$, define $\hat{\varphi}$ as

$$\hat{\varphi}(t) = (t-1)\varphi'(t).$$

Although $\hat{\varphi}$ may fail to belong to Φ_1, because it may not be convex, $d_{\hat{\varphi}}$, as defined by (3) with $\hat{\varphi}$ substituting for φ, shares most properties of φ-divergences. Also, since φ is convex and $\varphi(1) = 1$, we have

$$\varphi(t) = \varphi(t) - \varphi(1) \leq (t-1)\varphi'(t) = \hat{\varphi}(t),$$

which implies $d_\varphi(x,y) \leq d_{\hat{\varphi}}(x,y)$ for all $x, y > 0$.

RMIG is defined by (14)–(20) except that (19) is replaced by

$$\alpha_k = \operatorname*{argmin}_{\alpha \in [0,1]} \{f(x^k + \alpha(y^k - x^k)) + \sigma\alpha\}, \tag{37}$$

where

$$\sigma = \delta\lambda_k d_{\hat{\varphi}}(y^k, x^k) \tag{38}$$

with $\delta \in (0,1)$.

The motivation behind the choice of a linear regularization term is the following. For x in the segment between x^k and y^k we have

$$d_\varphi(x, x^k) = d_\varphi((1-\alpha)x^k + \alpha y^k, x^k) \leq \alpha d_\varphi(y^k, x^k) \leq \alpha d_{\hat{\varphi}}(y^k, x^k)$$

using convexity of d_φ in its first variable. Note that $\alpha d_{\hat{\varphi}}(y^k, x^k)$ is a linear function of α with positive slope. So we take a linear regularization term; i.e., we minimize $f((1-\alpha)x^k + \alpha y^k) + \sigma\alpha$ with $\alpha \in [0,1]$. It can be seen, using (3) and (18), that the derivative of this function with respect to α at $\alpha = 0$ is $\sigma - \lambda_k d_{\hat{\varphi}}(y^k, x^k)$. Since we want a descent direction, such derivative must be negative and so we take $\sigma = \delta\lambda_k d_{\hat{\varphi}}(y^k, x^k)$ with $\delta \in (0,1)$. With this regularized line search we prove in this section quasi-Fejér convergence of $\{x^k\}$ to the solution set of problem P for f convex and $\varphi \in \Psi$, and as a consequence full convergence to a solution, without any level set boundedness assumption.

Possibly the use of $\sigma\alpha$ instead of $d_\varphi(x, x^k)$ is not essential and due just to our proof techniques, but we mention that computationally a linear term is easier to handle than $d_\varphi(x, x^k)$. Note also that adding either $\sigma\alpha$ or the restriction of $d_\varphi(x, x^k)$ to the segment does not make the linear search any harder: for convex f the regularized minimand is still convex, therefore unimodal, and both nonderivative methods (e.g., Fibonacci search) and a derivative one (e.g., Newton's method) are equally easy to implement with or without the regularization term.

We proceed to the convergence analysis of this algorithm. From now on $\{x^k\}$ refers to the sequence generated by algorithm RMIG.

The next proposition extends the result of Propositions 3(i), 3(ii), and 4 to the sequence generated by algorithm RMIG. Define

$$\hat{h}_k(\alpha) = f((1-\alpha)x^k + \alpha y^k) + \sigma\alpha. \tag{39}$$

PROPOSITION 6. i) $f^* \leq f(x^{k+1}) + \alpha_k \delta \lambda_k d_{\hat{\varphi}}(y^k, x^k) \leq f(x^k)$,
ii) $\{f(x^k)\}$ *is decreasing and convergent,*
iii) $\lim_{k\to\infty}(x^{k+1} - x^k) = 0$,
iv) $\sum_{k=0}^{\infty} \alpha_k d_{\hat{\varphi}}(y^k, x^k) < \infty$.

Proof. i) By (37)–(39), $\hat{h}_k(\alpha_k) \leq \hat{h}_k(0)$, which gives the rightmost inequality. The leftmost one is trivial.

ii) The proof follows from (i), because $\alpha_k \delta \lambda_k d_{\hat{\varphi}}(y^k, x^k) \geq 0$.

iii) Use convexity of d_φ in its first variable, (17), (11), and (i) to get

$$\begin{aligned} 0 \leq \delta\hat{\lambda} d_\varphi(x^{k+1}, x^k) &= \delta\hat{\lambda} d_\varphi((1-\alpha_k)x^k + \alpha_k y^k, x^k) \leq \delta\hat{\lambda}\alpha_k d_{\hat{\varphi}}(y^k, x^k) \\ &\leq \delta\lambda_k\alpha_k d_{\hat{\varphi}}(y^k, x^k) \leq f(x^k) - f(x^{k+1}). \end{aligned} \tag{40}$$

Since $\delta\hat{\lambda} > 0$ and the right-hand side of (40) converges to 0 by (ii), we conclude that $\lim_{k\to\infty} d_\varphi(y^k, x^k) = 0$ and the result follows from Proposition 1(i).

iv) From (40), $\alpha_k d_{\hat{\varphi}}(y^k, x^k) \leq (1/\delta\hat{\lambda})(f(x^k) - f(x^{k+1}))$ so that

$$\sum_{k=0}^{\infty} \alpha_k d_{\hat{\varphi}}(y^k, x^k) \leq (1/\delta\hat{\lambda})(f(x^0) - \lim_{k\to\infty} f(x^k)) < \infty. \qquad \square$$

Next we introduce the concept of quasi-Fejér convergence. This notion was introduced in [5] and discussed for the case of φ-divergences in [8]. Take $\varphi \in \Phi$ and $U \subset \mathbf{R}^n_{++}$.

DEFINITION 2. *A sequence* $\{z^k\} \subset \mathbf{R}^n_{++}$ *is said to be quasi-Fejér convergent to* U *with respect to* d_φ *if for each* $u \in U$ *there exists a sequence of real numbers* $\varepsilon_k \geq 0$ *such that* $\sum_{k=0}^{\infty} \varepsilon_k < \infty$ *and* $d_\varphi(u, z^{k+1}) \leq d_\varphi(u, z^k) + \varepsilon_k$.

We will use the following result on quasi-Fejér convergence.

THEOREM 3. *If a sequence* $\{z^k\}$ *is quasi-Fejér convergent to a nonempty set* U, *then* $\{z^k\}$ *is bounded and* $\{d_\varphi(u, z^k)\}$ *is bounded for all* $u \in U$. *If furthermore at least one limit point of* $\{z^k\}$ *belongs to* U, *then* $\{z^k\}$ *converges.*

Proof. The proof can be found in [8, Thm. 4.1]. $\square$

We will prove quasi-Fejér convergence of the sequence $\{x^k\}$ generated by algorithm RMIG to a set which contains all solutions of problem P, not with respect to d_φ used in the algorithm but with respect to a specific φ-divergence, namely d_{φ_2}, with φ_2 as in Example 2 in §1. We will use the notation ψ instead of φ_2; i.e., $\psi(t) = t\log t - t + 1$, and

$$d_\psi(x, y) = \sum_{j=1}^{n} (x_j \log \frac{x_j}{y_j} + y_j - x_j) \tag{41}$$

is the Kullback–Leibler divergence and has several interesting properties. One of them is that it can be continuously extended to $\mathbf{R}^n_+ \times \mathbf{R}^n_{++}$; i.e., it admits vectors with zero components in its first variable. It can be seen that Theorem 3 holds in this case also for $U \subset \mathbf{R}^n_+$.

Let $T^* = \{x \in \mathbf{R}^n_+ : f(x) \leq \lim_{k\to\infty} f(x^k)\}$. T^* is well defined by Proposition 6(ii), and contains the solutions of problem P, so it is nonempty. Next we prove

quasi-Fejér convergence of $\{x^k\}$ to T^* with respect to d_ψ. We need first a technical result.

PROPOSITION 7. *Take $\varphi \in \Phi_1$ and let $\theta = 1/\varphi''(1)$. There exists $\nu > 0$ such that $(t-1) - \theta\varphi'(t) \leq \nu(t-1)\varphi'(t)$ for all $t > 0$.*

Proof. It suffices to prove that $\overline{\varphi}$, defined as

$$\overline{\varphi}(t) = \frac{t - 1 - \theta\varphi'(t)}{(t-1)\varphi'(t)} = \frac{1}{\varphi'(t)} - \frac{\theta}{t-1},$$

is bounded above. Since $\overline{\varphi}$ is trivially continuous at any $t \neq 1$, it is enough to prove that $\limsup_{t\to 0} \overline{\varphi}(t) < \infty$, $\limsup_{t\to\infty} \overline{\varphi}(t) < \infty$, $\lim_{t\to 1} \overline{\varphi}(t) < \infty$. Since φ' is increasing for all t, negative for $t < 1$, and positive for $t > 1$ we have

$$\limsup_{t\to 0} \overline{\varphi}(t) \leq -\lim_{t\to 0} \frac{\theta}{t-1} = \theta < \infty$$

$$\limsup_{t\to\infty} \overline{\varphi}(t) \leq \frac{1}{\varphi'(2)} - \lim_{t\to\infty} \frac{\theta}{t-1} = \frac{1}{\varphi'(2)} < \infty.$$

For $t = 1$, expanding φ' around 1, and using $\varphi'(1) = 0$, $\varphi''(1) = \frac{1}{\theta}$, we get

$$\lim_{t\to 1} \overline{\varphi}(t) = -\frac{\theta^2}{2}\varphi'''(1) < \infty. \qquad \square$$

PROPOSITION 8. *If f is convex and $\varphi \in \Psi$ (i.e., $\varphi'(t) \leq \varphi''(1)\log t$ for all $t > 0$) then $\{x^k\}$ is quasi-Fejér convergent to T^* with respect to d_ψ.*

Proof. Take $z \in T^*$ and let $\theta = 1/\varphi''(1) > 0$. Since $\{f(x^k)\}$ decreases by Proposition 6(i), we have $f(z) \leq f(x^k)$ for all k and therefore

$$0 \leq \theta\frac{\alpha_k}{\lambda_k}(f(x^k) - f(z)) \leq \theta\frac{\alpha_k}{\lambda_k}\nabla f(x^k)^t(x^k - z) = \theta\alpha_k \sum_{j=1}^n \varphi'(y_j^k/x_j^k)(z_j - x_j^k) \tag{42}$$

using convexity of f and Corollary 1. Use now convexity of d_ψ in its second variable and (41) to get

$$\begin{aligned} d_\psi(z, x^{k+1}) - d_\psi(z, x^k) &= d_\psi(z, (1-\alpha_k)x^k + \alpha_k y^k) - d_\psi(z, x^k) \\ &\leq \alpha_k(d_\psi(z, y^k) - d_\psi(z, x^k)) \\ &= \alpha_k \sum_{j=1}^n z_j(\log(x_j^k/y_j^k) + y_j^k - x_j^k). \end{aligned} \tag{43}$$

Let $t_j^k = y_j^k/x_j^k$ and add (42) and (43) together

$$\begin{aligned} d_\psi(z, x^{k+1}) - d_\psi(z, x^k) &\leq \alpha_k \left[\sum_{j=1}^n z_j(\theta\varphi'(t_j^k) - \log t_j^k) + \sum_{j=1}^n x_j^k(t_j^k - 1 - \theta\varphi'(t_j^k))\right] \\ &\leq \alpha_k \sum_{j=1}^n x_j^k(t_j^k - 1 - \theta\varphi'(t_j^k)) \leq \nu\alpha_k \sum_{j=1}^n x_j^k(t_j^k - 1)\varphi'(t_j^k) \\ &= \nu\alpha_k d_{\hat{\varphi}}(y^k, x^k), \end{aligned} \tag{44}$$

where we use $\varphi \in \Psi$ in the second inequality and Proposition 7 in the third inequality. From (44), $d_\psi(z, x^{k+1}) \leq d_\psi(z, x^k) + \nu\alpha_k d_{\hat{\varphi}}(y^k, x^k)$. The result follows from Definition 2 with $\varepsilon_k = \nu\alpha_k d_{\hat{\varphi}}(y^k, x^k)$ and Proposition 6(iv). $\square$

COROLLARY 3. *If f is convex and $\varphi \in \Psi$, then*
i) *the sequence $\{x^k\}$ converges,*
ii) *$\{\lambda_k\}$ is bounded,*
iii) *$\{y^k\}$ is bounded.*

Proof. (i) follows from Proposition 8 and Theorem 3, since all limit points of $\{x^k\}$ belong to T^*. For (ii) and (iii) we use the proofs of Proposition 3(iv) and 3(v), respectively, with a bounded set D containing $\{x^k\}$ in place of $L(\rho)$ in the definition of μ. □

We present next the main result of this section.

THEOREM 4. *If f is convex and $\varphi \in \Psi$, then the sequence generated by algorithm RMIG converges to a solution of Problem P.*

Proof. $\{x^k\}$ converges by Corollary 3(i). Let x^* be its limit. By convexity of f it suffices to check that x^* satisfies the Karush–Kuhn–Tucker conditions of problem P, namely,

$$x^* \geq 0, \tag{45}$$

$$\nabla f(x^*) \geq 0, \tag{46}$$

$$\nabla f(x^*)_j x_j^* = 0 \qquad (1 \leq j \leq n). \tag{47}$$

Proof. (45) follows from Proposition 2. We look now at (47). We could attempt a direct proof using the fact that $\{x^k\}$ is fully convergent, but we take advantage of the proof of Proposition 5, which holds for this case almost verbatim, using Corollary 3 and Proposition 6(iii) instead of Proposition 3(iii)–(v) and 4. The only difference is that we must use $\hat{h}_k$ defined in (39) instead of $\overline{h}_k$. Since $\hat{h}_k'(\alpha) = \overline{h}_k(\alpha) + \sigma$, we have, instead of (32), (33), $0 \leq \hat{h}_{\ell_k}'(\alpha_{\ell_k}) = \nabla f(x^{e_{k+1}})^t(y^{\ell_k} - x^{\ell_k}) + \sigma$, implying

$$\begin{aligned}(\nabla f(x^{\ell_k+1}) - \nabla f(x^{\ell_k}))^t(y^{\ell_k} - x^{\ell_k}) &\geq -\nabla f(x^{\ell_k})^t(y^{\ell_k} - x^{\ell_k}) - \sigma \\ &= \lambda_{\ell_k} d_{\hat{\varphi}}(y^{\ell_k}, x^{\ell_k}) - \delta\lambda_{\ell_k} d_{\hat{\varphi}}(y^{\ell_k}, x^{\ell_k}) \\ &= (1-\delta)\lambda_{\ell_k} d_{\hat{\varphi}}(y^{\ell_k}, x^{\ell_k}) \\ &\geq (1-\delta)\hat{\lambda} d_{\hat{\varphi}}(y^{\ell_k}, x^{\ell_k}) \geq 0\end{aligned}$$

using $\delta \in (0,1)$. The remainder of the proof is exactly as in Proposition 5 after (33). Finally we can establish (46) using Lemma 1 as in Theorem 2, but we prefer to give a direct proof to show to what extent Lemma 1 simplifies under the hypothesis of convergence of $\{x^k\}$. Since $\nabla f(x^*)_j = 0$, whenever $x_j^* > 0$ by (47) it suffices to prove that $\nabla f(x^*)_j \geq 0$ for j such that $x_j^* = 0$. Assume $\nabla f(x^*)_j < 0$ for any such j. Then there exists k_0 such that $\nabla f(x^k)_j < 0$ for $k > k_0$; i.e., we have $0 > \nabla f(x^k)_j = -\lambda_k \varphi'(y_j^k / x_j^k)$, implying $\varphi'(y_j^k/x_j^k) > 0$ so that $y_j^k > x_j^k$ and therefore $x_j^{k+1} \geq x_j^k$ for all $k > k_0$. Then $0 = x_j^* = \lim x_j^k \geq x_j^{k_0} > 0$, which is a contradiction. (This corresponds to the case $p_k = p$ for large enough k in the proof of Lemma 1.) The proof is complete. □

6. Final remarks. We discuss here three possible extensions of our results, which hold in fact under weaker hypotheses than those imposed in the previous sections. We have excluded these extensions from our exposition for the sake of clarity, since they demand more involved proofs, with technicalities which would obscure, perhaps, more substantial arguments. The first two extensions refer to the class of

admissible objective functions, which we have assumed to be convex and continuously differentiable. Both assumptions can be relaxed.

In the first place, all our results hold if we assume that f is pseudoconvex rather than convex. We recall that f is pseudoconvex if and only if $\nabla f(x)^t(y-x)$ implies $f(y) \geq f(x)$. The use of convexity in our proofs has been twofold. First we use it in Lemma 1 and Theorem 4 in the form of sufficiency of the Karush–Kuhn–Tucker conditions for optimality. This is true for pseudoconvex functions (see [1, p. 152]). Then we use it in (42) of Proposition 8 to prove that $\nabla f(x^k)^t(x^k - z) \geq 0$ for any $z \in T^*$. We claim that this is also true for pseudoconvex f. Assume on the contrary that $\nabla f(x^k)^t(z - x^k) > 0$. Then, by definition of pseudoconvexity, using Propositions 6(i) and (ii) and $z \in T^*$

$$f(z) \geq f(x^k) \geq f(x^{k+1}) + \lambda_k \alpha_k d_{\hat{\varphi}}(y^k, x^k) \geq f(x^{k+1}) \geq \lim_{k\to\infty} f(x^k) \geq f(z)$$

so that $\alpha_k d_{\hat{\varphi}}(y^k, x^k)) = 0$; i.e., either $\alpha_k = 0$ or $d_{\hat{\varphi}}(y^k, x^k) = 0$. If $\alpha_k = 0$, then $0 \leq \hat{h}'_k(0) = -\left\|\nabla f(x^k)\right\|^2$, implying $\nabla f(x^k) = 0$. If $d_{\hat{\varphi}}(y^k, x^k) = 0$, then $y^k = x^k$ and so $\nabla f(x^k) = 0$ by Corollary 1. In both cases $\nabla f(x^k) = 0$ in contradiction with the assumption above. The claim is established.

In the second place, we would like to assume just nonemptiness of the subdifferential of f at any x in its effective domain, rather than continuous differentiability, among other reasons, to be able to accommodate in our analysis the case of additional convex constraints besides nonnegativity, because by dropping the continuously differentiability assumption, such constraints (particularly linear equality constraints) can be transferred to the objective, in which case MIG transforms an inequality constrained problem into a sequence of equality constrained ones, a reduction which has proved to be quite useful in many instances. Extension of our results to the case of nondifferentiable f is more complicated. It is possible to go through the convergence results of §4 with a subgradient of f at x substituting for $\nabla f(x)$ up to Theorem 1, but we have no proof of Lemma 1 without differentiability of f and so Theorem 2 does not hold. The situation is better for the results of §5, because we have full convergence of $\{x^k\}$ to x^*. For the φ-divergence proximal algorithm given by iteration (7) it has been proved in [8, Prop. 4.3] that if the sequence $\{x^k\}$ converges to x^* and there exists a solution z of problem P satisfying $z_j = 0$ for all j such that $x^*_j = 0$, then x^* is a solution of problem P, without assumming differentiability of f. This proof can be extended to the sequence $\{x^k\}$ generated by algorithm RMIG. Since $\{x^k\}$ converges by Corollary 3, such a result is enough. The hypothesis on the null components of x^* holds because $d_\psi(z, x^k)$ is bounded, by Proposition 8 and Theorem 3, for any solution z, while if x^k_j approaches 0 for some j such that $z_j > 0$ we get $d_\psi(z, x^k)$ unbounded. There are several other involved technicalities to be dealt with in the nondifferentiable case.

The third extension is to admit overrelaxation in (19) or (37), i.e., replace $[0, 1]$ by an interval $[0, \pi_k]$ with π_k possibly larger than 1. In this case, some restrictions must be imposed upon π_k to guarantee $x^k > 0$ for all k. One way to incorporate these positivity-preserving safeguards is the following.

Take $\zeta \in (0, 1)$, a sequence $\{\zeta_k\} \subset (\zeta, 1)$, and $\tau > 0$. Let $J(x) = \{j : \nabla f(x)_j < 0\}$. Define after (18)

$$\omega_k = \begin{cases} \min\limits_{j \in J(x^k)} \{\frac{y^k_j}{x^k_j}\} & \text{if } J(x^k) \neq \emptyset, \\ 0 & \text{otherwise,} \end{cases} \tag{48}$$

$$\pi_k = \frac{\zeta_k}{1+\tau-\omega_k}, \tag{49}$$

and finally substitute $[0, \pi_k]$ for $[0, 1]$ in (19) or (37). It is easy to check that $1-\omega_k > 0$ and that this choice of π_k guarantees positivity of $\{x^k\}$. (In [7], we use this approach in a similar setting.) Regarding the convergence analysis, the situation is to some extent the opposite to the nondifferentiability extension discussed above: in this case the results of §4, but not those of §5, hold. For §4, there are only some technical complications (e.g., $\tau > 0$ is required to get $\{\pi^k\}$ bounded, which is required in the proof of Proposition 4). For §5, when x^{k+1} is not in the segment between x^k and y^k, (43) does not hold and we do not get quasi-Fejér convergence. The issue of full convergence under an overrelaxed line search requires further study.

Finally, we present the full form of the algorithms corresponding to the choices of φ introduced in the examples in §1.

1. $\varphi_1(t) = t - \log t - 1$,

$$x_j^{k+1} = x_j^k \left(1 - \frac{\alpha_k \nabla f(x^k)_j}{\lambda_k + \nabla f(x^k)_j}\right). \tag{50}$$

2. $\varphi_2(t) = t \log t - t + 1$,

$$x_j^{k+1} = x_j^k \left\{1 + \alpha_k \left[\exp\left(-\frac{\nabla f(x^k)_j}{\lambda_k}\right) - 1\right]\right\}. \tag{51}$$

3. $\varphi_3(t) = (\sqrt{t} - 1)^2$,

$$x_j^{k+1} = x_j^k \left[1 - \frac{\alpha_k \nabla f(x^k)_j (2\lambda_k + \nabla f(x^k)_j)}{(\lambda_k + \nabla f(x^k)_j)^2}\right]. \tag{52}$$

Since $\varphi_i \in \Psi$ $(i = 1, 2, 3)$ all the convergence results of §5 hold for iterations (50), (51), and (52) when λ_k satisfies (17) and α_k is given by (37). If α_k is given by (19), or when the interval $[0, 1]$ in (19) is replaced by $[0, \pi_k]$, with π_k as in (48), (49), the convergence results of section 4 hold for (50), (51), and (52). For $\varphi(t) = \varphi_4(t) = \frac{1}{2}(t-1)^2$ we get $x_j^{k+1} = x_j^k(1 - \frac{\alpha_k}{\lambda_k}\nabla f(x^k)_j)$. Our analysis does not include this case, because $\varphi_4 \notin \Phi_1$, and y^k given by (18) may fail to be positive. In this case the positivity-preserving safeguards are always needed and other adjustments are required. A convergence analysis tailored for this iteration can be found in [7].

REFERENCES

[1] M. AVRIEL, *Nonlinear Programming, Analysis and Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[2] I. CSISZÁR, *Information-type measures of difference of probability distributions and indirect observations*, Studia Sci. Math. Hungar., 2 (1967), pp. 299–318.

[3] I.I. DIKIN, *Iterative solutions of problems of linear and quadratic programming*, Soviet Math. Dokl., 8 (1967), pp. 674–675.

[4] P. P. B. EGGERMONT, *Multiplicative iterative algorithms for convex programming*, Linear Algebra Appl., 130 (1990), pp. 25–42.

[5] YU. M. ERMOL'EV, *On the method of generalized stochastic gradients and quasi-Fejér sequences*, Cybernetics, 5 (1969), pp. 208–220.

[6] C. G. GONZAGA, *Path following methods for linear programming*, SIAM Rev., 34 (1992), pp. 167–224.

[7] A. N. IUSEM, *Interior point multiplicative methods for optimization under positivity constraints*, Acta Appl. Math., to appear.

[8] A. N. Iusem, B. F. Svaiter, and M. Teboulle, *Entropy-like proximal methods in convex programming*, Math. Oper. Res., 19 (1994), pp. 790–814.

[9] N. Karmarkar, *A new polynomial time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.

[10] B. Lemaire, *The Proximal Algorithm*, Internat. Ser. Numer. Math. 87, Birkhauser-Verlag, Basel, 1989, pp. 73–87.

[11] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.

[12] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

[13] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Math. Oper. Res., 17 (1992), pp. 670–690.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 407–427, March 1996

EPSILON-MAXIMUM PRINCIPLE OF PONTRYAGIN TYPE AND PERTURBATION ANALYSIS OF CONVEX OPTIMAL CONTROL PROBLEMS*

MOHAMMED MOUSSAOUI† AND ALBERTO SEEGER†

Abstract. In this paper we study the first-order behaviour of the optimal-value function associated with a convex parametric problem of optimal control. A formula for the subdifferential of the optimal-value function is derived without assuming the existence of optimal solutions to the unperturbed problem. The so-called epsilon-maximum principle is a key ingredient in the writing of our main sensitivity result.

Key words. optimal control problem, sensitivity analysis, approximate subdifferential, Pontryagin's principle, transversality condition

AMS subject classifications. 49N99, 90C31

1. Introduction.

The sensitivity analysis of the optimal-value function associated with a convex parametric program has been the main concern of numerous papers in past decades. Formulas for the directional derivative and/or the subdifferential of the optimal-value function have been derived under various types of assumptions. As a standard rule, it is assumed that the unperturbed problem admits at least one optimal solution. This hypothesis is, however, quite restrictive in some cases, especially in the context of infinite dimensional programs. For this reason, Moussaoui and Seeger [MS1, MS2] have recently developed a sensitivity analysis theory well suited for dealing with convex parametric programs with possibly empty solution sets. In a subsequent paper, Seeger [Se1] has applied this general theory to the specific case of a convex parametric problem of calculus of variations written in the Bolza form

$$\underset{x\in X}{\text{minimize}} \left\{ H(x(0),x(1),\alpha) + \int_0^1 L(t,x(t),\dot{x}(t),\theta(t))\,dt \right\}. \tag{1.1}$$

Here α and θ are interpreted as perturbation parameters affecting the endpoint cost and the integral cost, respectively. The sensitivity results obtained in [Se1] are stated in terms of concepts such as approximate Euler–Lagrange inclusion and approximate transversality condition.

The purpose of the present paper is to apply the above-mentioned theory to the case of a convex parametric problem of optimal control. Our work is influenced by [Se1] and can be viewed as an extension of it. Although it is possible to write an optimal control problem as a Bolza problem of calculus of variations, we follow, however, a completely different approach. (See §5 for further discussion.)

Next we describe the basic model of parametric optimal control problem to be considered in this paper. The control $u \in U$ and the trajectory $x \in X$ are related by means of the state equation

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + \gamma(t) \quad \text{for a.e. } t \in [0,1]. \tag{1.2}$$

The space of controls is defined as

$$U = L_p^d := L_p([0,1];\mathbb{R}^d) \quad (\text{with } 1 \le p < +\infty),$$

and the space of trajectories

$$X = A_p^n := A_p([0,1];\mathbb{R}^n)$$

* Received by the editors March 3, 1993; accepted for publication (in revised form) November 2, 1994.
† Department of Mathematics, University of Avignon, 33, rue Louis Pasteur, 84000 Avignon, France.

consists of the absolutely continuous functions $x : [0,1] \to \mathbb{R}^n$ whose derivatives $\dot{x}$ belong to $L_p^n := L_p([0,1];\mathbb{R}^n)$. The velocity vector $\dot{x}(t)$ is defined of course for all $t \in [0,1]$ except on a set of measure zero. In (1.2), the term $\gamma \in \Gamma$ is regarded as an external perturbation function affecting the dynamics of the system.

In this work we are concerned with a convex parametric problem of optimal control written in the form

$$\text{minimize}\{J(x,u,\alpha,\theta) : (x,u) \in F(\gamma)\},$$

where the feasible set $F(\gamma)$ consists of all pairs $(x,u) \in X \times U$ satisfying the state equation (1.2). The cost functional

$$J(x,u,\alpha,\theta) := H(x(0),x(1),\alpha) + \int_0^1 L(t,x(t),u(t),\theta(t))\,dt \tag{1.3}$$

involves a perturbation function $\theta \in \Theta$ affecting the integral term and a perturbation vector $\alpha \in \mathbb{R}^k$ affecting the endpoint term.

The aim of this paper is to study the first-order behaviour of the optimal-value function

$$(\alpha,\theta,\gamma) \mapsto V(\alpha,\theta,\gamma) = \text{Inf}\{J(x,u,\alpha,\theta) : (x,u) \in F(\gamma)\} \tag{1.4}$$

around a given point, say, $(\alpha_0,\theta_0,\gamma_0) \in \mathbb{R}^k \times \Theta \times \Gamma$. This point represents the reference level or nominal value of the parameters. Thus,

$$\text{minimize}\{J(x,u,\alpha_0,\theta_0) : (x,u) \in F(\gamma_0)\} \tag{1.5}$$

is regarded as the "unperturbed" optimal control problem.

There is a long history behind the study of an optimal-value function like (1.4). For instance, sensitivity results for smooth parametric optimal control problems can be found in works by Oniki [On] and Tu [Tu, §10], to mention just a few names. Parametric optimal control problems involving locally Lipschitz data have been discussed by Clarke [C1, C2, C3], Loewen [Lo], and Clarke and Loewen [CL1, CL2], among others. In such a setting, the concept of Clarke's generalized subdifferential is a suitable tool for studying the first-order behaviour of the function V. In this paper all the data are assumed to be convex, so we will characterize the subdifferential of V in the sense of convex analysis. The precise definition of this concept will be recalled in §2. In contrast with the sensitivity results found in the literature, in our approach the unperturbed problem (1.5) is not required to be solvable. In fact, convex optimal control problems with empty solution sets are encountered quite often in practice. The following examples are just academic but serve to illustrate the range of applicability of our sensitivity results.

Example 1.1. For each $\theta \in L_1[0,1]$ and $\gamma \in L_1[0,1]$, consider the problem of minimizing the cost

$$\int_0^1 t^2[u_1^2(t) + u_2^2(t) + \theta^2(t)]^{1/2}\,dt$$

among all controls $u_1,\ u_2 \in L_1\,[0,1]$ and trajectories $x \in A_1\,[0,1]$ satisfying

$$\begin{cases} \dot{x}(t) = b_1\,u_1(t) + b_2\,u_2(t) + \gamma(t), & \text{for a.e. } t \in [0,1], \\ x(0) = 0, \quad x(1) = 1. \end{cases}$$

Here $b_1 \neq 0$ and $b_2 \neq 0$ are two given real numbers. Denote by $V(\theta,\gamma)$ the optimal value of the above problem. Setting the parameters θ and γ at the reference level

$$\theta_0(t) = 0 \quad \text{and } \gamma_0(t) = 0 \quad \text{for a.e. } t \in [0,1],$$

one gets a nonsmooth convex optimal control problem with optimal value equal to zero. To see this, consider the minimizing sequence $(u_1^{(k)}, u_2^{(k)})$ given by

$$u_i^{(k)} = \begin{cases} \dfrac{k}{2\,b_i} & \text{on} \left[0, \frac{1}{k}\right[\\ \\ 0 & \text{on} \left[\dfrac{1}{k}, 1\right] \end{cases} \quad \text{for } i = 1, 2.$$

However, the optimal value $V(\theta_0, \gamma_0) = 0$ is not attained. Observe that the controls

$$u_1(t) = u_2(t) = 0 \quad \text{for a.e. } t \in [0, 1]$$

do not steer the system from $x(0) = 0$ to $x(1) = 1$. How does the function V behave around (θ_0, γ_0)?

Example 1.2. Consider the problem of minimizing the endpoint cost

$$H(x(1)) = [\exp(x_1(1)) + x_2^2(1)]^{1/2}$$

among all controls $u \in L_1[0, 1]$ and trajectories $x \in A_1([0, 1]; \mathbb{R}^2)$ satisfying

$$\begin{cases} \dot{x}_1(t) = -2x_1(t) + 3x_2(t) + u(t) + \gamma_1(t), \\ \dot{x}_2(t) = x_2(t) + \gamma_2(t), \\ x_1(0) = 0, \quad x_2(0) = 0. \end{cases}$$

Setting the parameter function $\gamma = (\gamma_1, \gamma_2) \in L_1([0, 1]; \mathbb{R}^2)$ at the reference level

$$\gamma_0(t) = (0, 0) \quad \text{for a.e. } t \in [0, 1],$$

one gets a nonsmooth convex optimal control problem whose optimal value $V(\gamma_0) = 0$ is finite but not attained. How does one compute in this case the subdifferential of V at γ_0?

2. General results on subdifferentials of optimal-value functions. As is customary in the context of convex analysis, we work in the setting of two real linear spaces, say, Ξ and Ξ^*, which have been paired by means of a bilinear form $\langle \cdot, \cdot \rangle : \Xi \times \Xi^* \mapsto \mathbb{R}$. The topologies on Ξ and Ξ^* are supposed to be compatible with respect to the pairing (see [Ro1, §3] for details).

Given a convex function $V : \Xi \to \mathbb{R} \cup \{+\infty\}$, the *subdifferential* of V at $\xi_0 \in \Xi$ is defined by

$$\partial V(\xi_0) = \{\eta \in \Xi^* : V(\xi) \geq V(\xi_0) + \langle \xi - \xi_0, \eta \rangle \text{ for all } \xi \in \Xi\}. \tag{2.1}$$

Each element of (2.1) is called a *subgradient* of V at ξ_0. An equivalent definition of this set is

$$\partial V(\xi_0) = \{\eta \in \Xi^* : V^*(\eta) + V(\xi_0) - \langle \xi_0, \eta \rangle = 0\},$$

where

$$\eta \in \Xi^* \mapsto V^*(\eta) := \operatorname*{Sup}_{\xi \in \Xi} \{\langle \xi, \eta \rangle - V(\xi)\}$$

stands for the Fenchel conjugate of V.

The set $\partial V(\xi_0)$ reflects the first-order behaviour of the function V around the point $\xi_0 \in \Xi$. Calculus rules for computing subdifferentials can be found in standard references like Ioffe

and Tihomirov [IT], Laurent [La], and Rockafellar [Ro1]. The next calculus rule serves, for instance, for computing the subdifferential of the optimal-value function

$$\xi \in \Xi \mapsto f_K(\xi) := \operatorname*{Inf}_{z \in Z} \{f(z) : Kz = \xi\}.$$

The elements in the space Ξ are regarded here as parameters. The space Z of minimization variables is supposed to be paired with another real linear space, say, Z^*.

LEMMA 2.1. *Suppose that the following general assumption is true*:

$$\text{(2.2)} \qquad \begin{cases} f : Z \to \mathbb{R} \cup \{+\infty\} & \text{is a convex proper function,} \\ K : Z \to \Xi & \text{is a continuous linear operator,} \\ f_K & \text{is finite at } \xi_0 \in \Xi. \end{cases}$$

Let z_0 be any element in the solution set

$$S(\xi_0) := \{z \in Z : Kz = \xi_0,\ f(z) = f_K(\xi_0)\}.$$

Then

$$\text{(2.3)} \qquad \partial f_K(\xi_0) = \{\eta \in \Xi^* : K^*\eta \in \partial f(z_0)\},$$

where $K^* : \Xi^* \to Z^*$ stands for the adjoint operator of $K : Z \to \Xi$.

The above result can be found in Hiriart-Urruty [Hi] and Zalinescu [Za], for instance. The writing of formula (2.3) makes sense only if an element z_0 in the solution set $S(\xi_0)$ does exist, but here we want to evaluate the subdifferential mapping ∂f_K at a point ξ_0 at which the solution set is possibly empty. To handle this more complicated situation we invoke a recent result by Moussaoui and Seeger [MS1, MS2]. Recall that the *ε-subdifferential* of f at a point $z_0 \in Z$ is the set

$$\partial_\varepsilon f(z_0) = \{\omega \in Z^* : f(z) \geq f(z_0) + \langle z - z_0, \omega\rangle - \varepsilon \text{ for all } z \in Z\}.$$

The above set is also known as the approximate subdifferential of f at z_0. The (exact) subdifferential $\partial f(z_0)$ corresponds of course to the case $\varepsilon = 0$.

LEMMA 2.2 [MS1, Thm. 1]. *Suppose that the general assumption* (2.2) *is true. Then*

$$\text{(2.4)} \qquad \partial f_K(\xi_0) = \bigcap_{\varepsilon > 0} \bigcup_{Kz = \xi_0} \{\eta \in \Xi^* : K^*\eta \in \partial_\varepsilon f(z)\}.$$

Observe that if z satisfies $Kz = \xi_0$ and $\{\eta \in \Xi^* : K^*\eta \in \partial_\varepsilon f(z)\} \neq \emptyset$, then z belongs necessarily to the set

$$S_\varepsilon(\xi_0) := \{z \in Z : Kz = \xi_0,\ f(z) \leq f_K(\xi_0) + \varepsilon\}.$$

Thus formula (2.4) can also be written in the form

$$\partial f_K(\xi_0) = \bigcap_{\varepsilon > 0} \bigcup_{z \in S_\varepsilon(\xi_0)} \{\eta \in \Xi^* : K^*\eta \in \partial_\varepsilon f(z)\}.$$

For subsequent use we need to adjust Lemma 2.2 to the case in which f is defined as a sum of two convex functions, say, f_1 and f_2. We also incorporate a linear operator R defined over the space of parameters Ξ and with values in another linear space, say, Π. The notation $\operatorname{Im} R := \{R\xi : \xi \in \Xi\}$ refers to the range of R, and $C^\perp$ stands for the orthogonal complement of C. The indicator function of the set

$$K^{-1}(\operatorname{Im} R) := \{z \in Z : Kz \in \operatorname{Im} R\}$$

is by definition

$$z \in Z \mapsto \psi_{K^{-1}(\operatorname{Im} R)}(z) := \begin{cases} 0 & \text{if } Kz \in \operatorname{Im} R, \\ +\infty & \text{otherwise.} \end{cases}$$

LEMMA 2.3. *Let $f_1, f_2 : Z \to \mathbb{R} \cup \{+\infty\}$ be two proper convex functions and $K : Z \to \Pi$ and $R : \Xi \to \Pi$ be two continuous linear operators. Suppose that the optimal-value function*

$$\xi \in \Xi \mapsto V(\xi) = \inf_{z \in Z} \{f_1(z) + f_2(z) : Kz = R\xi\}$$

is finite at $\xi_0 \in \Xi$. Then the subdifferential $\partial V(\xi_0)$ admits the following inner estimate:

$$\partial V(\xi_0) \supset \bigcap_{\varepsilon > 0} \bigcup_{Kz = R\xi_0} \{R^*\varphi : K^*\varphi \in \partial_\varepsilon f_1(z) + \partial_\varepsilon f_2(z)\}. \tag{2.5}$$

Moreover, one can write the formula

$$\partial V(\xi_0) = \bigcap_{\varepsilon > 0} \bigcup_{Kz = R\xi_0} \{R^*\varphi : K^*\varphi \in \partial_\varepsilon f_1(z) + \partial_\varepsilon f_2(z)\} \tag{2.6}$$

if one adds the following constraint qualification hypotheses:

$$\operatorname{Im} R \quad \textit{and} \quad \operatorname{Im} R^* \textit{ are closed sets}, \tag{2.7}$$

$$K^*([\operatorname{Im} R]^\perp) \textit{ is a closed set}, \tag{2.8}$$

$$\begin{aligned} &(f_1 + f_2 + \psi_{K^{-1}(\operatorname{Im} R)})^*(w) \\ &\quad = \inf_{\omega_1 + \omega_2 + \omega_3 = \omega} \{f_1^*(\omega_1) + f_2^*(\omega_2) + \psi^*_{K^{-1}(\operatorname{Im} R)}(\omega_3)\} \quad \textit{for all } w \in Z^*. \end{aligned} \tag{2.9}$$

Proof. Denote by Ω the set appearing on the right-hand side of (2.5), and let η be an arbitrary element in Ω. To prove that $\eta \in \partial V(\xi_0)$, it suffices to show the inequality

$$V(\xi) \geq V(\xi_0) + \langle \xi - \xi_0, \eta \rangle - 2\varepsilon \quad \text{for all } \xi \in \Xi \text{ and } \varepsilon > 0. \tag{2.10}$$

Then fix $\xi \in \Xi$ and $\varepsilon > 0$. Since $\eta \in \Omega$, there exist $z_\varepsilon \in Z$ and $\varphi_\varepsilon \in \Pi^*$ such that

$$Kz_\varepsilon = R\xi_0, \quad R^*\varphi_\varepsilon = \eta, \quad \text{and} \quad K^*\varphi_\varepsilon \in \partial_\varepsilon f_1(z_\varepsilon) + \partial_\varepsilon f_2(z_\varepsilon). \tag{2.11}$$

Momentarily decompose $K^*\varphi_\varepsilon$ in the form

$$K^*\varphi_\varepsilon = \omega_1 + \omega_2 \quad \text{with } \omega_1 \in \partial_\varepsilon f_1(z_\varepsilon) \text{ and } \omega_2 \in \partial_\varepsilon f_2(z_\varepsilon).$$

Now take any $z \in Z$ satisfying $Kz = R\xi$. By summing up the inequalities

$$f_i(z) \geq f_i(z_\varepsilon) + \langle z - z_\varepsilon, \omega_i \rangle - \varepsilon, \qquad i = 1, 2,$$

one gets

$$f_1(z) + f_2(z) \geq f_1(z_\varepsilon) + f_2(z_\varepsilon) + \langle z - z_\varepsilon, K^*\varphi_\varepsilon \rangle - 2\varepsilon,$$

and therefore,

$$f_1(z) + f_2(z) \geq V(\xi_0) + \langle R\xi - R\xi_0, \varphi_\varepsilon \rangle - 2\varepsilon.$$

Since $z \in Z$ was an arbitrary vector satisfying $Kz = R\xi$, one gets finally

$$\operatorname*{Inf}_{Kz=R\xi}\{f_1(z) + f_2(z)\} \geq V(\xi_0) + \langle R\xi - R\xi_0, \varphi_\varepsilon \rangle - 2\varepsilon.$$

This completes the proof of inequality (2.10). We prove now the reverse inclusion $\partial V(\xi_0) \subset \Omega$. Take any $\eta \in \partial V(\xi_0)$ and $\varepsilon > 0$. We need to exhibit a pair $(z_\varepsilon, \varphi_\varepsilon) \in Z \times \Pi^*$ satisfying (2.11). Taking into account hypothesis (2.7) and Lemma A.1 (see Appendix), we can write $\eta = R^*\varphi'$ for some $\varphi' \in \Pi^*$. Since $\eta \in \partial V(\xi_0)$, this element $\varphi' \in \Pi^*$ satisfies

$$V^*(R^*\varphi') + V(\xi_0) - \langle \xi_0, R^*\varphi' \rangle = 0. \tag{2.12}$$

As a matter of computation one has

$$\begin{aligned} V^*(R^*\varphi') &= \operatorname*{Sup}_{\xi\in\Xi}\{\langle \xi, R^*\varphi' \rangle - \operatorname*{Inf}_{Kz=R\xi}\{f_1(z) + f_2(z)\}\} \\ &= \operatorname*{Sup}_{\xi\in\Xi} \operatorname*{Sup}_{Kz=R\xi}\{\langle R\xi, \varphi' \rangle - f_1(z) - f_2(z)\} \\ &= \operatorname*{Sup}_{Kz\in \operatorname{Im} R}\{\langle z, K^*\varphi' \rangle - f_1(z) - f_2(z)\} \\ &= [f_1 + f_2 + \psi_{K^{-1}(\operatorname{Im} R)}]^*(K^*\varphi'). \end{aligned}$$

Here the constraint qualifications (2.7)–(2.9) come into play. In fact, they allow us to write

$$\begin{aligned} V^*(R^*\varphi') &= \operatorname*{Inf}_{\omega_1+\omega_2+\omega_3=K^*\varphi'}\{f_1^*(\omega_1) + f_2^*(\omega_2) + \psi^*_{K^{-1}(\operatorname{Im} R)}(\omega_3)\} \\ &= \operatorname*{Inf}_{\omega_1,\omega_2}\{f_1^*(\omega_1) + f_2^*(\omega_2) : K^*\varphi' - \omega_1 - \omega_2 \in K^*([\operatorname{Im} R]^\perp)\}. \end{aligned}$$

So we can select $\omega_1 \in Z^*$, $\omega_2 \in Z^*$, and $\varphi'' \in [\operatorname{Im} R]^\perp$ in such a way that

$$K^*\varphi' - \omega_1 - \omega_2 = K^*\varphi'' \quad \text{and} \quad f_1^*(\omega_1) + f_2^*(\omega_2) \leq V^*(R^*\varphi') + \frac{\varepsilon}{2}. \tag{2.13}$$

Independently, we pick up an element $z \in Z$ satisfying

$$Kz = R\xi_0 \quad \text{and} \quad f_1(z) + f_2(z) \leq V(\xi_0) + \frac{\varepsilon}{2}. \tag{2.14}$$

The combination of (2.12)–(2.14) yields

$$\left\{f_1^*(\omega_1) + f_2^*(\omega_2) - \frac{\varepsilon}{2}\right\} + \left\{f_1(z) + f_2(z) - \frac{\varepsilon}{2}\right\} - \langle \xi_0, R^*\varphi' \rangle \leq 0.$$

Rearranging terms, one finally gets

$$[f_1^*(\omega_1) + f_1(z) - \langle z, \omega_1 \rangle] + [f_2^*(\omega_2) + f_2(z) - \langle z, \omega_2 \rangle] + [\langle z, \omega_1 + \omega_2 \rangle - \langle R\xi_0, \varphi' \rangle] \leq \varepsilon.$$

Observe that the last term

$$\langle z, \omega_1 + \omega_2 \rangle - \langle R\xi_0, \varphi' \rangle = -\langle z, K^*\varphi' - \omega_1 - \omega_2 \rangle = -\langle z, K^*\varphi'' \rangle$$

is equal to zero. Indeed, z and $K^*\varphi''$ are orthogonal with respect to the duality product $\langle \cdot, \cdot \rangle : Z \times Z^* \to \mathbb{R}$. Now, according to the classical Young–Fenchel inequality, the first two terms enclosed by the square brackets are nonnegative. This implies in particular that

$$\omega_1 \in \partial_\varepsilon f_1(z) \quad \text{and} \quad \omega_2 \in \partial_\varepsilon f_2(z).$$

Thus we have found elements $z \in Z$, $\varphi' \in \Pi^*$, and $\varphi'' \in \Pi^*$ satisfying $Kz = R\xi_0$, $R^*\varphi' = \eta$, $\varphi'' \in [\operatorname{Im} R]^\perp$, and $K^*(\varphi' - \varphi'') \in \partial_\varepsilon f_1(z) + \partial_\varepsilon f_2(z)$. From the condition $\varphi'' \in [\operatorname{Im} R]^\perp$, it follows that $R^*\varphi'' = 0$. Setting $\varphi = \varphi' - \varphi''$, one sees that the pair $(z, \varphi) \in Z \times \Pi^*$ solves the system

$$Kz = R\xi_0, \quad R^*\varphi = \eta, \quad \text{and} \quad K^*\varphi \in \partial_\varepsilon f_1(z) + \partial_\varepsilon f_2(z).$$

The proof is complete in this way. □

Remark 2.1. If $\operatorname{Im} K \subset \operatorname{Im} R$, then $K^{-1}(\operatorname{Im} R) = Z$ and $K^*([\operatorname{Im} R]^\perp) = \{0\}$. Thus, (2.8) holds trivially, and (2.9) takes the simpler form

$$(f_1 + f_2)^*(\omega) = \inf_{\omega_1 + \omega_2 = \omega} \{f_1^*(\omega_1) + f_2^*(\omega_2)\} \quad \text{for all } \omega \in Z^*.$$

The above equality holds, for instance, if the function f_2 is continuous at a point in which f_1 is finite. See also Attouch and Brezis [AB] and Rockafellar [Ro1, Thm. 20].

3. Sensitivity results for convex optimal control problems. In this section we study the first-order behaviour of the optimal-value function

$$(\alpha, \theta, \gamma) \mapsto V(\alpha, \theta, \gamma) := \operatorname{Inf}\{J(x, u, \alpha, \theta) : (x, u) \in F(\gamma)\} \tag{3.1}$$

around the reference level $(\alpha_0, \theta_0, \gamma_0) \in \mathbb{R}^k \times \Theta \times \Gamma$. We want to characterize the subdifferential $\partial V(\alpha_0, \theta_0, \gamma_0)$ in terms of the data of our optimal control problem. These data belong essentially to two different categories. First, one has the endpoint cost H and the Lagrangian L, that is to say, the terms appearing in the definition of the total cost

$$J(x, u, \alpha, \theta) := H(x(0), x(1), \alpha) + \int_0^1 L(t, x(t), u(t), \theta(t))\, dt. \tag{3.2}$$

And second, one has the matrix-valued functions A and B, which appear in the definition of the feasible set

$$F(\gamma) := \{(x, u) \in X \times U : \dot{x}(t) = A(t)\, x(t) + B(t)\, u(t) + \gamma(t) \quad \text{for a.e. } t \in [0, 1]\}. \tag{3.3}$$

The general assumptions on these data are presented next. First, a word on the space $\mathbb{R}^k \times \Theta \times \Gamma$ of parameters. The choice of the space Θ is dictated by the need of manipulating an integral function of the form

$$I_L(x, u, \theta) := \int_0^1 L(t, x(t), u(t), \theta(t))\, dt. \tag{3.4}$$

In what follows we suppose that Θ is a certain *decomposable* space of measurable functions $\theta : [0, 1] \to \mathbb{R}^m$ (see Rockafellar [Ro1, p. 59] for the precise definition of decomposability). Those not familiar with this concept can retain the particular choice

$$\Theta = L_s^m := L_s([0, 1]; \mathbb{R}^m) \qquad (1 \le s < +\infty). \tag{3.5}$$

By assumption, the decomposable space Θ is paired with Θ^* by means of a bilinear form $\langle \cdot, \cdot \rangle : \Theta \times \Theta^* \to \mathbb{R}$. As way of example, the decomposable space (3.5) is paired with $\Theta^* = L_r^m$ by means of the bilinear form

$$\langle \theta, \nu \rangle = \int_0^1 \theta(t) \cdot \nu(t)\, dt, \qquad \theta \in \Theta,\ \nu \in \Theta^*.$$

Here $r = s/(s-1)$ is the conjugate number of s, and the dot "$\cdot$" stands for the usual Euclidean product. Concerning the space of functions perturbing the dynamics of the system, we set

$$\Gamma = L_p^n := L_p([0,1]; \mathbb{R}^n).$$

This space is paired with $\Gamma^* = L_q^n$, where $q = p/(p-1)$ is the conjugate number of p, by means of the bilinear form

$$\langle \gamma, \lambda \rangle = \int_0^1 \gamma(t) \cdot \lambda(t)\, dt, \qquad \gamma \in \Gamma,\ \lambda \in \Gamma^*.$$

Since the space of trajectories $X = A_p^n$ is not decomposable, the integral function (3.4) is regarded as a function defined over the space $L_q^n \times U \times \Theta$.

The sensitivity results presented in this section rely upon the following basic hypotheses:

(3.6) $\begin{cases} \text{the matrix-valued functions } A : [0,1] \to M_{n,n}(\mathbb{R}) \text{ and } B : [0,1] \to M_{n,d}(\mathbb{R}) \\ \text{are measurable and essentially bounded;} \end{cases}$

(3.7) $\begin{cases} \text{the endpoint cost function } H : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\} \text{ is proper convex} \\ \text{lower-semicontinuous;} \end{cases}$

(3.8) $\begin{cases} \text{the Lagrangian } L : [0,1] \times \mathbb{R}^n \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\} \text{ is measurable, and} \\ L(t, \cdot, \cdot, \cdot) \text{ is a proper convex lower-semicontinuous function for a.e. } t \in [0,1]; \end{cases}$

(3.9) $\begin{cases} \text{for all } (x,u,\theta) \in L_q^n \times U \times \Theta, \text{ the negative part of the integrand} \\ t \mapsto L(t, x(t), u(t), \theta(t)) \text{ is summable over } [0,1]. \end{cases}$

Hypothesis (3.9) is introduced to avoid all confusion regarding the sum of $+\infty$ and $-\infty$. The integral (3.4) has a classical value, possibly $+\infty$ but never $-\infty$. We need to take into account also a constraint qualification condition on the cost functional, namely,

(3.10) $\begin{cases} \text{there exist } x \in X,\ u \in U,\ \alpha \in \mathbb{R}^k, \text{ and } \theta \in \Theta \\ \text{such that } H \text{ is finite at } (x(0), x(1), \alpha) \\ \text{and } I_L : L_q^n \times U \times \Theta \to \mathbb{R} \cup \{+\infty\} \text{ is continuous at } (x,u,\theta). \end{cases}$

Our first sensitivity result is somewhat standard. It deals with the "easy" case in which the unperturbed problem

(3.11) $$\text{minimize}\{J(x,u,\alpha_0,\theta_0) : (x,u) \in F(\gamma_0)\}$$

admits at least one optimal solution. Whether an element $(\beta_0, \nu_0, \lambda_0)$ belongs to the subdifferential $\partial V(\alpha_0, \theta_0, \gamma_0)$ will depend on the existence of a "dual" trajectory satisfying a given set of extremality conditions. The space of dual trajectories is defined as follows:

(3.12) $$Y := \begin{cases} A_p^n & \text{if } p \in [1,2] \text{ or } A = 0, \\ A_q^n & \text{otherwise.} \end{cases}$$

The above definition takes into account the choice of the space $X = A_p^n$ of "primal" trajectories and the matrix-valued function A appearing in the "autonomous" system

$$\dot{x}(t) = A(t)\,x(t) \quad \text{for a.e. } t \in [0,1].$$

The particular cases $p = 1$ and $p = 2$ deserve special mention since they are encountered quite often in practice. For these particular choices one has $X = Y = A_p^n$. Without further ado we write the following theorem.

THEOREM 3.1. *With assumptions* (3.6)–(3.10), *let the optimal-value function V be finite at $(\alpha_0, \theta_0, \gamma_0) \in \mathbb{R}^k \times \Theta \times \Gamma$ and $(x_0, u_0) \in X \times U$ be an optimal solution of the unperturbed problem* (3.11), *i.e.*,

$$(3.13)\quad \begin{cases} \dot{x}_0(t) = A(t)x_0(t) + B(t)u_0(t) + \gamma_0(t) \quad \textit{for a.e. } t \in [0,1], \\ V(\alpha_0, \theta_0, \gamma_0) = H(x_0(0), x_0(1), \alpha_0) + \displaystyle\int_0^1 L(t, x_0(t), u_0(t), \theta_0(t))\,dt. \end{cases}$$

Then $(\beta_0, \nu_0, \lambda_0) \in \mathbb{R}^k \times \Theta^* \times \Gamma^*$ is a subgradient of V at $(\alpha_0, \theta_0, \gamma_0)$ if and only if there exists a dual trajectory $y \in Y$ (necessarily unique) such that

$$(3.14)\qquad \lambda_0(t) = -y(t) \quad \textit{for a.e. } t \in [0,1],$$

$$(3.15)\qquad (y(0), -y(1), \beta_0) \in \partial H(x_0(0), x_0(1), \alpha_0) \quad (\textit{transversality condition}),$$

and

$$(3.16)\qquad (\dot{y}(t) + A^T(t)y(t), B^T(t)y(t), \nu_0(t)) \in \partial L(t, x_0(t), u_0(t), \theta_0(t)) \quad \textit{for a.e. } t \in [0,1]\ (\textit{Pontryagin's principle}).$$

Proof. To prove this result, the reader can adjust the proof of the next theorem. □

Remark 3.1. The symbol ∂L stands for the subdifferential mapping of $L(t,\cdot,\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$. (Subdifferentiation will never refer to the time variable.)

We are ready to state the main result of this paper. Now the existence of an optimal solution to the unperturbed problem (3.11) is no longer assumed. We emulate the technique described in Seeger [Se1, §3]; that is to say, we enlarge the subdifferential mappings ∂H and ∂L by introducing a parameter $\varepsilon > 0$. The extremality conditions (3.15) and (3.16) are now written in terms of the enlarged mappings $\partial_\varepsilon H$ and $\partial_{\sigma(t)} L(t,\cdot,\cdot,\cdot)$. Here σ refers to a nonnegative function with total weight equal to ε. More precisely, σ belongs to the set

$$(3.17)\qquad \Sigma(\varepsilon) := \left\{ \sigma \in L_1[0,1] : \int_0^1 \sigma(t)\,dt = \varepsilon,\ \sigma(t) \geq 0 \text{ for a.e. } t \in [0,1] \right\}.$$

THEOREM 3.2. *With the assumptions* (3.6)–(3.10), *let the optimal-value function V be finite at $(\alpha_0, \theta_0, \gamma_0) \in \mathbb{R}^k \times \Theta \times \Gamma$. Then the element $(\beta_0, \nu_0, \lambda_0) \in \mathbb{R}^k \times \Theta^* \times \Gamma^*$ is a subgradient of V at $(\alpha_0, \theta_0, \gamma_0)$ if and only if for all $\varepsilon > 0$ there exist a trajectory $x \in X$, a control $u \in U$, a dual trajectory $y \in Y$ (necessarily unique), and a function $\sigma \in \Sigma(\varepsilon)$ such that*

$$(3.18)\qquad \dot{x}(t) = A(t)x(t) + B(t)u(t) + \gamma_0(t) \quad \textit{for a.e. } t \in [0,1],$$

$$(3.19)\qquad \lambda_0(t) = -y(t) \quad \textit{for a.e. } t \in [0,1],$$

(3.20) $(y(0), -y(1), \beta_0) \in \partial_\varepsilon H(x(0), x(1), \alpha_0)$ (ε-*transversality condition*),

and

(3.21) $$(\dot{y}(t) + A^T(t)y(t), B^T(t)y(t), \nu_0(t)) \in \partial_{\sigma(t)} L(t, x(t), u(t), \theta_0(t))$$
for a.e. $t \in [0,1]$ (ε- *maximum principle*).

Proof. A trajectory in $X = A_p^n$ will be represented by an initial point, say, $a \in \mathbb{R}^n$, and a velocity vector, say, $v \in L_p^n$. Observe that the linear operator

$$G : Z \to X,$$
$$(a, v) \mapsto G(a, v) = a + \int_0^{(\cdot)} v(\tau)\, d\tau$$

maps $Z = \mathbb{R}^n \times L_p^n$ onto the space X. In terms of these new variables, the cost term

$$H(x(0), x(1), \alpha) + \int_0^1 L(t, x(t), u(t), \theta(t))\, dt$$

becomes

(3.22) $$H(a, a + Ev, \alpha) + \int_0^1 L(t, G(a, v)(t), u(t), \theta(t))\, dt,$$

where E stands for the linear operator

$$E : L_p^n \to \mathbb{R}^n$$
$$v \mapsto \int_0^1 v(\tau)\, d\tau.$$

If one writes the state equation

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + \gamma(t) \quad \text{for a.e. } t \in [0,1]$$

in integral form, one gets

$$x(\cdot) - x(0) = \int_0^{(\cdot)} A(\tau)x(\tau)\, d\tau + \int_0^{(\cdot)} B(\tau)u(\tau)\, d\tau + \int_0^{(\cdot)} \gamma(\tau)\, d\tau.$$

The above linear equation can be regarded as an equality

$$G(a, v) - a = R_A G(a, v) + R_B u + R_I \gamma$$

in the space

$$\mathcal{H} = \{h \in A_p^n : h(0) = 0\},$$

where the operators $R_A : X \to \mathcal{H}$, $R_B : U \to \mathcal{H}$, and $R_I : \Gamma \to \mathcal{H}$ are defined in the obvious way, namely,

$$R_A x = \int_0^{(\cdot)} A(\tau)x(\tau)\, d\tau,$$
$$R_B u = \int_0^{(\cdot)} B(\tau)u(\tau)\, d\tau,$$
$$R_I \gamma = \int_0^{(\cdot)} \gamma(\tau)\, d\tau.$$

Thus $V(\alpha, \theta, \gamma)$ is the optimal value of the convex parametric program

$$\text{(3.23)} \qquad \begin{cases} \text{minimize } H(a, a + Ev, \alpha) + \displaystyle\int_0^1 L(t, G(a,v)(t), u(t), \theta(t))\, dt, \\ (a, v, u) \in Z \times U, \\ M(a, v, u) = R_I \gamma, \end{cases}$$

where $M : Z \times U \to \mathcal{H}$ is the linear operator given by

$$M(a, v, u) = G(a, v) - a - R_A G(a, v) - R_B u.$$

We prefer to write (3.23) in the form

$$\text{(3.24)} \qquad \begin{cases} \text{minimize } H(a, a + Ev, \alpha') + \displaystyle\int_0^1 L(t, G(a,v)(t), u(t), \theta'(t))\, dt, \\ (a, v, u, \alpha', \theta') \in Z \times U \times \mathbb{R}^k \times \Theta, \\ \alpha' = \alpha, \\ \theta' = \theta, \\ M(a, v, u) = R_I \gamma, \end{cases}$$

because the later formulation fits exactly into the general scheme of Lemma 2.3. To see this, just introduce the convex functions $f_1, f_2 : Z \times U \times \mathbb{R}^k \times \Theta \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\begin{aligned} f_1(a, v, u, \alpha, \theta) &:= H(a, a + Ev, \alpha), \\ f_2(a, v, u, \alpha, \theta) &:= \int_0^1 L(t, G(a,v)(t), u(t), \theta(t))\, dt, \end{aligned}$$

and the linear operators

$$\begin{aligned} K : Z \times U \times \mathbb{R}^k \times \Theta &\to \mathbb{R}^k \times \Theta \times \mathcal{H}, \\ (a, v, u, \alpha, \theta) &\mapsto K(a, v, u, \alpha, \theta) = (\alpha, \theta, M(a, v, u)), \end{aligned}$$

$$\begin{aligned} R : \mathbb{R}^k \times \Theta \times \Gamma &\to \mathbb{R}^k \times \Theta \times \mathcal{H}, \\ (\alpha, \theta, \gamma) &\mapsto R(\alpha, \theta, \gamma) = (\alpha, \theta, R_I \gamma). \end{aligned}$$

The remaining part of the proof consists of applying Lemma 2.3 to the particular case

$$V(\alpha, \theta, \gamma) = \operatorname{Inf}\{(f_1 + f_2)(a, v, u, \alpha', \theta') : K(a, v, u, \alpha', \theta') = R(\alpha, \theta, \gamma)\},$$

where the infimum is taken with respect to $(a, v, u, \alpha', \theta') \in Z \times U \times \mathbb{R}^k \times \Theta$. If all the assumptions in Lemma 2.3 were fulfilled, then we could assert that $(\beta_0, \nu_0, \lambda_0) \in \partial V(\alpha_0, \theta_0, \gamma_0)$ if and only if for all $\varepsilon > 0$ there exist

$$(a, v, u, \alpha', \theta') \in Z \times U \times R^k \times \Theta \quad \text{and} \quad (\beta, \nu, \mu) \in \mathbb{R}^k \times \Theta^* \times \mathcal{H}^*$$

such that

$$\begin{aligned} K(a, v, u, \alpha', \theta') &= R(\alpha_0, \theta_0, \gamma_0), \\ (\beta_0, \nu_0, \lambda_0) &= R^*(\beta, \nu, \mu), \\ K^*(\beta, \nu, \mu) &\in \partial_\varepsilon f_1(a, v, u, \alpha', \theta') + \partial_\varepsilon f_2(a, v, u, \alpha', \theta'). \end{aligned}$$

This amounts to saying that for all $\varepsilon > 0$ there exist

$$(a, v, u) \in Z \times U \quad \text{and} \quad (\beta, \nu, \mu) \in \mathbb{R}^k \times \Theta^* \times \mathcal{H}^*$$

such that

$$\begin{cases} M(a, v, u) = R_I \gamma_0, \\ (\beta_0, \nu_0, \lambda_0) = R^*(\beta, \nu, \mu), \\ K^*(\beta, \nu, \mu) \in \partial_\varepsilon f_1(a, v, u, \alpha_0, \theta_0) + \partial_\varepsilon f_2(a, v, u, \alpha_0, \theta_0). \end{cases} \tag{3.25}$$

Our task now is to evaluate the adjoint operators K^* and R^* and the subdifferential mappings $\partial_\varepsilon f_1$ and $\partial_\varepsilon f_2$. For the sake of the exposition we divide this task into three steps.

Step 1 (computation of K^* and R^*). We start by pairing $Z = \mathbb{R}^n \times L_p^n$ with the space $Z^* = A_q^n$ by means of the bilinear form

$$\langle (a, v), w \rangle = a \cdot w(0) + \int_0^1 v(\tau) \cdot \dot{w}(\tau)\, d\tau, \qquad (a, v) \in Z,\ w \in Z^*.$$

The space $\mathcal{H} = \{h \in A_p^n : h(0) = 0\}$ is paired with $\mathcal{H}^* = L_q^n$ by means of

$$\langle h, \mu \rangle = \int_0^1 \dot{h}(\tau) \cdot \mu(\tau)\, d\tau, \qquad h \in \mathcal{H},\ \mu \in \mathcal{H}^*.$$

The spaces $U = L_p^d$ and $U^* = L_q^d$ are paired in the usual way, i.e.,

$$\langle u, \ell \rangle = \int_0^1 u(\tau) \cdot \ell(\tau)\, d\tau, \qquad u \in U,\ \ell \in U^*.$$

Similarly, the spaces $\Gamma = L_p^n$ and $\Gamma^* = L_q^n$ are paired by means of

$$\langle \gamma, \lambda \rangle = \int_0^1 \gamma(\tau) \cdot \lambda(\tau)\, d\tau, \qquad \gamma \in \Gamma,\ \lambda \in \Gamma^*.$$

With respect to these pairings, the linear operators K and R are continuous. Moreover, the adjoint operator K^* of K is given by

$$\begin{cases} K^* : \mathbb{R}^k \times \Theta^* \times \mathcal{H}^* \to Z^* \times U^* \times \mathbb{R}^k \times \Theta^*, \\ (\beta, \nu, \mu) \mapsto (w_\mu, \ell_\mu, \beta, \nu), \end{cases} \tag{3.26}$$

where

$$\begin{cases} w_\mu(0) = -\displaystyle\int_0^1 A^T(\tau)\mu(\tau)\, d\tau, \\ \dot{w}_\mu(t) = \mu(t) - \displaystyle\int_t^1 A^T(\tau)\mu(\tau)\, d\tau \quad \text{for a.e. } t \in [0, 1], \\ \ell_\mu(t) = -B^T(t)\mu(t) \quad \text{for a.e. } t \in [0, 1]. \end{cases} \tag{3.27}$$

The expression of the adjoint operator R^* of R is less involved. One has simply

$$\begin{aligned} R^* : \mathbb{R}^k \times \Theta^* \times \mathcal{H}^* &\to \mathbb{R}^k \times \Theta^* \times \Gamma^*, \\ (\beta, \nu, \mu) &\mapsto (\beta, \nu, R_I^* \mu), \end{aligned} \tag{3.28}$$

where $R_I^* : \mathcal{H}^* \to \Gamma^*$ is given by

$$[R_I^* \mu](t) = \mu(t) \quad \text{for a.e. } t \in [0,1]. \tag{3.29}$$

Step 2 (computation of $\partial_\varepsilon f_1$). Observe first that the convexity and the properness of the function f_1 are ensured by hypothesis (3.7). The function f_1 is just the composition $H \circ T$ of the endpoint cost function $H : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ and the continuous linear operator

$$\begin{aligned} T : Z \times U \times \mathbb{R}^k \times \Theta &\to \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^k, \\ (a, v, u, \alpha, \theta) &\mapsto (a, a + Ev, \alpha). \end{aligned}$$

Since we are working under appropriate constraint qualification hypotheses (cf. [Ro1, Thm. 19b]), we can apply the general formula [Hi, Thm. 2.2]

$$\partial_\varepsilon (H \circ T)(a, v, u, \alpha_0, \theta_0) = T^* \partial_\varepsilon H(T(a, v, u, \alpha_0, \theta_0)).$$

A simple calculus shows that the adjoint operator T^* of T is given by

$$\begin{aligned} T^* : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^k &\to Z^* \times U^* \times \mathbb{R}^k \times \Theta^*, \\ (c, e, \beta) &\mapsto \left(c + e + \int_0^{(\cdot)} \tilde{e}(\tau)\, d\tau, 0, \beta, 0 \right), \end{aligned}$$

where $\tilde{e} \in L_q^n$ is defined as

$$\tilde{e}(t) = e \quad \text{for a.e. } t \in [0,1].$$

One gets in this way

$$\partial_\varepsilon f_1(a, v, u, \alpha_0, \theta_0) = \left\{ \left(c + e + \int_0^{(\cdot)} \tilde{e}(\tau)\, d\tau, 0, \beta, 0 \right) : (c, e, \beta) \in \partial_\varepsilon H(a, a + Ev, \alpha_0) \right\}. \tag{3.30}$$

Step 3 (computation of $\partial_\varepsilon f_2$). Hypotheses (3.8) – (3.10) imply the convexity and properness of the function f_2. To evaluate the approximate subdifferential of f_2, we represent this function as the composition $I_L \circ P$ of the integral functional $I_L : L_q^n \times U \times \Theta \to \mathbb{R} \cup \{+\infty\}$ and the continuous linear operator

$$\begin{aligned} P : Z \times U \times \mathbb{R}^k \times \Theta &\to L_q^n \times U \times \Theta, \\ (a, v, u, \alpha, \theta) &\mapsto (G(a, v), u, \theta). \end{aligned}$$

In this case

$$\operatorname{Im} P = X \times U \times \Theta \subset L_q^n \times U \times \Theta.$$

Hypothesis (3.10) implies that there exists an element in the range of P at which the proper convex function I_L is continuous. Under this constraint qualification hypothesis, one can write [Hi, Thm. 2.2]

$$\partial_\varepsilon (I_L \circ P)(a, v, u, \alpha_0, \theta_0) = P^* \partial_\varepsilon I_L(P(a, v, u, \alpha_0, \theta_0)).$$

Here the adjoint operator $P^* : L_p^n \times U^* \times \Theta^* \to Z^* \times U^* \times \mathbb{R}^k \times \Theta^*$ takes the form

$$P^*(s, \ell, \nu) = (\delta_s, \ell, 0, \nu),$$

where $\delta_s \in Z^*$ is given by

$$\delta_s(0) = \int_0^1 s(\tau)\, d\tau, \qquad \dot{\delta}_s(t) = \int_t^1 s(\tau)\, d\tau \quad \text{for a.e. } t \in [0,1]. \tag{3.31}$$

Hence

$$\partial_\varepsilon f_2(a,v,u,\alpha_0,\theta_0) = \{(\delta_s,\ell,0,\nu) : (s,\ell,\nu) \in \partial_\varepsilon I_L(G(a,v),u,\theta_0)\}. \tag{3.32}$$

To compute the approximate subdifferential of the integral functional I_L, we invoke the decomposability principle stated in [Bu, Prop. 2.2.1]. We are working under hypotheses which allow us to apply this principle, and so we can write

(3.33)

$$(s,\ell,\nu) \in \partial_\varepsilon I_L(G(a,v),u,\theta_0) \Longleftrightarrow \begin{cases} \text{there exists } \sigma \in \Sigma(\varepsilon) \text{ such that} \\ (s(t),\ell(t),\nu(t)) \in \partial_{\sigma(t)} L(t, G(a,v)(t), u(t), \theta_0(t)) \\ \text{for a.e. } t \in [0,1]. \end{cases}$$

Step 3 is thus complete.

Before proceeding to join all the pieces together, we briefly check the hypotheses (2.7)–(2.9) in Lemma 2.3. First, observe that condition (2.7) holds if $\operatorname{Im} R_I = \{R_I\gamma : \gamma \in \Gamma\}$ is closed in $\mathcal{H}$ and $\operatorname{Im} R_I^* = \{R_I^*\mu : \mu \in \mathcal{H}^*\}$ is closed in Γ^*. But one can easily see that $\operatorname{Im} R_I$ is the whole space $\mathcal{H}$, and $\operatorname{Im} R_I^*$ is the whole space Γ^*. Second, condition (2.8) amounts to saying that $M^*([\operatorname{Im} R_I]^\perp)$ is a closed set in $Z^* \times U^*$. But since $\operatorname{Im} R_I$ is the whole space $\mathcal{H}$, the set $M^*([\operatorname{Im} R_I]^\perp)$ reduces to the origin in $Z^* \times U^*$. Finally, a word on condition (2.9). As mentioned in Remark 2.1, this condition holds, for instance, if f_2 is continuous at a point in which f_1 is finite. Of course, hypothesis (3.10) takes care of this constraint qualification requirement.

In short, we are allowed to apply Lemma 2.3. By combining (3.25), (3.26), (3.28), (3.30), and (3.32), we can assert that $(\beta_0,\nu_0,\lambda_0) \in \partial V(\alpha_0,\theta_0,\gamma_0)$ if and only if for all $\varepsilon > 0$ there exist

$$(a,v) \in Z, \quad u \in U, \quad \mu \in \mathcal{H}^*, \quad (c,e) \in \mathbb{R}^n \times \mathbb{R}^n, \quad s \in L_p^n$$

such that

$$\begin{cases} M(a,v,u) = R_I\gamma_0, \\ \lambda_0 = R_I^*\mu, \\ w_\mu = c + e + \displaystyle\int_0^{(\cdot)} \tilde{e}(\tau)\, d\tau + \delta_s, \\ (c,e,\beta_0) \in \partial_\varepsilon H(a, a+Ev, \alpha_0), \\ (s,\ell_\mu,\nu_0) \in \partial_\varepsilon I_L(G(a,v),u,\theta_0). \end{cases}$$

Now we write these conditions in full extent by incorporating the information given in (3.27), (3.29), (3.31), and (3.33). One can assert that $(\beta_0,\nu_0,\lambda_0) \in \partial V(\alpha_0,\theta_0,\gamma_0)$ if and only if for all $\varepsilon > 0$ there exist

$$(a,v) \in Z, \quad u \in U, \quad \mu \in \mathcal{H}^*, \quad (c,e) \in \mathbb{R}^n \times \mathbb{R}^n, \quad s \in L_p^n, \quad \sigma \in \Sigma(\varepsilon)$$

such that

$$M(a,v,u) = R_I\gamma_0, \tag{3.34}$$

$$\lambda_0(t) = \mu(t) \quad \text{for a.e. } t \in [0,1], \tag{3.35}$$

$$-\int_0^1 A^T(\tau)\,\mu(\tau)\,d\tau = c + e + \int_0^1 s(\tau)\,d\tau, \tag{3.36}$$

$$\mu(t) - \int_t^1 A^T(\tau)\mu(\tau)\,d\tau = e + \int_t^1 s(\tau)\,d\tau \quad \text{for a.e. } t \in [0,1], \tag{3.37}$$

$$(c, e, \beta_0) \in \partial_\varepsilon H(a, a + Ev, \alpha_0), \tag{3.38}$$

and

$$(s(t), -B^T(t)\mu(t), \nu_0(t)) \in \partial_{\sigma(t)} L(t, G(a,v)(t), u(t), \theta_0(t)) \quad \text{for a.e. } t \in [0,1]. \tag{3.39}$$

One obtains in this way a complete characterization of the subdifferential $\partial V(\alpha_0, \theta_0, \gamma_0)$. However, the above conditions are not easy to manipulate, and their meaning is somewhat obscure. These conditions can be stated in a simpler manner if one introduces a space of dual trajectories. To this end, it is convenient to inspect closely the following function:

$$t \in [0,1] \mapsto y(t) = c + \int_0^t [s(\tau) + A^T(\tau)\mu(\tau)]\,d\tau.$$

We regard y as a dual trajectory emanating from the point $y(0) = c$ and whose velocity $\dot{y}$ is given by

$$\dot{y}(t) = s(t) + A^T(t)\mu(t) \quad \text{for a.e. } t \in [0,1].$$

Observe that

$$\dot{y} \in \begin{cases} L_p^n & \text{if } p \in [1,2] \text{ or } A = 0, \\ L_q^n & \text{otherwise.} \end{cases}$$

Thus y belongs to the space Y introduced in (3.12). Now from (3.36) one gets

$$-\int_0^1 [s(\tau) + A^T(\tau)\mu(\tau)]\,d\tau = c + e, \tag{3.40}$$

and therefore $y(1) = -e$. By using (3.37) and (3.40), one obtains

$$\mu(t) = e - \left(c + e + \int_0^t [s(\tau) + A^T(\tau)\mu(\tau)]\,d\tau\right) \quad \text{for a.e. } t \in [0,1],$$

and thus

$$\mu(t) = -y(t) \quad \text{for a.e. } t \in [0,1].$$

In short, the elements $(c, e) \in \mathbb{R}^n \times \mathbb{R}^n$, $\mu \in \mathcal{H}^*$, and $s \in L_p^n$ are related to the dual trajectory $y \in Y$ in the manner described below:

$$\begin{cases} c = y(0), \quad e = -y(1), \\ \mu(t) = -y(t) \quad \text{for a.e. } t \in [0,1], \\ s(t) = \dot{y}(t) + A^T(t)y(t) \quad \text{for a.e. } t \in [0,1]. \end{cases} \tag{3.41}$$

Of course, the relation between $(a, v) \in Z$ and the primal trajectory $x \in X$ is simply

$$a = x(0), \quad v(t) = \dot{x}(t) \quad \text{for a.e. } t \in [0, 1]. \tag{3.42}$$

To complete the proof, it suffices to plug (3.41) and (3.42) into the conditions (3.34) – (3.35) and (3.38) – (3.39). □

The ε-transversality condition (3.20) and the ε-maximum principle (3.21) take various forms depending on the structure of the cost function H and the Lagrangian L, respectively. Let us illustrate this fact with the help of two important examples.

Example 3.1 (L does not depend on state variables). If the Lagrangian L is independent of the vector of state variables, then the ε-maximum principle (3.21) decomposes into the *dual-state equation*

$$\dot{y}(t) + A^T(t)y(t) = 0 \quad \text{for a.e. } t \in [0, 1] \tag{3.43}$$

plus the differential inclusion

$$(B^T(t)y(t), \nu_0(t)) \in \partial_{\sigma(t)} L(t, u(t), \theta_0(t)) \quad \text{for a.e. } t \in [0, 1]. \tag{3.44}$$

If L is also independent of external parameters, then (3.44) reduces further to

$$B^T(t)y(t) \in \partial_{\sigma(t)} L(t, u(t)) \quad \text{for a.e. } t \in [0, 1].$$

The above condition amounts to saying that, for almost every $t \in [0, 1]$, the control vector $u(t)$ solves the problem

$$\text{maximize } \{y(t) \cdot B(t)z - L(t, z) : z \in \mathbb{R}^d\} \tag{3.45}$$

within a tolerance $\sigma(t)$. The difference with respect to the classical maximum principle of Pontryagin is that $u(t)$ does not need to be an exact solution of (3.45).

Example 3.2 (a Mayer problem of optimal control). Suppose that the cost term to be minimized is simply $h(x(1))$; that is to say, it depends only on the final state $x(1)$. Consider also a perturbed and controlled system of the form

$$\begin{cases} \dot{x}(t) = A(t)x(t) + B(t)u(t) + \gamma(t) & \text{for a.e. } t \in [0, 1], \\ x(0) = 0. \end{cases}$$

As we have seen in the previous example, the ε-maximum principle (3.21) takes here the form

$$\begin{aligned} \dot{y}(t) + A^T(t)y(t) &= 0 \\ B^T(t)y(t) &= 0 \end{aligned} \qquad \text{for a.e. } t \in [0, 1].$$

In order to write the ε-transversality condition (3.20), one has to recognize first the form of the cost term H. In the present example one clearly has

$$H(x(0), x(1)) = \begin{cases} h(x(1)) & \text{if } x(0) = 0, \\ +\infty & \text{if } x(0) \neq 0. \end{cases}$$

Thus (3.20) reduces to

$$\begin{cases} -y(1) \in \partial_\varepsilon h(x(1)), \\ x(0) = 0, \quad y(0) \in \mathbb{R}^n \quad \text{(i.e., } y(0) \text{ is unconstrained)}. \end{cases}$$

To illustrate how Theorem 3.2 works in practice, consider the specific case of Example 1.2. Here (3.18) and (3.19) take, respectively, the forms

$$\dot{x}_1(t) = -2x_1(t) + 3x_2(t) + u(t), \quad \dot{x}_2(t) = x_2(t) \quad \text{for a.e. } t \in [0,1]$$

and

$$\lambda_1(t) = -y_1(t), \quad \lambda_2(t) = -y_2(t) \quad \text{for a.e. } t \in [0,1].$$

The ε-maximum principle (3.21) yields simply

$$\begin{aligned} &\dot{y}_1(t) = 2y_1(t) \\ &\dot{y}_2(t) = -3y_1(t) - y_2(t) \quad \text{for a.e. } t \in [0,1], \\ &y_1(t) = 0, \end{aligned}$$

and the ε-transversality condition (3.20) reads

$$\begin{cases} -(y_1(1), y_2(1)) \in \partial_\varepsilon h(x_1(1), x_2(1)), \\ x_1(0) = 0, \quad x_2(0) = 0, \end{cases} \tag{3.46}$$

where $h : \mathbb{R}^2 \to \mathbb{R}$ is given by

$$h(x_1(1), x_2(1)) = [\exp(x_1(1)) + x_2^2(1)]^{1/2}.$$

We can draw a lot of information from the above conditions. Writing $e = \exp(1)$, we get

$$x_2(t) = 0, \quad y_1(t) = 0, \quad y_2(t) = y_2(0)e^{-t} \quad \text{for all } t \in [0,1]$$

and

$$\lambda_1(t) = 0, \quad \lambda_2(t) = -y_2(0)e^{-t} \quad \text{for a.e. } t \in [0,1].$$

Plugging $y_1(1) = 0$, $y_2(1) = y_2(0)e^{-1}$, and $x_2(1) = 0$ into (3.46), one obtains

$$-(0, y_2(0)e^{-1}) \in \partial_\varepsilon h(x_1(1), 0). \tag{3.47}$$

Now we take into account the specific structure of h and write (3.47) in the more explicit form

$$\exp\left[\frac{1}{2}x_1(1)\right] \le \varepsilon, \qquad |y_2(0)| \le e.$$

Summarizing, $(\lambda_1, \lambda_2) \in \partial V(0,0)$ if and only if for all $\varepsilon > 0$ there exist $x_1 \in A_1[0,1], u \in L_1[0,1]$, and $c \in \mathbb{R}$ such that

$$\begin{aligned} &\dot{x}_1(t) = -2x_1(t) + u(t) \\ &\lambda_1(t) = 0 \qquad\qquad\qquad \text{for a.e. } t \in [0,1] \\ &\lambda_2(t) = ce^{-t} \end{aligned}$$

and

$$\exp\left[\frac{1}{2}x_1(1)\right] \le \varepsilon, \quad x_1(0) = 0, \quad c \in [-e, e].$$

This amounts to saying that

$$(\lambda_1, \lambda_2) \in \partial V(0,0) \Longleftrightarrow \begin{cases} \text{for some constant} \quad c \in [-e, e], \\ \lambda_1(t) = 0 \quad \text{and} \quad \lambda_2(t) = ce^{-t} \quad \text{for a.e. } t \in [0,1]. \end{cases}$$

In the present example, one can also compute $\partial V(0,0)$ by evaluating first the optimal-value function V. As a matter of computation one obtains

$$V(\gamma_1, \gamma_2) = e \left| \int_0^1 e^{-t} \gamma_2(t)\, dt \right|.$$

4. Extensions. In this section we explore the case in which the optimal control problem

$$\text{minimize } \{J(x, u, \alpha, \theta) : (x, u) \in F(\gamma)\}$$

involves explicit constraints on the control function $u \in U$. Such a case is important in applications and deserves further discussion. In what follows, U_{ad} denotes the set of "admissible" controls, i.e., those functions u in U such that

$$u(t) \in \Omega(t) \quad \text{for a.e. } t \in [0,1]. \tag{4.1}$$

The feasible set $F(\gamma)$ is understood as the set of all pairs $(x, u) \in X \times U_{ad}$ satisfying the state equation (1.2).

To remain within the realm of convex analysis, we suppose that

(4.2)
$\Omega : [0,1] \to \mathbb{R}^d$ is a measurable multifunction with nonempty closed convex values.

In principle, one can take care of the admissibility concern (4.1) by working with the modified Lagrangian

$$\tilde{L}(t, x, z, \theta) := L(t, x, z, \theta) + \psi_{\Omega(t)}(z), \tag{4.3}$$

where $\psi_{\Omega(t)}$ stands for the indicator function of the set $\Omega(t)$. In this case the ε-maximum principle (3.21) is stated in terms of the approximate subdifferential mapping $\partial_{\sigma(t)} \tilde{L}$. But since $\tilde{L}$ is only an auxiliary tool, we should express $\partial_{\sigma(t)} \tilde{L}$ in terms of L and Ω. By applying Kutateladze's rule (cf. [Hi, Thm. 2.1]) on the approximate subdifferential of the sum of two functions, one gets

$$\partial_{\sigma(t)} \tilde{L}(t, x, z, \theta) = \bigcup_{\substack{\varepsilon_1 \geq 0, \varepsilon_2 \geq 0 \\ \varepsilon_1 + \varepsilon_2 = \sigma(t)}} \{\partial_{\varepsilon_1} L(t, x, z, \theta) + \{0\} \times \partial_{\varepsilon_2} \psi_{\Omega(t)}(z) \times \{0\}\}. \tag{4.4}$$

Formula (4.4) not only requires an additional constraint qualification assumption but also leads to the formulation of an ε-maximum principle that is very cumbersome and of little practical interest. For this reason we prefer to avoid the use of Kutateladze's rule and return to the proof of Theorem 3.2. In the next result we keep the same assumptions as in Theorem 3.2, except that the constraint qualification condition (3.10) is replaced with

$$\begin{cases} \text{there exist } x \in X,\ u \in U_{ad},\ \alpha \in \mathbb{R}^k, \text{ and } \theta \in \Theta \text{ such that } H \text{ is continuous at} \\ (x(0), x(1), \alpha) \text{ and } I_L : L_q^n \times U \times \Theta \to \mathbb{R} \cup \{+\infty\} \text{ is continuous at } (x, u, \theta). \end{cases} \tag{4.5}$$

THEOREM 4.1. *With the assumptions* (3.6)–(3.9), (4.2), *and* (4.5), *let the optimal-value function* V *be finite at* $(\alpha_0, \theta_0, \gamma_0) \in \mathbb{R}^k \times \Theta \times \Gamma$. *Then* $(\beta_0, \nu_0, \lambda_0) \in \mathbb{R}^k \times \Theta^* \times \Gamma^*$ *is a subgradient of* V *at* $(\alpha_0, \theta_0, \gamma_0)$ *if and only if for all* $\varepsilon > 0$ *there exist a trajectory* $x \in X$, *a control* $u \in U$, *a function* $r \in U^*$, *a dual trajectory* $y \in Y$ *(necessarily unique), and functions* σ *and* δ *in* $\Sigma(\varepsilon)$ *such that the conditions* (3.18)–(3.20) *hold, together with the* ε*-maximum principle*

(4.6)
$$(\dot{y}(t) + A^T(t)y(t), B^T(t)y(t) - r(t), \nu_0(t)) \in \partial_{\sigma(t)} L(t, x(t), u(t), \theta_0(t)) \quad \textit{for a.e. } t \in [0,1]$$

and the ε*-normality condition*

(4.7)
$$r(t) \in \partial_{\delta(t)} \psi_{\Omega(t)}(u(t)) \quad \textit{for a.e. } t \in [0,1].$$

Proof. We follow the same steps as in the proof of Theorem 3.2, except that now we have the extra function

$$f_3(a, v, u, \alpha, \theta) := \int_0^1 \psi_{\Omega(t)}(u(t))\, \mathrm{dt} = F_\Omega(u).$$

One can modify Lemma 2.3 and write formula (2.6) with the extra term $\partial_\varepsilon f_3(z)$. Of course this requires an adjustment in the constraint qualification hypotheses (2.9). This explains why we have replaced the assumption (3.10) with (4.5).

But as a matter of computation one has

$$\partial_\varepsilon f_3(a, v, u, \alpha_0, \theta_0) = \{0\} \times \partial_\varepsilon F_\Omega(u) \times \{0\} \times \{0\},$$

with

$$\partial_\varepsilon F_\Omega(u) = \bigcup_{\delta \in \Sigma(\varepsilon)} \{r \in U^* : r(t) \in \partial_{\delta(t)} \psi_{\Omega(t)}(u(t)) \quad \text{for a.e. } t \in [0,1]\}.$$

The rest of the proof is now routine. □

Remark 4.1. The ε-normality condition (4.7) says that $u \in U$ is admissible in the sense of (4.1) and that for a.e. $t \in [0,1]$, the vector $r(t)$ is $\delta(t)$-normal to $\Omega(t)$ at $u(t)$, i.e.,

$$r(t) \cdot (z - u(t)) \leq \delta(t) \quad \text{for all } z \in \Omega(t).$$

Remark 4.2. An inclusion like (4.6) appears already in a paper by Rockafellar [Ro2, p. 217]. However, that inclusion does not involve parameters and is stated in terms of the exact subdifferential mapping ∂L.

5. Final remarks. This paper follows as close as possible the methodology used by the second author [Se1] in the context of a Bolza problem of calculus of variations. It is shown in [Se1] that if $V(\alpha, \theta)$ denotes the optimal-value of the Bolza problem (1.1), then a subgradient (β_0, ν_0) of V at (α_0, θ_0) is characterized in terms of the ε-transversality condition (3.20) and the approximate Euler–Lagrange inclusion

$$(\dot{y}(t), y(t), \nu_0(t)) \in \partial_{\sigma(t)} L(t, x(t), \dot{x}(t), \theta_0(t)) \quad \text{for a.e. } t \in [0,1].$$

In principle it is possible to write our optimal control problem as a Bolza problem of calculus of variations (see, for instance, [Ro2, §4] or [IT, Chap. 2]). This can be done by introducing a suitable Lagrangian $\tilde{L}$ that incorporates the dynamics of the system, that is to

say, the state equation (1.2). Although this approach seems quite natural, there are several reasons why we do not recommend it.

(a) First, it is of no interest to state the approximate Euler–Lagrange inclusion in terms of the modified Lagrangian $\tilde{L}$. One has of course to evaluate $\partial_{\sigma(t)}\tilde{L}$ in terms of the original Lagrangian L and of the data appearing in the state equation. This can be done only at a very heavy price. One encounters the same kind of difficulties as with the modified Lagrangian (4.3); in particular, one needs to introduce additional constraint qualification assumptions.

(b) Second, this approach leads to the formulation of an ε-maximum principle that is not as simple as the one established in Theorem 3.2.

(c) Third, one should keep in mind that writing the approximate Euler–Lagrange inclusion in terms of $\tilde{L}$ is just the starting point of an alternative proof. Most of the heavy work is left in the evaluation of $\partial_{\sigma(t)}\tilde{L}$. It is not without reason that the proof of Theorem 3.2 took us around five pages.

Finally, we would like to mention that our proof of Theorem 3.2 cannot be obtained from that of [Se1, Thm. 2] by minor modifications. Our proof contains several steps and results proper to the framework of an optimal control problem.

Appendix. The following result is probably known, but we have not been able to find it in the literature. It has to do with the subdifferential of a composite function like

$$\xi \in \Xi \mapsto (\Phi \circ R)(\xi) := \Phi(R\xi),$$

where

$$\text{(A)} \qquad \begin{cases} R : \Xi \to \Pi & \text{is a continuous linear operator,} \\ \Phi : \Pi \to \overline{\mathbb{R}} & \text{is a convex function.} \end{cases}$$

The spaces Ξ and Π are as in §2.

LEMMA A.1. *With assumption* (A), *let* $\Phi \circ R$ *be finite at the point* $\xi_0 \in \Xi$. *Then the subdifferential* $\partial(\Phi \circ R)(\xi_0)$ *is contained in the closure of* $\operatorname{Im} R^*$.

Proof. Take $\eta \in \partial(\Phi \circ R)(\xi_0)$, and suppose that η does not belong to the closure of

$$\operatorname{Im} R^* = \{R^*\varphi : \varphi \in \Pi^*\}.$$

By a separation argument, there exists $\tilde{\xi} \in \Xi$ such that

$$\langle \tilde{\xi}, \eta \rangle > \langle \tilde{\xi}, R^*\varphi \rangle \quad \text{for all } \varphi \in \Pi^*.$$

Thus, $R\tilde{\xi} = 0$ and $\langle \tilde{\xi}, \eta \rangle > 0$. Since η is supposed to be in $\partial(\Phi \circ R)(\xi_0)$, one can write

$$\Phi(R\xi) \geq \Phi(R\xi_0) + \langle \xi - \xi_0, \eta \rangle \quad \text{for all } \eta \in \Xi.$$

Setting $\xi = \xi_0 + t\tilde{\xi}$ $(t > 0)$, one gets

$$\Phi(R\xi_0) \geq \Phi(R\xi_0) + t\langle \tilde{\xi}, \eta \rangle,$$

that is to say, $\langle \tilde{\xi}, \eta \rangle \leq 0$. This is clearly a contradiction. □

The above lemma provides only very rough information on the subdifferential $\partial(\Phi \circ R)(\xi_0)$. However, it does not require any constraint qualification hypothesis.

Acknowledgment. The original title of this paper was "Approximate Pontryagin's principle and perturbation analysis of convex optimal control problems." As kindly pointed out by one of the referees, the name "approximate Pontryagin's principle" has been already used, although in a different context (see, e.g., the works of B. Mordukhovich [M1, M2, M3]). Following the recommendation of the referees, we have added §§4 and 5. We are grateful to them for their criticism and remarks.

REFERENCES

[AB] H. Attouch and H. Brezis, *Duality for the sum of convex functions in general Banach spaces*, in Aspects of Mathematics and its Applications, J. Barroso, ed., North-Holland, Amsterdam, 1986, pp. 125–133.

[Bu] M. Bustos, *Conditions d'optimalité à ε-près dans un problème d'optimisation non différentiable*, Thesis, Université Paul Sabatier, Toulouse, 1989.

[C1] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[C2] ———, *Perturbed optimal control problems*, IEEE Trans. Automat. Control. AC-31, (1986), pp. 535–542.

[C3] ———, *Methods of Dynamic and Nonsmooth Optimization*, CBMS-NSF Regional Conf. Series 57, Society for Industrial and Applied Mathematics, Philadelphia, 1989.

[CL1] F. H. Clarke and P. D. Loewen, *The value function in optimal control: Sensitivity, controllability, and time-optimality*, SIAM J. Control Optim., 24 (1986), pp. 243–263.

[CL2] ———, *State constraints in optimal control: A case study in proximal normal analysis*, SIAM J. Control Optim., 25 (1987), pp. 1440–1456.

[Hi] J.-B. Hiriart-Urruty, *ε-subdifferential calculus*, in Convex Analysis and Optimization, Notes in Math. 57, J. P. Aubin and R. B. Vinter, eds., Pitman, Boston, 1982, pp. 43–92.

[IT] A. D. Ioffe and V. M. Tihomirov, *Theory of Extremal Problems*, Studies in Mathematics and its Applications, Vol. 6, North-Holland, Amsterdam, 1979.

[La] P. L. Laurent, *Approximation et optimisation*, Herman, Paris, 1972.

[Lo] P. D. Loewen, *Perturbed differential inclusion problems*, in Nonsmooth Optimization and Related Topics, F. H. Clarke, V. F. Demyanov, and F. Giannessi, eds., Plenum Press, New York, 1989, pp. 255–263.

[M1] B. Mordukhovich, *An approximate maximum principle for finite difference control systems*, U.S.S.R. Comput. Maths. Math. Phys., 26 (1988), pp. 106–114.

[M2] ———, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988. (In Russian; English transl. by Wiley-Interscience, to appear.)

[M3] ———, *Maximum principle for nonconvex finite difference control systems*, In Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Control and Inform. Sci. 144, Springer-Verlag, New York, 1990, pp. 539–548.

[MS1] M. Moussaoui and A. Seeger, *Sensitivity analysis of optimal-value functions of convex parametric programs with possibly empty solution sets*, SIAM J. Optim., 4 (1994), pp. 659–675.

[MS2] ———, *Etude de sensibilité de la fonction valeur optimale en programmation convexe*, C. R. Acad. Sci. Paris Ser. I. Math., 321 (1995).

[On] H. Oniki, *Comparative dynamics in optimal control theory*, Journal of Economic Theory, 6 (1973), pp. 265–283.

[Ro1] R. T. Rockafellar, *Conjugate duality and optimization*, Regional Conference Series in Applied Mathematics, Vol. 16, Society for Industrial and Applied Mathematics, Philadelphia, 1974.

[Ro2] ———, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.

[Se1] A. Seeger, *Approximate Euler–Lagrange inclusion, approximate transversality condition and sensitivity, analysis of convex parametric problems of calculus of variations*, Set-Valued Analysis, 2 (1994), pp. 307–325.

[Tu] P. N. V. Tu, *Introductory Optimization Dynamics*, Springer-Verlag, Berlin, 1991.

[Za] C. Zalinescu, *Stability for a class of nonlinear optimization problems and applications*, in Nonsmooth Optimization and Related Topics, F. H. Clarke, V. F. Demayanov, and F. Giannessi, eds., Plenum Press, New York, 1989, pp. 437–458.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 428–436, March 1996

002

SOLAR CARS AND VARIATIONAL PROBLEMS EQUIVALENT TO SHORTEST PATHS*

D. J. GATES† AND M. WESTCOTT†

Abstract. A classical theorem of Hardy, Littlewood, and Pólya on rearrangements of functions is used to prove the equivalence of a class of variational problems. As a consequence, solutions are shortest paths and can be computed numerically via quadratic programming. Applications include solar cars and reservoirs.

Key words. optimization, solar-powered car, storage, quadratic program, rearrangements of functions, convex envelopes, variational problem

AMS subject classifications. 49J05, 49J40, 49K05, 49N10

1. Introduction.

During our study of driving strategy for the 1993 Darwin-to-Adelaide solar-powered car race, we encountered variational problems of the following form: minimize

$$\Gamma(F) = \int_0^1 \phi[F'(t)]dt \tag{1}$$

with respect to F, subject to

$$A(t) \le F(t) \le B(t), \tag{2}$$

$$F(0) = F_0, \qquad F(1) = F_1, \tag{3}$$

where ϕ is a convex function, F' is the derivative of F (assumed to exist almost everywhere), $A(t)$ and $B(t)$ are given functions on $[0,1]$, and F_0 and F_1 are given constants satisfying $A(0) \le F_0 \le B(0)$, $A(1) \le F_1 \le B(1)$. There are many applications of this problem.

Example 1. One wishes to travel by the shortest two-dimensional path $(t, F(t))$ starting from $(0, F_0)$ and reaching destination $(1, F_1)$ while remaining between boundaries defined by $(t, A(t))$ and $(t, B(t))$. These boundaries might be the banks of a river, if travelling by boat, or the edges of a track or road or gully, if travelling by land. In this case

$$\phi(f) = (1 + f^2)^{1/2}$$

and $\Gamma(F)$ is the length of path F. Of course, one knows intuitively the solution to this problem (Figure 1).

Example 2. Replace the traveller in Example 1 by a piece of string, and pull the string tight. This problem is mathematically identical to Example 1, and one understands intuitively why the solutions are the same.

Example 3. A solar-powered car has a battery of capacity K which stores electrical energy perfectly efficiently. The speed of the car is $v[g(t)]$, where $g(t)$ is the power delivered to the motor at time t, so

$$\int_0^1 v[g(t)]dt \tag{4}$$

* Received by the editors December 15, 1993; accepted for publication (in revised form) October 17, 1994.

† CSIRO, Divisions of Mathematics and Statistics, GPO Box 1965, Canberra, A.C.T. 2601, Australia.

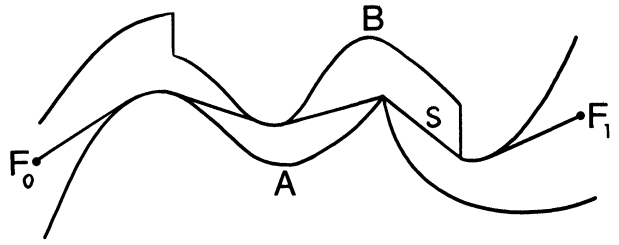


FIG. 1.

is the distance travelled in time interval $[0, 1]$. Here, v is an increasing, *concave* function, a typical example being

$$v(g) = cg^{1/3}$$

for constant c, this implying that the car faces a resistive force proportional to the square of the speed. Then maximizing the distance travelled is equivalent to minimizing Γ with $\phi = -v$.

The constraints arise from the battery capacity. Let $P(t)$ be the power generated by the solar panel, assumed known. Then the energy content of the battery at time t is

$$E(t) = E(0) + \int_0^t [P(s) - g(s) - h(s)]\, ds, \tag{5}$$

where $h(s)$ is the power *overflow*, i.e., the power which is available at the panel but is rejected because the battery is full and the motor is not drawing enough power to use this energy. Then the constraints are

$$0 \le E(t) \le K. \tag{6}$$

Because $h(s)$ depends on $P(s)$ and $g(s)$, the constraints are not of the form (2). But putting

$$f(t) \equiv g(t) + h(t) \ge g(t) \tag{7}$$

implies that

$$\int_0^1 v[f(t)]dt \ge \int_0^1 v[g(t)]dt; \tag{8}$$

i.e., it is always better to divert the overflow through the motor. Then (6) reduces to (2) with $F(t) = \int_0^t f(s)ds$ and

$$\begin{aligned} B(t) &= E(0) + \int_0^t P(s)ds, \\ A(t) &= B(t) - K, \end{aligned} \tag{9}$$

while $F_0 = 0$ and

$$F_1 = E(0) - E(1) + \int_0^1 P(t)dt. \tag{10}$$

Thus the variational problem is of the form (1) to (3).

Example 4. Water flow from a reservoir of capacity K (litres, say) drives a generator and produces power output $\Pi[g(t)]$, where $g(t)$ is the flow rate (litres per second, say) of water through the turbine at time t. Due to decreased efficiency of the system at higher flow rates, $\Pi(g)$ is concave, although it is increasing. Then

$$\int_0^1 \Pi[g(t)]dt \tag{11}$$

is the total energy generated during $[0, 1]$. Putting $\phi = -\Pi$ and minimizing Γ amount to maximizing the energy output of the system.

Now let $P(t)$ be the rate of input (litres per second) to the reservoir (from runoff and direct rainfall). If $E(t)$ is the water content of the reservoir and $h(t)$ is the overflow rate, then we again have equations (5) to (10) and a variational problem of the form (1) to (3).

Example 5. A commodity, such as salt, is supplied to a stockpile of fixed capacity and sold at a time-varying, irregular but known rate (on arrival of ships) at a fixed price per ton. The total cost is the supply to the stockpile, and this cost is an increasing, convex function of the supply rate. (Higher supply rates are increasingly costly due to overtime pay and the like.) What is the optimal supply rate? The mathematical formulation follows previous examples and is left to the reader.

Example 6. Suppose an unknown distribution function $F(t)$ satisfies conditions (2) and (3) with $F_0 = 0$ and $F_1 = 1$. These might arise from a model or perhaps from confidence intervals constructed from data. Then the maximum entropy (information) $F(t)$ is the solution of the variational problem with

$$\phi(f) = f \log f.$$

For a related problem, see [6, §3a.6].

Example 7. Make the path in Example 1 as straight as possible in the sense that

$$\int_0^1 [F'(t)]^2 dt$$

is minimized, or equivalently

$$\int_0^1 [F'(t) - \bar{F}'(t)]^2 dt$$

is minimized, where

$$\bar{F}(t) = F_0 + (F_1 - F_0)t.$$

Thus one is making the direction as uniform as possible or the route as direct as possible by penalizing large deviations from the direct route.

The main result of this paper is that the solutions to all these problems are independent of ϕ, provided that it is convex. Consequently the shortest path (Example 1) solves all of them.

For the solar car, this means that, in the sense of Example 7, the power consumption of the motor should be as uniform as possible over time. Similarly, the power generation in the hydroelectric example should be as uniform as possible, the supply

to the stockpile should be as uniform as possible, and the probability distribution in Example 6 should be as uniform as possible.

Another useful consequence follows from Example 7; since the problem with

$$\phi(f) = f^2$$

has the common solution, all such problems can be solved by quadratic programming, for which very fast algorithms exist. This fact is exploited in our computer program SOLARMAX, which was used by the Aurora Q1 solar-car team to study optimal driving strategies for the 1993 Darwin-to-Adelaide solar-powered car race.

Problems resembling Example 1 have, of course, a long history and were the motivating force behind the early development of the calculus of variations. For example, the problems of Jakob Bernoulli and Fermat take the following form: minimize

$$\int_0^1 \phi[F'(t)]\psi[F(t), t]dt$$

subject to fixed $F(0)$ and $F(1)$. Evidently, our problem is broadly of this type if we take

$$\psi(F, t) = \begin{cases} 1 & \text{if } A(t) \leq F \leq B(t) \\ \infty & \text{if not.} \end{cases}$$

However, our task is not to find explicit solutions but to prove the equality of solutions of a class of problems. Here lies the crux of our problem: to prove that a class of functionals have their minima at the same point without explicitly knowing that point.

2. Preliminary results.

DEFINITION. *A function C is called* piecewise hollow *(PWH) if there is a finite set of points $0 < t_1 < \cdots < t_n < 1$ such that C is differentiable and either convex or concave in (t_{i-1}, t_i) $(i = 1, \ldots, n)$. We call these* intervals of hollowness.

Note that PWH functions are continuous, except perhaps at $t_1, \ldots, t_n$, and have one-sided limits everywhere [3, §3.18].

Assumptions. 1. A, B are PWH.

2. $F \in \mathcal{F}$, the class of absolutely continuous functions on $[0, 1]$.

We use the notation $f \equiv F'$ for $F \in \mathcal{F}$, which exists almost everywhere (a.e.).

The conditions on A and B are appropriate for our applications and simplify the proofs somewhat. We now proceed to prove the main result in the absence of one or other of the bounds A, B and then use these special cases to establish the main theorem.

DEFINITION. *The* convex envelope *of a function C, denoted $C_\cup$, is the* maximal *convex function not exceeding C. (Equivalently $C_\cup$ is the envelope of convex functions not exceeding C).*

The definition is not vacuous because the set of convex functions $\leq C$ obviously has a supremum which is unique. Its convexity is easily proved directly [4, p. 103].

DEFINITION. *The* concave envelope *of C, denoted by $C_\cap$, is the* minimal *concave function not exceeded by C.*

DEFINITION. *Let C stand for either A or B. Then C^* is given by*

$$\begin{aligned} C^*(t) &= C(t) \quad \text{if } 0 < t < 1, \\ C^*(0) &= F_0, \\ C^*(1) &= F_1. \end{aligned} \tag{12}$$

By $A^*_\cap$ we mean $(A^*)_\cap$, not $(A_\cap)^*$. Note that $A^* \le F \le B^*$ and $A^*_\cap(0) = B^*_\cup(0) = F_0$, $A^*_\cap(1) = B^*_\cup(1) = F_1$. Further, $A^*_\cap, B^*_\cup \in \mathcal{F}$ since they are concave/convex [3, p. 130].

DEFINITION. *For any function $g(t)$, its* (decreasing) rearrangement *is*

$$\bar{g}(t) \equiv \sup\{y \ : \ m(y) > t\},$$

where $m(y)$ is the measure of the set

$$\{t \ : \ g(t) > y\}.$$

The notion of rearrangement is discussed in [3, §10.12] and [5, p. 15].

THEOREM 1. *If A is PWH and $B \equiv \infty$,*

$$\inf_{F \in \mathcal{F}} \Gamma(F) = \Gamma(A^*_\cap) \tag{13}$$

for any convex ϕ. If ϕ is strictly convex, $A^*_\cap$ is the unique minimum.

Proof. For notational simplicity, write $Q(t) \equiv A^*_\cap(t)$. Form the rearrangements of f, for any $F \in \mathcal{F}$, and of Q'; both derivatives exist a.e. Then, using [3, eq. (10.12.2)], and $F \in \mathcal{F}$,

$$G_1(t) \equiv F_0 + \int_0^t \bar{f}(s)ds \ge F_0 + \int_0^t f(s)ds = F(t). \tag{14}$$

Obviously $G_1(0) = F_0$, while $G_1(1) = F(1) = F_1$ by a basic property of rearrangements [3, p. 277]. Since $\bar{f}$ is decreasing, $G_1(t)$ is concave, with $G_1 \ge A^*$ by (14), whence $G_1 \ge Q$. Because Q' is decreasing, $\overline{Q'} = Q'$, so

$$Q(t) = F_0 + \int_0^t Q'(s)ds = F_0 + \int_0^t \overline{Q'}(s)ds \tag{15}$$

since $Q \in \mathcal{F}$. From (14) and (15),

$$\int_0^t \bar{f}(s)ds \ge \int_0^t \overline{Q'}(s)ds \quad \text{for} \quad 0 \le t < 1,$$

$$\int_0^1 \bar{f}(s)ds = \int_0^1 \overline{Q'}(s)ds. \tag{16}$$

So by a classical theorem of Hardy, Littlewood and Pólya [2] (see also [5, p. 15]),

$$\int_0^1 \phi[Q'(t)]dt \le \int_0^1 \phi[f(t)]dt \tag{17}$$

for any $F \in \mathcal{F}$ and any convex ϕ. Since $Q \in \mathcal{F}$, (14) follows.

To prove uniqueness, suppose $F \in \mathcal{F}$ is any other function with $A \le F$. Then clearly the convex combination $F_\theta \equiv \theta F + (1-\theta)Q$, $0 < \theta < 1$, is also in $\mathcal{F}$ and $A \le F_\theta$. Now Γ is a *strictly* convex functional if ϕ is strictly convex, so $\Gamma(F_\theta) < \theta\Gamma(F) + (1-\theta)\Gamma(Q)$ and hence $\Gamma(F_\theta) - \Gamma(Q) < \theta\{\Gamma(F) - \Gamma(Q)\}$. But the left side is ≥ 0, since Q is a minimum, and $\theta > 0$, so $\Gamma(F) > \Gamma(Q)$. Since F is arbitrary, uniqueness follows and the theorem is proved.

THEOREM 2. *If B is PWH and $A \equiv -\infty$,*

$$\inf_{F \in \mathcal{F}} \Gamma(F) = \Gamma(B^*_{\cup}) \tag{18}$$

for any convex ϕ. If ϕ is strictly convex, $B^*_{\cup}$ is the unique minimum.

Proof. Let $R(t) \equiv B^*_{\cup}(t) \in \mathcal{F}$. Then, as before, for any $F \in \mathcal{F}$,

$$G_2(t) \equiv \int_0^t \overline{(-f')}(s)ds - F_0 \geq -F(t), \tag{19}$$

$G_2(0) = -F_0$, $G_2(1) = -F_1$, $G_2(t)$ is concave, and $G_2 \geq -B^*$. Hence $G_2 \geq -R$, since the convex envelope of B^* is the concave envelope of $-B^*$. So we again get

$$\int_0^1 \phi[-R'(t)]dt \leq \int_0^1 \phi[-f(t)]dt \tag{20}$$

for any $F \in \mathcal{F}$ and any convex ϕ. Since $\phi(-x)$ is convex if $\phi(x)$ is convex, and $R \in \mathcal{F}$, (20) proves (18). Uniqueness follows as above.

3. The main results. To prove our main theorem, we need to suitably characterize the shortest path $S(t)$, which, as mentioned in the introduction, turns out to give the minimum of Γ. We call it the "string function," since it is the mathematical equivalent of a piece of string fixed at $P_0 \equiv (0, F_0)$, threaded through the "tube" between A and B, and pulled tight at $P_1 \equiv (1, F_1)$ (cf. Example 2).

DEFINITION. *$S(t)$ is the unique solution to the variational problem when $\phi(f) = (1 + f^2)^{1/2}$.*

Clearly, $S(t)$ is the shortest path between P_0 and P_1 lying between the curves A and B, because $\mathcal{L}(F) \equiv \int_0^1 [1 + f^2(t)]^{1/2} dt$ is the length of $F \in \mathcal{F}$. Note that any $F \in \mathcal{F}$ has a length.

It is evident from Theorems 1 and 2 that the minimizing envelope functions $A^*_{\cap}, B^*_{\cup}$ are the (unique) shortest paths between P_0 and P_1 above A and below B, respectively. For $A^*_{\cap}$ minimizes $\Gamma(F)$ for $F \geq A$ and *any* convex ϕ, hence for $\Gamma = \mathcal{L}$.

We must first establish that the definition is not vacuous in general. The *existence* of *a* solution to the shortest path problem is intuitively obvious but requires a little effort to prove formally. Initially we work with the wider class of functions of bounded variation, which also have lengths.

LEMMA 1. *Suppose $F \in \mathcal{B}$, the class of functions of bounded variation on $[0, 1]$ which satisfy* (2) *and* (3), *with the length $L(F)$ of F defined as the total variation of F over $[0, 1]$* [1, §7.3]. *Then there is an $F_\star \in \mathcal{B}$ which minimizes L.*

Proof. Let K be the closure of the subset of the plane lying between A and B. It is clear that there is at least one path of finite length between P_0 and P_1 lying in K; call its length λ. Now define

$$\mathcal{C} = \{F : F \in \mathcal{B}, L(F) \leq \lambda\}.$$

The graph of F, $(t, F(t))$ for $t \in [0, 1]$, is in K.

By the Hilbert compactness theorem [1, Thm. 7.10] $\mathcal{C}$ is sequentially compact, while L is lower semicontinuous [1, Thm. 7.6]. The lemma now follows from [1, Thm. 7.1].

Remark. Theorems 1 and 2 also hold in this wider context. (Use the inscribed polygons whose limits give the variational length L.)

If in fact $F_\star \in \mathcal{F}$, then we can take $S = F_\star$, because then $L = \mathcal{L}$ [1, Thm. 9.7]. For general PWH A and B this is not true (see §4). We now prove $F_\star \in \mathcal{F}$ for suitable A and B.

DEFINITION. *$A \prec B$, for PWH A, B, means that $A(t+0) < B(t-0)$ for all $t \in (0,1)$.*

The next lemma establishes the required properties of $F_\star$, which, although intuitively obvious, need to be deduced from its definition.

LEMMA 2. *Suppose A, B are PWH and $A \prec B$. Then*

(a) *There is a* finite *partition of $[0,1]$ into intervals I_j such that, throughout any I_j, either $F_\star = A$ or $F_\star = B$, or $A < F_\star < B$.*

(b) *If $F_\star = A$ ($F_\star = B$) throughout I_j, $F_\star$ is concave (convex) throughout I_j.*

(c) *If $A < F_\star < B$ throughout I_j, $F_\star$ is a straight line thoughout I_j.*

Proof. Define $T_A \equiv \{t : F_\star(t) = A^*(t)\}$. Any $t' \in T_A$ either is one of the finite number of points of nondifferentiability of A^* or belongs to one of the finite number of intervals of hollowness of A^* by properties of PWH functions. If $t', t'' \in T_A$ are in the *same* interval of hollowness, then

(i) A is concave in this interval;

(ii) $[t', t''] \subset T_A$.

If $t' \in T_A$ is interior to an interval of convexity of A, we can modify $F_\star$ into the straight line between two points on opposite sides of t', keeping above A (by convexity) and below B, thereby producing a shorter path, contrary to the definition of $F_\star$. So (i) holds, while (ii) follows obviously from the above remark. We conclude that T_A consists of a finite number of isolated points (which are degenerate intervals) and a finite number of disjoint intervals each a subinterval of an interval of concavity of A.

Next define $T_B \equiv \{t : F_\star(t) = B^*(t)\}$. Since $A \prec B$, $T_A \cap T_B = \{0,1\}$. Similar reasoning shows that T_B consists of a finite number of isolated points and a finite number of disjoint intervals, each a subinterval of an interval of convexity of B.

The set $T \equiv [0,1] - T_A - T_B$, where $A < F_\star < B$, is made up of a finite number of disjoint intervals (which may have end points in common). Since $F_\star$ is effectively unconstrained over each of these intervals, it must be a straight line. This proves (a)–(c) and hence the lemma.

It is clear that a function with the properties deduced in Lemma 2 is in $\mathcal{F}$. So $F_\star$ is *a* solution to the variational problem; its uniqueness follows easily since ϕ is strictly convex here. Hence we may now write $S = F_\star$.

THEOREM 3. *If A, B are PWH and $A \prec B$,*

$$\inf_{F \in \mathcal{F}} \Gamma(F) = \Gamma(S) \tag{21}$$

for any convex ϕ. If ϕ is strictly convex, S is the unique minimum.

Proof (see Figure 2). Consider the partition $\{I_j\}$ in the preceding lemma for S, and combine adjacent I's if necessary to produce a new partition $\{J_j\}$ with the following properties. Over J_1, $S \neq$ one boundary—say, B—without loss of generality, and $S = A$ at each end of J_1. Over J_2, $A < S < B$, so S is a straight line. Over J_3, $S \neq A$ and $S = B$ at each end of J_3. Over J_4, $A < S < B$, and so on.

Now choose any other $F \in \mathcal{F}$. At the right-hand end of J_1, $F \geq A = S$, while at the left end of J_3, $F \leq B = S$. So since F and S are continuous over J_2, there is at least one point in J_2—t_2, say—where $S = F$. Then over the interval $[0, t_2]$ we have a one-boundary problem like that of Theorem 1, because we have ensured that both F

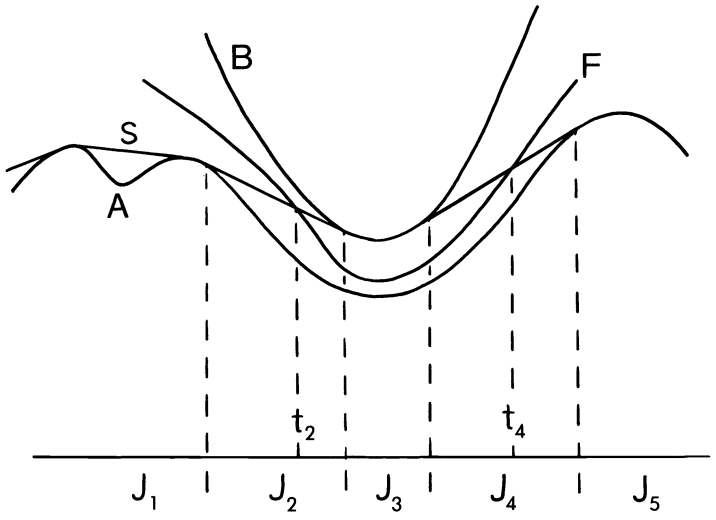


FIG. 2.

and S have the same end points. By Theorem 1, then,

$$\int_0^{t_2} \phi[F'(s)]ds \geq \int_0^{t_2} \phi[S'(s)]ds \tag{22}$$

for any convex ϕ, since we have already deduced that $S = A^*_\cap$ over this interval.

Similarly, there will be a t_4 in J_4 where $S = F$, and then we have a one-boundary problem like that of Theorem 2 over $[t_2, t_4]$. So

$$\int_{t_2}^{t_4} \phi[F'(s)]ds \geq \int_{t_2}^{t_4} \phi[S'(s)]ds. \tag{23}$$

We can continue to produce a finite number of such intervals $[t_{2n}, t_{2n+2}]$ over which a result like (23) holds, finishing with an interval $[t_{2n}, 1]$. Combining (22) and all those other inequalities proves (21), since F is arbitrary. The proof of uniquness follows that in Theorem 1.

The following functions provide alternative characterizations of S, more analogous to $A^*_\cap$ and $B^*_\cup$.

DEFINITION. *$V(t)$ is the* minimal *function that satisfies $A^* \leq V \leq B^*$ on $[0,1]$ and is* concave *on any interval where it is not equal to B^*.*

DEFINITION. *$U(t)$ is the* maximal *function that satisfies $A^* \leq U \leq B^*$ on $[0,1]$ and is* convex *on any interval where it is not equal to A^*.*

THEOREM 4. *Suppose A, B are PWH and $A \prec B$. Then $S = V = U$.*

Proof. From Lemma 2, S is concave when it is not equal to B^*, and clearly $A^* \leq S \leq B^*$. So $V \leq S$. Now it is an easy deduction from Theorem 1 that $S = A^*_\cap$ over the intervals $J_1, J_5, \ldots$ defined in the proof of Theorem 3, so $S = V$ over these intervals. Again it follows from Lemma 2 that S is convex over the interval $J_2 \cup J_3 \cup J_4$ while $V(\leq S)$ is concave unless it is equal to B. But we have just shown that $S = V$ at each end of this interval, so $S = V$ throughout. Hence $S = V$ on $[0,1]$. The proof that $S = U$ is similar.

4. Extensions. An extension which sometimes proves useful is the relaxation of the condition $A \prec B$. Without this condition, $F_\star \in \mathcal{F}$ may not hold, as is clear from inspection of Figure 3. However, we can proceed as follows.

Suppose that we allow $A = B$ over a finite number of intervals (which can degenerate into points) or require only that $A(t-0) \leq B(t-0)$ and $A(t+0) \leq B(t+0)$ at jump points. These restrictions mean that if A, B have a jump-point in common the jumps can partially overlap or even be exactly equal (cf. Figure 3). In this case the

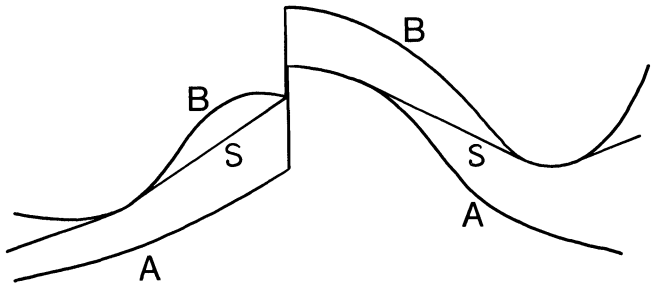


FIG. 3.

string function $S(t)$ is still formally the unique solution, provided that we recognize that it may be completely constrained at points where $A = B$ or where jumps overlap. A proof proceeds by breaking up the problem into a finite number of problems where these constraints operate only at the ends of intervals as before. For example, if $A \equiv B$ over $[a, b] \subset [0, 1]$ but nowhere else, we get the solution as $S(t)$ between P_0 and $(a, A(a))$ and between $(b, A(b))$ and P_1, and $A(t)$ over $[a, b]$, provided that A and B have the necessary properties outside $[a, b]$.

Acknowledgments. We are grateful to colleagues at Aurora Q1 Vehicles, the University of South Australia, and CSIRO for helpful discussions and to Aurora Q1 Vehicles for financial support. We thank the referees and an associate editor for constructive comments.

REFERENCES

[1] G. W. EWING (1985), *Calculus of Variations with Applications*, Dover, New York.
[2] G. H. HARDY, J. E. LITTLEWOOD, AND G. PÓLYA (1929), *Some simple inequalities satisfied by convex functions*, Messenger of Math., 58, pp. 145–152.
[3] ——— (1952), *Inequalities*, 2nd ed., Cambridge University Press, Cambridge, UK.
[4] J. L. LEBOWITZ AND O. PENROSE (1966), *Rigorous treatment of the Van Der Waals–Maxwell theory of the liquid-vapour transition*, J. Math. Phys., 7, pp. 98–113.
[5] A. W. MARSHALL AND I. OLKIN (1979), *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York.
[6] C. R. RAO (1973), *Linear Statistical Inference and Its Applications*, 2nd ed., Wiley, New York.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 437–472, March 1996

003

A DYNAMICAL SYSTEM APPROACH TO STOCHASTIC APPROXIMATIONS*

MICHEL BENAIM†

Abstract. It is known that some problems of almost sure convergence for stochastic approximation processes can be analyzed via an ordinary differential equation (ODE) obtained by suitable averaging. The goal of this paper is to show that the asymptotic behavior of such a process can be related to the asymptotic behavior of the ODE without any particular assumption concerning the dynamics of this ODE. The main results are as follows: a) The limit sets of trajectory solutions to the stochastic approximation recursion are, under classical assumptions, almost surely nonempty compact connected sets invariant under the flow of the ODE and contained in its set of chain-recurrence. b) If the gain parameter goes to zero at a suitable rate depending on the *expansion rate* of the ODE, any trajectory solution to the recursion is almost surely asymptotic to a forward trajectory solution to the ODE.

Key words. stochastic approximations, ordinary differential equations, chain-recurrence, neural networks

AMS subject classifications. 62L20, 34D05, 34C29

Introduction. The classical theory of stochastic approximations, born with the papers of Robbins and Monro (1951) and Kiefer and Wolfowitz (1952), concerns the study of stochastic algorithms whose general form can be written as

$$w_{n+1} - w_n = \gamma_n H(w_n, \xi_n), \tag{1}$$

where $H : \mathbf{R}^m \times \mathbf{R}^d \mapsto \mathbf{R}^m$ is a measurable function that characterizes the algorithm, $\{w_n\}_{n\geq 0} \in \mathbf{R}^m$ is the sequence of parameters to be recursively updated, $\{\xi_n\}_{n\geq 0} \in \mathbf{R}^d$ is a sequence of random inputs where $H(w_n, \xi_n)$ is observable, and $\{\gamma_n\}_{n\geq 0}$ is a sequence of "small" nonnegative scalar gains.

At each time step, the vector ξ_n is a new observation that causes w_n to be updated to take new information into account. The gain sequence $\{\gamma_n\}_{n\geq 0}$ can be chosen to be constant or decreasing. In this paper we restrict attention to algorithms with *decreasing gain sequence.* More precisely, we shall always assume that $\{\gamma_n\}_{n\geq 0}$ is a decreasing sequence of positive numbers which satisfies the classical relations

$$\lim_{n\to\infty} \gamma_n = 0$$

and

$$\sum_{n\geq 0} \gamma_n = +\infty.$$

To analyze the asymptotic behavior of the algorithm (1) it is convenient to introduce the averaged ordinary differential equation (ODE)

$$\frac{dw}{dt} = \overline{H}(w), \tag{2}$$

* Received by the editors August 9, 1993; accepted for publication (in revised form) October 18, 1994. This research was supported by a grant from the Centre National de la Recherche Scientifique (Programme Cogniscience).

† Department of Mathematics, University of California at Berkeley, Berkeley, CA 94720.

where

$$\overline{H}(w) = \lim_{n\to\infty} E(H(w, \xi_n))$$

and $E(.)$ denotes the mathematical expectation.

This method, called the *method of ordinary differential equation*, was introduced by Ljung (1977) and Kushner and Clark (1978) and widely studied thereafter. It has inspired a number of important works, such as the book by Kushner and Clark (1978), numerous articles by Kushner, and, more recently, the book by Benveniste, Métivier, and Priouret (1990). The main idea of the method is to describe the asymptotic behavior of the algorithm in terms of the behavior of the ODE. For stochastic algorithms having a decreasing gain sequence, the classical result stating the relationship between the algorithm (1) and the ODE (2) has the following form:

> *Let $w*$ be a stable equilibrium for the ODE. If $\{\gamma_n\}_{n\geq 0}$ goes to zero at a suitable rate and if the sequence $\{w_n\}_{n\geq 0}$ enters infinitely often a compact subset of the domain of attraction of $w*$, then $\{w_n\}_{n\geq 0}$ converges almost surely toward w^*.*

This kind of result has been obtained by Ljung (1977); Kushner and Clark (1978); Métivier and Priouret (1984, 1987); Benveniste, Métivier, and Priouret (1990); and Kuan and White (1992), among others, under fairly general conditions. It relies the asymptotic behavior of the algorithm with a strong notion of recurrence for the ODE: the notion of *fixed point.*

With increasing interest in *artificial neural networks* and due to some limitations of the standard *backpropagation* algorithm, "heuristic" learning rules for feedforward neural networks have been recently proposed and experimentally studied. The ODE associated with these algorithms is not given by a gradient vectorfield (as is the case for backpropagation), and the classical convergence results on stochastic gradient algorithms cannot be successfully applied. The consideration of these algorithms led us to formulate the following problem:

> *Without any particular assumption on the dynamics of $\overline{H}$, is it again possible to describe the asymptotic behavior of* (1) *in terms of the asymptotic behavior of* (2)*?*

The main goal of this paper is to address this question.

In §§1 and 2 we relate the behavior of the algorithm to a weak notion of recurrence for the ODE: the notion of *chain recurrence.* We state a theorem which asserts that under the assumptions of the Kushner and Clark lemma (1978) the limit sets of the trajectory solutions to (1) are nonempty compact connected sets invariant under the flow of the ODE and contained in its set of chain-recurrence.

This result shows that the limit sets of (1) look like the omega limit sets of (2), and we ask the question of their exact relationship. We address this question in §5. It is shown that it may happen that the limit set of a trajectory solution to (1) never coincides with an omega limit set of (2), but that it always does if the gain parameter goes to zero at a suitable rate depending on the vectorfield $\overline{H}$. Our approach, in this section, is essentially based on "shadowing" results recently proved by Morris W. Hirsch together with L^q estimates of the distance between the trajectory solutions to (1) and (2).

In §8 we apply the results of §§1–5 to prove some convergence theorems for the neural network learning algorithms mentioned above.

Main theorems are proved in §§4 and 7. Several applications are considered in §§3 and 6.

1. A deterministic theorem. In order to introduce the main result of this section we begin with a few notations and classical definitions from dynamical systems.

Notation and definitions. Let Γ be a topological space and $\Phi : \mathbf{R} \times \Gamma \mapsto \Gamma$ be a continous map denoted by $\Phi(t, x) = \Phi_t(x)$. The family $\{\Phi_t\}_{t \in \mathbf{R}}$ is called a *flow* on Γ if it satisfies the group property

$$\Phi_0 = \text{Identity},$$

$$\forall (t, s) \in \mathbf{R}^2,\ \Phi_t \circ \Phi_s = \Phi_{t+s}.$$

Let $\overline{H}$ denote a continous vectorfield defined on $\mathbf{R}^m$ with unique integral curves. The *flow* of $\overline{H}$ is the family of mappings defined on $\Gamma = \mathbf{R}^m$ by

$$\frac{d}{dt}\Phi_t(w) = \overline{H}(\Phi_t(w)).$$

A set X is said to be *invariant* (respectively, *positively invariant*) under the flow Φ if for all $t \in \mathbf{R}, \Phi_t(X) \subset X$ (respectively, for all $t \geq 0$). In this case we let $\Phi|X$ denote the restricted flow (respectively, semiflow).

A point x is an *equilibrium* if $\Phi_t(x) = x$ for all $t \in \mathbf{R}$. When Φ is induced by the vectorfield $\overline{H}$, equilibria coincide with zeros of $\overline{H}$. A point x is a *periodic point* if there exists $T > 0$ such that $\Phi_T(x) = x$. Equilibria and periodic points are clearly recurrent points. In general, we may say that a point is recurrent if it somehow returns near where it was under time evolution.

A notion of recurrence related to slightly perturbed orbits is the notion of *chain recurrence*. Suppose Γ is a metric space with a metric d. Let $\delta > 0$ and $T > 0$. A point x is said to be (δ, T) *recurrent* if there exist an integer k, some points y_i in Γ, and numbers $t_i, 0 \leq i \leq k-1$, such that

$$t_i \geq T; \quad d(y_0, x) < \delta; \quad d(\Phi_{t_i}(y_i),\ y_{i+1}) < \delta \quad \text{for } i = 0, \ldots, k-1; \quad x = y_k.$$

Intuitively (δ, T) recurrent points are points that one would take to be periodic if the position of points were only known with a finite accuracy δ. If x is (δ, T) recurrent for any $\delta > 0$ and $T > 0$, x is said to be *chain-recurrent*. We denote by $CR(\Phi)$ the set of chain-recurrent points. If Φ is induced by the vectorfield $\overline{H}$, we may also use the notation $CR(\overline{H})$ for $CR(\Phi)$. The set $CR(\Phi)$ has the property to be closed and invariant.

A subset $X \subset \Gamma$ is said *internally chain-recurrent* if X is a nonempty compact invariant set of which every point is chain-recurrent for the restricted flow $\Phi|X$ (i.e., $CR(\Phi|X) = X$).

For example, if Γ is compact, Conley (1978) proved that $CR(\Phi)$ is internally chain-recurrent.

The sets which describe the asymptotic behavior of the orbits of the flow Φ are the *omega limit sets*. The omega limit set of $w \in \Gamma$, denoted by $\omega(w)$, is the set of $x \in \Gamma$ such that $\lim_{k \to \infty} \Phi_{t_k}(w) = x$ for some sequence $t_k > 0$ with $\lim_{k \to \infty} t_k = +\infty$. If the forward trajectory $\{\Phi_t(w); t \geq 0\}$ has compact closure, $\omega(w)$ is a nonempty compact connected set internally chain-recurrent. The alpha limit set $\alpha(w)$ of w is defined as the omega limit set of w for the reversed flow $\{\Phi_{-t}\}_{t \geq 0}$.

To recapitulate, if we note $Per(\Phi)$ the set of periodic points (including the equilibria) and $\mathcal{L}^+(\Phi) = \bigcup_{w \in \Gamma} \omega(w)$, the following inclusions hold:

$$Per(\Phi) \subset \mathcal{L}^+(\Phi) \subset CR(\Phi).$$

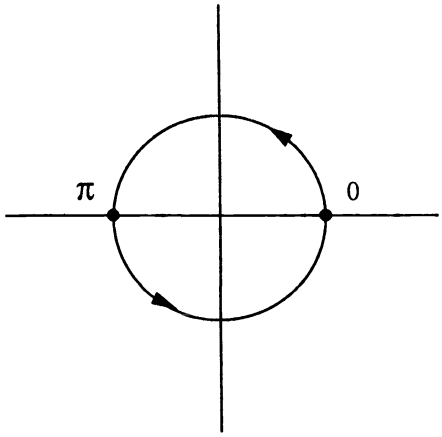


FIG. 1.

EXAMPLE 1.1. *Consider the flow on the unit circle* $S^1 = \mathbf{R}/2\pi\mathbf{Z}$ *induced by the differential equation*

$$\frac{d\theta}{dt} = f(\theta),$$

where f *is a* 2π*-periodic smooth nonnegative function such that*

$$f^{-1}(0) = \{k\pi : k \in \mathbf{Z}\}.$$

See Fig. 1.
We have

$$Per(\Phi) = \{0, \pi\} = \mathcal{L}^+(\Phi)$$

and

$$CR(\Phi) = S^1.$$

Internally chain-recurrent sets are $\{0\}, \{\pi\}$, *and* S^1. *Note that the set* $X = [0, \pi]$ *is a compact invariant set consisting of chain-recurrent points. However,* X *is not internally chain-recurrent.*

A deterministic theorem. To describe the asymptotic behavior of the algorithm (1) we introduce the limit set of the sequence $\{w_n\}_{n\geq 0}$. We denote this limit set by $L(\{w_n\}_{n\geq 0})$. It is the set of $x \in \mathbf{R}^m$ such that $\lim_{k\to\infty} w_{n_k} = x$ for some subsequence $\{n_k\}_{k\geq 0}$ with $\lim_{k\to\infty} n_k = +\infty$.

The following theorem is a deterministic result that will be applied in §2 to show that the limit sets of the trajectories solutions to the algorithm (1) have basically the same properties as the omega limit sets of the trajectories solution to the ODE (2). The assumptions A1, A2, and A3 of this theorem are the assumptions of the Kushner and Clark lemma (1978).

We use the following notation:

$$\tau_0 = 0,$$

$$\tau_n = \sum_{i=0}^{n-1} \gamma_i.$$

We let $\|.\|$ denote a norm on $\mathbf{R}^m$.

THEOREM 1.2. *Let $\overline{H} : \mathbf{R}^m \mapsto \mathbf{R}^m$ be a continuous vectorfield with unique integral curves. Let $\{w_n\}_{n\geq 0}$ be solution to the recursion*

$$w_{n+1} - w_n = \gamma_n(\overline{H}(w_n) + u_n + b_n), \tag{3}$$

where $\{\gamma_n\}_{n\geq 0}$ is a decreasing gain sequence. Assume that

- A1) *$\{w_n\}_{n\geq 0}$ is bounded.*
- A2) *$\lim_{n\to\infty} b_n = 0$.*
- A3) *For each $T > 0$,*

$$\lim_{n\to\infty} \left(\sup_{\{k\,;\,0\leq \tau_k - \tau_n \leq T\}} \left\| \sum_{i=n}^{k-1} \gamma_i . u_i \right\| \right) = 0.$$

Then $L(\{w_n\}_{n\geq 0})$ is a connected set internally chain-recurrent for the flow Φ induced by $\overline{H}$.

The next theorem shows that Theorem 1.2 gives the best result that can be expected under the Kushner and Clark assumptions. It justifies the fact that the chain recurrence is a notion well suited to the description of the asymptotic behavior of (1).

Assume given a locally Lipschitz vectorfield $\overline{H} : \mathbf{R}^m \mapsto \mathbf{R}^m$ and a decreasing gain sequence $\{\gamma_n\}_{n\geq 0}$.

THEOREM 1.3. *Let $L \subset \mathbf{R}^m$ be a connected set internally chain-recurrent for the flow induced by $\overline{H}$. There exist sequences $\{b_n\}_{n\geq 0}$, $\{u_n\}_{n\geq 0}$, and $\{w_n\}_{n\geq 0}$ such that*

(a) *Conditions* A1, A2, *and* A3 *of Theorem* 1.2 *are satisfied.*

(b) *The sequence $\{w_n\}_{n\geq 0}$ is the solution to* (3) *and admits L as a limit set.*

Theorem 1.3 follows easily from the following proposition (see Benaim and Hirsch (1995b)).

PROPOSITION 1.4. *Let $L \subset \mathbf{R}^m$ be a connected set internally chain-recurrent for the flow induced by $\overline{H}$. There exists a continuous function $u : \mathbf{R}_+ \mapsto \mathbf{R}^m$ and a point $w_0 \in \mathbf{R}^m$ such that*

(a) *$\lim_{t\to\infty} u(t) = 0$.*

(b) *The solution to the nonautonomous system*

$$\frac{dw}{dt} = \overline{H}(w) + u(t)$$

with initial condition $w(0) = w_0$ is bounded and admits L as a limit set.

To prove Theorem 1.3 we let $w_n = w(\tau_n)$ and $u_n = u(\tau_n)$, where $w(.)$ and $u(.)$ are the functions of Proposition 1.4. Then we have

$$w_{n+1} - w_n = \gamma_n(\overline{H}(w_n) + u_n) + O(\gamma_n^2)$$

and Theorem 1.3 follows from Proposition 1.4.

REMARK 1.5. *Throughout this paper the process $\{w_n\}_{n\geq 0}$ will be assumed to be bounded. Several conditions ensuring that this assumption is fulfilled are discussed in the literature on stochastic approximations. They usually rely on the existence of*

some convergent supermartingale for the process (1) (*see, e.g., Theorem* 5.2, *chapter* 2, *of Nevel'son and Has'minskii* (1974) *or Theorem* 8 *of Fort and Pagès* (1994)).

In the spirit of this section, we give a simple condition which is purely deterministic.

PROPOSITION 1.6. *Assume that* $\overline{H}$ *is globally Lipschitz. Assume the existence of a function* $V : \mathbf{R}^m \to \mathbf{R}_+$ *uniformly continuous such that*

(i) $\lim_{||x||\to\infty} V(x) = \infty$.

(ii) *There exist positive numbers* δ, r, *and* T *such that*

$$\forall x \in \mathbf{R}^m,\ ||x|| \geq r \Longrightarrow V(\Phi_T(x)) - V(x) \leq -\delta.$$

Then conditions A2 *and* A3 *of Theorem* 1.2 *imply condition* A1.

The proof of this result follows easily from Lemma 4.4 and is left to the reader.

Note that if V is smooth, condition (ii) holds if the following more easily checked condition is satisfied: there exists $\delta' > 0$ such that for all $||x|| \geq r$

$$\langle \nabla V(x), \overline{H}(x) \rangle \leq -\delta',$$

where ∇ denotes the gradient.

2. Limit sets of stochastic approximation processes. In this section, we assume that $\{\xi_n\}_{n\geq 0}$ is a sequence of $\mathbf{R}^d$-valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$. We note $\mathcal{F}_n^m$, the σ field generated by $\{\xi_i;\ n \leq i \leq m\}$ for $m \geq n$. For $q \in [1, \infty]$ we let $\|.\|_q$ denote the $L^q(\Omega)$ norm for random variables ($\|X\|_q = E(\|X\|^q)^{1/q}$) and $\|.\|_\infty$ the $L^\infty(\Omega)$ norm ($\|X\|_\infty = \operatorname{ess\,sup}\|X\|$).

In applications of Theorem 1.2 to the stochastic approximation (1) one may choose

$$\overline{H}(w) = \lim_{n\to\infty} E(H(w, \xi_n)),$$

$$u_n = H(w_n, \xi_n) - \int H(w_n, \xi)\mu_n(d\xi),$$

and

$$b_n = \int H(w_n, \xi)\mu_n(d\xi) - \overline{H}(w_n),$$

where μ_n is the distribution of ξ_n. Then we try to verify assumptions A2 and A3 by use of some regularity properties of H and maximal inequalities for sum of random variables. Let us mention two examples.

Independent inputs. The first example is a classical Robbins–Monro algorithm in which the observations are assumed to be independent and identically distributed. This yields a simple martingale access to condition A3 as in Gladyshev (1965) and Hall and Heyde (1980).

We let M denote a given subset of $\mathbf{R}^m$ (not necessarily compact).

PROPOSITION 2.1. *Let* $\{w_n\}_{n\geq 0}$ *be the solution to* (1). *Assume that*

- A1) $\{\xi_n\}_{n\geq 0}$ *is a sequence of independent and identically distributed random variables.*
- A2) $P(\{w_n\}_{n\geq 0}$ *is bounded*$) = 1$ *and* $P(\forall n \in N, w_n \in M) = 1$.
- A3) $w \mapsto \overline{H}(w) = E(H(w, \xi_0))$ *is continuous with a unique flow.*

There exists $q \geq 2$ *such that*

- A4) $w \mapsto \|H(w,\xi_0)\|_q$ *is bounded on* M.
- A5) $\sum_{n=0}^{\infty} \gamma_n^{1+q/2} < +\infty$.

Then the conclusions of Theorem 1.2 *hold with probability one.*

Proof. To see that, we let $b_n = 0$ and $u_n = H(w_n,\xi_n) - \overline{H}(w_n)$. Then $E(u_n/\mathcal{F}_0^{n-1}) = 0$. For $q = 2$, assumptions A4 and A5 imply $\sum_n \gamma_n^2 . \|u_n\|_2^2 < +\infty$, and condition A3 of Theorem 1.2 is a direct consequence of the L^2-bounded martingale convergence theorem. For $q > 2$, it follows from a result of Métivier and Priouret (1987, Cor. 11). (See also Benveniste, Métivier, and Priouret (1990, Cor. 8, p. 297).) Note that in this case the sequence $\{\sum_n \gamma_n . u_n\}_{n\geq 0}$ is not necessarily convergent. □

Mixing inputs. The following example extends this result to situations in which the observable inputs are nonindependent and nonstationary random variables which satisfy a strong mixing condition. Such situations arise naturally in some applications of feedforward neural networks as forecasting, prediction of time series, or chaos modelling.

Here our approach is motivated by the work of Kuan and White (1992), who have proved some convergence results for stochastic approximation procedures by using the theory of *mixingales* developed by McLeish (1975). Conditions A1–A6 can be compared with conditions of Kuan and White's theorems (Thm. 2.2.1 and Cors. 2.2.3 and 2.3.5). The condition A6′ gives a generalization which allows a gain parameter of the order of $\frac{1}{n^\alpha}$ with $\alpha < 1$. The price for this is a strengthening of the boundness condition.

For $n \geq 0, m \geq 0$ define

$$\phi_{n,m} = \sup_{\{A\in\mathcal{F}_0^n, B\in\mathcal{F}_{n+m}^{n+m}\}} | P(B/A) - P(B) |,$$

$$\alpha_{n,m} = \sup_{\{A\in\mathcal{F}_0^n, B\in\mathcal{F}_{n+m}^{n+m}\}} | P(B\cap A) - P(B)P(A) |,$$

$$\phi_m = \sup_{n\geq 0} \phi_{n,m},$$

$$\alpha_m = \sup_{n\geq 0} \alpha_{n,m}.$$

We shall say that the process $\{\xi_n\}_{n\geq 0}$ is ϕ mixing (respectively, α mixing) if $\lim_{n\to\infty} \phi_n = 0$ (respectively, $\lim_{n\to\infty} \alpha_n = 0$). Observe, however, that this condition is a weakening of the classical ϕ mixing (respectively, α mixing) definition (see, for instance, Billingsley (1968, §20, p. 166)). It would be the same if $\mathcal{F}_{n+m}^{n+m}$ were replaced by $\mathcal{F}_{n+m}^{+\infty}$. This weaker definition is motivated by our use of McLeish's results (1975).

PROPOSITION 2.2. *Let* $\{w_n\}_{n\geq 0}$ *be the solution to* (1). *Assume that*

- A1) $\{\xi_n\}_{n\geq 0}$ *is a* ϕ *mixing (respectively,* α *mixing) process.*
- A2) $\{w_n\}_{n\geq 0}$ *is bounded with probability one.*
- A3) $\overline{H}(w) = \lim_{n\to\infty} E(H(w,\xi_n))$ *exists.*
- A4) *There exists a measurable function* $k(.)$ *such that*

$$\forall x, y \in \mathbf{R}^m \|H(x,\xi) - H(y,\xi)\| \leq k(\xi)\|x - y\|.$$

There exists $r \in [2,\infty]$ *such that*

- A5) *The map* $w \mapsto \sup_{n\geq 0} \|H(w,\xi_n)\|_r$ *is bounded on any bounded set and* $\sup_{n\geq 0} \|k(\xi_n)\|_r < +\infty$.
- A6) *(L^r case). If* $r < \infty$, $\phi_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{r}{2r-2}$ *(respectively,* $\alpha_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{r}{r-2}$*) and*

$$\sum_{n=0}^{\infty} \gamma_n^2 < +\infty.$$

- A6′) *(L^∞ case). If* $r = \infty$, $\phi_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{1}{2}$ *(respectively,* $\alpha_n = O(\frac{1}{n^\beta})$ *for some* $\beta > 1$*) and*

$$\sum_{n=0}^{\infty} \gamma_n^{1+q/2} < +\infty$$

for some $q \in [2, 2\beta + 1[$.

Then the conclusions of Theorem 1.2 *hold with probability one.*

The proof of this result is given in the appendix (§9).

In view of the fact that the assumptions of Theorem 1.2 are the assumptions of the Kushner and Clark lemma, several other examples of application can be found in the literature. We refer the reader to the book by Kushner and Clark (1978, Chap. II) for such examples. In the case where the input process $\{\xi_n\}_{n\geq 0}$ is a Markov process or, more generally, a Markov process controlled by the parameter w, condition A3 of Theorem 1.2 can be derived from the analysis provided in the articles by Ljung (1977) and Métivier and Priouret (1987, Cor. 11) (see also Benveniste, Métivier, and Priouret (1990, Cor. 8, p. 297)).

3. Applications. In this section we give a few examples to illustrate how results of §§1 and 2 can be used to describe the global asymptotic behavior of stochastic approximation processes.

In the remainder of this section $\overline{H}$ is a vectorfield on $\mathbf{R}^m$ with unique integral curves. The sequence $\{w_n\}_{n\geq 0}$ denotes either a deterministic sequence solution to (3) under assumptions of Theorem 1.2 or a random sequence solution to (1) under assumptions of Proposition 2.1 or 2.2. In this last case, all the properties stated below have to be understood as "almost sure" properties.

Local behavior. First, note that Theorem 1.2 generalizes the classical result mentioned in the introduction.

An equilibrium w^* of $\overline{H}$ is said *asymptotically stable* if there exists an open neighborhood U of w^* such that

$$\lim_{t\to\infty} \Phi_t(w) = w^*$$

uniformly in $w \in U$. The *domain of attraction* of w^* is the set of all points whose forward trajectories are attracted by w^*.

PROPOSITION 3.1. *Let* w^* *be an asymptotically stable equilibrium of* $\overline{H}$. *Assume that* $\{w_n\}_{n\geq 0}$ *enters infinitely often a compact subset—say,* Q*—of the domain of attraction of* w^*. *Then*

$$\lim_{n\to\infty} w_n = w^*.$$

Proof. According to Theorem 1.2, $L(\{w_n\}_{n\geq 0}) \cap Q$ is nonempty and is contained in $CR(\overline{H}) \cap Q$. On the other hand, it is not difficult[1] to show that $CR(\overline{H}) \cap Q = \{w^*\}$. Thus $\{w^*\} = L(\{w_n\}_{n\geq 0}) \cap Q$ and, as $L(\{w_n\}_{n\geq 0})$ is connected, $L(\{w_n\}_{n\geq 0}) = \{w^*\}$. □

Gradientlike systems. Let Φ be a flow on a metric space Γ and $\Lambda \subset \Gamma$ be an invariant set.

A C^0 map $V : \Gamma \mapsto \mathbf{R}$ is said to be a *Lyapunov function* for Λ if for all $x \in \Gamma$ the function $t \in \mathbf{R}_+ \mapsto V(\Phi_t(x))$ is constant for $x \in \Lambda$ and strictly decreasing for $x \notin \Lambda$.

If Λ equals the equilibria set, V is called a *strict Lyapunov function* and Φ is called a *gradientlike system.*

PROPOSITION 3.2. *Assume that Γ is compact. Let $\Lambda \subset \Gamma$ be a compact invariant set and $V : \Gamma \mapsto \mathbf{R}$ a Lyapounov function for Λ. Assume that the cardinal of $V(\Lambda)$ is finite. Then*

$$CR(\Phi) \subset \Lambda.$$

COROLLARY 3.3. *Assume that $\overline{H}$ admits a strict Lyapunov function and isolated equilibria. Then $\{w_n\}_{n\geq 0}$ converges toward an equilibrium.*

Proof. We apply Proposition 3.2 to the flow induced by $\overline{H}$ on $\Gamma = L(\{w_n\}_{n\geq 0})$. It follows from Theorem 1.2 that $L(\{w_n\}_{n\geq 0})$ consists of equilibria. As it is a connected set and equilibria are isolated, $L(\{w_n\}_{n\geq 0})$ is an equilibrium. □

REMARK 3.4. *Note that Corollary 3.3 applies to stochastic gradient algorithms for which $\overline{H}$ is the gradient of a cost function $C : \mathbf{R}^m \mapsto \mathbf{R}$,*

$$\overline{H}(w) = \nabla C(w).$$

In §8 we will give another application of Proposition 3.2 to a class of learning processes which are not given by a stochastic gradient.

Proof of Proposition 3.2. Let $V(\Lambda) = \{v_1, \ldots, v_l\}, v_1 < v_2 < \cdots < v_l$. Choose real numbers $v_1', v_2', \ldots, v_l'$ such that $v_1 < v_1' < v_2 < \cdots < v_{l-1}' < v_l < v_l'$, and define $M_i = \{x \in \Gamma / V(x) \leq v_i'\}$.

Let $\Lambda_i = \Lambda \cap V^{-1}(v_i)$; Λ_i is the largest invariant set contained in $M_i - M_{i-1}$. Indeed, let $A \subset M_i - M_{i-1}$ be an invariant set and let $x \in A$. By a standard theorem on Lyapunov functions, $\alpha(x) \cup \omega(x) \subset \Lambda$. So $V(\alpha(x)) = V(\omega(x)) = v_i$, and as V is strictly decreasing along any trajectory outside Λ, x is necessarily in Λ_i.

Let $T > 0$. By compactness of the sets M_j, there exists $\epsilon > 0$ such that

$$\forall x \in M_j, V(\Phi_T(x)) \leq v_j' - \epsilon.$$

Pick $\delta > 0$ such that

$$\forall (x, y) \in \Gamma \times \Gamma, d(x, y) \leq \delta \Rightarrow | V(x) - V(y) | \leq \epsilon.$$

It follows that any (δ, T) chain $\{y_0, y_1, \ldots, y_k\}$ (i.e., $d(\Phi_{t_i}(y_i), y_{i+1}) < \delta$ for some $t_i \geq T$) with $y_0 \in M_j$ is included in M_j. Therefore, the set $CR_j = CR(\Phi) \cap (M_j - M_{j-1})$ is invariant. Hence, $CR_j \subset \Lambda_j$ and $CR(\Phi) \subset \Lambda$. □

[1] This follows, for example, from Proposition 3.10.

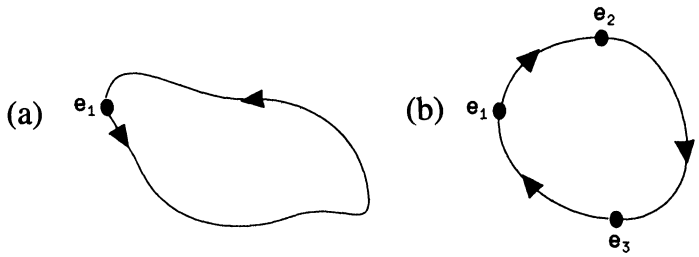


FIG. 2. (a) *A one-equilibrium cycle.* (b) *A three-equilibrium cycle.*

No-cycle systems. Let Φ be a flow on a metric space Γ. We say that Φ has *simple dynamics* if for every $x \in \Gamma$ the alpha and omega limit sets of x are equilibria. This means that every backward and forward trajectory converges toward an equilibrium. If Φ is induced by the vectorfield $\overline{H}$, we say that $\overline{H}$ has simple dynamics if $\Phi|L$ has simple dynamics for each compact invariant set $L \subset \mathbf{R}^m$.

For a flow with simple dynamics, we say that the equilibrium e_1 goes to the equilibrium e_2 if there exists a nonequilibrium orbit $\gamma \subset \Gamma$ such that $\alpha(\gamma) = e_1$ and $\omega(\gamma) = e_2$. γ is called a *connecting* orbit. To indicate that e_1 goes to e_2, we write $e_1 \rightsquigarrow e_2$. To indicate that γ is the connecting orbit from e_1 to e_2, we write $\gamma : e_1 \rightsquigarrow e_2$.

A *cycle of equilibria* is an union

$$A = \bigcup_{j=1}^{n} (\{e_j\} \cup \gamma_j)$$

consisting of equilibria e_j, $j = 1, \ldots, n$, and connecting orbits γ_j, $j = 1, \ldots, n$, such that

(i) $\gamma_j : e_j \rightsquigarrow e_{j+1}$, $j = 1, \ldots, n-1$.

(ii) $\gamma_n : e_n \rightsquigarrow e_1$.

REMARK 3.5. *A cycle of equilibria is connected internally chain-recurrent (Fig. 2).*

PROPOSITION 3.6. *Assume that Γ is compact and Φ has a finite number of equilibria, simple dynamics, and no cycle. Then $CR(\Phi)$ is the equilibria set.*

COROLLARY 3.7. *Assume that $\overline{H}$ has isolated equilibria, simple dynamics, and no cycle. Then $\{w_n\}_{n\geq 0}$ converges toward an equilibrium.*

Proof. We apply Proposition 3.6 to the flow induced by $\overline{H}$ on $\Gamma = L(\{w_n\}_{n\geq 0})$ and conclude exactly as in the proof of Corollary 3.3. □

Fort and Pagès (1994) recently proved a result similar to Corollary 3.7 by using the Kushner and Clark lemma. Systems with cycle of equilibria will be considered in §6.

The notion of *simple dynamics* and *no-cycle property* can be extended to nonconvergent situations. Denote by $\mathcal{L}(\Phi)$ the union of all alpha and omega limit sets of Φ. Assume that there exist nonempty compact disjoint invariant subsets $\Lambda_j \subset \Gamma$, $j = 1, \ldots, n$, such that

$$\mathcal{L}(\Phi) \subset \Lambda = \bigcup_{j=1}^{n} \Lambda_j.$$

If there exists $x \notin \Lambda$ such that $\alpha(x) \subset \Lambda_1$ and $\omega(x) \subset \Lambda_2$, we write $\Lambda_1 \rightsquigarrow \Lambda_2$ and define cycles among the Λ_j exactly as in the simple dynamics case.

PROPOSITION 3.8. *Assume* Γ *is compact and there is no cycle among the* Λ_j. *Then* $CR(\Phi) \subset \Lambda$.

Proof. Let $\hat{\Gamma}$ be the topological quotient space obtained by collapsing each Λ_i to a point. It is not difficult to check that $\hat{\Gamma}$ is a regular space with a countable basis. Therefore, by the Urysohn theorem, $\hat{\Gamma}$ is metrizable. Let π denotes the quotient map $\pi : \Gamma \to \hat{\Gamma}$. The flow Φ induces a flow $\hat{\Phi}$ on $\hat{\Gamma}$ defined by $\hat{\Phi} \circ \pi = \pi \circ \Phi$, which has simple dynamics, no cycle, and the Λ_j as equilibria. Therefore, by Proposition 3.6, chain-recurrent points of $\hat{\Phi}$ are equilibria. If $x \in \Gamma$ is chain-recurrent for Φ it is clear, by definition of chain-recurrence and uniform continuity of π, that $\pi(x)$ is chain-recurrent for $\hat{\Phi}$. Thus $CR(\Phi) \subset \Lambda$. □

COROLLARY 3.9. *Assume there exist nonempty compact disjoint subsets* $\Lambda_j \subset \mathbf{R}^m$, $j = 1, \ldots, n$, *invariant under the flow of* $\overline{H}$ *such that every alpha or omega limit point belongs to* $\Lambda = \bigcup_{j=1}^n \Lambda_i$. *Assume there is no cycle among the* Λ_j. *Then there exists* $j \in \{1, \ldots, n\}$ *such that* $L(\{w_n\}_{n \geq 0}) \subset \Lambda_j$.

Proof of Proposition 3.6. There are several ways to prove Proposition 3.6. For example it can be easily deduced from the "filtration theory" exposed in Shub (1986). Here, for simplicity we decided to deduce it from elementary properties of chain-recurrent sets. On the other hand these properties are very useful and give a good understanding of the notion of chain-recurrence.

Let $Y \subset \Gamma$. The forward trajectory of Y is the set $Y \cdot [0, \infty) = \Phi([0, \infty) \times Y) = \{\Phi_t(y);\ t \geq 0;\ y \in Y\}$.

The omega limit set (respectively, alpha limit set) of Y, denoted by $\omega(Y)$ (respectively, $\alpha(Y)$) is defined as the maximal invariant set in $\text{clos}(Y \cdot [0, \infty))$ (respectively, $\text{clos}(Y \cdot (-\infty, 0])$), where "clos" denotes closure.

A nonempty compact invariant set $A \subset X$ is an *attractor* if A has an open neighborhood U in X such that $\omega(U) = A$ or a *repeller* if $\alpha(U) = A$. An attractor or repeller is *proper* provided that it is not open in X.

The following proposition follows from §§5 and 6 of Conley (1978, Chap. 2).

PROPOSITION 3.10 (Conley (1978)).

(a) *Let* $N \subset \Gamma$ *be a compact set. Let* $A \subset \Gamma$ *be the maximal invariant set contained in* N. *If* A *is nonempty and* not *an attractor, there exists* $p \in \partial N \subset \Gamma$ *such that the backward orbit* $\gamma_-(p) \subset N$ *and* $\alpha(p)$ *is a nonempty subset of* A.

(b) *A internally chain-recurrent set has no proper attractor or repeller.*

(c) *The chain-recurrent set is internally chain-recurrent.*

Let us now prove Proposition 3.6. Let X be a connected component of $CR(\Phi)$. By assertion (c) of Proposition 3.10, X is internally chain-recurrent. Consider the flow $\Psi = \Phi | X$ and let $Equ(\Psi) = \{e_i,\ i = 1, \ldots, n\}$ denote the equilibria set of Ψ. Since Φ has simple dynamics and no cycle the relation $\rightsquigarrow$ induces a partial ordering on $Equ(\Psi)$.

Assume e_n is minimal for this partial ordering. We claim that e_n is an attractor for Ψ. It follows from assertion (b) of Proposition 3.10 that e_n is open and closed in X. Thus $X = \{e_n\}$.

It remains to prove that e_n is an attractor for Ψ. Let N be a compact neighborhood of e_n which separates e_n from other equilibria. The maximal invariant set in N is e_n; otherwise it would exist a entire orbit disjoint from e_n inside N. The dynamics being simple, this orbit would have to connect e_n to itself. Since we assume that there is no cycle, this is impossible.

Now we use assertion (a) of Proposition 3.10. If e_n is not an attractor, there

exists $p \in \partial N$ with $\alpha(p) = e_n$, but this contradicts the fact that e_n is minimal for the partial ordering $\rightsquigarrow$. □

Morse–Smale systems. In this subsection we mention briefly an application of the previous results to a class of stochastic approximation processes and urn models which have been recently considered by Benaim and Hirsch (1995a). For more details the reader is referred to that paper.

Assume $\overline{H}$ is C^r $(r \geq 1)$. $\overline{H}$ is called *Morse–Smale* if

(i) $\overline{H}$ has a global compact attractor (i.e., the point at infinity is a source);

(ii) all periodic orbits and equilibria are hyperbolic;

(iii) stable and unstable manifolds of periodic orbits (and equilibria) intersect only transversely;

(iv) every alpha or omega limit set is a periodic orbit or an equilibrium.

It is known that these conditions imply that there are only finitely many periodic orbits.

Suppose $\overline{H}$ is a Morse–Smale vector field. Denote by $\mathcal{L}(\overline{H})$ the union of all alpha and omega limit sets of $\overline{H}$, and by $Per(\overline{H})$ the union of all periodic orbits and equilibria. If $\overline{H}$ is Morse–Smale, $\mathcal{L}(\overline{H})$ decomposes as

$$\mathcal{L}(\overline{H}) = Per(\overline{H}) = \Lambda_1 \cup \cdots \cup \Lambda_n,$$

where the Λ_i are the distinct hyperbolic periodic orbits and equilibria. On the other hand, it follows from the transversal condition (iii) that there is no cycle among the Λ_i (see, e.g., Proposition 3.2 of Palis (1969)). Thus, we have the following corollary.

COROLLARY 3.11. *Assume $\overline{H}$ is Morse–Smale. Then $L(\{w_n\}_{n\geq 0})$ is an equilibrium or a periodic orbit.*

Proof. By Corollary 3.9, $L(\{w_n\}_{n\geq 0}) \subset \Lambda_i$ for some i. Since $L(\{w_n\}_{n\geq 0})$ is invariant and Λ_i is a periodic orbit or an equilibrium, we must have $L(\{w_n\}_{n\geq 0}) = \Lambda_i$. □

Nonconvergence toward unstable periodic orbits is considered in Benaim and Hirsch (1995a).

Planar systems. For planar systems it is possible to give a complete description of $L(\{w_n\}_{n\geq 0})$. A planar flow is a flow defined on a open subset of $\mathbf{R}^2$. The following theorem is proved in Benaim and Hirsch (1995c).

THEOREM 3.12. *Let Φ be a planar flow with isolated equilibria and L be an internally chain-recurrent set for Φ. Every point $x \in L$ satisfies one of the following conditions:*

(i) *x is an equilibrium.*

(ii) *x is a periodic point (i.e., x belongs to a periodic orbit).*

(iii) *There exists a cycle of equilibria in L which contains x.*

COROLLARY 3.13. *If $\overline{H}$ is a planar vectorfield with isolated equilibria, $L(\{w_n\}_{n\geq 0})$ is a connected union of equilibria, periodic orbits, and cycles of equilibria.*

Using the same kind of result, a Poincaré–Bendixson theorem for a class of stochastic differential equations is given in Benaim (1995b).

4. Proof of Theorem 1.2. We denote by 1_A the indicator function of the set A (i.e., $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ if $x \notin A$).

For any sequence $\{z_n\}_{n\geq 0} \in \mathbf{R}^m$ we denote by $Z(.)$ the function defined for all $t \geq 0$ by

$$Z(t) = \sum_{n\geq 0} z_n 1_{[\tau_n, \tau_{n+1}[}(t)$$

and by $Z^0(.)$ the interpolated process defined for all $t \geq 0$ by

$$Z^0(t) = \sum_{n\geq 0} \left[(z_{n+1} - z_n).\frac{(t-\tau_n)}{\gamma_n} + z_n \right] 1_{[\tau_n, \tau_{n+1}[}(t).$$

With these notations, the recursion satisfied by $\{w_n\}_{n\geq 0}$ can be rewritten as

$$W^0(t) - W^0(0) = \int_0^t \overline{H}(W(s))ds + \int_0^t U(s)ds + \int_0^t B(s)ds. \tag{4}$$

Remark that the assumptions A1, A2, and A3 are equivalent to

- A1′) $\{W^0(t), t \geq 0\}$ is bounded.
- A2′) $\lim_{t\to\infty} B(t) = 0$.
- A3′) For each $T > 0$,

$$\lim_{t\to\infty} \left(\sup_{h\in[0,T]} \left\| \int_t^{t+h} U(s)ds \right\| \right) = 0.$$

The function $t \mapsto W^0(t)$ is uniformly continous. This follows easily from the integral formula (4) and conditions A1′, A2′, A3′. This can also be deduced from the Kushner and Clark lemma (1978) (see Theorem 4.5).

We denote by $L(W^0)$ the limit set of $\{W^0(t), t \geq 0\}$ and let Q denote a compact subset of $\mathbf{R}^m$ which contains $\{W^0(t), t \geq 0\}$.

LEMMA 4.1. $L(\{w_n\}_{n\geq 0}) = L(W^0)$.

Proof. It is clear that $L(\{w_n\}_{n\geq 0}) \subset L(W^0)$. Conversely, let

$$w* = \lim_{t_k\to+\infty} W^0(t_k),$$

a limit point of W^0. Define the map $m : \mathbf{R}_+ \mapsto \mathbb{N}$ by

$$m(t) = \sup\{p \in \mathbb{N}/\tau_p \leq t\}. \tag{5}$$

One has $\lim_{t\to+\infty}(t - \tau_{m(t)}) = 0$ because $\lim_{n\to+\infty} \gamma_n = 0$. The uniform continuity of W^0 implies $\lim_{t_k\to+\infty} W^0(\tau_{m(t_k)}) = w*$. This proves the lemma. □

LEMMA 4.2. *For all* $T > 0$,

$$\lim_{t\to+\infty} \sup_{h\in[-T,T]} ||W^0(t+h) - \Phi_h(W^0(t))|| = 0.$$

For convenience, the proof of this lemma is postponed to the end of the section.

COROLLARY 4.3. $L(\{w_n\}_{n\geq 0})$ *is internally chain-recurrent.*

Proof. Since W^0 is continuous and bounded, $L(W^0)$ is a nonempty compact connected set.

Let us verify that $L(W^0)$ is invariant under Φ. Let $p \in L(W^0)$, $p = \lim_{t_i\to\infty} W^0(t_i)$ for some sequence $t_i \to \infty$. Let $T \in \mathbf{R}$. If $T > 0$, then

$$\lim_{t_i\to\infty} d(\Phi_T(W^0(t_i)), W^0(t_i + T)) = 0$$

by Lemma 4.2. Therefore $\Phi_T(p) = \lim_{t_i\to\infty} W^0(t_i + T) \in L(W^0)$. If $T < 0$, the proof is analogous.

It remains to prove that $L(W^0)$ is chain-recurrent for the restricted flow $\Phi|L(W^0)$. Here we adopt a method used by Robinson (1977) to show that a diffeomorphism on

a compact manifold is chain-recurrent on the set of chain-recurrence. Recall that $Q \subset \mathbf{R}^m$ denotes a compact set which contains $\{W^0(t), t \geq 0\}$.

Claim 1. Let $n \in \mathbf{N}$, $T > 0$, $p \in L(W^0)$. There exists a finite sequence

$$n \leq a_0^n \leq \cdots \leq a_{k(n)}^n$$

such that, with the notations

$$y_i^n = W^0(a_i^n), \qquad i = 0, \ldots, k(n),$$

and

$$t_i^n = a_{i+1}^n - a_i^n, \qquad i = 0, \ldots, k(n) - 1,$$

the following hold:

(a) $d(y_0^n, p) \leq \frac{1}{n}$ and $d(y_{k(n)}^n, p) \leq \frac{1}{n}$.

(b) $T \leq t_i^n \leq 2T$, $i = 0, \ldots, k(n) - 1$.

(c) $d(\Phi_{t_i^n}(y_i^n), y_{i+1}^n) \leq \frac{1}{n}$, $i = 0, \ldots, k(n) - 1$.

Proof. Let $n \in \mathbf{N}$. Lemma 4.2 shows that there exists $A_n > 0$ such that for any $t \geq A_n$ and for all $0 \leq h \leq 2T$, $d(\Phi_h(W^0(t)), W^0(t+h)) \leq \frac{1}{n}$.

As $p \in L(W^0)$ there exists $a_0^n \geq \sup(A_n, n)$ such that $d(W^0(a_0^n), p) \leq \frac{1}{n}$ and there exists $T' > T$ such that $d(W^0(a_0^n + T'), p) \leq \frac{1}{n}$. Write $a_0^n + T' = kT + r$, where $k \in \mathbf{N}$ and $0 \leq r < T$. Then define $a_i^n = a_0^n + i(T + \frac{r}{k})$, $i = 0, \ldots, k$. □

Let $C_n = \{y_i^n, i = 0, \ldots, k(n)\}$, where y_i^n is defined as in Claim 1. As C_n is a compact set, we may extract from $\{C_n\}_{n \geq 0}$ a subsequence which converges toward a compact set C for the Haussdorf metric in Q. It is clear that $C \subset L(W^0)$.

Claim 2. Let $\delta > 0$ and $T > 0$; then p is (δ, T) recurrent for the restricted flow $\Phi | L(W^0)$.

Proof. By uniform continuity of the flow on Q there exists $\alpha > 0$ such that $d(x, y) \leq \alpha$ implies $d(\Phi_t(x), \Phi_t(y)) \leq \delta/3$ uniformly in $t \in [0, 2T]$. We may always assume $\alpha \leq \delta/3$. Choose n large enough such that $1/n \leq \delta/3$ and $d(C_n, C) \leq \alpha$. Then we construct a finite sequence $Z_0, \ldots, Z_{k(n)} \in C$ such that $d(Z_i, y_i^n) \leq \alpha$ for $i = 0, \ldots, k(n)$. Then

$$d(Z_0, p) \leq \alpha + 1/n \leq \delta,$$

$$d(Z_{k(n)}, p) \leq \alpha + 1/n \leq \delta,$$

and

$$d(\Phi_{t_i^n}(Z_i), Z_{i+1}) \leq d(\Phi_{t_i^n}(Z_i), \Phi_{t_i^n}(y_i^n)) + d(\Phi_{t_i^n}(y_i^n), y_{i+1}^n) + d(y_{i+1}^n, Z_{i+1})$$

$$\leq \delta/3 + \frac{1}{n} + \alpha \leq \delta. \qquad \square$$

Proof of Lemma 4.2.

The Lipschitz case. Here we assume that $\overline{H}$ is locally Lipschitz. We let $L(Q)$ denote the Lipschitz constant of $\overline{H}$ on Q and $\|\overline{H}\|_Q$ the uniform norm of $\overline{H}$ on Q. The next lemma proves Lemma 4.2 with an estimate. This estimate will be useful to prove the main result of §5.

LEMMA 4.4. *For all* $T > 0$ *and all* $t \geq 0$,

$$\sup_{h\in[0,T]} \|W^0(t+h) - \Phi_h(W^0(t))\| \leq e^{L(Q)\,T}[2\epsilon(t,T)(1+TL(Q)) + TL(Q)\|\overline{H}\|_Q\gamma_{m(t)}],$$

where

$$\epsilon(t,T) = \sup_{\{k;\ 0\leq\tau_k-\tau_{m(t)}\leq T+1\}} \left\| \sum_{i=m(t)}^{k-1} \gamma_i.u_i \right\| + (T+1)\left[\sup_{\{k;0\leq\tau_k-\tau_{m(t)}\leq T+1\}} \|b_k\|\right].$$

Proof. We begin with a simple inequality:

$$\forall u, v \in [\tau_{m(t)}, \tau_{m(t+T)+1}[\ \left\| \int_u^v (U(s)+B(s))ds \right\| \leq 2\epsilon(t,T). \tag{6}$$

To prove (6) we note that for any $u \geq \tau_{m(t)}$ there exists $\alpha \in [0,1]$ for which

$$\int_{\tau_{m(t)}}^{u} (U(s)+B(s))ds = \alpha \int_{\tau_{m(t)}}^{\tau_{m(u)}} (U(s)+B(s))ds + (1-\alpha)\int_{\tau_{m(t)}}^{\tau(m(u)+1)} (U(s)+B(s))ds.$$

As for $u, v \in [\tau_{m(t)}, \tau_{m(t+T)+1}[$,

$$\int_u^v (U(s)+B(s))ds = -\int_{\tau_{m(t)}}^{u} (U(s)+B(s))ds + \int_{\tau_{m(t)}}^{v} (U(s)+B(s))ds.$$

Inequality (6) follows.

According to (4),

$$\begin{aligned} W^0(t+h) - \Phi_h(W^0(t)) = \int_0^h \overline{H}(W(t+s))ds - \int_0^h \overline{H}(\Phi_s(W^0(t))ds \\ + \int_t^{t+h} (U(s)+B(s))ds. \end{aligned} \tag{7}$$

Let

$$A(h) = \|W^0(t+h) - \Phi_h(W^0(t))\|.$$

Equation (7) implies

$$\begin{aligned} A(h) \leq L(Q)\int_0^h A(s)ds + \int_0^h \|\overline{H}(W^0(t+s)) - \overline{H}(W(t+s))\|ds \\ + \left\| \int_t^{t+h} (U(s)+B(s))ds \right\|. \end{aligned} \tag{8}$$

On the other hand, for any $h \in [0,T]$

$$\|W^0(t+h) - W(t+h)\| = \left\| \int_{\tau_{m(t+h)}}^{t+h} [\overline{H}(W(s)) + U(s) + B(s)]ds \right\|,$$

and inequality (6) implies

$$\|W^0(t+h) - W(t+h)\| \leq \gamma_{m(t+h)} \|\overline{H}\|_Q + 2\epsilon(t,T). \tag{9}$$

From inequalities (8) and (9), we deduce that for any $h \in [0,T]$

$$A(h) \leq L(Q) \int_0^h A(s)ds + 2\epsilon(t,T)(1 + TL(Q)) + TL(Q)\gamma_{m(t)} \|\overline{H}\|_Q,$$

and we conclude by using Gronwall's inequality. □

The non-Lipschitz case. Here we prove Lemma 4.2, assuming only that $\overline{H}$ is continous with unique integral curves. The key of the proof is to use the Kushner and Clark lemma (1978). Let $W^s()$ be the function defined for any $s \geq 0$ by

$$\forall t \geq -s, W^s(t) = W^0(t+s)$$

and

$$\forall t < -s, W^s(t) = w_0.$$

The Kushner and Clark lemma is the following.

THEOREM 4.5 (Kushner and Clark (1978)). *Under the assumptions* A1, A2, *and* A3 *of Theorem* 1.2, $\{W^s(.)\}_{s\geq 0}$ *is relatively compact in* $C^0(\mathbf{R}, \mathbf{R}^m)$ *with respect to the topology of uniform convergence on bounded intervals (i.e., from every sequence of the set* $\{W^s(.)\}_{s\geq 0}$ *it is possible to select a subsequence which converges uniformly on bounded intervals), and the limit of each convergent subsequence is the solution to the ODE.*

What we want to prove (i.e., Lemma 4.2) is equivalent to

$$\lim_{s\to\infty} sup_{h\in[-T,T]} \|W^s(h) - \Phi_h(W^s(0))\| = 0 \tag{10}$$

for all $T > 0$. Let D denote a distance on $C^0(\mathbf{R}, \mathbf{R}^m)$ induced by the topology of uniform convergence on bounded intervals; then (10) can be rewritten as

$$\lim_{s\to\infty} D(W^s(.), \Phi(., W^s(0))) = 0. \tag{11}$$

Let W^* be an arbitrary limit point of $\{W^s(.)\}_{s\geq 0}$. By Theorem 4.5 W^* is a solution to the ODE, and by uniqueness of integral curves $W^*(t) = \Phi(t, W^*(0))$ for all t. Thus, $W^*(.) = \Phi(., W^*(0))$. This proves (11). □

5. L^q estimates and shadowing. In this section we consider the following question:

Given $\{w_n\}_{n\geq 0}$, *a trajectory solution to* (1), *does there exist a solution to* (2) *whose omega limit set is* $L(\{w_n\}_{n\geq 0})$*?*

Theorem 1.3 shows that (at least under assumptions of Theorem 1.2) the answer is generally negative since $L(\{w_n\}_{n\geq 0})$ can be an arbitrary internally chain-recurrent set. However, it is useful to understand what kind of conditions ensure a positive answer to this question. A case of particular interest in applications is given by the following problem:

Assume that each solution to (2) *converges toward an equilibrium.*

Does every solution to (1) *converge also toward an equilibrium?*

We saw in §3 several examples for which $CR(\overline{H})$ is the set of equilibria and the theorems of §§1 and 2 were applied to answer positively. But it may happen that $CR(\overline{H})$ contains nonequilibrium points (see Example 6.3) and further conditions are required.

We begin with a simple example.

EXAMPLE 5.1. *Consider the recursion which is defined in polar coordinates* $\rho \geq 0$, $\theta \in \mathbf{R}/(2\pi\mathbf{Z})$ *by*

$$\rho_{n+1} - \rho_n = \gamma_n(g(\rho_n) + 1_{[0.5\,,3]}(\rho_n).\xi_n),$$

$$\theta_{n+1} - \theta_n = -\gamma_n,$$

where $\{\xi_n\}_{n\geq 0}$ *is a sequence of independently and identically distributed random variables with uniform distribution on* $[-\frac{1}{2}, \frac{1}{2}]$, $\gamma_n = \frac{1}{n^\alpha}$ *for some* $0 < \alpha \leq 1$, *and* $g : \mathbf{R}_+ \to \mathbf{R}$ *is a smooth function which is zero on* $\{0\} \cup [1, 2]$, *positive on* $]0, 1[$, *and negative on* $]2, \infty)$. *The ODE associated with this recursion is defined by*

$$\frac{d\rho}{dt} = g(\rho),$$

$$\frac{d\theta}{dt} = -1.$$

The phase portrait of this ODE is given by Fig. 3.

We see that any connected internally chain recurrent set of this ODE is either the equilibrium $0_{\mathbf{R}^2}$ *or a cylinder of periodic orbits*

$$C_{a,b} = \{\rho : a \leq \rho \leq b\} \times \{\theta \in \mathbf{R}/(2\pi\mathbf{Z})\},\ 1 \leq a \leq b \leq 2.$$

Assume that the initial condition of the process is not $0_{\mathbf{R}^2}$. *Therefore, according to Proposition* 2.1, *the limit set* $L(\{w_n\}_{n\geq 0})$ *of the process has to be a cylinder. In fact, it is not difficult to show that*

(a) *if* $\alpha > \frac{1}{2}$, $L(\{w_n\}_{n\geq 0})$ *is almost surely a periodic orbit* $L(\{w_n\}_{n\geq 0}) = C_{a,a}$ *for some* $1 \leq a \leq 2$;

(b) *if* $\alpha < \frac{1}{2}$, $L(\{w_n\}_{n\geq 0}) = C_{1,2}$.

The main reason is that the sum $\sum_n \gamma_n \xi_n$ *converges for* $\alpha > \frac{1}{2}$, *while*

$$\limsup_{n\to\infty} \sum_n \gamma_n \xi_n = -\liminf_{n\to\infty} \sum_n \gamma_n \xi_n = +\infty$$

for $\alpha < \frac{1}{2}$ (*see, e.g., Neveu* (1964, *p.* 138)).

In case (a) $L(\{w_n\}_{n\geq 0})$ *is an omega limit set of the ODE. Case* (b) *gives an example for which the asymptotic behavior of* (1) *is quite different from the asymptotic behavior of* (2).

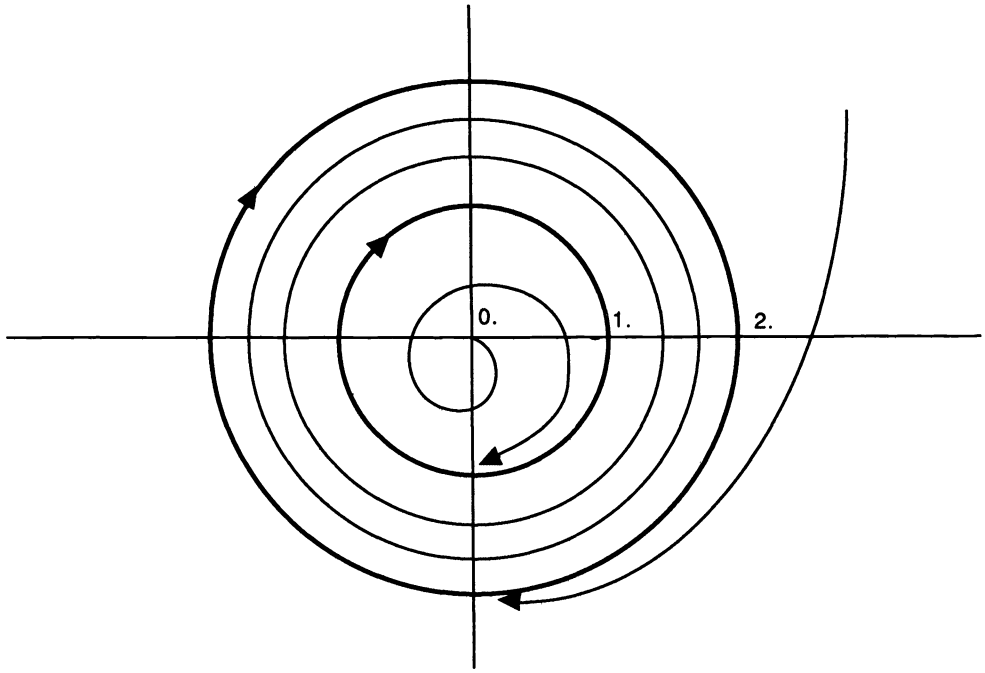


FIG. 3.

Expansion rate. In the previous example, the condition $\alpha < 1$ means, intuitively, that the convergence of $\{\gamma_n\}_{n\geq 0}$ to zero is not fast enough to ensure the convergence of $\{w_n\}_{n\geq 0}$ toward the omega limit sets of the ODE. We now formalize this idea and show that, conversely, if $\{\gamma_n\}_{n\geq 0}$ goes to zero at a suitable rate depending on the *expansivity* of the ODE, then $\{w_n\}_{n\geq 0}$ is in some sense asymptotic to a forward trajectory of (2).

Here we make crucial use of the ideas and methods introduced by Morris W. Hirsch in a recent paper (1993). The main idea of what follows is to use a shadowing theorem proved in Hirsch (1993) together with L^q estimates of the error which is made when (1) is replaced by (2).

To avoid technicalities, we will assume throughout the remainder of this section that $\overline{H}$ is a C^1 vectorfield on $\mathbf{R}^m$ with the point at infinity as a source. By ∞ as a source we mean that there exists a bounded nonempty open set $U \subset \mathbf{R}^m$ such that for all $w \in \mathbf{R}^m$

$$\lim_{t\to\infty} d(\Phi_t(w), \operatorname{clos}(U)) = 0$$

and for some $T > 0$

$$\Phi_T(\operatorname{clos}(U)) \subset U.$$

Let K denote a nonempty compact set positively invariant under the flow of $\overline{H}$. The *expansion rate* of $\overline{H}$ in K is defined in Hirsch (1993) (see also Hirsch and Pugh (1970)). For convenience, we introduce it in a logarithmic form:

$$l_{exp}(\overline{H}, K) = \lim_{t\to+\infty} \left[\min_{w\in K} \frac{\log(\|(D\Phi_t(w))^{-1}\|^{-1})}{t}\right],$$

where $D\Phi_t(w)$ denote the differential of Φ_t at w. The limit exists by subadditivity. We call $l_{exp}(\overline{H}, K)$ the "log-expansion rate" of $\overline{H}$ in K. This real number measures

the expansivity of the dynamical system induced by $\overline{H}$. It is zero if the flow is isometric, positive if the flow has a tendency to be expansive, and negative otherwise. Let $\mathrm{clos}(\mathcal{L}(\overline{H}))$ be the closure of all alpha and omega limit points of the trajectories solution to (2). As we assume that "∞" is a source, $\mathrm{clos}(\mathcal{L}(\overline{H}))$ is a compact nonempty invariant set. We define the "log-expansion rate" of $\overline{H}$ as

$$l_{exp}(\overline{H}) = l_{exp}(\overline{H}, \mathrm{clos}(\mathcal{L}(\overline{H}))).$$

This definition makes sense and is motivated by the following important property (Hirsch (1993)): If L is a compact invariant subset of K containing all alpha and omega limit points in K, then $l_{exp}(\overline{H}, K) = l_{exp}(\overline{H}, L)$. Some properties of $l_{exp}(\overline{H})$ are given in §6.

A shadowing theorem. Let $\{\alpha_n\}_{n\geq 0}$ denote a sequence of nonnegative real numbers.

Define the "log-convergence rate" of $\{\alpha_n\}_{n\geq 0}$ with respect to the time scale $\tau_n = \sum_{i=0}^{n-1} \gamma_i$ as

$$l_\tau(\alpha) = \limsup_{n\to+\infty} \frac{\log(\alpha_n)}{\tau_n}.$$

Now consider the same recursion as in Theorem 1.2 in a probabilistic framework:

$$w_{n+1} - w_n = \gamma_n \overline{H}(w_n) + \gamma_n u_n + \gamma_n b_n, \tag{12}$$

where $\{\gamma_n\}_{n\geq 0}$ is a decreasing gain sequence, $\{u_n\}_{n\geq 0}$ and $\{b_n\}_{n\geq 0}$ are two sequences of $\mathbf{R}^m$-valued random variables defined on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and $\overline{H} : \mathbf{R}^m \mapsto \mathbf{R}^m$ is a C^1 vectorfield with ∞ as a source.

Recall that $\|.\|_q$ denotes the $L^q(\Omega)$ norm. For each $T > 0$ and each $q \in [1, +\infty[$ let

$$\alpha_n^{q,T} = \left\| \sup_{\{k\,;0\leq\tau_k-\tau_n\leq T\}} \left\| \sum_{i=n}^{k-1} \gamma_i u_i \right\| \right\|_q$$

and

$$\beta_n^{q,T} = \left\| \sup_{\{k\,;0\leq\tau_k-\tau_n\leq T\}} \|b_k\| \right\|_q .$$

THEOREM 5.2. *Let $\{w_n\}_{n\geq 0}$ be solution to the recursion* (12). *Assume that there exists $q \geq 1$ such that*

- A1) $E(\sup_{n\geq 0} \|\{w_n\}_{n\geq 0}\|^q) < +\infty$.
- A2) *For each $T > 0$,*

$$l_\tau(\beta^{q,T}) < \min(0, l_{exp}(\overline{H})),$$

$$l_\tau(\alpha^{q,T}) < \min(0, l_{exp}(\overline{H})),$$

$$l_\tau(\gamma) < \min(0, l_{exp}(\overline{H})).$$

Then

- a) *there exists a random vector* w' *such that*

$$\lim_{n\to+\infty} \|w_n - \Phi_{\tau_n}(w')\| = 0$$

almost surely.

- b) *If there exists a compact* $Q \subset \mathbf{R}^m$ *such that* $\{w_n\}_{n\geq 0}$ *remains in* Q *almost surely (in which case A1 is obviously satisfied), then the following estimate holds:*

$$l_\tau(\{\|w_n - \Phi_{\tau_n}(w')\|_q\}_{n\geq 0}) \leq \sup(l_\tau(\beta^{q,T}), l_\tau(\alpha^{q,T}), l_\tau(\gamma)).$$

REMARK 5.3. *Conclusion* (a) *of Theorem* 5.2 *implies that* $L(\{w_n\}_{n\geq 0}) = \omega(w')$ *almost surely.*

As in §2 we apply the previous result to the stochastic approximation (1), where $\{\xi_n\}_{n\geq 0}$ is a sequence of random variables defined on $(\Omega, \mathcal{F}, \mathcal{P})$. Maximal inequalities for sum of random variables reduce condition A2 to a simple condition on $l_\tau(\gamma)$. First of all, note that for any $\lambda > 0$,

$$l_{\lambda\tau}(\lambda\gamma) = \frac{1}{\lambda} l_\tau(\gamma).$$

If $\gamma_n = f(n)$ for some positive decreasing function f with $\int_1^{+\infty} f(s)ds = +\infty$, then

$$l_\tau(\gamma) = \limsup_{x\to+\infty} \frac{\log(f(x))}{\int_1^x f(s)ds}.$$

For example, if

$$\gamma_n = \frac{1}{n^\alpha \log(n)^\beta},$$

then $l_\tau(\gamma) = 0$ for $0 < \alpha < 1$ and $\beta \geq 0$, $l_\tau(\gamma) = -1$ for $\alpha = 1$ and $\beta = 0$, and $l_\tau(\gamma) = -\infty$ for $\alpha = 1$ and $0 < \beta \leq 1$.

Independent inputs. As in Proposition 2.1, we let M denote a subset of $\mathbf{R}^m$.

PROPOSITION 5.4. *Let* $\{w_n\}_{n\geq 0}$ *be the solution to* (1). *Assume that*

- A1) $\{\xi_n\}_{n\geq 0}$ *is a sequence of independent and identically distributed random variables.*
- A2) $P(\forall n \in N, w_n \in M) = 1$.
- A3) $w \mapsto \overline{H}(w) = E(H(w, \xi_0))$ *is* C^1 *with* ∞ *as a source.*

There exists $q \geq 2$ *such that*

- A4) $E(\sup_{n\geq 0} \|\{w_n\}_{n\geq 0}\|^q) < +\infty$ *and* $w \mapsto \|H(w, \xi_0)\|_q$ *is bounded on* M.
- A5) $l_\tau(\gamma) < 2\min(0, l_{exp}(\overline{H}))$.

Then

- a) *The conclusion* a) *of Theorem* 5.2 *holds.*
- b) *If* M *is compact,* $l_\tau(\{\|w_n - \Phi_{\tau_n}(w')\|_q\}_{n\geq 0}) \leq \frac{1}{2} l_\tau(\gamma)$.

Proof. Let $b_n = 0$ and $u_n = H(w_n, \xi_n) - \overline{H}(w_n)$. As already noted, $\{u_n\}_{n\geq 0}$ is a martingale difference. For $q = 2$, Doob's inequality for L^2 martingales gives

$$\alpha_n^{2,T} \leq \left[C(M) \sum_{i=n}^{m(\tau_n+T)-1} \gamma_i^2 \right]^{\frac{1}{2}},$$

where $C(M)$ is a positive constant and $m(T)$ is defined by (5). So

$$\alpha_n^{2,T} \leq [C(M)\gamma_n T]^{\frac{1}{2}}.$$

Then $l_\tau(\alpha^{2,T}) \leq \frac{1}{2} l_\tau(\gamma)$ and the condition A2 of Theorem 5.2 is satisfied. For $q > 2$,

$$\alpha_n^{q,T} \leq \left[C(M,T) \sum_{i=n}^{m(\tau_n+T)-1} \gamma_i^{1+q/2} \right]^{\frac{1}{2}}$$

for some constant $C(M,T)$. This inequality is proved, in a more general context, in Métivier and Priouret (1987, Prop. 8). Therefore,

$$\alpha_n^{q,T} \leq [C(M,T)T]^{\frac{1}{q}} \gamma_n^{1/2},$$

and the result follows. □

REMARK 5.5. *It is interesting to note that the condition* A5 *of Proposition* 5.4 *is always satisfied for* $\gamma_n = \frac{1}{n \log(n)}$. *For* $\gamma_n = \frac{\epsilon}{n+n_0}$, *it reduces to the condition* $\epsilon < -\frac{1}{2 l_{exp}(\overline{H})}$.

Mixing inputs. In the case corresponding to Proposition 2.2, in which the observations are given by a mixing process, our approach of condition A2 in Theorem 5.2 is based on some kind of uniform maximal inequalities. Unfortunately, these estimates depend on the dimension of the parameter space and the condition we obtain presents the "curse of dimensionality." Here we shall assume that $\{\xi_n\}_{n\geq 0}$ is stationary to facilitate the verification of assumption A2 of Theorem 5.2.

PROPOSITION 5.6. *Let* $\{w_n\}_{n\geq 0}$ *be the solution to* (1). *Assume that*

- A1) $\{\xi_n\}_{n\geq 0}$ *is a stationary* ϕ *mixing (respectively,* α *mixing) process.*
- A2) *There exists a compact set* $Q \subset \mathbf{R}^m$ *such that* $P(\forall n \geq 0, w_n \in Q) = 1$.
- A3) $\overline{H}(w) = \lim_{n\to\infty} E(H(w,\xi_n))$ *exists.*
- A4) *There exists a measurable function* $k(.)$ *such that*

$$\forall x, y \in \mathbf{R}^m \|H(x,\xi) - H(y,\xi)\| \leq k(\xi)\|x - y\|.$$

There exists $r \in [2,\infty]$ *such that*

- A5) *The map* $w \mapsto \sup_{n\geq 0} \|H(w,\xi_n)\|_r$ *is bounded on* Q *and* $\sup_{n\geq 0} \|k(\xi_n)\|_r < +\infty$.
- A6) *(L^r case). If* $r < \infty$, $\phi_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{r}{2r-2}$ *(respectively,* $\alpha_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{r}{r-2}$*).*
- A6′) *(L^∞ case). If* $r = \infty$, $\phi_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{1}{2}$ *(respectively,* $\alpha_n = O(\frac{1}{n^\beta})$ *for some* $\beta > 1$*).*
- A7) $l_\tau(\gamma) < 2(m+1)\min(0, l_{exp}(\overline{H}))$.

Then the conclusions of Theorem 5.2 *hold with probability one.*

The proof is given in appendix (§9).

6. Applications. Here again $\overline{H}$ denotes a C^1 vectorfield with ∞ as a source.

Convergent systems. We say that $\overline{H}$ is a *convergent system* if $\overline{H}$ admits a finite number of equilibria $\{e_1,\ldots,e_n\}$ and

$$\mathcal{L}(\overline{H}) = \{e_1,\ldots,e_n\}.$$

Equivalently, this means that the flow induced by $\overline{H}$ on $S^m = \mathbf{R}^m \cup \{\infty\}$ (the compactification of $\mathbf{R}^m$) has simple dynamics and finitely many equilibria.

Let $\{\lambda_i^j,\ j = 1, \ldots, m\}$ denote the set of eigenvalues of the matrix $D\overline{H}(e_i)$. Define

$$\beta(e_i) = \min\{\mathrm{Re}(\lambda_i^j) : j = 1, \ldots, m\},$$

where Re denotes the real part. Since e_i is a fixed point, $D\Phi_t(e_i) = exp(tD\overline{H}(e_i))$. Therefore

$$\lim_{t\to\infty} \frac{\log(\|(D\Phi_t(e_i))^{-1}\|^{-1})}{t} = \beta(e_i),$$

and by definition of the log-expansion rate we deduce the following proposition.

PROPOSITION 6.1. *If $\overline{H}$ is a convergent system with equilibria $\{e_1, \ldots, e_n\}$, then*

$$l_{exp}(\overline{H}) = \min\{\beta(e_i) : i = 1, \ldots, n\}.$$

COROLLARY 6.2. *Let $\{w_n\}_{n\geq 0}$ be the solution to* (1). *Assume that conditions* A1–A4 (*respectively,* A1–A6, A6′) *of Proposition* 5.4 (*respectively,* 5.6) *are satisfied. Assume that the averaged vectorfield $\overline{H}$ defined by* A3 *is convergent and that*

$$\forall i \in \{1, \ldots, n\},\ l_\tau(\gamma) < 2\min(0, \beta(e_i))$$

(respectively, $l_\tau(\gamma) < 2(m+1)\min(0, \beta(e_i))$). Then $\{w_n\}_{n\geq 0}$ converges almost surely toward an equilibrium.

EXAMPLE 6.3. *Consider the following stochastic approximation process defined on $\mathbf{R}^2$ by*

$$x_{n+1} - x_n = \frac{\epsilon}{n} H_1(x_n, y_n, \xi_n),$$

$$y_{n+1} - y_n = \frac{\epsilon}{n} H_2(x_n, y_n, \xi_n),$$

where $\{\xi_n\}_{n\geq 0}$ is a sequence of independently and identically distributed random variables uniformly distributed on $[-1, 1]$.

$$H_1(x, y, \xi) = (1 - (x^2 + y^2))x - yf(y) + \xi,$$

$$H_2(x, y, \xi) = (1 - (x^2 + y^2))y + xf(y) + \xi,$$

where $f(y) = y^2$. The phase portrait of the averaged ODE is given by Fig. 4.

We see that this system is convergent but admits $S^1 = \{(x, y) : x^2 + y^2 = 1\}$ as a cycle of equilibria. Therefore, the theorems of §§2 and 3 *are not sufficent to ensure the convergence of the process $\{x_n, y_n\}_{n\geq 0}$.*

The equilibria of this system are $e_1 = (0, 0)$, $e_2 = (1, 0)$, and $e_3 = (-1, 0)$. A simple computation shows that 0 *and* -2 *are the eigenvalues of the linearized ODE at points e_2 and e_3, and* 1 *is a double eigenvalue at point e_1. Thus,*

$$\beta(e_1) = 1,\ \beta(e_2) = \beta(e_3) = -2.$$

For $\gamma_n = \frac{\epsilon}{n}$ we have $l_\tau(\gamma) = -\frac{1}{\epsilon}$. Therefore, according to Corollary 6.2, *if $\epsilon < \frac{1}{4}$, the sequence $\{x_n, y_n\}_{n\geq 0}$ converges almost surely toward an equilibrium. Furthermore, a theorem of Pemantle* (1990) *can be used to show that this equilibrium cannot be the hyperbolic unstable equilibrium e_1.*

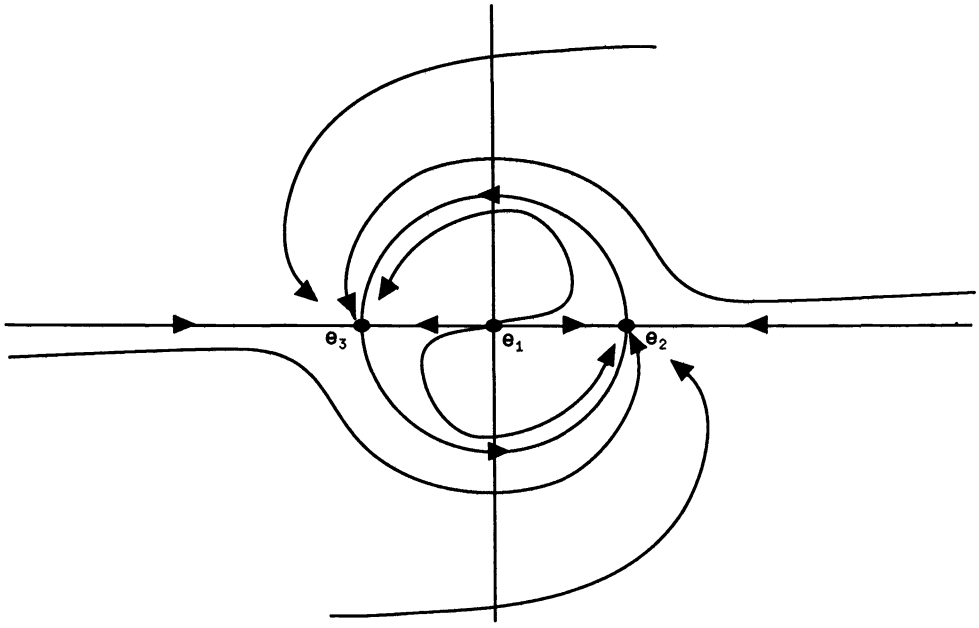


FIG. 4.

Globally convergent systems. A convergent system $\overline{H}$ with one unique equilibrium $\{e_1\}$ is said *globally convergent.*

The L^q estimate given by assertion (b) of Theorem 5.2 can be used to bound the L^q rate of convergence of algorithms associated to globally convergent ODEs. We mention here a corollary based on Proposition 5.4. Other estimates based on Proposition 5.6 or Theorem 5.2 are possible. Define

$$\rho(e_1) = \sup\{\mathrm{Re}(\lambda_1^j) : j = 1, \ldots, m\},$$

where $\{\lambda_1^j : j = 1, \ldots, m\}$ are the eigenvalues of $D\overline{H}(e_1)$.

Note that $\beta(e_1) \leq \rho(e_1) \leq 0$.

COROLLARY 6.4. *Let* $\{w_n\}_{n\geq 0}$ *be the solution to* (1). *Assume that conditions* A1–A4 *of Proposition* 5.4 *are satisfied and* M *is compact. Assume that the averaged vectorfield* $\overline{H}$ *defined by* A3 *is globally convergent and that* $l_\tau(\gamma) < 2\min(0, \beta(e_1))$. *Then*

$$l_\tau(||w_n - e_1||_q) \leq \rho(e_1).$$

Proof. Proposition 6.1 and (b) of Proposition 5.4 imply that $l_\tau(||w_n - \Phi_{\tau_n}(w')||_q) \leq \beta(e_1)$ for some random variable $w' \in M$. Since $\overline{H}$ is globally convergent and M is compact, we have the estimate $l_\tau(||\Phi_{\tau_n}(w') - e_1||) \leq \rho(e_1)$. Thus $l_\tau(||w_n - e_1||_q) \leq l_\tau(||\Phi_{\tau_n}(w') - e_1||_q + ||w_n - \Phi_{\tau_n}(w')||_q) \leq \sup(\rho(e_1), \beta(e_1)) = \rho(e_1)$. □

For $\gamma_n = \frac{\epsilon}{n+n_0}$ and $\epsilon < \frac{1}{2|\beta(e_1)|}$ this gives the following estimate: For all $\delta > 0$

there exists $n(\delta) \geq 0$ such that

$$\forall n \geq n(\delta),\ \|w_n - e_1\|_q \leq \frac{1}{n^{\epsilon(|\rho(e_1)|-\delta)}}.$$

This estimate can be compared with L^2 upper bounds given in Eweda and Macchi (1983) and Benveniste, Métivier, and Priouret (1990, Thms. 22 and 24, pp. 244, 246). It is slightly weaker but requires a weaker condition on the vectorfield.

Nonconvergent systems. For a general vectorfield $\overline{H}$ the log-expansion rate can be difficult to compute. The following proposition is useful to estimate it.

PROPOSITION 6.5 (Hirsch (1993)).

(a) *Let $\beta_s(w)$ be the smallest eigenvalue of the symmetric matrix*

$$\frac{1}{2}(D\overline{H}(w) + D\overline{H}(w)^T),$$

where ^{T} denotes the transpose operation. Then

$$l_{exp}(\overline{H}) \geq \min\{\beta_s(w) : w \in \operatorname{clos}(\mathcal{L}(\overline{H}))\}.$$

(b) *$l_{exp}(\overline{H})$ is invariant by C^1 change of coordinates.*

Assertion (a) is proved in Hirsch (1993). Assertion (b) is easy to check from the definition.

EXAMPLE 6.6. *Consider the stochastic approximation process defined in Example 6.3, where the function $f(.)$ which appears in the definition of H_1 and H_2 is now chosen to be the function $f(y) = 1$. The averaged ODE admits two internally chain-recurrent sets: the unstable equilibrium $0_{\mathbf{R}^2}$ and the stable limit cycle $S^1 = \{(x, y) : x^2 + y^2 = 1\}$. By a theorem of Pemantle already mentioned, $L(\{w_n\}_{n\geq 0})$ cannot be $0_{\mathbf{R}^2}$. Thus, according to Proposition 2.1, $L(\{w_n\}_{n\geq 0}) = S^1$.*

Note that this result is true for all values of $\epsilon > 0$. Let us now show how it can be sharpened by use of the log-expansion rate. To compute the log-expansion rate we use (b) of Proposition 6.5. In polar coordinates, the averaged ODE takes the simple form

$$\frac{d\rho}{dt} = \rho(1 - \rho^2), \qquad \frac{d\theta}{dt} = 1$$

from which it is easy to deduce that the log-expansion rate is given as $l_{exp}(\overline{H}) = \min\{-2, 1, 0\} = -2$. Let θ_n be the angular variable which measures the angle between the x-axis and the vector (x_n, y_n). If $\epsilon < \frac{1}{4}$, Proposition 5.4 applies and we deduce the "asymptotic phase property":

$$\lim_{n\to\infty} \theta_n - [\epsilon log(n)] \bmod 2\pi = \theta^*,$$

where θ^* is a random variable taking values in $[0, 2\pi]$.

7. Proof of Theorem 5.2.

LEMMA 7.1. *Under the assumptions of Theorem 5.2, the conditions of Theorem 1.2 are satisfied.*

Proof. Assumption A1 of Theorem 5.2 implies condition A1 of Theorem 1.2. Now check conditions A2 and A3. Let $\{n_j\}_{j\geq 0}$ be the sequence defined by $n_j = m(jT)$ for $j \geq 0$, where $m(.)$ is defined by (5). For any integer $n \in [n_j, n_{j+1}[$,

$$\|b_n\| \leq \sup_{\{k;\ 0\leq\tau_k-\tau_{n_j}\leq T\}} \|b_k\|.$$

Thus

$$E\left(\sup_{n\geq n_p}\|b_n\|^q\right)\leq E\left(\sup_{j\geq p}\sup_{\{k;\,0\leq\tau_k-\tau_{n_j}\leq T\}}\|b_k\|^q\right)\leq\sum_{j\geq p}\beta_{n_j}^{q,T}.$$

Now, assumption A2 of Theorem 1.2 implies

$$\beta_{n_j}^{q,T}\leq Ce^{-\lambda\tau_{n_j}}\leq C'e^{-(\lambda T)j}$$

for some constants $C, C', \lambda > 0$. It follows that

$$E\left(\sup_{n\geq n_p}\|b_n\|^q\right)\leq\sum_{j\geq p}C'e^{-(\lambda T)j}<+\infty$$

and the Cauchy criterion implies that condition A2 of Theorem 1.2 holds almost surely.

For A3, remark that for any integer $n\in[n_j,n_{j+1}[$,

$$\sum_{i=n_j}^{k-1}\gamma_iu_i=\sum_{i=n_j}^{n_{j+1}-1}\gamma_iu_i-\sum_{i=n_j}^{n-1}\gamma_iu_i+\sum_{i=n_{j+1}}^{k-1}\gamma_iu_i$$

with the convention $\sum_i^j=-\sum_j^i$. Working as previously, we deduce

$$E\left(\sup_{n\geq n_p}\sup_{\{k;\,0\leq\tau_k-\tau_n\leq T\}}\left\|\sum_{i=n}^{k-1}\gamma_iu_i\right\|^q\right)\leq 3^q\sum_{j\geq p}\alpha_{n_j}^{q,T}$$

and conclude exactly as for A2. □

The following definitions and theorem are due to Morris W. Hirsch. The main result of §5 will be derived from this theorem.

Let (E,d) be a metric space and $G: E\mapsto E$ be a map. Let $0\leq\lambda<1$. A sequence $\{Y_k\}_{k\geq 0}$ in E is called a *λ-pseudoorbit* for G if

$$\limsup_{k\to+\infty}d(G(Y_k);Y_{k+1})^{\frac{1}{k}}\leq\lambda.$$

A point $Z\in E$ is said to *λ-shadow* the sequence $\{Y_k\}_{k\geq 0}$ if

$$\limsup_{k\to+\infty}d(G^k(Z);Y_{k+m})^{\frac{1}{k}}\leq\lambda$$

for some integer m.

The following theorem is a consequence of Hirsch (1993, Thm. 3.2) (more precisely, a consequence of its proof).

THEOREM 7.2 (Hirsch (1993)). *Assume E is a complete metric space. Assume there exists $\rho_*>0$ and $\mu>0$ such that for all $0\leq\rho\leq\rho_*$*

$$\forall X\in E\ B(G(X),\rho\mu)\subset G(B(X,\rho)).$$

Let $\{Y_k\}_{k\geq 0}$ a λ-pseudoorbit *for G in E such that*

$$0<\lambda<\min(1,\mu).$$

Then

- a) *there exists* $Z \in E$ *which* λ*-shadows* $\{Y_k\}_{k\geq 0}$;
- b) *if* $Z, Z' \in E$ *both* λ*-shadow* $\{Y_k\}_{k\geq 0}$, *then there exists natural numbers* l, r *such that* $G^l(Z) = G^r(Z')$.

Now we prove Theorem 5.2. Consider the recursion

$$v_{n+1} - v_n = \gamma_n.(f(v_n)\overline{H}(v_n) + u_n + b_n), \tag{13}$$

where $\{u_n\}_{n\geq 0}$ and $\{b_n\}_{n\geq 0}$ are the sequences of recursion (12), v_0 is in $L^q(\Omega)$, and f is a smooth function which is 1 on a closed ball $\overline{B(0,r)}$ which contains $CR(\overline{H})$ and is zero outside $B(0, r+1)$.

The proof decomposes in two steps. The first step is to prove that, under the assumptions of Theorem 5.2, $\{v_n\}_{n\geq 0}$ is asymptotic to a trajectory solution to (2). The second step is to show that any trajectory solution to (12) is asymptotically a solution to (13).

Step 1. Let $\{\Psi\}_{t\in\mathbf{R}}$ be the flow of the vectorfield $f\overline{H}$. The set of all α and ω limit points for $f\overline{H}$ is the disjoint union of $\mathcal{L}(\overline{H})$ and $\{x \in \mathbf{R}^m; \|x\| \geq r+1\}$. Therefore,

$$l_{exp}(f\overline{H}) = \min(0, l_{exp}(\overline{H})).$$

Note $\nu = l_{exp}(f\overline{H})$. Assumption A2 of Theorem 5.2 allows us to choose two real numbers ν'', ν' such that $0 < \nu'' < \nu' < \nu$ and

$$\sup(l_\tau(\beta^{q,T}), l_\tau(\alpha^{q,T}), l_\tau(\gamma)) < \nu''. \tag{14}$$

Because $\nu' < \nu$, there exists $T > 0$ such that for all $x \in \mathbf{R}^m$ and all $t \geq T$

$$\|D\Psi_t(x)^{-1}\|^{-1} \geq e^{\nu' t}. \tag{15}$$

Let $E = L^q(\Omega)$ and G be the map defined by $G(X) = \Psi_T(X)$. First we note that G is well defined (G maps E into E). Indeed, for any random vector $X \in L^q(\Omega)$,

$$E(\|G(X)\|^q) = E(\|G(X)\|^q 1_{X\in B(0,r+1)}) + E(\|G(X)\|^q 1_{X\notin B(0,r+1)}).$$

The first term on the right of this equality is bounded by continuity of the flow and the second term is finite because $G(x) = x$ outside $B(0, r+1)$.

Let $\mu = e^{-\nu' T}$. Inequality (15) implies

$$\forall x, y \in \mathbf{R}^m \|\Psi_T^{-1}(y) - x\| \leq \mu\|y - \Psi_T(x)\|,$$

so, for all X, Y in E,

$$\|G^{-1}(Y) - X\|_q \leq \mu\|Y - G(X)\|_q.$$

Therefore,

$$B(G(X), \mu\rho) \subset G(B(X, \rho))$$

for any $\rho > 0$.

In order to apply Theorem 7.2, it remains to construct a λ-pseudoorbit for G in E with $0 < \lambda < \min(1, \mu)$. With this purpose let $V^0(.)$ be the interpolated process associated to the sequence $\{v_n\}_{n\geq 0}$ (see §4 for the definition of the interpolated process) and define $Y_n = V^0(nT)$ for all integers n. As v_0, u_n, and b_n are in $L^q(\Omega)$, a

simple induction shows that v_n is in $L^q(\Omega)$ for all n. Hence $\{Y_n\}_{n\geq 0}$ is a sequence of L^q.

As $\overline{H}$ is C^1, $f\overline{H}$ is Lipschitz and bounded. Let L denote a Lipschitz constant for $f\overline{H}$ and K denote a bound for $\|f\overline{H}(x)\|$. Lemma 4.4, applied to V^0 and $f\overline{H}$, gives

$$\|V^0((n+1)T) - \Psi_T(V^0(nT))\| \leq e^{L.T}[2\epsilon(nT,T)(1+TL) + TLK\gamma_{m(nT)}], \tag{16}$$

where $\epsilon(t,T)$ is defined as in Lemma 4.4.

Let $\lambda = e^{-\nu''T}$. From (14) and (16) $\{Y_n\}_{n\geq 0}$ is a λ-pseudoorbit for G. As $0 < \lambda < \min(1,\mu)$, the Theorem 7.2 applies. Thus, there exists $Z \in E$ such that for any $\lambda < \lambda_1 < \inf(1,\mu)$ and k large enough

$$\|G^k(Z) - Y_{k+m}\|_q \leq \lambda_1^k \tag{17}$$

for some integer m. The Borel–Cantelli lemma now implies

$$\lim_{k\to+\infty} (G^k(Z) - Y_{k+m}) = 0$$

almost surely. Let $Z' = \Psi_{-mT}(Z)$. We have

$$\lim_{k\to+\infty} (G^k(Z') - Y_k) = 0. \tag{18}$$

For any $t > 0$, write $t = kT + r$ with $k \in N$ and $0 \leq r < T$. Thus

$$\begin{aligned}\Psi_t(Z') - V^0(t) = [\Psi_r(\Psi_{kT}(Z')) - \Psi_r(V^0(kT))] \\ +[\Psi_r(V^0(kT)) - V^0(kT+r)].\end{aligned} \tag{19}$$

Uniform continuity of the flow on $[0,T]$ and relation (18) imply that the first term on the right of equality (19) goes to zero. The second term goes to zero by Lemma 4.2. Then

$$\lim_{t\to+\infty} \Psi_t(Z') - V^0(t) = 0$$

almost surely. This concludes step 1.

The equality (19) also proves part (b) of Theorem 5.2. Indeed, if $\{w_n\}_{n\geq 0}$ remains in a compact Q almost surely, the ball $B(0,r)$ can be chosen large enough to contain Q and $\{w_n\}_{n\geq 0}$ is solution to (13). The first term on the right side of equality (19) can be bounded in $L^q(\Omega)$, using (17) and the fact that $x \mapsto \Psi_r(x)$ is Lipschitz uniformly in $r \in [0,T]$. The second term can be bounded in $L^q(\Omega)$ by using (14) and (16). □

Step 2. Let x be a vector arbitrary chosen outside $B(0,r)$. For any integer k define

$$\Omega_k = \cap_{n\geq k}\{w_n \in B(0,r)\} \cap \{w_{k-1} \notin B(0,r)\}$$

with the convention $w_1 = x$. Let $\Omega' = \cup_{k\in N}\Omega_k$. Because $CR(\overline{H}) \subset B(0,r)$ and $L(\{w_n\}_{n\geq 0}) \subset CR(\overline{H})$ almost surely (Theorem 1.2), we have $P(\Omega') = 1$.

Let $\{v_n^k\}_{n\geq k}$ be the sequence solution to (13) defined by the initial condition $v_k^k = w_k$. As $w_k \in L^q(\Omega)$, $v_k^k \in L^q(\Omega)$. Therefore, we deduce from Step 1 the existence of a vector $Z_k \in L^q(\Omega)$ such that

$$\lim_{n\to+\infty} (v_n^k - \Psi_{\tau_n}(Z_k)) = 0$$

almost surely. Define $Z = \sum_{k\in N} 1_{\Omega_k} Z_k$. Because $v_n^k = w_n$ on Ω_k for $n \geq k$, we have

$$\lim_{n\to+\infty} (w_n - \Phi_{\tau_n}(Z)) = 0$$

almost surely. □

8. An application to neural network learning. In this section we show briefly how the previous result can be applied to prove the consistency of some "hybrid" learning rules recently proposed in the neural network area.

A feedforward neural network can be seen as a function $G : I \times W \mapsto O$ mapping the Cartesian product of an *input space* I and a *weight space* W into an *output space* O. The dimension of I and O are the number of input units and the number of output units, respectively. Without loss of generality we take $O \subset \mathbf{R}$, $I \subset \mathbf{R}^{d-1}$, and $W \subset \mathbf{R}^m$. The function G embodies the network architecture. Given an input $x \in I$ and a weight vector $w \in W$, the network's output is given as $G(x, w)$. At this level of description, the form of G is not of particular importance. We only assume that G is smooth enough. For more details and an in-depth presentation of feedforward neural networks in the framework of approximation theory we refer the reader to the excellent book by Halbert White and co-workers (1992).

The goal of learning is to adapt the weight vector w in such a way that the network realizes some specific relationship between the input space and the output space. This relationship is generally expressed by an "environmental" probability law μ defined on $I \times O$. A "training set" is a sequence $\{\xi_n\}_{n\geq 0} \subset I \times O$ asymptotically stationary with μ as limiting law. We let $\xi_n = (x_n, y_n)$, x_n is referred to as the "input vector" and y_n is referred to as the "desired output" or "target." A general learning rule for feedforward net can be written as (1). The gain γ_n is called the "learning rate" in the connectionist jargon.

The most popular example is the classical "backpropagation algorithm." Given a pair of input and target $\xi = (x, y)$ and weight w, the network error is given as $Er(w, \xi) = e(G(x, w), y)$ where $e : \mathbf{R} \times \mathbf{R} \mapsto \mathbf{R}^+$ is a smooth "error function" (usually, $e(o, y) = (y - o)^2$). The algorithm is given by (1) with $H(w, \xi) = -\nabla_w Er(w, \xi)$. Here $\nabla_w Er(w, \xi)$ denotes the gradient of the map $w \mapsto Er(w, \xi)$. Therefore, assuming that interchange of derivative and expectation is possible, the ODE associated with the backpropagation is a gradient vectorfield:

$$\frac{dw}{dt} = -\nabla \overline{Er}(w), \tag{20}$$

where

$$\overline{Er}(w) = \int Er(w, \xi)\mu(d\xi).$$

Convergence of the backpropagation can be analyzed by using classical results on stochastic gradients (see, e.g., Nevel'son and Has'minskii (1974). See also Benveniste, Métivier, and Priouret (1990, p. 91) for a presentation of the backpropagation as a stochastic gradient). It is also a direct application of Corollary 3.3 restated here for convenience.

PROPOSITION 8.1. *Let* $\{w_n\}_{n\geq 0}$ *be the solution to* (1). *Assume that the assumptions of Proposition* 2.1 *or* 2.2 *hold with* $\overline{H}$ *given by* (20). *Assume that critical points of* $\overline{Er}$ (*i.e., the zeros of* (20)) *are isolated. Then* $\{w_n\}_{n\geq 0}$ *converges almost surely toward a critical point of* $\overline{Er}$.

"Hybrid" learning rules have been considered by Moody and Darken (1989), Poggio and Girosi (1990), Nowlan (1990), Benaim and Tomasini (1991, 1992), and Benaim (1995c) among others for neural architectures with "nonsigmoid" units. The main idea of these algorithms is to train each layer according to different learning rules.

Consider, for simplicity, a single hidden-layer network. (Extension to multilayers is easy.) Formally, $G(x, w) = G_2(G_1(x, w_1), w_2)$, where $w = (w_1, w_2) \in \mathbf{R}^{m_1} \times \mathbf{R}^{m_2}$,

$m_1 + m_2 = m$, $G_1 : \mathbf{R}^{d-1} \times \mathbf{R}^{m_1} \mapsto \mathbf{R}^k$, $G_2 : \mathbf{R}^k \times \mathbf{R}^{m_2} \mapsto \mathbf{R}$. The integer k is the number of hidden units, G_1 embodies the architecture of the (input-layer, hidden-layer) subnet and G_2 the architecture of the (hidden-layer, output-unit) subnet. The subnet G_1 is trained according to an "unsupervised" learning rule (for example, a data clustering algorithm (Moody and Darken (1989)) or a maximum likelihood algorithm (Nowlan (1990), Benaim and Tomasini (1992)), and the subnet G_2 is trained according to a "supervised" algorithm (backpropagation). This leads to an ODE of the form

$$\frac{dw_1}{dt} = -\nabla \overline{E_1}(w_1), \tag{21}$$

$$\frac{dw_2}{dt} = -\nabla_{w_2} \overline{E_2}(w_1, w_2) \tag{22}$$

for some smooth functions $E_1 : \mathbf{R}^{m_1} \mapsto \mathbf{R}$, $E_2 : \mathbf{R}^{m_1} \times \mathbf{R}^{m_2} \mapsto \mathbf{R}$. Remark that such an ODE is not a gradient vectorfield. It is a *cascade* of gradients.

It is often assumed that the output unit is linear: $G_2(G_1, w_2) = \langle w_2, G_1 \rangle$, and the performance of the subnet G_2 is measured by the squared error function, $e(o, y) = (y - o)^2$. In that case the equation (22) has the particular form

$$\frac{dw_2}{dt} = -A(w_1)w_2 + B(w_1), \tag{23}$$

where $A(w_1)$ is the $k \times k$ matrix defined by

$$A(w_1) = \int G_1(x, w_1).G_1(x, w_1)^T \nu(dx)$$

with $\nu(.) = \int \mu(., dy)$ and $B(w_1)$ is the k-dimensional vector

$$B(w_1) = \int y G_1(x, w_1) \mu(dx, dy).$$

PROPOSITION 8.2. *Let $\{w_n\}_{n \geq 0}$ be the solution to* (1). *Assume that assumptions of Proposition* 2.1 *or* 2.2 *hold with $\overline{H}$ given by the system* (21), (23). *Assume that equilibria of* (21) *are isolated. Then $L(\{w_n\}_{n \geq 0})$ is almost surely a connected compact subset of the equilibria set of $\overline{H}$.*

Proof. Write $w_n = (w_{1,n}, w_{2,n}) \in \mathbf{R}^{m_1} \times \mathbf{R}^{m_2}$. Proposition 8.1 shows that $\{w_{1,n}\}_{n \geq 0}$ converges almost surely toward an equilibrium of (21), say, ${w_1}^*$. Thus, $L(\{w_n\}_{n \geq 0}) = \{{w_1}^*\} \times L'$ for some set $L' \subset \mathbf{R}^{m_2}$. According to Proposition 2.1 (or 2.2), L' is compact and invariant under the dynamics

$$\frac{dw_2}{dt} = A({w_1}^*)w_2 - B({w_1}^*). \tag{24}$$

Since $A({w_1}^*)$ is a symmetric matrix, any compact invariant set for (24) is contained in the equilibria set of (24). This concludes the proof. □

For the more general system (21), (22), we shall use the fact that $L(\{w_n\}_{n \geq 0})$ is *internally* chain-recurrent combined with Proposition 3.2.

PROPOSITION 8.3. *Let $\{w_n\}_{n \geq 0}$ be a solution to* (1). *Assume that assumptions of Proposition* 2.1 *or* 2.2 *hold with $\overline{H}$ given by the system* (21), (22). *Assume that equilibria of* (21) *and $\overline{H}$ are isolated. Then, $L(\{w_n\}_{n \geq 0})$ is almost surely a equilibrium of $\overline{H}$.*

Proof. We begin exactly as in the proof of Proposition 8.2. Write $w_n = (w_{1,n}, w_{2,n}) \in \mathbf{R}^{m_1} \times \mathbf{R}^{m_2}$. Using Proposition 8.1 we see that $L(\{w_n\}_{n\geq 0}) = \{w_1{}^*\} \times L'$, where $w_1{}^*$ is an equilibrium of (21) and $L' \subset \mathbf{R}^{m_2}$ is a compact connected set invariant under the dynamics of

$$\frac{dw_2}{dt} = -\nabla \overline{E_2}(w_1^*, w_2). \tag{25}$$

Since $L(\{w_n\}_{n\geq 0})$ is *internally* chain-recurrent, every point of L' has to be chain-recurrent for the flow induced by (25). Since (25) is a gradient vectorfield with isolated equilibria, it follows from Proposition 3.2 that L' consists of equilibria. By connectedness, L' is an equilibrium of (25). □

9. Appendix.

Proof of Proposition 2.2. We denote by $\mathcal{F}_n^m$ the σ field generated by $\{\xi_i; n \leq i \leq m\}$ for $m \geq n \geq 0$ and let $\mathcal{F}_0^n = \{\emptyset, \Omega\}$ for $n < 0$.

DEFINITION 9.1. *Let $\{X_n\}_{n\geq 0}$ be a sequence of random variables belonging to $L^2(\Omega)$. $\{X_n\}_{n\geq 0}$ is said to be a mixingale process if there are sequences of finite nonnegative constants $\{c_n\}_{n\geq 0}$ and $\{\psi_m\}_{m\geq 0}$, where $\lim_{n\to\infty} \psi_m = 0$, such that for all $n \geq 1$ and $m \geq 0$*

- a) $\|E(X_n/\mathcal{F}_0^{n-m})\|_2 \leq c_n.\psi_m$,
- b) $\|X_n - E(X_n/\mathcal{F}_0^{n+m})\|_2 \leq c_n.\psi_{m+1}$.

Throughout this section we will only consider sequence of random variables $\{X_n\}_{n\geq 0}$ such that each X_n is measurable $\mathcal{F}_0^n$ so that condition b) holds automatically.

The following lemma relates the concept of mixing process to that of mixingale. It is due to McLeish (1975, Lem. 2.1).

LEMMA 9.2 (McLeish (1975)). *Suppose that $\{\xi_n\}_{n\geq 0}$ is a ϕ mixing (respectively, α mixing) process. Let $\{X_n\}_{n\geq 0}$ be a sequence of random variables such that each X_n is measurable $\mathcal{F}_0^n$ and $E(X_n) = 0$.*

Then, for $2 \leq r \leq +\infty$, $n, m \geq 0$,

- a) $\|E(X_n/\mathcal{F}_0^{n-m}\|_2 \leq 2\phi_m^{1-1/r}\|X_n\|_r$,
- b) $\|E(X_n/\mathcal{F}_0^{n-m}\|_2 \leq 2(1+\sqrt{2})\alpha_m^{1/2-1/r}\|X_n\|_r$.

REMARK. *It follows from this lemma that $\{X_n\}_{n\geq 0}$ is a mixingale with $c_n = \|X_n\|_r$ and $\psi_m = 2\phi_m^{1-1/r}$ in the ϕ mixing case or $\psi_m = 2(1+\sqrt{2})\alpha_m^{1/2-1/r}$ in the α mixing case.*

The following lemma is the main result of this section. The proof of the lemma follows closely the proof of McLeish's Theorem 1.6 (1975), but instead of using Doob's inequality for martingales (as McLeish does) we use the Burkholder inequality together with the ideas involved in the proof of Métivier and Priouret's Proposition 8 (1987). Note that for $q = 2$ the lemma is a direct consequence of McLeish's Theorem 1.6 (1975).

LEMMA 9.3. *Let $\{X_n\}_{n\geq 0}$ be a sequence of real random variables. Let $S_n = \sum_{k=0}^n \gamma_k.X_k$. We assume that*

- *X_n is measurable $\mathcal{F}_0^n$;*
- *$X_n \in L^q(\Omega)$ for some $q \geq 2$;*
- *$\{X_n\}_{n\geq 0}$ is a mixingale with sequences $\{c_n\}_{n\geq 0}$ and $\{\psi_m\}_{m\geq 0}$.*

We assume that there exists a positive decreasing sequence $\{a_n\}_{n\geq -1}$ such that

- $\sum_{k\geq 0} \psi_k^2(a_k^{-1} - a_{k-1}^{-1}) < +\infty$;
- $\sum_{k\geq 0} a_k^{1/(q-1)} < +\infty$.

Then

$$E\left(\sup_{m\le n} \mid S_m \mid^q\right) \le \tau_{n+1}^{q/2-1} D(q) \sum_{i=0}^{n} \gamma_i^{1+q/2} c_i^2 \| X_i^{q-2} \|_\infty,$$

where $D(q) > 0$ *is a constant.*

Proof. Let $Z_{i,k} = E(X_i/\mathcal{F}_0^{i-k}) - E(X_i/\mathcal{F}_0^{i-k-1})$ for $i \ge 0, k \ge 0$. Let $Y_{n,k} = \sum_{i=0}^{n} \gamma_i Z_{i,k}$. Since $E(X_i/\mathcal{F}_0^i) = X_i$ and $E(X_i/\mathcal{F}_0^{i-k}) = E(X_i) = 0$ for $k > i$, it is clear that

$$X_i = \sum_{k=0}^{i} Z_{i,k} = \sum_{k\ge 0} Z_{i,k}.$$

Thus $S_n = \sum_{k\ge 0} Y_{n,k}$ and by Hölder's inequality we have

$$\mid S_n \mid^q \le \sum_{k\ge 0} \frac{\mid Y_{n,k} \mid^q}{a_k} \left(\sum_{k\ge 0} a_k{}^l\right)^{1/l}, \tag{26}$$

where $l = \frac{1}{q-1}$.

Observe that $E(Z_{n,k}/\mathcal{F}_0^{n-k}) = 0$. So by Burkholder and Rosenthal's inequalities (Hall and Heyde (1980, Thms. 2.10–2.12)):

$$E\left(\sup_{m\le n} \mid Y_{m,k} \mid^q\right) \le C(q) E\left(\left(\sum_{i=0}^{n} \gamma_i^2 Z_{i,k}^2\right)^{q/2}\right). \tag{27}$$

Now, exactly as in Métivier and Priouret (1987), we shall apply the following form of Hölder's inequality:

$$\left(\sum_i \mid \alpha_i \beta_i \mid\right)^u \le \left(\sum_i \alpha_i^{\delta u/(u-1)}\right)^{u-1} \left(\sum_i \alpha_i^{(1-\delta)u} \mid \beta_i \mid^u\right) \tag{28}$$

for $\alpha_i \ge 0, \beta_i \in \mathbf{R}, u > 1, 0 < \delta < 1$.

Applying (28) to (27)) with $\alpha_i = \gamma_i^2, \beta_i = \mid Z_{i,k} \mid^2, u = \frac{q}{2}$, and $\delta = \frac{q-2}{2q}$ we obtain

$$E\left(\sup_{m\le n} \mid Y_{m,k} \mid^q\right) \le C(q) \tau_{n+1}^{(q/2-1)} \sum_{i=0}^{n} E(\mid Z_{i,k} \mid^q) \gamma_i^{1+q/2}, \tag{29}$$

and by using (26), we deduce

$$E\left(\sup_{m\le n} \mid S_m \mid^q\right) \le \sum_{k\ge 0} (a_k^l)^{1/l} C(q) \tau_{n+1}^{(q/2-1)} \sum_{k\ge 0} a_k^{-1} \sum_{i=0}^{n} E(\mid Z_{i,k} \mid^q) \gamma_i^{1+q/2}. \tag{30}$$

On the other hand we have

$$E(\mid Z_{i,k} \mid^q) \le \| Z_{i,k}^{q-2} \|_\infty E(\mid Z_{i,k} \mid^2),$$

and from the Pythagorean theorem in $L^2(\Omega)$,

$$E(\mid Z_{i,k} \mid^2) = \| E(X_i/\mathcal{F}_0^{i-k}) \|_2^2 - \| E(X_i/\mathcal{F}_0^{i-k-1}) \|_2^2.$$

It follows that

$$\sum_{k\geq 0} a_k^{-1} \sum_{i=0}^{n} E(|\, Z_{i,k}\,|^q)\gamma_i^{1+q/2} \leq$$

$$\sum_{k\geq 0} \psi_k^2(a_k^{-1} - a_{k-1}^{-1}).\sum_{i=0}^{n} \gamma_i^{1+q/2} c_i^2 \|X_i^{q-2}\|_\infty. \tag{31}$$

Putting together (30) and (31) and letting $D(q) = \sum_{k\geq 0}(a_k^l)^{1/l} C(q) \sum_{k\geq 0} \psi_k^2(a_k^{-1} - a_{k-1}^{-1})$ conclude the proof. □

From this lemma we shall deduce the following lemma, which is analogous to Corollary 11 of Métivier and Priouret (1987).

LEMMA 9.4. *Let* $\{X_n\}_{n\geq 0}$ *be as in Lemma* 9.3. *Then*

- a) *for all* $T \geq 0$,

$$E\left(\sup_{\tau_n\geq\tau_p}\left(\sup_{0\leq\tau_k-\tau_n\leq T}\left|\sum_{i=n}^{k-1}\gamma_i X_i\right|^q\right)\right) \leq T^{q/2-1} D(q) \sum_{i\geq p} \gamma_i^{1+q/2} c_i^2 \|X_i^{q-2}\|_\infty;$$

- b) *if* $\sum_{i\geq 0} \gamma_i^{1+q/2} c_i^2 \|X_i^{q-2}\|_\infty < +\infty$, *then*

$$\lim_{n\to\infty}\left(\sup_{\{0\leq\tau_k-\tau_n\leq T\}}\left|\sum_{i=n}^{k-1}\gamma_i X_i\right|\right) = 0$$

with probability one.

The proof is exactly the same as the proof of Corollary 11 in Métivier and Priouret (1987).

COROLLARY 9.5. *Suppose that* $\{\xi_n\}_{n\geq 0}$ *is a* ϕ *mixing (respectively,* α *mixing) process. Let* $\{X_n\}_{n\geq 0}$ *be a sequence of random variables such that*

- *each* X_n *is measurable* $\mathcal{F}_0^n$ *and* $E(X_n) = 0$;
- $\sup_n \|X_n\|_r < +\infty$ *for some* $r \in [0, +\infty]$;
- *if* $r < \infty$, $\phi_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{r}{2r-2}$ *(respectively,* $\alpha_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{r}{r-2}$*) and* $\sum_{n=0}^{\infty} \gamma_n^2 < +\infty$;
- *if* $r = \infty$, $\phi_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{1}{2}$ *(respectively,* $\alpha_n = O(\frac{1}{n^\beta})$ *for some* $\beta > 1$*) and* $\sum_{n=0}^{\infty} \gamma_n^{1+q/2} < +\infty$ *for some* $q \in [2, 2\beta + 1[$.

Then

- a) $\lim_{n\to\infty}(\sup_{k;\, 0\leq\tau_k-\tau_n\leq T} |\sum_{i=n}^{k-1} \gamma_i X_i\,|) = 0$;
- b) *there exists a constant* $D(q, r)$ *such that*

$$E\left(\sup_{k;\, 0\leq\tau_k-\tau_n\leq T}\left|\sum_{i=n}^{k-1}\gamma_i X_i\right|^q\right) \leq T^{q/2-1} D(q,r) \sum_{i=n}^{m(\tau_n+T)-1} \gamma_i^{1+q/2},$$

where $q = 2$ *for* $r < +\infty$ *and* $q \in [2, 2\beta + 1[$ *for* $r = +\infty$.

Proof. The proof follows from Lemma 9.2, the remark which follows Lemma 9.2, and Lemma 9.3 for part (a) and the Lemma 9.4 for part (b). □

We are now able to prove Proposition 2.2. Let $\{w_n\}_{n\geq 0}$ be a solution to (1). Let

$$u_n(w) = H(w, \xi_n) - E(H(w, \xi_n)),$$

$$u_n = u_n(w_n),$$

$$b_n(w) = E(H(w, \xi_n)) - \overline{H}(w),$$

$$b_n = b_n(w_n).$$

We suppose that conditions A1 to A6 of Proposition 2.2 hold. Let $K = \sup_n \|k(\xi_n)\|_r$. We remark that $\overline{H}$ is K Lipschitz. Indeed, for any x, y,

$$\|\overline{H}(x) - \overline{H}(y)\| = \lim_{n\to\infty} \|E(H(x, \xi_n) - H(y, \xi_n))\|$$

but

$$\|E(H(x, \xi_n) - H(y, \xi_n))\| \leq E(k(\xi_n))\|x - y\| \leq K\|x - y\|$$

by application of Jensen's inequality. Using the same kind of argument we see that

$$\|b_n(x) - b_n(y)\| \leq 2K\|x - y\|, \tag{32}$$

$$\|u_n(x) - u_n(y)\| \leq K\|x - y\| + k(\xi_n)\|x - y\|. \tag{33}$$

Let us now verify the conditions A1, A2, and A3 of Theorem 1.2. A1 is true by assumption (assumption A2 of Proposition 2.2). The inequality (33) shows that the sequence $\{b_n(.)\}_{n\geq 0}$ is equicontinuous, and assumption A3 of the Proposition 2.2 means that $\{b_n(w)\}_{n\geq 0}$ converges to zero for any $w \in \mathbf{R}^m$. It follows that $\{b_n(.)\}_{n\geq 0}$ converges to zero uniformly on any compact set of $\mathbf{R}^m$, and as $\{w_n\}_{n\geq 0}$ is assumed to be bounded, $\{b_n\}_{n\geq 0}$ converges to zero. This proves assumption A2 of Theorem 1.2.

We now check the condition A3 of Theorem 1.2. Let $T \geq 0$ and let

$$S_n^T(w) = \sup_{\{k,\ 0\leq \tau_k - \tau_n \leq T\}} \left| \sum_{i=n}^{k} \gamma_i u_i(w) \right|,$$

$$Z_n^T = \sup_{\{k,\ 0\leq \tau_k - \tau_n \leq T\}} \sum_{i=n}^{k} \gamma_i k(\xi_i).$$

From inequalities (32) and (33) it follows that for any $w, w' \in \mathbf{R}^d$

$$\left| \sum_{i=n}^{k} \gamma_i u_i(w) \right| \leq \left| \sum_{i=n}^{k} \gamma_i u_i(w') \right| + K|w - w'| \sum_{i=n}^{k} \gamma_i + |w - w'| \sum_{i=n}^{k} \gamma_i k(\xi_i).$$

Thus

$$0 \leq S_n^T(w) \leq S_n^T(w') + TK|w - w'| + Z_n^T|w - w'|. \tag{34}$$

Lemma 9.4(b) shows that under the assumptions A1 and A6, A6′ of Proposition 2.2, $\lim_{n\to\infty} Z_n^T = 0$ and $\lim_{n\to\infty} S_n^T(w) = 0$ for any $w \in \mathbf{R}^m$. From inequality (34) it is easy to see that $\{S_n^T(w)\}_{n\geq 0}$ converges to zero uniformly on any compact set of $\mathbf{R}^m$ so that $\lim_{n\to\infty} S_n^T(w_n) = 0$. □

Proof of Proposition 5.6. Let Q be a compact set. Let $\epsilon_n > 0$ such that $\lim_{n\to+\infty} \epsilon_n = 0$ and let $\{B(x_i, \epsilon_n)\}_{i\in I_n}$ be a finite cover of Q by balls of radius ϵ_n. From (34), we deduce

$$\sup_{w\in Q} S_n^T(w) \leq \sum_{i\in I_n} S_n^T(x_i) + KT\epsilon_n + \epsilon_n Z_n^T.$$

Therefore, Corollary 9.5(b) implies

$$\left\| \sup_{w\in Q} S_n^T(w) \right\|_q \leq (A.\#(I_n) + B\epsilon_n) \left(\sum_{i=n}^{m(\tau_n+T)-1} \gamma_i^{1+q/2} \right)^{1/q} + KT\epsilon_n \tag{35}$$

for some constants A and B depending on $Q, q, T, \overline{H}$, and $\#(I_n)$ denotes the cardinal of I_n.

Since Q is a compact subset of $\mathbf{R}^m$, the family $\{x_i\}_{i\geq 0}$ can be chosen such that $\#(I_n) \leq C.\epsilon_n^{-m}$ for some $C > 0$. Noting that

$$\sum_{i=n}^{m(\tau_n+T)-1} \gamma_i^{1+q/2} \leq T\gamma_n^{q/2},$$

inequality (35) gives

$$\left\| \sup_{w\in Q} S_n^T(w) \right\|_q \leq (A.C.\epsilon_n^{-m} + B\epsilon_n)T\gamma_n^{q/2} + KT\epsilon_n.$$

Therefore, with α_q^T as defined in §5,

$$\begin{aligned} l_\tau(\alpha^{q,T}) &\leq \sup\left(\frac{1}{2}l_\tau(\gamma) - ml_\tau(\epsilon), l_\tau(\epsilon), \frac{1}{2}l_\tau(\gamma) + l_\tau(\epsilon)\right) \\ &\leq \sup\left(\frac{1}{2}l_\tau(\gamma) - ml_\tau(\epsilon), l_\tau(\epsilon)\right). \end{aligned} \tag{36}$$

Now we choose a sequence $\{\epsilon_n\}_{n\geq 0}$ such that

$$l_\tau(\epsilon) = \min_{\{l<0\}} \sup\left(l, \frac{1}{2}l_\tau(\gamma) - ml\right).$$

This easily gives

$$l_\tau(\epsilon) = \frac{l_\tau(\gamma)}{2(1+m)}.$$

For example, one may choose $\epsilon_n = \gamma_n^{1/(2+2m)}$.

Acknowledgment. I am particulary grateful to Morris W. Hirsch for many comments and fruitful discussions on the topics in this paper. A large part of this work makes a crucial use of some of his ideas and results on "shadowing."

REFERENCES

M. BENAIM (1994a), *Un théorème de Poincaré–Bendixson pour une classe d'equations différentielles stochastiques*, C. R. Acad. Sci. Paris Sér. I, 318, pp. 837–839.

——— (1994b), *On functional approximation with normalized Gaussian units*, Neural Computation, 6, pp. 319–333.

M. BENAIM AND M. W. HIRSCH (1995a), *Dynamics of Morse–Smale urns processes*, Ergodic Theory Dynamical Systems, to appear.

——— (1995b), *Asymptotic pseudo-trajectories, chain-recurrent flows and stochastic approximations*, J. Dynamics Differential Equations, to appear.

——— (1995c), *Chain recurrence in surface flows*, Discrete and Continuous Dynamics Systems, 1 (1995), pp. 1–17.

M. BENAIM AND L. TOMASINI (1991), *Competitive and self-organizing algorithms based on the minimization of an information criterion*, in Artificial Neural Networks I, Vol. 1, T. Kohonen, ed., North-Holland, Amsterdam, pp. 391–396.

——— (1992), *Approximating function and predicting time series with multisigmoidal basis functions*, in Artificial Neural Networks II, Vol. 1, I. Aleksander and J. Taylor, eds., Elsevier, Amsterdam, pp. 407–411.

A. BENVENISTE, M. MÉTIVIER, AND P. PRIOURET (1990), *Stochastic Approximations and Adaptive Algorithms*, Springer-Verlag, Berlin, Heidelberg, New York. Translated from *Algorithmes adaptatifs et approximations stochastiques*, Masson, Paris, 1987.

P. BILLINGSLEY (1968), *Convergence of Probability Measures*, Wiley, London, New York.

C. C. CONLEY (1978), *Isolated Invariant Sets and the Morse Index*, Regional Conference Series in Mathematics 38, American Mathematical Society, Providence, RI.

E. EWEDA AND O. MACCHI (1983), *Quadratic mean and almost sure convergence of unbounded stochastic approximation algorithms with correlated observations*, Ann. Inst. H. Poincaré, 19, pp. 235–255.

J. C. FORT AND G. PAGÈS (1994), *Convergences d'algorithmes stochastiques: le théorème de Kushner et Clark revisité*, pré-pub. labo. de proba., Univ. Paris 6.

E. G. GLADYSHEV (1965), *On stochastic approximation*, Theory Probab. Appl., 10, pp. 275–278.

D. HALL AND C. C. HEYDE (1980), *Martingale Limit Theory and Applications*, Academic Press, New York.

M. W. HIRSCH (1993), *Asymptotic phase, shadowing and reaction-diffusion systems*, in Control Theory, Dynamical Systems and Geometry of Dynamics, K. D. Elworthy, W. N. Everitt, and E. B. Lee, eds., Marcel Dekker, New York.

M. W. HIRSCH AND C. PUGH (1970), *Stable manifolds for hyperbolic sets*, in Global Analysis, Proceedings of Symposia in Pure Mathematics 14, American Mathematical Society, Providence, RI, pp. 133–163.

J. KIEFER AND J WOLFOWITZ (1952), *Stochastic estimation of the maximum of a regression function*, Ann. Math. Statist., 23, pp. 462–466.

C. M. KUAN AND H. WHITE (1992), *Artificial neural networks: An econometric perspective*, Econometric Reviews, to be published.

H. J. KUSHNER AND D. S. CLARK (1978), *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, Berlin, Heidelberg, New York.

L. LJUNG (1977), *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22, pp. 551–575.

D. L. MCLEISH (1975), *A maximal inequality and dependent strong laws*, Ann. Probab., 3, pp. 829–839

M. MÉTIVIER AND P. PRIOURET (1984), *Application of a Kushner and Clark lemma to general classes of stochastic algorithms*, IEEE Trans. Inform. Theory, IT-30, pp. 140–150

——— (1987), *Théorèmes de convergence presque sure pour une classe d'algorithmes stochastiques à pas décroissant*, Probab. Theory Related Fields, 74, pp. 403–428.

J. MOODY AND C. DARKEN (1989), *Fast learning in networks of locally-tuned processing units*, Neural Computation, 1, pp. 281–582.

M. B. NEVEL'SON AND R. Z. HAS'MINSKII (1974), *Stochastic Approximation and Recursive Estimation*, Translation of Math. Monographs 47, American Mathematical Society, Providence, RI.

J. NEVEU (1964), *Bases mathematiques du calcul des probabilites*, Masson, Paris.

S. NOWLAN (1990), *Maximum likelihood competitive learning*, in Proceedings of Neural Information Processing Systems, pp. 574–582.

J. PALIS (1969), *On Morse–Smale dynamical systems*, Topology, 8, pp. 385–405.

R. PEMANTLE (1990), *Nonconvergence to unstable points in urn models and stochastic approximations*, Ann. Probab., 18, pp. 698–712.

T. POGGIO AND F. GIROSI (1990), *Regularization algorithms for learning that are equivalent to multilayer networks*, Science, 247, pp. 979–982.

H. ROBBINS AND S. MONRO (1951), *A stochastic approximation method*, Ann. Math. Statist., 22, pp. 400–407.

C. ROBINSON (1977), *Stability theorems and hyperbolicity in dynamical systems*, Rocky J. Math., 7, pp. 425–434.

M. SHUB (1986), *Global Stability of Dynamical Systems*, Springer-Verlag, New York.

H. WHITE (1992), *Artificial Neural Networks, Approximation and Learning Theory*, Blackwell, Oxford, Cambridge.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 473–490, March 1996

004

NONLINEAR BOUNDARY CONTROL OF SEMILINEAR PARABOLIC SYSTEMS*

N. U. AHMED† AND X. XIANG‡

Abstract. Nonlinear boundary control problems for a class of semilinear parabolic systems are considered, from the point of view of semigroup theory. The method is based on some recent general results on parabolic evolution equations with nonlinear boundary conditions. Existence of optimal (boundary) controls is proved using the theory of measurable selections and the Cesari property for multifunctions. Three results are presented covering relaxed controls and controls with state constraints. This generalizes, in a substantial way, existing results on linear boundary control problems [M.C. Delfour and M. Sorine, *Control of Distributed Parameter Systems*, Pergamon Press, Oxford, 1983, pp. 87–90], [I. Lasiecka, *Appl. Math. Optim.*, 6 (1980), pp. 287–383], [P. Acquistapace, et al., *SIAM J. Control Optim.*, 29 (1991), pp. 89–118]. The result presented can be further extended to differential inclusions. Two examples are presented for illustration.

Key words. semilinear parabolic system, nonlinear boundary controls, optimal control, existence, multifunctions, measurable selections

AMS subject classifications. 49A, 49B, 34H, 35K

Introduction. Boundary control of systems governed by partial differential equations is one of the most important problems in control theory; see [1]–[7], [11]. In the case of parabolic equations, both variational and semigroup methods have been successfully applied (see, for instance, [4]–[7]). Most of these papers deal with linear quadratic regulator problems for autonomous and nonautonomous linear parabolic equations giving feedback controls via the associated Riccati equations.

This paper is concerned with the question of the existence of optimal boundary controls for semilinear problems with nonlinear boundary operators containing controls. For motivation we present Example 2 (see §3) arising in steel manufacturing, where specific power nonlinearities appear in the boundary data. Nonlinear boundary control problems are difficult, in general. To the best of the authors' knowledge, nonlinear boundary control problems in this very general setting have not been widely considered in the literature. Some results in this direction are available in [11], where both identification and relaxed control problems have been considered using different techniques. Here we prove the existence of optimal boundary controls using selection theorems for measurable multifunctions by following an approach similar to that in [2], [12]. In the case of nonlinear problems, the classical technique based on lower semicontinuity and compactness arguments does not apply unless restrictive assumptions are imposed on the nonlinear boundary operator, such as control linearity or weak continuity. In fact for nonlinear problems, weak compactness of solution trajectories, compact embeddings, and selection arguments have become the standard tools (see [2], [12], [16], [17]). In contrast for linear problems with quadratic cost, existence results and necessary conditions of optimality can be obtained simultaneously and relatively easily through the study of Riccati equations [5]–[7]. For nonlinear problems also, existence of optimal feedback controls and necessary conditions of optimality can

*Received by the editors September 2, 1993; accepted for publication (in revised form) October 31, 1994.

†Department of Electrical Engineering and Department of Mathematics, University of Ottawa, Ottawa K1N 6N5, Canada.

‡Department of Mathematics, Guizhu University, Guizhu, People's Republic of China.

be obtained provided viscosity solutions of the associated Hamilton–Jacobi–Bellman equation on infinite-dimensional spaces are better understood. This is a very difficult problem and not well developed as yet.

In §1, using some recent results of Amann [8], we derive a representation formula for the solution of abstract semilinear evolution equations with nonlinear boundary operators containing controls. Since the notation of Amann [8] is rather involved, we review only some relevant results for the convenience of our readers. Section 2 is devoted to the question of the existence of optimal boundary controls, which is the main concern of this paper. Here we present three existence results: Theorem 2.1 for ordinary controls, Theorem 2.2 for controls with state constraints, and Theorem 2.3 for relaxed controls. We give a complete proof of the first result. In the last section, we first present an example of a system of second-order semilinear parabolic equations with nonlinear boundary controls illustrating the applicability of our results. Then we present a practical example that arises in steel manufacturing.

1. Abstract parabolic systems with nonlinear boundary controls. In this section, we study a class of abstract evolution equations corresponding to semilinear parabolic systems with nonlinear boundary operators containing controls. Based on the results of Amann [8], we establish a representation formula for the solutions which is useful from the point of view of control theory.

Throughout this paper all vector spaces are assumed over the complex field.

Let X and Y be locally convex Hausdorff topological vector spaces. We denote by X' the dual of X endowed with the strong topology and by $\langle \bullet, \bullet \rangle_X : X' \times X \to C$, the corresponding duality pairing. Observe that $\langle u, v \rangle_X = \langle v, u \rangle_{X'}$ if X is reflexive.

We write $X \hookrightarrow Y (X \hookrightarrow\hookrightarrow Y)$ to denote continuous (dense, compact) embeddings.

If $X \hookrightarrow Y$ and $B : D(B) \subset Y \to Y$ is a linear operator in Y, we define the X-realization B_X of B to be the linear operator in X given by

$$D(B_X) = \{x \in D(B) \cap X, Bx \in X\}, \quad B_X x = Bx, \text{ for } x \in D(B_X).$$

In general we denote by $\mathcal{L}(X, Y)$ the vector space of all continuous linear operators from X to Y. Moreover $\mathcal{L}(X) = \mathcal{L}(X, X)$ and $\mathcal{L}(X, Y)$ are given the usual norm topologies if X and Y are normed vector spaces. Further, $\mathcal{L}^2(X \times Y, C)$ denotes the class of complex-valued bilinear forms on $X \times Y$.

We use the notation of Amann [8]. Suppose $W, W^1, W_\#, W^1_\#$ are Banach spaces. W is reflexive, $W_\# = W'$. ∂W, ∂W^1, $\partial_1 W^1$ as well as $\partial W_\#$, $\partial W^1_\#$, $\partial_1 W^1_\#$ are reflexive Banach spaces such that $\partial W_\# = (\partial W)'$. The topological vector spaces mentioned above satisfy the assumptions (A_1)–(A_4) of Amann [8] (see also [1, pp. 206–209]). Formally the system is governed by the following semilinear initial boundary value problem:

$$\begin{aligned} \dot{x} + \mathcal{A}(t)x &= f(t, x), \\ \mathcal{B}(t)x &= g(t, x, u), \quad 0 \le t \le T \equiv I, \\ x(0) &= x_0, \end{aligned} \tag{1.1}$$

where x denotes the state and u denotes the control to be defined precisely later. The basic assumptions are as follows.

The spatial operator and its formal adjoint satisfy the following assumption.

Assumption A_1.

$$\begin{aligned} &\mathcal{A}(\bullet) \in C([0, T], \mathcal{L}(W^1, W)), \\ &\mathcal{A}^\#(\bullet) \in C([0, T], \mathcal{L}(W^1_\#, W_\#)). \end{aligned} \tag{1.2}$$

The boundary operator $\mathcal{B}(.)$ along with its complimentary counterpart $\ell(.)$, as defined below, and the corresponding duals satisfy the following basic assumptions.

Assumption B_1.

$$\begin{aligned} &\mathcal{T}(\bullet) \equiv (\mathcal{B}(\bullet), \ell(\bullet)) \in C([0,T], \mathcal{L}(W^1, \partial W^1 \times \partial_1 W^1)), \\ &\mathcal{T}^{\#}(\bullet) = (\mathcal{B}^{\#}(\bullet), \ell^{\#}(\bullet)) \in C([0,T], \mathcal{L}(W^1_{\#}, \partial W^1_{\#} \times \partial_1 W^1_{\#})). \end{aligned} \tag{1.3}$$

The operators $\mathcal{T}(\bullet), \mathcal{T}^{\#}(\bullet)$ are retractions such that the following Green's formula holds for all $u \in W^1, v \in W^1_{\#}$:

$$\langle v, \mathcal{A}(t)u\rangle + \langle \ell^{\#}(t)v, \mathcal{B}(t)u\rangle_{\partial} = \langle \mathcal{A}^{\#}(t)v, u\rangle + \langle \mathcal{B}^{\#}(t)v, \ell(t)u\rangle_{\partial'}, \quad t \geq 0. \tag{1.4}$$

Define

$$\begin{aligned} &\mathring{W}^1 \equiv \operatorname{Ker}(\mathcal{T}) \quad \text{and} \quad \mathring{W}^1_{\#} \equiv \operatorname{Ker}(\mathcal{T}^{\#}), \\ &W^1_{\mathcal{B}} \equiv \operatorname{Ker}(\mathcal{B}) \quad \text{and} \quad W^1_{\#\mathcal{B}^{\#}} \equiv \operatorname{Ker}(\mathcal{B}^{\#}). \end{aligned} \tag{1.5}$$

Introduce the operators

$$A(t) = \mathcal{A}(t)|_{W^1_{\mathcal{B}}} \quad \text{and} \quad A^{\#}(t) = \mathcal{A}^{\#}(t)|_{W^1_{\#\mathcal{B}^{\#}}} \tag{1.6}$$

and consider these as (unbounded) linear operators in W and $W^{\#}$, respectively. We assume that these operators satisfy all the Assumptions AP1–AP3 of Amann [8, §11].

We present some of the relevant assumptions of Amann.

Assumption A_2. $A(t), t \in I$, is a family of closed and densely defined linear operators in $E = W$ such that there exist constants M_0 and $\sigma \in R$ with $\rho(-A(t)) \supset \sigma + \Sigma_0$, and $\|(\lambda + A(t))^{-1}\| \leq M_0/(1+|\lambda|)$. It is well known that (see [3]) under this assumption, for each fixed $t \geq 0$, $A(t)$ is the infinitesimal generator of an analytic semigroup on E.

In the following we let

$$\|x\|_{k,t} = \|A^k(t)x\|, \quad x \in D(A^k(t)), \quad k \in Z,$$

and define $E_k(t) = (D(A^k(t)), \|\bullet\|_{k,t})$ if $k \geq 0, E \equiv E_0$, and let $E_k(t)$ be the completion of $(E, \| \bullet \|_{k,t})$ if $k < 0$. Moreover we denote by $A_k(t)$ the $E_k(t)$-realization of $A(t)$ if $k \geq 0$, and the closed extension of $A(t)$ in E_k if $k < 0$. Finally we let $E_\alpha(t) = (E_k(t), E_{k+1}(t))_{\alpha-k}$, $k < \alpha < k+1$, $k \in Z$, where $(\bullet, \bullet)_{\alpha-k}$ are standard interpolation functors (see [1]).

We denote by $A_\alpha(t)$ the $E_\alpha(t)$-realization of $A_k(t)$ for $k < \alpha < k+1$ and $k \in Z$.

For uniform notation we set

$$W^{\alpha}_{\mathcal{B}(t)} = E_\alpha(t), \quad W^{\alpha}_{\#\mathcal{B}^{\#}(t)} = E^{\#}_\alpha(t), \quad \alpha \in R, \ t \in I.$$

Assumption A_3. There exists $\beta \in [0,1]$ such that

$$E_\beta(t) = E_\beta(0) = E_\beta, \quad E_{\beta-1}(t) = E_{\beta-1}.$$

Assumption A_4. There exist $\rho \in (0,1)$ and

$$a(\bullet) \in C^\rho([0,T], \mathcal{L}^2(E^{\#}_{1-\beta} \times E_\beta, C)), \quad \beta \in [0,1],$$

such that

$$a(t)(v,u) = \langle v, A(t)u\rangle, \quad (u,v) \in E_\beta(t) \times E^{\#}_{1-\beta}(t). \tag{1.7}$$

Observe that $E^{\#}_{1-\beta} = E^{\#}_{1-\beta}(t)$ is independent of t for $\beta \in [0,1]$ since $E^{\#}_{1-\beta} = (E_{\beta-1})'$ (see Theorem 8.1 of [8]). Also, we assume that $a(\bullet)$ is independent of $\beta \in [0,1]$ (in an obvious sense). The assumptions above imply, in particular, that the following condition is also satisfied independently of $t \geq 0$.

Assumption A_5. There exists $\sigma \in C$ such that $\sigma \in \rho(-A) \cap \rho(-A^{\#})$, where $\rho(\bullet)$ denotes the resolvent set.

It follows from Theorem 6.3 of [8] that, for $0 \leq \theta \leq 1$, the Dirichlet map $R_\theta(\bullet)$ has the following regularity property:

$$R_\theta(\bullet) \equiv (\sigma + \mathcal{A}(\bullet), \mathcal{B}(\bullet))^{-1}|_{\{0\}\times\partial W^{-1+2\theta}} \in C([0,T], \mathcal{L}(\partial W^{-1+2\theta}, W^\theta)). \tag{1.8}$$

Further it is clear that $R_\alpha(t) = R_\beta(t)|_{\partial W^{-1+2\alpha}}$, $\beta < \alpha \leq 1$, where $\partial W^{-1+2\theta} = (\partial W^{-1}, \partial W^1)_\theta$, $0 \leq \theta \leq 1$. We now fix for each $\theta \in [0,1]$ a closed linear subspace $\partial_0 W^{-1+2\theta}$ of $\partial W^{-1+2\theta}$ such that $\partial_0 W^{-1+2\alpha} \subset \partial_0 W^{-1+2\beta}$, for $\alpha > \beta$, and that $R_\beta(t)(\partial_0 W^{-1+2\beta}) \subset W^\beta_{\mathcal{B}}$, for $t \in I$.

Assumption B_2. $[(y,z) \to \langle \ell^{\#}(t)y, z\rangle_\partial] \in \mathcal{L}^2(W^{1-\beta}_{\#\mathcal{B}^\#} \times \partial_0 W^{-1+2\beta}, C)$ for each $\beta \in [0,1]$ and $t \in I$.

Assumption F_1. For $0 \leq \beta < \alpha \leq 1$, $f \in C^{0,1-}([0,T] \times W^\beta_{\mathcal{B}}, W^{\alpha-1}_{\mathcal{B}})$.

Assumption G_1. The control space U is a Banach space and

$$g(\bullet,\bullet,\bullet) : [0,T] \times W^\beta_{\mathcal{B}} \times U \to \partial_0 W^{-1+2\alpha}$$

is a map so that, for each $u \in U$,

$$g(\bullet,\bullet,u) \in C^{0,1-}([0,T] \times W^\beta_{\mathcal{B}}, \partial_0 W^{-1+2\alpha}).$$

We consider, for each $x_0 \in W^\beta_{\mathcal{B}}$ and $u \in L_\infty([0,T], U)$, the state equation given by the abstract semilinear initial boundary value problem

$$\begin{aligned} \dot{x} + \mathcal{A}(t)x &= f(t,x), \\ \mathcal{B}(t)x &= g(t,x,u), \quad t \in I, \\ x(0) &= x_0. \end{aligned} \tag{1.9}$$

By a solution x of (1.9) on $I \equiv [0,T]$ we mean a function $x \in C([0,T], W^\beta_{\mathcal{B}})$ such that

$$\begin{aligned} &\int_0^{T'} \{-\langle \dot{\varphi}, x\rangle + a(t)(\varphi, x)\}dt \\ &\quad = \int_0^{T'} \{-\langle \varphi, f(t,x)\rangle + \langle \ell^{\#}(t)(\varphi), g(t,x,u)\rangle_\partial\}dt + \langle \varphi(0), x_0\rangle \end{aligned} \tag{1.10}$$

for every $T' \in (0,T]$ and every $\varphi \in C([0,T'], W^{1-\beta}_{\#\mathcal{B}^\#}) \cap C^1([0,T'], W^{-\beta}_{\#\mathcal{B}^\#})$ satisfying $\varphi(T') = 0$.

LEMMA 1.1 (For proof see Theorems 9.1 and 12.1 of [8]). *Suppose $\mathcal{A}, \mathcal{B}, f, g$ satisfy all the assumptions stated above.*

(a) *Then, for each fixed* $u \in U$, *the mapping* $F^u_{\beta-1}$, *defined by*

$$F^u_{\beta-1}(\bullet,\bullet) = f(\bullet,\bullet) + (\sigma + A_{\beta-1}(\bullet))R_\beta(\bullet)g(\bullet,\bullet,u), \tag{1.11}$$

has the regularity property: $F^u_{\beta-1}(\bullet,\bullet) \in C^{0,1-}([0,T] \times W^\beta_{\mathcal{B}}, W^{\alpha-1}_{\mathcal{B}})$.

(b) *There exists a unique parabolic fundamental solution* $U_{\beta-1}$ *on* $E_{\beta-1}$ *for* $\{A_{\beta-1}(t): t \in I\}$ *and it possesses* E_β *as regularity subspace.*

(c) *For each* $u \in L_\infty([0,T],U)$, *an element* $x \in C([0,T],W^\beta_{\mathcal{B}})$ *is a* $W^\beta_{\mathcal{B}}$*-weak solution of* (1.9) *on* I *iff* x *is a solution of the following integral equation:*

$$x(t) = U_{\beta-1}(t,0)x_0 + \int_0^t U_{\beta-1}(t,\tau)F^u_{\beta-1}(\tau,x(\tau))d\tau \tag{1.12}$$

in $C([0,T],W^\beta_{\mathcal{B}})$.

The next theorem gives a useful sufficient condition that guarantees the existence of a solution of the state equation (1.9).

THEOREM 1.2 (For proof see Theorem 12.1 and Proposition 12.6 of [8]). *Suppose the assumptions of Lemma* 1.1 *hold and further there exists a constant* $C > 0$, *possibly dependent on* u, *such that* f *and* g *satisfy the following growth condition:*

$$\|f(t,y)\|_{W^{\alpha-1}_{\mathcal{B}}} + \|g(t,y,u)\|_{\partial W^{-1+2\alpha}} \leq C(1 + \|y\|_{W^\beta_{\mathcal{B}}}) \tag{1.13}$$

for all $(t,y) \in$ *graph* $x(\bullet,x_0,u)$, *which is the maximal solution of* (1.9). *Then* $x(\bullet,x_0,u)$ *is a global solution.*

2. Existence of optimal boundary controls. In this section we are concerned with the boundary control problem for the semilinear abstract evolution equation (1.9). Since our approach is based on the properties of multifunctions, we present below some basic definitions and facts. Let $\mathcal{Z}$ be any locally convex topological vector space and let $c(\mathcal{Z})$, $(cc(\mathcal{Z})$, $cbc(\mathcal{Z})$, $wkc(\mathcal{Z}))$ denote the class of nonempty closed (closed convex, closed bounded convex, weakly compact convex) subsets of $\mathcal{Z}$. A multifunction F mapping a Hausdorff topological space X to $c(Y)$, Y any locally convex topological vector space, is said to be upper (lower) semicontinuous with respect to inclusion if for every $x_0 \in X$ and every open set $V \subset Y$ satisfying $[F(x_0) \subset V]$ $(V \cap F(x_0) \neq \emptyset)$, there exists an open set $U \subset X$ containing x_0 such that $[F(x) \subset V]$ $(F(x) \cap V \neq \emptyset)$ for all $x \in U$. If Y is a metric space with metric d, then one can introduce a metric d_H, called the Hausdorff metric, on $c(Y)$ as follows:

$$d_H(C,D) \equiv \mathrm{Max}\{\mathrm{Sup}_{y\in D}d(C,y), \mathrm{Sup}_{z\in C}d(z,D)\}$$

for $C, D \in c(Y)$. If (Y,d) is complete, then so is $(c(Y),d_H)$. $F : X \mapsto c(Y)$ is said to be continuous in the Hausdorff metric if, whenever $x_n \longrightarrow x$ in the topology of X,

$$\mathrm{Lim}_{n\to\infty}d_H(F(x_n),F(x)) = 0.$$

It is said to be quasi upper semicontinuous if for each $x \in X$

$$d^*(F(x_n),F(x)) \equiv \mathrm{Sup}\{d(y,F(x)), y \in F(x_n)\} \longrightarrow 0,$$

whenever $x_n \longrightarrow x$. In [10] quasi upper semicontinuity is called mild upper semicontinuity. More precisely, if Y is a metric space the two notions are equivalent. For different types of continuity, see [2], [10], [12].

Now, we introduce the class of admissible controls. Let U be a Banach space and let $wkc(U)$ denote the class of nonempty weakly compact convex subsets of U.

Assumption U1. (a) U is a separable Banach space; (b) $V : I \to wkc(U)$ is an integrably bounded measurable multifunction, and

$$\mathcal{U}_{\text{ad}} = \{u : [0,T] \longrightarrow U \text{ weakly (strongly) measurable such that } u(t) \in V(t) \text{ a.e.}\}$$

We consider the following optimal (boundary) control problem which we shall denote by (P):

Minimize

$$J(u) = \int_0^T L(t, x(t), u(t))dt \tag{2.1}$$

over all controls $u \in \mathcal{U}_{\text{ad}}$ subject to the evolution equation (1.9).

We introduce the following assumptions for the function L.

Assumption L. $L : [0,T] \times E_\beta \times U \to R_\infty = R \cup \{+\infty\}$ be a mapping satisfying

(1) $(t,e,u) \to L(t,e,u)$ is Borel measurable,

(2) $(e,u) \to L(t,e,u)$ is lower semicontinuous for each $t \in I = [0,T]$,

(3) $u \to L(t,e,u)$ is convex for all $(t,e) \in I \times E_\beta$,

(4) $\varphi(t) - \lambda(\|e\|_{E_\beta} + \|u\|_U) \leq L(t,e,u)$ a.e. with $\varphi \in L_1(I), \lambda \geq 0$.

Now we are prepared to deal with the question of existence of optimal controls for the problem (P). The following theorem is our main result.

THEOREM 2.1. *Suppose the hypotheses on $\mathcal{A}, \mathcal{B}, f, g, L$, and* U1 *hold, $A^{-1}(t)$ is compact for each $t \in I$, $x_0 \in E_\alpha$, and $g : I \times E_\beta \times U \longrightarrow \partial W^{-1+2\alpha}$ is measurable in the first variable continuous with respect to the last two variables. Suppose also that the multifunction G given by $G(t,z) \equiv g(t,z,V(t))$ maps $I \times E_\beta$ to $cc(\partial_0 W^{-1+2\alpha})$. Then there exists an optimal control $u_0 \in \mathcal{U}_{\text{ad}}$ for problem* (P).

Proof. Note that by virtue of Assumptions U1 and L(4), $J(u) > -\infty$. Let $x(u)$ denote the solution of the initial boundary value problem (1.9) corresponding to an admissible control u and $\mathcal{X} \equiv \{x(u), u \in \mathcal{U}_{\text{ad}}\}$ denote the family of attainable trajectories of the corresponding control system. Define

$$\Xi \equiv \{(u,x) \in \mathcal{U}_{\text{ad}} \times C(I, E_\beta) : x = x(u)\}. \tag{2.2}$$

For $(u,x) \in \Xi$, define $\eta(u,x) \equiv J(u)$. Let $\{u_n, x_n\} \subset \Xi$ be a minimizing sequence, that is, $\lim_{n\to\infty} \eta(u_n, x_n) = m = \inf\{\eta(u,x) : (u,x) \in \Xi\}$.

We show that, through a subsequence if necessary, x_n converges to x^* and that there exists a control $u^* \in \mathcal{U}_{\text{ad}}$ such that $x^* = x(u^*)$ and that $\eta(u^*, x^*) = m$. With this in mind, first we show that the sequence $\{x_n\}$ is compact in $C([0,T], E_\beta)$. Theorem 1.2 shows that for every $u \in L_\infty([0,T], U)$, the controlled system (1.9) has at least one solution $x(\bullet, u) \in C([0,T], E_\beta)$. By virtue of Theorem 8.1 of [8] and a generalized version of the Gronwall inequality, we can prove that

$$\|x(\bullet, u)\|_{C([0,T],E_\beta)} \leq b < \infty,$$

where b is a constant dependent only on $\|x_0\|_{E_\beta}$ and $\|u\|_{L_\infty([0,T],U)}$. Thus it follows from growth condition (1.13), and Assumption U1, that there exists a constant $\tilde{b}$ such that

$$\text{Sup}\{\|x\|_{C([0,T],E_\beta)}, x \in \mathcal{X}\} \leq \tilde{b} < \infty.$$

In other words, the solution set $\mathcal{X}$ is a bounded subset of $C(I, E_\beta)$.

Take γ such that $\beta < \gamma < \alpha < 1 + \gamma$. Let $x_n \equiv x(u_n)$ be any solution of the following integral equation:

$$x_n(t) = U_{\beta-1}(t,0)x_0 + \int_0^t U_{\beta-1}(t,\tau)F_{\beta-1}^{u_n}(\tau, x_n(\tau))d\tau \tag{2.3}$$

in $C([0,T], E_\beta)$. Using the growth assumption on f, g as stated in Theorem 1.2, it follows from Theorem 8.1 of [8] that

$$\begin{aligned}
\|x_n(t)\|_{E_\gamma} &\le \|U_{\beta-1}(t,0)x_0\|_{E_\gamma} + \int_0^t \|U_{\beta-1}(t,s)F_{\beta-1}^{u_n}(s,x_n(s))\|_{E_\gamma} ds \\
&\le \|U_{\beta-1}(t,0)x_0\|_{E_\gamma} + C_1 \int_0^t (t-s)^{\alpha-1-\gamma} \|F_{\beta-1}^{u_n}(s,x_n(s))\|_{E_{\alpha-1}} ds \\
&\le C_2 e^{\sigma t}\|x_0\|_{E_\alpha} + C_3 \int_0^t (t-s)^{\alpha-1-\gamma}[1 + \|x_n(s)\|_{E_\beta}]ds \\
&\le C_4\|x_0\|_{E_\alpha} + C_5 \int_0^t (t-s)^{\alpha-1-\gamma}[1 + \|x_n(s)\|_{E_\gamma}]ds,
\end{aligned}$$

where C_1, C_2, C_3, C_4, C_5 are suitable constants depending only on $(\alpha, \beta, \gamma, \sigma, T, M_0, \mathcal{U}_{\text{ad}})$. Then it follows from a slight generalization of the Gronwall inequality that there exists a constant M^* such that

$$\|x_n(t)\|_{E_\gamma} \le M^*, \quad t \in [0,T],$$

for all positive integers n. Thus the set $\{x_n\}$ is bounded in $C([0,T], E_\gamma)$.

Further, for $0 \le t_2 < t_1 \le T$, we have

$$\begin{aligned}
\| x_n(t_1) - x_n(t_2) \|_{E_\beta} \le & \| (U_{\beta-1}(t_1,0) - U_{\beta-1}(t_2,0))x_0 \|_{E_\beta} \\
&+ \int_{t_2}^{t_1} \| U_{\beta-1}(t_1,s)F_{\beta-1}^{u_n}(s,x_n(s)) \|_{E_\beta} ds \\
&+ \int_0^{t_2} \| (U_{\beta-1}(t_1,s) - U_{\beta-1}(t_2,s))F_{\beta-1}^{u_n}(s,x_n(s)) \|_{E_\beta} ds \\
&\le C_6(t_1-t_2)^\sigma + C_7(t_1-t_2)^\sigma + C_8(t_1-t_2)^\delta,
\end{aligned}$$

where $0 < \sigma < \alpha - \beta, 0 < \delta \le \gamma - \beta$. The constant C_6 depends on M_0 and $\| x_0 \|_{E_\alpha}$, and the constants C_7 and C_8 depend on M_0 and ess-sup$\{\|F_{\beta-1}^{u_n}(s,x_n(s))\|_{\alpha-1}, s \in [0,T], n \in N\}$. From this one can easily verify that there exists a constant C_9 such that

$$\|x_n(t_1) - x_n(t_2)\|_{E_\beta} \le C_9(t_1-t_2)^\delta.$$

Since $E_\gamma \hookrightarrow\hookrightarrow E_\beta$ (that is, the injection is compact), for each $t \in I$, the set $X(t) \equiv \{x_n(t), n \in N\}$ is a compact subset of E_β. Summarizing, the family $\{x_n\}$ is a bounded and equicontinuous subset of $C(I, E_\beta)$ with each t-section, $X(t)$, being compact. Thus it follows from the Ascoli–Arzela theorem that the sequence $\{x_n\}$ is compact in $C([0,T], E_\beta)$. Therefore, there exists a convergent subsequence, relabeled $\{x_n\}$, and an $x^* \in C(I, E_\beta)$ such that $x_n \to x^*$ in $C([0,T], E_\beta)$. Denote by $\{h_n(\bullet)\}$ the corresponding sequence $\{g(\bullet, x_n(\bullet), u_n(\bullet))\}$. By virtue of the growth condition (see Theorem 1.2) and the boundedness of the sequence $\{x_n\}$ it follows that the sequence

$\{h_n\}$ is contained in a bounded subset of $L_\infty(I, \partial_0 W^{-1+2\alpha}) \subset L_p(I, \partial_0 W^{-1+2\beta})$ for any $p \geq 1$. Since for $1 < p < \infty$, $L_p(I, \partial_0 W^{-1+2\alpha})$ is a closed subspace of a reflexive Banach space, there exists a subsequence of $\{h_n\}$, relabeled $\{h_n\}$, and an element $h^* \in L_p(I, \partial_0 W^{-1+2\alpha})$ such that $h_n \xrightarrow{w} h^*$ in $L_p(I, \partial_0 W^{-1+2\alpha})$.

By the definition of E_β-weak solution, for all $t \in I$ and every $\varphi \in C([0,T), W^{1-\beta}_{\#\mathcal{B}^\#}) \cap C^1([0,T'], W^{\beta}_{\#\mathcal{B}^\#})$ satisfying $\varphi(T') = 0$, we have

$$\int_0^{T'} \{-\langle \varphi, x_n \rangle + a(t)\langle \varphi, x_n \rangle\} dt = \int_0^{T'} \{\langle \varphi, f(t, x_n) \rangle + \langle \ell^\#(t)\varphi, h_n(t) \rangle_\partial\} dt + \langle \varphi(0), x_0 \rangle.$$

Letting $n \to \infty$, through a subsequence if necessary, it follows from the facts that $x_n \xrightarrow{s} x^*$ in $C([0,T], E_\beta)$, the bilinear functional a is continuous, continuity of $x \longrightarrow f(t,x)$, and weak convergence of $\{h_n\}$ to h^* that

$$\int_0^{T'} \{-\langle \varphi, x^* \rangle + a(t)\langle \varphi, x^* \rangle\} dt = \int_0^{T'} \{\langle \varphi, f(t, x^*) \rangle + \langle \ell^\#(t)\varphi, h^*(t) \rangle_\partial\} dt + \langle \varphi(0), x_0 \rangle. \tag{2.4}$$

This implies that x^* is a weak solution of (1.9) with the boundary data g replaced by h^*. We must show that there exists an admissible control u^* such that $h^*(t) = g(t, x^*(t), u^*(t))$ a.e. and that $\eta(u^*, x^*) = m$. In view of this we note that by Mazur's theorem there exists a finite convex combination of $\{h_n\}$ that converges strongly to h^* in $L_p(I, \partial_0 W^{-1+2\alpha})$. In particular, for each integer k, there exists an integer n_k, a set of integers $\{i = 1, 2, \ldots, m(k)\}$, and $\{\alpha_{k,i} \geq 0, i = 1, 2, \ldots, m(k)\}$ such that

$$\begin{aligned} &\sum_{i=1}^{m(k)} \alpha_{k,i} = 1, \quad \text{for all integers } k, \\ &g_k(t) \equiv \sum_{i=1}^{m(k)} \alpha_{k,i}\, h_{n_k+i}(t), \quad t \in I, \end{aligned} \tag{2.5}$$

and

$$g_k \xrightarrow{s} h^* \text{ in } L_p(I, \partial_0 W^{-1+2\alpha}). \tag{2.6}$$

Corresponding to the above sequence, define the sequence $\{\ell_k\}$ as follows

$$\begin{aligned} &L_{n_k+i}(t) \equiv L(t, x_{n_k+i}(t), u_{n_k+i}(t)), \\ &\ell_k(t) \equiv \sum_{i=1}^{m(k)} \alpha_{k,i}\, L_{n_k+i}(t). \end{aligned} \tag{2.7}$$

Define

$$\ell^*(t) \equiv \operatorname{Lim\,inf}_{k\to\infty} \ell_k(t), \quad t \in I.$$

By virtue of Assumption L(4), it follows from boundedness of the set $\mathcal{X}$ that $\operatorname{Lim\,inf} \ell_k(t)$ is well defined on I. Hence by Fatou's lemma we have

$$\int_I \ell^*(t)dt \leq \operatorname{Lim\,inf}_{k\to\infty} \int_I \ell_k(t)dt. \tag{2.8}$$

Clearly

$$\operatorname{Lim}_{k\to\infty} \eta(u_{n_k+i}, x_{n_k+i}) \longrightarrow m,$$

and hence it follows from (2.7) that

$$\operatorname{Lim}_{k\to\infty} \int_I \ell_k(t)dt = m,$$

which in turn leads to the inequality

$$\int_I \ell^*(t)dt \leq m. \tag{2.9}$$

Again by virtue of Assumption L(4) and boundedness of the solution set $\mathcal{X}$, there exists a $\tilde{\phi} \in L_1(I)$ dependent on ϕ such that

$$\ell^*(t) \geq \tilde{\phi}(t) \text{ a.e. on } I.$$

This along with (2.9) implies that $\ell^* \in L_1(I)$. Define the set-valued map $\mathcal{Q}$ from $I \times E_\beta$ to $2^{R_\infty \times \partial_0 W^{-1+2\alpha}} \setminus \emptyset$ as follows:

$$\mathcal{Q}(t,z) \equiv \{(\gamma,\beta) \in R_\infty \times \partial_0 W^{-1+2\alpha} : \gamma \geq L(t,z,v), \beta = g(t,z,v), v \in V(t)\},$$

for $t \in I, z \in E_\beta$. We prove that

$$(\ell^*(t), h^*(t)) \in \mathcal{Q}(t, x^*(t)) \text{ a.e. on } I. \tag{2.10}$$

By following techniques similar to those in [12, Thm. 3.1, p. 225], we can find a set $I_0 \subset I$ with Lebesgue measure $\lambda(I \setminus I_0) = 0$ such that

$$(\ell^*(t), h^*(t)) \in \bigcap_{\epsilon>0} C\ell Co\mathcal{Q}(t, N_\epsilon(x^*(t))), \quad t \in I_0.$$

By virtue of convexity and lower semicontinuity of the map $v \longrightarrow L(t,e,v)$ and the assumption that $G(t,x) \in cc(\partial_0 W^{-1+2\alpha})$, it follows that the multifunction $\mathcal{Q}$ is closed convex valued. Further, it follows from lower semicontinuity Assumption L(2) and the continuity of the map $e \longrightarrow g(t,e,v)$ from E_β to $\partial_0 W^{-1+2\alpha}$, that $e \longrightarrow \mathcal{Q}(t,e)$ is quasi upper semicontinuous. A quasi upper semicontinuous multifunction with closed convex values satisfies the weak Cesari property [10, Thm. 5.5, p. 52]. Thus

$$\bigcap_{\epsilon>0} C\ell Co\mathcal{Q}(t, N_\epsilon(x^*(t))) \subset \mathcal{Q}(t, (x^*(t)), \quad t \in I_0. \tag{2.11}$$

Hence we have $(\ell^*(t), h^*(t)) \in \mathcal{Q}(t, (x^*(t)), \ t \in I_0$. Since I_0 has full Lebesgue measure, this proves (2.10). From (2.10) and the definition of $\mathcal{Q}$, it follows that, for almost all $t \in I$, there exists an element $\tilde{u}(t) \in V(t)$ such that

$$\begin{aligned} \ell^*(t) &\geq L(t, x^*(t), \tilde{u}(t)), \\ h^*(t) &= g(t, x^*(t), \tilde{u}(t)). \end{aligned} \tag{2.12}$$

In view of (2.9), (2.12), and the definition of admissible controls, it is sufficient to show that there exists a measurable substitute for $\tilde{u}$. Define for $t \in I_0$ the set-valued map

$$\Lambda(t) \equiv \{v \in V(t) : \ell^*(t) \geq L(t, x^*(t), v) \text{ and } h^*(t) = g(t, x^*(t), v)\}. \tag{2.13}$$

Clearly this set is nonempty. We prove that it has a measurable selection. A general result in this direction states that a (weakly) measurable multifunction with closed values, from an arbitrary measurable space to a Polish space, has measurable selections [9, Thm. 4.1, p. 867]. Since $V(t) \in wkc(U)$ and U is separable, the relative weak topology on $V(t)$ is metrizable [13, Thm. V.3, p. 434] and with respect to this metric topology, it is a separable complete metric space and hence a Polish space. Thus it is sufficient to verify that Λ is closed valued and measurable. The closedness of this set follows immediately from the lower semicontinuity of L and continuity of g in the control variable. Since G is closed convex valued, it follows from convexity of L (Assumption L(3)) that $\Lambda(t)$ is also convex and hence it is a closed convex-valued multifunction. For the proof of measurability we can follow the same technique as that in [12, Thm. 3.1, p. 228]. Here we give a simpler and direct proof. Define

$$\begin{aligned} \Lambda_1(t) &\equiv \{v \in U : L(t, x^*(t), v) - \ell^*(t) \leq 0\}, \quad t \in I, \\ \Lambda_2(t) &\equiv \{v \in U : h^*(t) - g(t, x^*(t), v) = 0\}. \end{aligned} \tag{2.14}$$

Then clearly

$$\Lambda(t) = \big(\Lambda_1(t) \cap V(t)\big) \bigcap \big(\Lambda_2(t) \cap V(t)\big).$$

We show that each component is measurable. First we show that for any closed subset $\Sigma \subset U$, $\Lambda_1^-(\Sigma) \equiv \{t \in I : \Lambda_1(t) \cap \Sigma \neq \emptyset\}$ is measurable. Since U is a separable Banach space, there exists a countable dense subset Σ_0 of the set Σ such that

$$\begin{aligned} \Lambda_1^-(\Sigma) &\equiv \{t \in I : \Lambda_1(t) \cap \Sigma \neq \emptyset\} \\ &= \bigcup_{v \in \Sigma_0} \{t \in I : L(t, x^*(t), v) \leq \ell^*(t)\}. \end{aligned}$$

Since ℓ^* is measurable and for each $v \in U$, $t \longrightarrow L(t, x^*(t), v)$ is measurable, the union is measurable. By assumption, V is measurable and hence $\Lambda_1 \cap V$ is measurable. Similarly we can write

$$\Lambda_2^-(\Sigma) \equiv \bigcup_{v \in \Sigma_0} \{t \in I : g(t, x^*(t), v) = h^*(t)\}.$$

Hence $\Lambda_2 \cap V$ is also measurable. Thus $t \longrightarrow \Lambda(t)$ is a measurable multifunction taking values from $cc(U)$. Hence by the selection theorem mentioned above, Λ has a measurable selection u^* which is a measurable substitute for $\tilde{u}$. This completes the proof.

The result of the above theorem can be extended to the case where the control constraint set is also state dependent (feedback) given by $V(t, x)$. We can prove the following result.

THEOREM 2.2. *Suppose the basic hypotheses on $\mathcal{A}, \mathcal{B}, f, g$ hold, $A^{-1}(t)$ is compact for each $t \in I$, $x_0 \in E_\alpha$, and L satisfies Assumptions* L(1)–L(3) *and* L(4*) *defined below.*

L(4*) *There exists* $\lambda \geq 0$ *such that* $\phi(t) - \lambda \parallel e \parallel_{E_\beta} \leq L(t,e,v), a.e.$ *for all* $v \in V(t,e)$. *The multifunction* $V : I \times E_\beta \longrightarrow wkc(U)$ *is graph measurable and the set-valued map*

$$\mathcal{Q}(t,e) \equiv \{(\gamma,\beta) \in R_\infty \times \partial_0 W^{-1+2\alpha} : \gamma \geq L(t,e,v), \beta = g(t,e,v), v \in V(t,e)\}$$

satisfies the weak Cesari property.

Then there exists an optimal control $u_0 \in \mathcal{U}_{\text{ad}}$ *for problem* (P).

Remark. Following the same procedure, the existence result presented here can be extended to relaxed control problems thereby relaxing the convexity assumptions on L and g. Further the result can be extended also to (boundary) control differential inclusions. We will provide the details in another paper. Here we present only a brief discussion.

Let Γ be a compact Polish space and $M(\Gamma)$ be the space of probability measures on Borel subsets of Γ. Let $\mathcal{U}_{\text{ad}}$ denote the class of w^*-measurable functions on $I \equiv [0,T]$ with values in $M(\Gamma)$ furnished with the Young topology (see [2], [3], [12]). For more general relaxed controls see [17]. Let

$$g : I \times E_\beta \times M(\Gamma) \longrightarrow \partial_0 W^{-1+2\alpha},$$
$$L : I \times E_\beta \times M(\Gamma) \longrightarrow R_\infty$$

be maps measurable in $t \in I$, continuous in $x \in E_\beta$, and w^*-continuous in $\mu \in M(\Gamma)$. In particular, as a function of μ, these maps may be given by

$$h(t,x,\mu) \equiv \int_\Gamma \tilde{h}(t,x,\xi)\mu(d\xi),$$

where $\xi \longrightarrow \tilde{h}(t,x,\xi)$ is continuous and bounded on Γ, with h denoting either of the maps g, L.

As in the preceding theorem, define the multifunction G on $I \times E_\beta$ with values

$$G(t,x) \equiv \{\zeta \in \partial_0 W^{-1+2\alpha} : \zeta = g(t,x,\mu),\ \mu \in M(\Gamma)\}.$$

Thus the relaxed control problem (P_r) can be stated as follows: Find $\nu \in \mathcal{U}_{\text{ad}}$ such that

$$J(\nu) \equiv \int_0^T L(t,x,\nu)dt \longrightarrow \text{Inf},$$
$$\dot{x}(t) + A_{\beta-1}(t)x(t) = \mathcal{F}^\nu_{\beta-1}(t,x(t)), \quad t \in I,$$

where $\mathcal{F}^\nu_{\beta-1}$ is defined accordingly.

Define the multifunction

$$Q(t,x) \equiv \{(\lambda,\xi) : (\lambda,\xi) \in R_\infty \times \partial_0 W^{-1+2\alpha}, \lambda \geq L(t,x,\mu), \xi = g(t,x,\mu), \mu \in M(\Gamma)\}.$$

Now using the weak Cesari property for the multifunction Q and the theory of measurable selections, we can prove, as in [2], [12], the existence of optimal relaxed controls under a much milder hypothesis on L and g.

THEOREM 2.3. *Suppose the basic hypotheses on* $\mathcal{A}, \mathcal{B}, f, g, L$ *hold,* $A^{-1}(t)$ *is compact for each* $t \in I$, $x_0 \in E_\alpha$, *and the multifunction* G *given by* $G(t,z) \equiv g(t,z,M(\Gamma))$ *mapping* $I \times E_\beta$ *to* $cc(\partial_0 W^{-1+2\alpha})$ *is quasi upper semicontinuous. Then there exists an optimal control* $u_0 \in \mathcal{U}_{\text{ad}}$ *for problem* (P_r).

Now we present a necessary condition of optimality. Let λ denote the Lebesgue measure and $M(I,U)$ denote the space of strongly measurable functions (equivalence classes) from I to U, furnished with the metric topology

$$\rho(u,v) \equiv \lambda\{t \in I : u(t) \neq v(t)\}.$$

Since U is a Banach space, (M,ρ) is a complete metric space. For admissible controls we choose

$$\mathcal{U}_{\text{ad}} \equiv \{u \in M : u(t) \in V(t) \text{ a.e.}\},$$

where V satisfies Assumption U1.

Assumption C. Let $\hat{F}(t) \equiv DF_{\beta-1}(t,x(t),u(t))$, $\hat{L}(t) \equiv DL(t,x(t),u(t))$ denote the Fréchet differentials of $F_{\beta-1}$ and L, respectively, along any admissible state control pair (u,x), satisfying $\hat{F} \in L_1(I,\mathcal{L}(E_\beta,E_{\alpha-1}))$, $\hat{L} \in L_1(I,E_\beta^*)$.

Following an approach similar to that in [15], [17], using Ekland's variational principle with respect to the metric topology ρ, we can prove the following necessary conditions of optimality.

THEOREM 2.4. *Suppose the assumptions of Theorem* 2.1 *hold and let* (u,x) *be an admissible pair. Suppose* f,g, *and* L *are continuously Fréchet differentiable in the state variable on* E_β *and continuous with respect to the control variable on* U *satisfying Assumption* C. *Then for the pair* $(u,x) \in \mathcal{U}_{\text{ad}} \times C(I,E_\beta)$ *to be optimal, it is necessary that there exists a* $\psi \in C(I,E_{1-\alpha}^{\#})$ *such that*

$$\begin{aligned} &\int_I \{\langle F_{\beta-1}(t,x(t),u(t)),\psi(t)\rangle + L(t,x(t),u(t))\}dt \\ &\quad \leq \int_I \langle F_{\beta-1}(t,x(t),v(t)),\psi(t)\rangle + L(t,x(t),v(t))dt \quad \textit{for all} \quad v \in \mathcal{U}_{\text{ad}}, \end{aligned} \tag{2.15}$$

where x *is the weak solution of equation* (1.9) *corresponding to* u *and* ψ *satisfies the adjoint equation*

$$-\dot{\psi} + A_{\beta-1}^*(t)\psi = \hat{F}^*(t)\psi + \hat{L}(t), \psi(T) = 0, \tag{2.16}$$

also in the weak sense.

Note that a pointwise necessary condition of optimality easily follows from (2.15).

3. Examples. In this section we give two examples to demonstrate applicability of our abstract results.

Example 1 (system of second-order PDEs). The system is governed by a coupled system of N second-order PDEs with nonlinear interactions both in the interior and on the boundary of a spatial domain in n-space. This is an appropriate model for reaction–diffusion processes. Let Ω be a bounded domain in R^n of class C^2, that is, Ω is an n-dimensional C^2-submanifold of R^n with boundary $\partial\Omega$.

For $1 < p < +\infty$ and $s \in R^1$ we let $L_p \equiv L_p(\Omega,C^N)$ and set $W_p^s \equiv W_p^s(\Omega,C^N)$, $W_p^s(\partial\Omega) \equiv W_p^s(\partial\Omega,C^N)$, where C^N is the N-dimensional complex Euclidean space.

We define the operators $-D\Delta$ and $-D\Delta^{\#}$ (in the sense of distributions) by

$$\begin{aligned} &-D\Delta: \quad W_p^2 \to L_p, \quad 1 < p < +\infty, \\ &-D\Delta^{\#} = -D\Delta: \quad W_{p'}^2 \to L_{p'}, \quad p' = p/(p-1), \end{aligned}$$

where $D = \operatorname{diag}(d_1, d_2, \ldots, d_N)(d_i > 0, i = 1, 2, \ldots, N)$, Δ is the Laplacian. Define the boundary operators B, $B^{\#}$ (in the sense of trace) by

$$By \equiv D(\partial y/\partial \nu) + b(x)y : \quad W_p^2 \to W_p^{1-1/p}(\partial\Omega), \quad 1 < p < +\infty,$$
$$B^{\#}y \equiv D(\partial y/\partial \nu) + b(x)y : \quad W_{p'}^2 \to W_{p'}^{1-1/p'}(\partial\Omega), \quad p' = p/(p-1),$$

where $b(x) = \operatorname{diag}(b_1(x), b_2(x), \ldots, b_N(x))$, $b_i(x) \in W_q^{1-1/q}(\partial\Omega, C^1)$ $(i = 1, 2, \ldots, N)$, $q > \max\{np, np'\}$. We set $W_B' \equiv \{y \in W_p^2 | By = 0\}$, $W_{B^{\#}}'^{\#} \equiv \{y \in W_{p'}^2 | B^{\#}y = 0\}$, and let $-D\Delta_B \equiv -D\Delta|_{W_B'}$, $-D\Delta_{B^{\#}}^{\#} \equiv -D\Delta^{\#}|_{W'^{\#}_{B^{\#}}}$.

We have the following important lemma (see [8]).

LEMMA 3.1. *$-D\Delta_B, -D\Delta_{B^{\#}}^{\#}$ are closed and densely defined linear operators in L_p and $L_{p'}$, respectively, and there exist constants $M_0 > 0$ and $c \in R$ with*

$$\rho(D\Delta_B) \supset c + \Sigma_0, \ \rho(D\Delta_{B^{\#}}^{\#}) \supset c + \Sigma_0,$$
$$\|(\lambda - D\Delta_B)^{-1}\| \le M_0/(1 + |\lambda - c|),$$
$$\|(\lambda - D\Delta_{B^{\#}}^{\#})^{-1}\| \le M_0/(1 + |\lambda - c|),$$

for $\lambda \in c + \Sigma_0$, where $\rho(\bullet)$ denotes the resolvent set, $\Sigma_0 \equiv \{z \in C^* | \ |\arg z| \le \pi/2\} \cup \{0\}$, $C^* \equiv C - \{0\}$.

Define $\mathcal{A} \equiv -D\Delta + c$, $\mathcal{A}^{\#} \equiv -D\Delta^{\#} + c$, $A \equiv -D\Delta_B + c$, and $A^{\#} \equiv -D\Delta_{B^{\#}}^{\#} + c$, then $A(A^{\#})$ is the infinitesimal generator of a strongly continuous analytic semigroup $\{e^{tA}\}$ $(\{e^{tA^{\#}}\})$ on L_p $(L_{p'})$ and there exist constants $M > 0$, $\omega > 0$ such that

$$\|e^{tA}\| \le Me^{-\omega t}, \quad t \ge 0,$$

$$\|e^{tA^{\#}}\| \le Me^{-\omega t}, \quad t \ge 0.$$

We write $A \in \mathcal{G}(L_p, M, -\omega) \cap \mathcal{H}(L_p)$, $A^{\#} \in \mathcal{G}(L_{p'}, M, -\omega) \cap \mathcal{H}(L_{p'})$.

Let $E \equiv L_p$, $E^{\#} = L_{p'}$, $1 < p < +\infty$, $p' = p/(p-1)$. Then $E^{\#} = (E)'$, $A^{\#} = (A)'$. Let $\|x\|_k = \|A^k x\|$, $x \in D(A^k)$, $k \in Z$. Put $E_k = (D(A^k), \| \bullet \|_k)$ if $k \ge 0$ and E_k is the completion of $(E, \| \bullet \|_k)$ if $k < 0$.

For $k < \beta < k+1$ and $k \in Z$, define the interpolation space $E_\beta \equiv (E_k, E_{k+1})_{\beta-k,p}$ as before and denote by A_β the E_β-realization of A_k.

LEMMA 3.2. *The "scale" $\{(E_\beta, A_\beta) : k < \beta < k+1\}$ is well defined, each E_β is a Banach space, and we have $A_\beta \in \mathcal{G}(E_\beta, M, -\omega) \cap \mathcal{H}(E_\beta)$, $A_\beta \in Isom\ (E_{\beta+1}, E_\beta)$. For $-\infty < \beta < \alpha < +\infty$, we have $E_\alpha \hookrightarrow\hookrightarrow E_\beta$, and for $t > 0, \sigma < \omega$,*

$$e^{-tA_\alpha} = e^{-tA_\beta}|_{E_\alpha}, \quad t \ge 0,$$
$$\|e^{-tA_\beta}\|_{\mathcal{L}(E_\beta, E_\alpha)} \le C(\alpha, \beta, \sigma, M)t^{\beta-1}e^{-\sigma t}.$$

Finally, let $\{(E_\beta^{\#}, A_\beta^{\#}) : \beta \in R\}$ denote the scale as constructed above starting with $E^{\#}$, $A^{\#}$. Then we have

$$(E_\beta)' = E_{-\beta}^{\#}, \ (A_\beta)' = A_{-\beta}^{\#};$$

$$E_0 = E = L_p, \ E_\beta = W_p^{2\beta} \ \text{ if } \ 1/p < 2\beta < 1 + 1/p, \ 1 < p < +\infty,$$

$$E_0^{\#} = E^{\#} = L_{p'},\ E_\beta^{\#} = W_{p'}^{2\beta} \text{ if } 1/p' < 2\beta < 1 + 1/p',\ p' = p/(p-1).$$

Let a denote the bilinear form on $E_{1-\beta}^{\#} \times E_\beta = W_{p'}^{2-2\beta} \times W_p^{2\beta}$ defined by

$$a(\varphi, y) \equiv \int_\Omega \left(\sum_{k=1}^{N} d_k (\nabla \varphi_k, \nabla y_k)_{R^n} + c(\varphi, y)_{R^N} \right) dx,$$

which satisfies

$$a(\varphi, y) = \langle \varphi, Ay \rangle, \quad \text{for } (\varphi, y) \in E_{1-\beta}^{\#} \times E_1,$$

$$a(\varphi, y) = \langle \varphi, A_{\beta-1} y \rangle, \quad \text{for } (\varphi, y) \in E_{1-\beta}^{\#} \times E_\beta.$$

LEMMA 3.3. *The operator* $(\mathcal{A}, B) \in Isom\ (W_p^2, L_p \times W_p^{1-(1/p)}(\partial\Omega))$, $1 < p < +\infty$, *and* $(\mathcal{A}, B)^{-1} \in \mathcal{L}(L_p \times W_p^{1-(1/p)}(\partial\Omega), W_p^2)$.

Further, for any θ $(0 \le \theta \le 1)$, $(\mathcal{A}, B)^{-1}$ *has a unique extension, which we denote also by* $(\mathcal{A}, B)^{-1} \in \mathcal{L}(L_p \times W_p^{2\theta-1-(1/p)}(\partial\Omega), W_p^{2\theta})$.

Defining $R_\theta = (\mathcal{A}, B)^{-1}|_{\{0\} \times W_p^{2\theta-1-(1/p)}(\partial\Omega)}$ we have $R_\theta \in \mathcal{L}(W_p^{2\theta-1-(1/p)}(\partial\Omega), W_p^{2\theta})$ and for $g \in W_p^{2\theta-1-(1/p)}(\partial\Omega)$, R_θ satisfies

$$\|R_\theta g\|_{W_p^{2\theta}} \le K(\theta) \|g\|_{W_p^{2\theta-1-(1/p)}(\partial\Omega)}.$$

We define the admissible controls $\mathcal{U}_{\text{ad}}$ as follows.

(U2) Let $U \equiv L_s(\partial\Omega), 1 < s < \infty$, $r \in L_1([0,T], L_s(\partial\Omega))$, and

$$V(t) \equiv \left\{ v \in U : |v|_U^s \le \int_{\partial\Omega} |r(t,\xi)|^s d\xi \right\}.$$

Take

$$\mathcal{U}_{\text{ad}} = \{u \in L_1(I, U) : u(t) \in V(t) \text{ a.e.}\}.$$

Clearly V is a measurable multifunction.

Define

$$f(t,\phi)(\xi) \equiv \hat{f}(t,\xi,\phi(\xi)), \quad \phi \in C((\bar{\Omega}, C^N)),\ t > 0,$$
$$g(t,\phi,v)(\xi) \equiv \hat{g}(t,\xi,\phi(\xi),v(\xi)), \quad \phi \in C(\partial\Omega, C^N), v \in U.$$

Define the operators $F(t,y) \equiv f(t,y) + cy$ and $G(t,y,u) = g(t,y,u)$, where c is the same constant as in Lemma 3.1. Suppose the operators F, G satisfy the following conditions.

(F,G) Let $p > n$, $1 < 2\beta < 1 + 1/p$, then there exists α with $2\beta < 2\alpha < 1 + 1/p$ such that

$$F \in C^{0,1-}([0,T] \times E_\beta, E_{\alpha-1})$$

and for each $v \in U$

$$G_v \equiv G(.,.,v) \in C^{0,1-}([0,T] \times E_\beta, W_p^{2\alpha-1-1/p}(\partial\Omega)),$$

where C^{1-} denotes local Lipschitz continuity and the Lip-constant for G_v depends only on $\|v\|_U$. Using the operators defined above, one can write the following controlled system:

$$\begin{aligned} y_t - D\,\triangle\, y &= f(t,\xi,y), \\ D(\partial y/\partial\nu) + b(\xi)y &= g(t,\xi,y,u), \\ y(0) &= y_0 \end{aligned} \tag{3.1}$$

in its abstract form

$$\begin{aligned} \dot{y} + \mathcal{A}y &= F(t,y), \\ By &= G(t,y,u), \\ y(0) &= y_0. \end{aligned} \tag{3.2}$$

As defined earlier, for each $y_0 \in E_\alpha = W_p^{2\alpha}$, the function $y \in C([0,T], E_\beta)$ is a E_β ($W_p^{2\beta}$)-weak solution of (3.2) if y satisfies

$$\begin{aligned} &\int_0^{T'} \{-\langle \dot{\varphi}, y\rangle + a(\varphi, y)\}dt \\ &\quad = \int_0^{T'} \{\langle \varphi, F(t,y)\rangle + \langle \varphi|_{\partial\Omega}, g(t,y,u)\rangle_\partial\}dt + \langle \varphi(0), y_0\rangle \end{aligned}$$

for every $T' \in (0,T)$ and every $\varphi \in C([0,T], E_{1-\beta}^{\#}) \cap C^1((0,T], E_{-\beta}^{\#})$ satisfying $\varphi(T') = 0$.

Here the duality pairings have the following specific meaning: $\langle \dot{\varphi}, y\rangle$ ($\langle \varphi, F(t,y)\rangle$, $\langle \varphi|_{\partial\Omega}$, $G_u(t,y)\rangle_\partial$, $\langle \varphi(0), y_0\rangle$) is the dual pairing of $E_{-\beta}^{\#}$ and $E_{\beta-1}$ ($E_{1-\beta}^{\#}$ and $E_{-\beta}$, $W_{p'}^{2-2\beta-1/p'}(\partial\Omega)$ and $W_p^{2\beta-1-1/p}(\partial\Omega)$, $E_{1-\beta}^{\#}$ and $E_{\beta-1}$, respectively).

Further we assume that for each $y \in E_\beta$,

$$\|F(t,y)\|_{E_{\alpha-1}} \le K_1(1 + \|y\|_{E_\beta}),$$

$$\|G(t,y(t),u)\|_{W_p^{2\alpha-1-1/p}(\partial\Omega)} \le K_2(1 + \|y\|_{E_\beta}),$$

where K_1 is constant, K_2 is only dependent on $\|u\|_{L_\infty([0,T],U)}$.

Combining Lemma 1.1 and Theorem 1.2, we have the following result for system (3.2).

THEOREM 3.4. *Suppose F, G satisfy all the assumptions mentioned above, $y_0 \in E_\beta \equiv W_p^{2\beta}$ $(1 < 2\beta < 1 + 1/p, p > n)$. Then we have*

(a) *for each $y \in C([0,T], E_\beta) = C([0,T], W_p^{2\beta})$, and $u \in \mathcal{U}_{\rm ad}$,*

$$t \longrightarrow F(t, y(t)) \in C([0,T], E_{\alpha-1})$$

$$\text{and } t \longrightarrow A_{\beta-1}R_\beta G(t,y(t),u) = A_{\alpha-1}R_\alpha G(t,y(t),u) \in L_\infty([0,T], E_{\alpha-1}),$$

where α is the number appearing in condition (F,G).

(b) *$y \in C([0,T], W_p^{2\beta})$ is a $W_p^{2\beta}$-weak solution of equation* (3.2) *on $[0,T]$, iff y is a solution of the following integral equation:*

$$y(t) = e^{-tA_{\beta-1}}y_0 + \int_0^t e^{-(t-s)A_{\beta-1}}[F(s,y(s)) + A_{\beta-1}R_\beta G(s,y(s),u(s))]ds$$

in $C([0,T],W_p^{2\beta})$ $(1<2\beta<1+1/p,\ p>n)$.

(c) *equation* (3.1) *has a unique global* $W_p^{2\beta}$*-weak solution.*

The basic assumptions on the cost integrand L are the same as those of Assumption L given for our main existence Theorem 2.1 with reference to $E_\beta \equiv W_p^{2\beta}, U \equiv L_s(\partial\Omega)$.

We consider the optimal boundary control problem $(\tilde{\mathrm{P}})$. Minimize

$$J(u)=\int_0^T L(t,x(t),u(t))dt \tag{3.3}$$

over all controls $u\in\mathcal{U}_{\mathrm{ad}}$ subject to the state equation (3.1) or, equivalently, (3.2).

THEOREM 3.5. *Under the assumptions of Theorem* 3.4, L, *and* U2, *there exists an optimal control for the problems* (3.2) *and* (3.3).

Proof. This is a special case of our general Theorem 2.1.

Example 2. In steel factories, ingots are raised to high temperature in a furnace and then stored in a ceramic kiln for a soaking process for a suitable period of time. The purpose of soaking is to allow time for the ingots to attain uniform temperature throughout the body before they are transported to the rolling mills. During this process heat loss by radiation and convection is controlled by maintaining the surrounding temperature in the kiln by an auxiliary heat source. Let Ω_o denote the interior of the kiln and $\Omega\subset\Omega_o$ denote the space occupied by the ingot. Loss of heat by radiation to the surrounding medium $\Omega_o\setminus\bar{\Omega}$ is governed by the Steffan–Boltzman law. The corresponding control system can then be described as follows:

$$\begin{aligned} y_t - K\,\triangle\, y &= 0, \quad \text{in } (0,T]\times\Omega,\\ K(\partial y/\partial\nu) &= -g(y,u), \quad \text{in } (0,T]\times\partial\Omega,\\ y(0) &= y_0, \quad \text{in } \Omega, \end{aligned} \tag{3.4}$$

where

(i) for purely radiative heat transfer g is given by

$$g(y,u)=\begin{cases}\sigma E(y^4-u^4)|_{\partial\Omega}, & \text{for } 0\le y\le\gamma_d,\ 0\le u\le\gamma_d,\\ \sigma E(\gamma_d^4-u^4)|_{\partial\Omega}, & \text{for } 0\le u\le\gamma_d,\ y\ge\gamma_d,\end{cases}$$

(ii) for a combination of radiative and convective heat transfer, g is given by

$$g(y,u)=\begin{cases}(\sigma E(y^4-u^4)+\alpha(y-u))|_{\partial\Omega}, & \text{for } 0\le y\le\gamma_d,\ 0\le u\le\gamma_d,\\ (\sigma E(\gamma_d^4-u^4)+\alpha(y-u))|_{\partial\Omega}, & \text{for } 0\le u\le\gamma_d,\ y\ge\gamma_d.\end{cases}$$

The parameters are as follows: K is the conductivity of the ingot material, σ is the Stefan–Boltzman constant, E is the emissivity of the ingot surface, α is the heat transfer coefficient due to convection, and γ_d is the maximum (attainable) temperature of the furnace.

(iii) In the case of hydrodynamics of liquid helium, Lin [18] has used for g the function

$$g(y,u)=\beta_1(y-u)|_{\partial\Omega}+\beta_2(y-u)^3|_{\partial\Omega},$$

where β_1 and β_2 are suitable constants determined by experiments.

Note that only case (i) can be transformed into a problem with control $v\equiv u^4$ appearing linearly; whereas cases (ii) and (iii) do not admit such simplification.

For the state space we choose $E_\beta \equiv W_p^{2\beta}$ so that $\beta > (3/p)$. Given that Ω has the cone property, by the Sobolev embedding theorem this last condition guarantees that $W_p^{2\beta}$ is an algebra (see [14, Thm. 5.23, p. 115]) and hence for $\phi \in W_p^{2\beta}, \phi^4 \in W_p^{2\beta}$, and its trace $\phi^4|_{\partial\Omega} \in W_p^{2\beta-1-1/p}(\partial\Omega)$ also. For admissible controls define

$$V \equiv \{v \in U \equiv W_p^{2\beta-1-1/p}(\partial\Omega) : v = \psi|_{\partial\Omega}, \psi \in W_p^{2\beta}(\Omega_o \setminus \bar{\Omega}) \text{ and } 0 \le v(\xi) \le \gamma_d \text{ a.e.}\}$$

and $\mathcal{U}_{\text{ad}} \equiv \{u \in L_\infty(I, U) : u(t) \in V \text{ a.e.}\}$. In particular we can choose $p = 4, \beta > 3/4$. This also implies that the embedding $W_p^{2\beta-1-1/p} \hookrightarrow L_p(\partial\Omega)$ is continuous and hence $V \subset L_p(\partial\Omega)$. Thus our boundary operator g is locally Lipschitz, satisfies the growth condition, and maps $E_\beta \times V$ to $W_p^{2\beta-1-1/p}$.

The cost integrand may be taken as

$$L(t, y(t), u(t)) \equiv \int_\Omega |y(t,\xi) - y^0 t, \xi)|^p d\xi + \int_{\partial\Omega} |u(t,\zeta)|^p d\zeta.$$

Thus all the assumptions of Theorem 3.5 are satisfied, and hence an optimal control exists.

Remark. We note that g is locally Lipschitz and that the (linear) growth condition is satisfied because of the physical limitation of the furnace temperature. Hence for arbitrary time interval I the equations have unique solutions and the optimal control is defined for the entire interval. If one has to deal with the polynomial growth our results will apply only for the maximal interval of existence of solutions of the evolution equations.

REFERENCES

[1] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, Heidelberg, 1971.

[2] N. U. AHMED AND K. L. TEO, *Optimal Control of Distributed Parameter Systems*, North-Holland, New York, Oxford, 1981.

[3] N. U. AHMED, *Semigroup Theory with Applications to Systems and Control*, Pitman Res. Notes Math. Ser. 246, Longman Scientific and Technical, Harlow, U.K., 1991.

[4] H. O. FATTORINI, *Boundary control systems*, SIAM J. Control Optim., 6 (1968), pp. 349–385.

[5] M. C. DELFOUR AND M. SORINE, *The linear quadratic optimal control problem for parabolic systems with boundary control through a Dirichlet condition*, in Control of Distributed Parameter Systems, Proc. 3rd IFAC symposium, Toulouse 1982, J. P. Babary and L. Le Lett, eds., Pergamon Press, Oxford, 1983, pp. 87–90.

[6] I. LASIECKA, *Unified theory for abstract parabolic boundary problems—a semigroup approach*, Appl. Math. Optim., 6 (1980), pp. 287–383.

[7] P. ACQUISTAPACE, F. FLANDONI, AND B. TERRENI, *Initial boundary value problems and optimal control for nonautonomous parabolic systems*, SIAM J. Control Optim., 29 (1991), pp. 89–118.

[8] H. AMANN, *Parabolic evolution equations and nonlinear boundary conditions*, J. Differential Equations, 72 (1988), pp. 201–269.

[9] D. H. WAGNER, *Survey of measurable selection theorems*, SIAM J. Control Optim., 15 (1977), pp. 859–903.

[10] S. H. HOU, *On property (Q) and other semicontinuity properties of multifunctions*, Pacific J. Math., 103 (1982), pp. 39–56.

[11] N. U. AHMED, *Optimization and Identification of Systems Governed by Evolution Equations on Banach Space*, Pitman Res. Notes Math. Ser. 184, Longman Scientific and Technical, Harlow, U.K., 1988.

[12] ———, *Existence of optimal controls for a class of systems governed by differential inclusions on a Banach space*, J. Optim. Theory Appl., 50 (1986), pp. 213–237.

[13] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part* I, InterScience Publishers, John Wiley, New York, London, 1964.

[14] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, San Francisco, London, 1975.

[15] N. U. AHMED, *Optimal control of infinite dimensional systems governed by integro differential equations*, in Differential Equations, Dynamical Systems and Control Science, A Festschrift in Honor of Lawrence Markus, Lecture Notes in Pure Appl. Math 152, K. D. Elworthy, et al., eds., Marcel Dekker, New York, Basel, Hong Kong, 1994, pp. 383–402.

[16] ———, *Optimal control of infinite dimensional systems governed by functional differential inclusions*, Discuss. Math.—Differential Inclusions, 15 (1995), pp. 75–94.

[17] H. O. FATTORINI, *Existence theory and the maximum principle for relaxed infinite dimensional optimal control problems*, SIAM J. Contr. Optim., 32 (1994), pp. 311–331.

[18] C. C. LIN, *Hydrodynamics of liquid helium* II*, Phys. Rev. Lett., 2 (1959), pp. 245–246.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 491–506, March 1996

005

A SIMPLICIAL ALGORITHM FOR COMPUTING ROBUST STATIONARY POINTS OF A CONTINUOUS FUNCTION ON THE UNIT SIMPLEX*

ZAIFU YANG†

Abstract. A simplicial algorithm is proposed to compute a robust stationary point of a continuous function f from the $(n-1)$-dimensional unit simplex S^{n-1} into R^n. The concept of robust stationary point is a refinement of the concept of stationary point of f on S^{n-1}. Starting from an arbitrarily chosen interior point v in S^{n-1}, the algorithm generates a piecewise linear path of points in S^{n-1}. This path is followed by alternating linear programming pivot steps and replacement steps in a specific simplicial subdivision of the relative interior of S^{n-1}. In this way an approximate robust stationary point of any a priori chosen accuracy is reached within a finite number of steps. The algorithm leaves the starting point along one out of $n!$ rays. When the path approaches the boundary of S^{n-1}, the mesh size of the triangulation along the path automatically goes to zero. This makes the algorithm different from all simplicial restart algorithms and homotopy algorithms known so far. Roughly speaking, the algorithm is a blend of a restart and a homotopy algorithm and maintains the basic properties of both. However, the algorithm does not need an extra dimension as homotopy algorithms do. Some examples are discussed.

Key words. robust stationary point, variational inequality, simplicial algorithm, subdivision, piecewise linear approximation, stability

AMS subject classifications. Primary, 49D35; Secondary, 90C30, 90C33

1. Introduction.

Let the $(n-1)$-dimensional unit simplex S^{n-1} be defined by

$$S^{n-1} = \left\{ x \in R^n_+ \mid \sum_{i=1}^{n} x_i = 1 \right\},$$

where R^n_+ is the nonnegative orthant of the n-dimensional Euclidean space. Let us assume that $f : S^{n-1} \mapsto R^n$ is a continuous function. Then the stationary point problem or variational inequality problem for f on S^{n-1} is to find a point $x^* \in S^{n-1}$ such that

$$(x^* - x)^\top f(x^*) \geq 0$$

for any point x in S^{n-1}. We call x^* a stationary point of f on S^{n-1}. It is well known that this problem is equivalent to the Brouwer fixed point problem on S^{n-1} (see, e.g., Eaves [6]).

To compute a fixed point or a stationary point of a continuous function on S^{n-1}, several simplicial algorithms have been developed (Scarf [17], [18], Kuhn [11], Eaves [7], Kuhn and MacKinnon [12], van der Laan and Talman [13], [14], Doup and Talman [4], and Doup, van der Laan, and Talman [5]). Todd [23] and Doup [3] presented excellent surveys on the development of simplicial algorithms. In a simplicial subdivision of S^{n-1} such algorithms search for a simplex which provides an approximate solution by generating a sequence of adjacent simplices. The simplex with which the algorithm terminates is found within a finite number of steps. The so-called variable dimension restart algorithm, originated in van der Laan and Talman [13], can be

* Received by the editors November 12, 1993; accepted for publication (in revised form) November 7, 1994. This research is part of the VF-program "Competition and Cooperation."

† Department of Econometrics and Center for Economic Research, Tilburg University, Postbox 90153, 5000 LE Tilburg, the Netherlands.

started in an arbitrarily chosen grid point of the subdivision and generates a sequence of adjacent simplices of varying dimension. When the end simplex does not yield an approximate solution with satisfactory accuracy, the algorithm can be restarted at the approximate solution with a finer triangulation in the hope of finding a better approximate solution within a small number of iterations.

The concept of robust stationary point is a refinement of the concept of stationary point on the unit simplex and is essentially motivated by economic equilibrium problems, noncooperative games, as well as biology and engineering applications (see, e.g., Myerson [16], Yamamoto [25], van Damme [2]). Because a continuous function from S^{n-1} into R^n may have multiple stationary points and some of them are undesirable from a point of view of stability, we need to refine the concept of stationary point.

In this paper we propose a simplicial algorithm to compute a robust stationary point. Starting from an arbitrarily chosen interior point v in S^{n-1}, the algorithm generates a piecewise linear path of points in S^{n-1}. This path is traced by alternating linear programming pivot steps, to follow a linear piece of the path and replacement steps in a simplicial subdivision of the relative interior of S^{n-1}. Within a finite number of function evaluations and linear programming pivot steps, the algorithm finds an approximate robust stationary point of any a priori, chosen accuracy. The path generated by the algorithm corresponds to a sequence of θ-robust stationary points of the piecewise linear approximation $\bar{f}$ of f with respect to the underlying simplicial subdivision, where $0 < \theta \leq 1$. This simplicial subdivision differs from other simplicial subdivisions of S^{n-1}. We call it the P-triangulation. When the path generated by the algorithm approaches the boundary of S^{n-1}, the mesh size of the triangulation along the path automatically converges to zero. This makes the algorithm different from all other simplicial algorithms. Roughly speaking, the algorithm is a blend of a simplicial restart algorithm and a homotopy algorithm and maintains the basic properties of both. This can be interpreted as follows. If the algorithm converges to a solution on the boundary of S^{n-1}, it shares the property with a homotopy algorithm that the variable θ can be considered as a homotopy parameter (see Eaves [7]), in the sense that when θ tends to zero, the mesh size of the triangulation also tends to zero. However it should be emphasized that the algorithm does not need an extra dimension which is required by homotopy algorithms. While the algorithm converges to a solution in the interior of S^{n-1}, it behaves exactly as a variable dimension algorithm does.

Although it may not be apparent from the arguments of this paper, the algorithm is implicitly related to the procedure proposed by Yamamoto [25] for the determination of a proper Nash equilibrium of finite-person games. Our algorithm can be seen as a constructive combinatorial analog of his continuous procedure when the starting point of our algorithm is chosen to be the barycenter of S^{n-1}.

This paper is organized as follows. In §2 we introduce the definition of a robust stationary point and prove the existence of a robust stationary point for a continuous function on the unit simplex. In §3 we specify the P-triangulation of the unit simplex. In §4 we give a detailed description of the algorithm. Section 5 is devoted to some numerical examples.

2. The concept of robust stationary point. In this section we first give the definition of a robust stationary point and then show the nonemptiness of the set of robust stationary points of a continuous function on the unit simplex. Let a function $f : S^{n-1} \mapsto R^n$ be given and N be the set of the integers $\{1, \ldots, n\}$.

DEFINITION 2.1. *For given $\theta > 0$ a point $x \in S^{n-1}$ is a θ-robust stationary point of f if*

(1) *x is a relative interior point of S^{n-1},*

(2) *$x_k \leq \theta x_l$ if $f_k(x) < f_l(x)$, for $k, l, 1 \leq k, l \leq n$.*

DEFINITION 2.2. *A point $x^* \in S^{n-1}$ is a robust stationary point of f on S^{n-1} if there exist sequences $\{\theta_t\}_1^\infty$ of positive numbers and $\{x(\theta_t)\}_1^\infty$ of θ_t-robust stationary points $x(\theta_t)$ of f such that*

$$\lim_{t\to\infty} \theta_t = 0 \quad \text{and} \quad \lim_{t\to\infty} x(\theta_t) = x^*.$$

We remark that if a stationary point x^* of f lies in the relative interior of S^{n-1}, then x^* must be a robust stationary point of f with equal values of the components. Some examples given in §5 will demonstrate that the concept of robust stationary point is a refinement of the concept of stationary point.

LEMMA 2.3. *Let $f : S^{n-1} \mapsto R^n$ be a continuous function. If $x^* \in S^{n-1}$ is a robust stationary point of f, then x^* is also a stationary point of f.*

Proof. We only need to consider two cases. If x^* lies in the relative interior of S^{n-1}, it implies that $f_i(x^*) = f_j(x^*)$ for $i, j \in N$. Hence we have

$$(x^* - x)^\top f(x^*) = \sum_{i=1}^{n} (x_i^* - x_i) f_i(x^*) = 0$$

for any $x \in S^{n-1}$, which means that x^* is a stationary point of f. On the other hand, if x^* is on the boundary of S^{n-1}, there exists a proper subset J of N such that $x_j^* = 0$ for $j \in J$. It follows from Definitions 2.1 and 2.2 that $f_i(x^*) = f_j(x^*)$ for $i, j \in N \backslash J$ and $f_i(x^*) \geq f_j(x^*)$ for $i \in N \backslash J$ and $j \in J$. Now for given $l \in N \backslash J$, we have

$$(x^* - x)^\top f(x^*) = \sum_{i \in N\backslash J} (x_i^* - x_i) f_i(x^*) - \sum_{j \in J} x_j f_j(x^*) \geq \sum_{i=1}^{n} (x_i^* - x_i) f_l(x^*) = 0$$

for any $x \in S^{n-1}$. This also implies that x^* is a stationary point of f. □

THEOREM 2.4. *Let $f : S^{n-1} \mapsto R^n$ be a continuous function. Then f has at least one robust stationary point in S^{n-1}.*

Proof. We first show that there exists at least one θ-robust stationary point, for any θ, $0 < \theta < 1$. Given such a θ, let $\delta = \frac{1}{n}\theta^{n-1}$ and define

$$S(\theta) = \{ x \in S^{n-1} \mid x_i \geq \delta,\ i = 1, \ldots, n \}.$$

It is clear that $S(\theta)$ is a nonempty, convex, compact subset of S^{n-1}. We further define a set-valued correspondence F on $S(\theta)$ by

$$F(x) = \{ y \in S(\theta) \mid \text{ if } f_i(x) < f_j(x) \text{ then } y_i \leq \theta y_j \text{ for any } i, j \},\ x \in S(\theta).$$

$F(x)$ is obviously a closed convex set for every $x \in S(\theta)$. Given $x \in S(\theta)$ and $i \in \{1, \ldots, n\}$, let $\Delta(i)$ be the number of j's such that $f_i(x) < f_j(x)$ and let

$$y_i^* = \theta^{\Delta(i)} \Big/ \sum_{l=1}^{n} \theta^{\Delta(l)}.$$

Then $y_i^* \geq \delta$ for $i = 1, \ldots, n$. Hence $y^* \in F(x)$ and therefore $F(x)$ is nonempty. Moreover the continuity of f guarantees that F is upper semicontinuous. Thus F

satisfies all conditions of the Kakutani fixed point theorem and so there exists a point $x(\theta) \in S(\theta)$ such that $x(\theta) \in F(x(\theta))$. It is easily seen that $x(\theta)$ is a θ-robust stationary point of f.

So for every $0 < \theta < 1$, f has a θ-robust stationary point $x(\theta)$. Now let us take a sequence $\{\theta_t\}_1^\infty$ of numbers between 0 and 1 converging to zero and a sequence of θ_t-robust stationary points of f. Since S^{n-1} is a compact set, there exists a subsequence converging to a cluster point $x^* \in S^{n-1}$. It is now clear that x^* is a robust stationary point of f. □

In the subsequent sections we will design an algorithm to compute a robust stationary point.

3. The P-triangulation of the unit simplex. We first introduce some notation to be used below. Z_+ and Z_0 represent the set of positive integers and the set of nonnegative integers, respectively. The ith unit vector in R^n is denoted by $e(i)$, $i \in N$. Moreover, $J \subset N$ denotes a proper subset J of N. Let v be a point in the relative interior of S^{n-1}. The point v will be the starting point of the algorithm. We rearrange the components of v in decreasing order to obtain a vector $p = (p_1, \ldots, p_n)^\top \in S^{n-1}$ represented by

$$p_i = v_{j_i}, \text{ for } i \in N$$
$$p_l \geq p_m, \text{ for } l \leq m, \text{ and } l, m \in N$$

where $(j_1, j_2, \ldots, j_n)$ is a permutation of $(1, 2, \ldots, n)$. For $t \in (0, 1]$, let

$$p_i(t) = p_i t^{i-1} \Big/ \sum_{j \in N} p_j t^{j-1}, \text{ for } i \in N,$$

and define

$$p_i(0) = \lim_{t \to 0^+} p_i(t) = \begin{cases} 1 & \text{for } i = 1, \\ 0 & \text{for } i \neq 1. \end{cases}$$

It is readily seen that $p_1(t) \geq p_2(t) \geq \cdots \geq p_n(t)$ for $t \in [0, 1]$.

DEFINITION 3.1. *For $t \in [0, 1]$, the set $A(t)$ is defined by*

$$A(t) = \left\{ x \in R^n \mid \sum_{i \in N} x_i = 1, \ \sum_{j \in J} x_j \leq \sum_{j=1}^{k} p_j(t) \text{ for any } J \subset N \text{ with } k = |J| \right\}.$$

It is easily seen that $A(0) = S^{n-1}$ and that if v is the barycenter of S^{n-1}, then $A(1) = \{v\}$. More generally, for every $t \in [0, 1]$ we have that $v \in A(t)$ and v is a vertex of $A(1)$. Moreover $A(t)$ is a polytope for every $t \in [0, 1]$.

For $J \subset N$ and $t \in [0, 1]$, we define $b(J)$ and $c_J(t)$ by

$$b(J) = \sum_{j \in J} e(j),$$
$$c_J(t) = \sum_{j=1}^{l} p_j(t) \text{ with } l = |J|.$$

Let $\mathcal{I} = \{ I = (I_1, I_2, \ldots, I_m) \mid \emptyset \neq I_1 \subset \cdots \subset I_m \subset N \}$. We say that $I \in \mathcal{I}$ conforms to $J \in \mathcal{I}$, if it holds that every component of I is also a component of J. For $I \in \mathcal{I}$ and a positive integer k, let

$$F(k, I) = \{ x \in A(2^{-k}) \mid b^\top(I_i)x = c_{I_i}(2^{-k}) \text{ for every } i \in \{1, 2, \ldots, m\} \}.$$

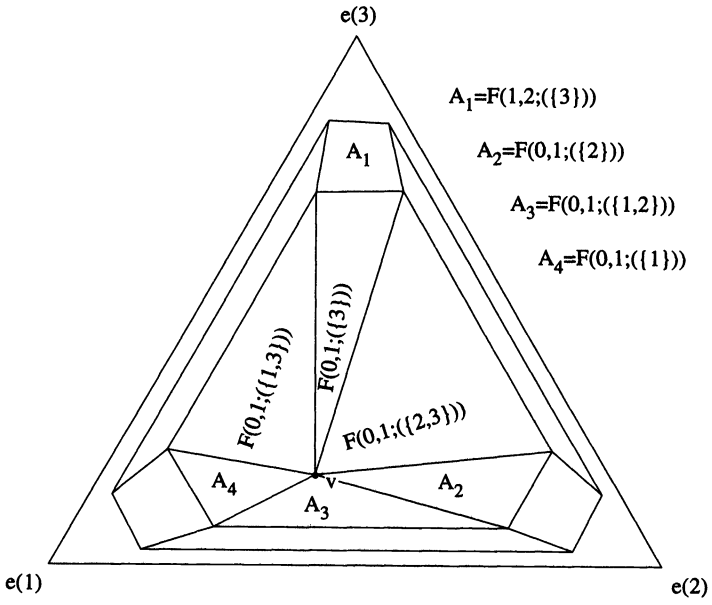


FIG. 1. *The subdivision of* S^{n-1} *for* $n = 3$ *and* $v = (1/2, 1/3, 1/6)^\top$.

Then $F(k, I)$ is a face of $A(2^{-k})$ with dimension equal to $n - 1 - m$. For $I \in \mathcal{I}$, let

$$F(0, 1; I) = \{ x \mid x = av + (1 - a)z \text{ for some } z \in F(1, I) \text{ and some } a \in [0, 1] \}$$

and for $k \in Z_+$

$$F(k, k+1; I) = \{ x \mid x = ay + (1 - a)z \text{ for some } y \in F(k, I), \text{ some } z \in F(k+1, I), \text{ and some } a \in [0, 1] \}.$$

Figure 1 shows the subdivision of S^{n-1} for $n = 3$ and $v = (1/2, 1/3, 1/6)^\top$.

For $I \in \mathcal{I}$ and $k \in Z_+$, we denote the union of $F(i-1, i; I)$ over $i = 1, 2, \ldots, k$ by $\mathcal{F}(k, I)$. We remark that $\mathcal{F}(k, I)$ is not necessarily a convex set. For $k \in Z_0$, we denote the union of $F(k, k+1; I)$ over all $I \in \mathcal{I}$ by $\mathcal{F}(k, k+1)$. For $I \in \mathcal{I}$, we denote the union of $F(k, k+1; I)$ over all $k = 0, 1, \ldots$ by $\mathcal{F}(I)$. Notice that the dimension of $\mathcal{F}(I)$ is equal to $t = n - m$ and that the union of $\mathcal{F}(I)$ over all $I \in \mathcal{I}$ is the relative interior of S^{n-1}. A simplicial subdivision underlying the algorithm must be such that every set $F(k, k+1; I)$ is subdivided into t-dimensional simplices. Such a triangulation can be described as follows. For $I \in \mathcal{I}$, we denote $v(0, I) = v$ and for $k \in Z_+$, let $v(k, I)$ be a relative interior point (e.g., the barycenter) of $F(k, I)$. For $I \in \mathcal{I}$, if I consists of $n-1$ components, then $F(k, I)$ is a vertex of $A(2^{-k})$. For general $I \in \mathcal{I}$, let $F(k, I(n-1))$ be a vertex of $F(k, I)$, i.e., $I(n-1)$ has $n-1$ components and I conforms to $I(n-1)$. Moreover let $(J_1, J_2, \ldots, J_t) = \gamma(I, I(n-1))$ be a conformation of I and $I(n-1)$, where $t = n - m$, i.e., $J_1 = I(n-1)$, $J_k \in \mathcal{I}$ for $k = 2, \ldots, t-1$, $J_t = I$, J_k conforms to J_{k-1} and has one component less than J_{k-1} for $k = 2, \ldots, t$. For given $k \in Z_0$, $I \in \mathcal{I}$, and $\gamma(I, I(n-1))$, the subset $F(k, k+1; I, \gamma(I, I(n-1)))$ of $F(k, k+1; I)$ is defined to be the convex hull of $v(k, J_1)$, $v(k, J_2), \ldots, v(k, J_t)$, $v(k+1, J_1)$, $v(k+1, J_2), \ldots$, and $v(k+1, J_t)$, so

$$F(k, k+1; I, \gamma(I, I(n-1))) = \left\{ x \in S^{n-1} \mid x = v(k, I(n-1)) + \alpha q_0 + \textstyle\sum_{j=1}^{t-1} \alpha_j q_j(\alpha), \right.$$

$$0 \le \alpha \le 1, \text{ and } 0 \le \alpha_{t-1} \le \cdots \le \alpha_1 \le 1 \Big\},$$

where $q_0 = (v(k+1, J_1) - v(k, J_1))$, and for $j = 1, \ldots, t-1$, $0 \le \alpha \le 1$,

$$q_j(\alpha) = \alpha(v(k+1, J_{j+1}) - v(k+1, J_j)) + (1-\alpha)(v(k, J_{j+1}) - v(k, J_j)).$$

The dimension of $F(k, k+1; I, \gamma(I, I(n-1)))$ is equal to t and $F(k, k+1; I)$ is the union of $F(k, k+1; I, \gamma(I, I(n-1)))$ over all conformations $\gamma(I, I(n-1))$ and over all index sets $I(n-1)$ conformed by I.

Let d be an arbitrary positive integer.

DEFINITION 3.2. *For given $k \in Z_0$, $I \in \mathcal{I}$, and $\gamma(I, I(n-1))$, the set $G^d(k, k+1; I, \gamma(I, I(n-1)))$ is the collection of t-simplices $\sigma(a, \pi)$ with vertices $y^1, \ldots, y^{t+1}$ in $F(k, k+1; I, \gamma(I, I(n-1)))$ such that*

(1) *$y^1 = v(k, I(n-1)) + a(0)d^{-1}q_0 + \sum_{j=1}^{t-1} a(j)q_j(a(0)/d)/(a(0)+kd)$, where $a = (a(0), a(1), \ldots, a(n-2))^\top$ is a vector of integers such that $0 \le a(0) \le d-1$ and $a(n-2) = \cdots = a(t) = 0 \le a(t-1) \le \cdots \le a(2) \le a(1) \le a(0) + kd$;*

(2) *$\pi = (\pi_1, \ldots, \pi_t)$ is a permutation of $(0, 1, \ldots, t-1)$ such that $s < s'$ if for some $q \in \{0, 1, \ldots, t-2\}$, it holds that $\pi_s = q$, $\pi_{s'} = q+1$, $a(q) = a(q+1)$ in the case where $q \ge 1$, and $a(0) + kd = a(1)$ in the case where $q = 0$;*

(3) *Let i be such that $\pi_i = 0$. Then*

$$\begin{aligned}
y^{j+1} &= y^j + q_{\pi_j}(a(0)/d)/(a(0)+kd), \ j = 1, \ldots, i-1, \\
y^{i+1} &= v(k, I(n-1)) + (a(0)+1)d^{-1}q_0 \\
&\quad + \sum_{j=1}^{t-1} a(j)q_j((a(0)+1)/d)/(a(0)+1+kd) \\
&\quad + \sum_{j=1}^{i-1} q_{\pi_j}((a(0)+1)/d)/(a(0)+1+kd), \\
y^{j+1} &= y^j + q_{\pi_j}((a(0)+1)/d)/(a(0)+1+kd), \ i < j \le t.
\end{aligned}$$

The set $G^d(k, k+1; I, \gamma(I, I(n-1)))$ is a simplicial subdivision of $F(k, k+1; I, \gamma(I, I(n-1)))$ with grid size d^{-1}. Moreover, the union $G^d(k, k+1; I)$ of $G^d(k, k+1; I, \gamma(I, I(n-1)))$ over all conformations $\gamma(I, I(n-1))$ and $I(n-1)$ conformed by I is a simplicial subdivision of $F(k, k+1; I)$, and the union $G^d(k, k+1)$ of $G^d(k, k+1; I)$ over all sets $I \in \mathcal{I}$ induces a triangulation of $\mathcal{F}(k, k+1)$. Taking the union $G^d(k)$ of $G^d(j, j+1)$ over $j = 0, 1, \ldots, k-1$, we obtain a simplicial subdivision of $A(2^{-k})$ with grid size d^{-1}. The union of $G^d(k)$ over all $k \in Z_0$ is a simplicial subdivision of the relative interior of S^{n-1} and is called the P-triangulation of S^{n-1} with grid size d^{-1}. Observe that for $I \in \mathcal{I}$ and $k \in Z_+$, the union $G^d(k, I)$ of $G^d(i-1, i; I)$ over $i = 1, 2, \ldots, k$, is a simplicial subdivision of the set $\mathcal{F}(I, k)$, and for $I \in \mathcal{I}$, the union $G^d(I)$ of $G^d(k, k+1; I)$ over $k = 0, 1, \ldots$, is a simplicial subdivision of the set $\mathcal{F}(I)$. The P-triangulation of S^{n-1} for $n = 3$, $d = 2$, and $v = (1/3, 1/3, 1/3)^\top$ is illustrated in Fig. 2.

As norm we use the Euclidean norm $||\cdot||$ in R^n. For a set B in R^n, we define the diameter of B by

$$diam(B) = \sup\{\, ||y^1 - y^2|| \,|\, y^1, y^2 \in B \,\}.$$

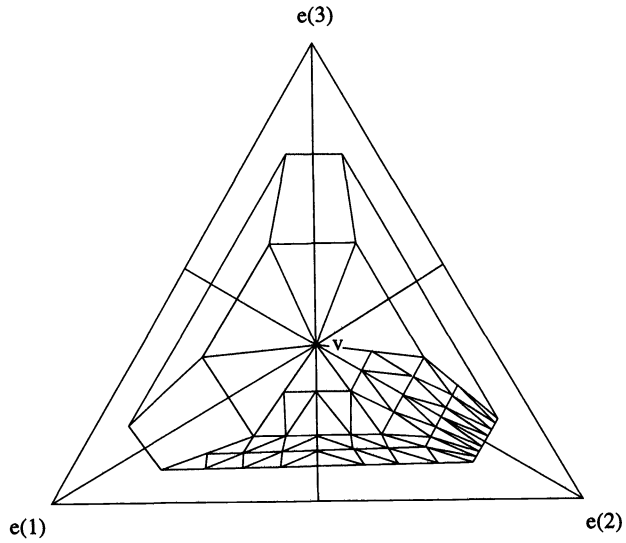


FIG. 2. *The P-triangulation of* S^{n-1} *for* $n = 3$, $d = 2$, *and* $v = (1/3, 1/3, 1/3)^\top$.

Then for given $k \in Z_0$ the mesh size of $G^d(k, k+1)$ is equal to

$$\delta_{k,d} = \sup\{\, diam(\sigma) \mid \sigma \in G^d(k, k+1) \,\}.$$

Now we have the following property.

LEMMA 3.3. *Let* d *be a given positive integer. For the P-triangulation of* S^{n-1} *with grid size* d^{-1}, *it holds that*

$$\lim_{k \to \infty} \delta_{k,d} = 0.$$

4. The algorithm. In this section we discuss how to operate the algorithm in the P-triangulation of S^{n-1} to approximate a robust stationary point of a continuous function on S^{n-1}. Starting at the point v, the algorithm will generate a sequence of adjacent simplices of the P-triangulation in the set $\mathcal{F}(I)$ having I-complete common facets, for varying $I \in \mathcal{I}$.

DEFINITION 4.1. *Let the function* $f : S^{n-1} \mapsto R^n$ *be given. For given* $I = (I_1, \ldots, I_m) \in \mathcal{I}$ *and* $s = t - 1$ *or* t, *where* $t = n - m$, *an* s*-simplex* σ *with vertices* $y^1, \ldots, y^{s+1}$ *is* I*-complete if the system of linear equations*

$$\sum_{i=1}^{s+1} \lambda_i \begin{pmatrix} f(y^i) \\ 1 \end{pmatrix} - \sum_{j=1}^{m} \mu_j \begin{pmatrix} b(I_j) \\ 0 \end{pmatrix} - \beta \begin{pmatrix} e \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{4.1}$$

where e *is an* n*-vector of* 1*'s, has a solution* λ_i^*, $i = 1, \ldots, s+1$, μ_j^*, $j = 1, \ldots, m$, *and* β^* *with* $\lambda_i^* \geq 0$, $i = 1, \ldots, s+1$, *and* $\mu_j^* \geq 0$, $j = 1, \ldots, m$.

A solution λ_i^*, $i = 1, \ldots, s+1$, μ_j^*, $j = 1, \ldots, m$, and β^* will be denoted by $(\lambda^*, \mu^*, \beta^*)$. For $s = t - 1$ we assume that the system (4.1) has a unique solution with $\lambda_i^* > 0$, $i = 1, \ldots, t$, and $\mu_j^* > 0$, $j = 1, \ldots, m$, and that for $s = t$ at most one variable of (λ^*, μ^*) is equal to zero (nondegeneracy assumption). We remark that this assumption can be dropped if we use the lexicographic pivoting method in linear programming to solve system (4.1), see e.g., Todd [23].

The algorithm starts to leave the point v in one out of $n!$ directions. This direction is uniquely determined by $f(v)$. Because of the nondegeneracy assumption, all components of the vector $f(v)$ are different. Let $(i_1,\dots,i_n)$ be a permutation of the set $(1,\dots,n)$ such that $f_{i_1}(v) > \cdots > f_{i_n}(v)$. Then the 0-dimensional simplex $\{v\}$ is I^0-complete with $I^0 = (I^0_1,\dots,I^0_{n-1})$, where $I^0_j = \{i_1,\dots,i_j\}$ for $j = 1,\dots,n-1$. Moreover, $\{v\}$ is a facet of a unique 1-simplex σ^0 in $\mathcal{F}(I^0)$, where $\sigma^0 = \sigma(a,\pi)$ with $a = 0$ and $\pi = (0)$. Since for given $I \in \mathcal{I}$ an I-complete t-simplex has at most two I-complete facets and a facet of a t-simplex in $\mathcal{F}(I)$ is a facet of at most one other t-simplex in $\mathcal{F}(I)$, we obtain that the I-complete t-simplices $\sigma(a,\pi)$ in $\mathcal{F}(I)$ determine sequences of adjacent t-simplices in $\mathcal{F}(I)$ with I-complete common facets. As described below, the sequences of the I-complete t-simplices in $\mathcal{F}(I)$ can be uniquely linked together for varying $I \in \mathcal{I}$ to obtain sequences of adjacent simplices of varying dimension. Under the nondegeneracy assumption, one of these sequences starts with σ^0 in $\mathcal{F}(I^0)$ and is followed by the algorithm. Thus, starting at the point v, the algorithm generates a unique sequence of I-complete adjacent t-simplices in $\mathcal{F}(I)$ of varying dimension. In this way within a finite number of steps either the algorithm reaches a point $\bar{x}$ in an $(n-1)$-dimensional simplex for which $\bar{f}_i(\bar{x}) = \bar{f}_j(\bar{x})$ for every i and $j \in N$, where $\bar{f}$ is the piecewise linear (PL) approximation of f with respect to the P-triangulation, or for $k = 1,2,\dots$ the algorithm finds an I-complete $(t-1)$-simplex in $F(k,I)$ for some $I \in \mathcal{I}$. Suppose the latter case holds, then we have the following result.

LEMMA 4.2. *For $k \in Z_+$ and $I \in \mathcal{I}$, let σ with vertices $y^1,\dots,y^t$ be an I-complete $(t-1)$-simplex lying in $F(k,I)$. Let $(\lambda^*,\mu^*,\beta^*)$ be the corresponding unique solution of system (4.1). Then $x = \sum_{i=1}^t \lambda_i^* y^i$ is a 2^{-k}-robust stationary point of the PL approximation $\bar{f}$ of f with respect to the P-triangulation. Moreover, x is a stationary point of $\bar{f}$ on $A(2^{-k})$.*

Proof. Since $I = (I_1, I_2,\dots, I_m) \in \mathcal{I}$, there exist $l_1 < l_2 < \cdots < l_m$ such that

$$\begin{aligned} I_1 &= \{\, i_1,\dots,i_{l_1} \,\},\\ I_2 &= \{\, i_1,\dots,i_{l_1}, i_{l_1+1},\dots,i_{l_2} \,\},\\ &\vdots\\ I_m &= \{\, i_1,\dots,i_{l_m} \,\},\\ N\backslash I_m &= \{\, i_{l_m+1},\dots,i_n \,\}. \end{aligned}$$

Then it follows from equation (4.1) that

$$\begin{aligned} \bar{f}_{i_1}(x) = \cdots = \bar{f}_{i_{l_1}}(x) &= \mu_1^* + \cdots + \mu_m^* + \beta^*\\ > \bar{f}_{i_{l_1+1}}(x) = \cdots = \bar{f}_{i_{l_2}}(x) &= \mu_2^* + \cdots + \mu_m^* + \beta^*\\ &\vdots\\ > \bar{f}_{i_{l_{m-1}+1}}(x) = \cdots = \bar{f}_{i_{l_m}}(x) &= \mu_m^* + \beta^*\\ > \bar{f}_{i_{l_m+1}}(x) = \cdots = \bar{f}_{i_n}(x) &= \beta^*, \end{aligned}$$

where $\mu_i^* > 0$ for $i = 1,\dots,m$. Now it is not difficult to check that

$$x_i \le 2^{-k} x_j \quad \text{whenever } \bar{f}_i(x) < \bar{f}_j(x).$$

This means that x is a 2^{-k}-robust stationary point of the PL approximation $\bar{f}$ of f with respect to the P-triangulation.

Moreover, for each face $F(k, I)$, $I \in \mathcal{I}$, let $F^*(I)$ be the set of all n-dimensional vectors y such that every point of $F(k, I)$ is a solution of the linear programming problem

$$\max\ y^\top \hat{x} \quad \text{subject to } \hat{x} \in A(2^{-k}).$$

Then the stationary point problem for $\bar{f}$ on $A(2^{-k})$ is the problem of finding a point x in $A(2^{-k})$ such that $\bar{f}(x) \in F^*(I)$ for a minimum face $F(k, I)$ of $A(2^{-k})$ containing x. Duality theory implies that $F^*(I) = \{\, y \mid y = \sum_{i=1}^{m} \mu_i b(I_i) + \beta e,\ \mu_i \geq 0 \text{ for } i = 1, \ldots, m, \text{ and } \beta \in R \,\}$. It follows from equation (4.1) that $\bar{f}(x) \in F^*(I)$. Hence x is a stationary point of $\bar{f}$ on $A(2^{-k})$. □

To extend the domain of the PL approximation $\bar{f}$ of f, we recall that for a given positive integer d, the mesh size $\delta_{k,d}$ converges to zero as k goes to infinity. We can therefore take $\bar{f}(x)$ to be $f(x)$ if x lies on the boundary of S^{n-1}, since f is a continuous function. Hence $\bar{f}$ is also a continuous function from S^{n-1} into R^n.

For each $t \in [0, 1/2]$, let $V(t)$ denote the set of stationary points of f on $A(t)$. We summarize the following observations from the above discussions:

(P1) For each $t \in [0, 1/2]$, the set $V(t)$ is a nonempty closed set.

(P2) For each $t \in (0, 1/2]$, $x \in A(t)$ is a t-robust stationary point of f on S^{n-1} if and only if x belongs to the set $V(t)$.

(P3) For each $t \in [0, 1/2]$, if $x \in V(t)$ lies in the interior of $A(t)$, then all the components of $f(x)$ must be the same.

The next lemma shows that a 2^{-k}-robust stationary point of $\bar{f}$ on S^{n-1} is an approximate 2^{-k}-robust stationary point of f on S^{n-1}.

LEMMA 4.3. *Let* $\eta_{k,d} = \sup\{\, diam(f(\sigma)) \mid \sigma \in G^d(k-1, k) \,\}$. *Let* x *be a* 2^{-k}*-robust stationary point of the PL approximation* $\bar{f}$ *of* f *with respect to the P-triangulation obtained by the algorithm, so that* $x \in F(k, I_k)$ *for some* $I_k \in \mathcal{I}$. *Then* $f(x)$ *lies in the* $\eta_{k,d}$*-neighborhood of* $F^*(I_k)$, *i.e., there is a* $y \in F^*(I_k)$ *such that* $||y - f(x)|| \leq \eta_{k,d}$.

Proof. Let $y^1, \ldots, y^t$ be the vertices of a $(t-1)$-simplex of $G^d(k-1, k)$ in $F(k, I_k)$ containing x. Then $\bar{f}(x) = \sum_{j=1}^{t} \lambda_j^* f(y^j)$ lies in $F^*(I_k)$, where $\lambda_1^*, \ldots, \lambda_t^*$ are convex combination coefficients such that $x = \sum_{j=1}^{t} \lambda_j^* y^j$. Therefore

$$\begin{aligned} ||\bar{f}(x) - f(x)|| &= \left\| \sum_{j=1}^{t} \lambda_j^* f(y^j) - f(x) \right\| \\ &= \left\| \sum_{j=1}^{t} \lambda_j^* (f(y^j) - f(x)) \right\| \\ &\leq \sum_{j=1}^{t} \lambda_j^* ||f(y^j) - f(x)|| \\ &\leq \eta_{k,d}. \end{aligned}$$

Note that the following inequality also holds:

$$||\bar{f}(z) - f(z)|| \leq \eta_{k,d}, \text{ for any } z \in \mathcal{F}(k-1, k). \qquad \square$$

Next we discuss the case where the algorithm converges to a boundary point of S^{n-1}. Since S^{n-1} is compact and f is continuous on S^{n-1}, the error $\eta_{k,d}$ in Lemma

4.3 tends to zero as d is fixed and $\delta_{k,d}$ goes to zero when k goes to infinity. Let x^k be a 2^{-k}-robust stationary point of $\bar{f}$ and let $\eta_{k,d}$ be the error in Lemma 4.3. We can therefore consider x^k as an approximate 2^{-k}-robust stationary point of f. Then the algorithm generates a sequence $\{ x^k \mid k = 1, 2, \dots \}$ of approximate 2^{-k}-robust stationary points of f which therefore has a cluster point x^*. For simplicity of notation we can assume that this sequence itself converges to x^*. We are now ready to state the following result.

THEOREM 4.4. *Suppose that for a given positive integer d the vector x^k is an approximate 2^{-k}-robust stationary point generated by the algorithm, for $k = 1, 2, \dots$, i.e., for each $k \in Z_+$, $x^k \in F(k, I_k)$ with $I_k \in \mathcal{I}$ is a 2^{-k}-robust stationary point of $\bar{f}$. Then the sequence $\{ x^k \mid k = 1, 2, \dots \}$ has a cluster point x^* which is a robust stationary point of f on S^{n-1}.*

Proof. By definition, x^* is a robust stationary point of $\bar{f}$ on S^{n-1}. Notice that x^* lies on the boundary of S^{n-1}. We shall demonstrate that for any given $\epsilon > 0$, there exists a positive integer M, such that for $k \in Z_+$ with $k > M$, there is a 2^{-k}-robust stationary point $y^k \in A(2^{-k})$ of f on S^{n-1} which is in the ϵ-neighborhood of x^k.

Let

$$U(t) = \begin{cases} A(t) & \text{for } t \in [0, 1/2], \\ 2(1-t)A(2^{-1}) + (2t-1)\{ v \} & \text{for } t \in [1/2, 1], \end{cases}$$

and denote the set of stationary points of f on $U(t)$ by $Y(t)$ for $t \in [0, 1]$. Observe that $Y(t) = V(t)$ for $t \in [0, 1/2]$, and that $U(t)$ is contained in $A(2^{-1})$ for $t \in [1/2, 1]$. As t decreases from 1 to 0, $U(t)$ expands from the starting point v to the set $A(2^{-1})$ and then to the whole set S^{n-1}.

Now we define a function g_t (see, e.g., Doup [4] and Yamamoto [25]) from $U(t)$ into itself by

$$g_t(x) = \operatorname{argmin}\{ \| x + f(x) - y \| \mid y \in U(t) \}, x \in U(t).$$

Since $U(t)$ is a convex set and f is continuous, g_t is nonexpansive and hence is a Lipschitz continuous function. It is readily seen that $x \in Y(t)$ if and only if $x = g_t(x)$. Let

$$H(x, t) = x - g_t(x).$$

This is a Lipschitz continuous homotopy defined on $S^{n-1} \times [0, 1]$ between $H(x, 1) = x - v$ and $H(x, 0) = x - g_0(x)$. Now set

$$H^{-1}(0) = \{ (x, t) \in S^{n-1} \times [0, 1] \mid H(x, t) = 0 \}.$$

Let $W(t)$ denote the set of stationary points of $\bar{f}$ on $U(t)$ for $t \in [0, 1]$. Similarly, we can construct a Lipschitz continuous function with respect to $\bar{f}$, i.e.,

$$G : S^{n-1} \times [0, 1] \mapsto R^n,$$

such that

$$W(t) = \{ x \in S^{n-1} \mid G(x, t) = 0 \}, \ t \in [0, 1].$$

Set

$$G^{-1}(0) = \{ (x, t) \in S^{n-1} \times [0, 1] \mid G(x, t) = 0 \}.$$

For each $k \in Z_+$, let

$$\xi^k = (x^k, 2^{-k}).$$

It is clear that $\lim_{k\to\infty} \xi^k = \xi^* = (x^*, 0)$, since $\lim_{k\to\infty} x^k = x^*$. We define

$$N(\epsilon) = \{\, (x,t) \in S^{n-1} \times [0,1] \,|\, ||(x,t) - (z,s)|| < \epsilon \text{ for some } (z,s) \in H^{-1}(0) \,\}.$$

Clearly, it holds that

$$||H(\psi)|| > 0 \text{ for any } \psi \in S^{n-1} \times [0,1] \backslash N(\epsilon).$$

But $N(\epsilon)$ is open, so the set $S^{n-1} \times [0,1]\backslash N(\epsilon)$ is compact. The compactness means that the minimum can be attained and for some $\nu > 0$

$$\min\{\, ||H(\psi)|| \,|\, \psi \in S^{n-1} \times [0,1]\backslash N(\epsilon) \,\} > \nu.$$

Hence, if $\psi \in S^{n-1} \times [0,1]$ satisfies

$$||H(\psi)|| \leq \nu, \tag{4.2}$$

then ψ must be in $N(\epsilon)$. Because H is uniformly continuous on $S^{n-1} \times [0,1]$ and $\bar{f}$ is the PL approximation of f, it follows that

$$||H(\psi) - G(\psi)|| < \nu \tag{4.3}$$

for any $\psi = (x,t) \in S^{n-1} \times [0,1]$ under the condition that the diameter of simplices in which x lies is small enough, say, smaller than $\Delta > 0$.

Lemma 3.3 states that given a positive integer d, as k goes to infinity, the mesh size $\delta_{k,d}$ converges to zero. It implies that there exists a positive integer M such that for every $k \in Z_+$ with $k > M$, it holds that

$$\delta_{k,d} < \Delta.$$

Since for any $k \in Z_+$ with $k > M$, $\xi^k \in G^{-1}(0)$, i.e., $G(\xi^k) = 0$, it follows from (4.3) that

$$||H(\xi^k)|| < \nu.$$

By (4.2) ξ^k must be in $N(\epsilon)$. This implies that for any $k \in Z_+$ with $k > M$, there is $\psi^k \in H^{-1}(0)$ which is in the ϵ-neighborhood of ξ^k. Without loss of generality we may assume that $\psi^k = (y^k, 2^{-k})$. This is what we claimed.

On the other hand, since $\lim_{k\to\infty} x^k = x^*$, it immediately follows that

$$\lim_{k\to\infty} y^k = x^*.$$

Hence x^* is a robust stationary point of f on S^{n-1}. □

In the case where the algorithm terminates with an $(n-1)$-dimensional simplex σ with vertices $y^1, \ldots, y^n$, $\bar{x} = \sum_{i=1}^n \lambda_i^* y^i$ is a robust stationary point of $\bar{f}$. If the accuracy of approximation is not satisfactory, the algorithm can be restarted at the point $\bar{x}$ with a smaller grid size d^{-1} to find a better approximate robust stationary point, hopefully within a small number of steps. Without loss of generality we assume that the algorithm generates a sequence $\{\, \bar{x}^h \,|\, h = 1, 2, \ldots \}$, where $\bar{x}^h$ is the robust

stationary point of $\bar{f}$ corresponding to the grid size d_h^{-1} for an increasing sequence of positive integers $\{ d_h \mid h = 1, 2, \dots \}$. It is readily seen that for every $k \in Z_0$, the mesh size δ_{k,d_h} tends to zero as h goes to infinity. Therefore the sequence $\{ \bar{x}^h \mid h = 1, 2, \dots \}$ has a subsequence converging to a cluster point x^*. Clearly, x^* is a robust stationary point of f on S^{n-1}.

As described above, starting at the point v, the algorithm generates a unique sequence of adjacent t-simplices $\sigma(a, \pi)$ in $\mathcal{F}(I)$ for varying $I \in \mathcal{I}$ of varying dimension $t = n - m$. When, with respect to some $\sigma(a, \pi)$ with vertices $y^1, \dots, y^{t+1}$ in some $G^d(k, k+1; I, \gamma(I, I(n-1)))$ for some $k \in Z_0$ and $\gamma(I, I(n-1))$, the variable λ_q, for some q, $1 \leq q \leq t+1$, becomes zero through a linear programming (LP) pivot step in (4.1), then the replacement step determines the unique t-simplex $\bar{\sigma}(\bar{a}, \bar{\pi})$ in $F(k, k+1; I, \gamma(I, I(n-1)))$ sharing with σ the common facet τ opposite vertex y^q, unless this facet lies in the boundary of $F(k, k+1; I, \gamma(I, I(n-1)))$. If τ does not lie in the boundary of the set $F(k, k+1; I, \gamma(I, I(n-1)))$, then $\bar{\sigma}(\bar{a}, \bar{\pi})$ can be obtained from a and π as given in Table 1, where $E(j-1)$ is the jth unit vector in R^{n-1}, $j = 1, \dots, n-1$.

TABLE 1
Parameters of $\bar{\sigma}$ if the vertex y^q of $\sigma(a, \pi)$ is replaced.

	$\bar{\pi}$	$\bar{a}$
$q = 1$	$(\pi_2, \dots, \pi_t, \pi_1)$	$a + E(\pi_1)$
$1 < q < t+1$	$(\pi_1, \dots, \pi_{q-2}, \pi_q, \pi_{q-1}, \pi_{q+1}, \dots, \pi_t)$	a
$q = t+1$	$(\pi_t, \pi_1, \dots, \pi_{t-1})$	$a - E(\pi_t)$

The algorithm continues with $\bar{\sigma}$ by making an LP pivot step in (4.1) with $(f^\top(\bar{y}), 1)^\top$, where $\bar{y}$ is the vertex of $\bar{\sigma}$ opposite the facet τ. In the case where a facet τ of a simplex in $G^d(k, k+1; I, \gamma(I, I(n-1)))$ is not a facet of another simplex in $G^d(k, k+1; I, \gamma(I, I(n-1)))$, τ lies in the boundary of $F(k, k+1; I, \gamma(I, I(n-1)))$. According to Definition 3.2 we have the following lemma.

LEMMA 4.5. *Let $\sigma(a, \pi)$ be a t-simplex in $F(k, k+1; I, \gamma(I, I(n-1)))$. The facet τ of σ opposite the vertex y^q, $1 \leq q \leq t+1$, lies in the boundary of this set if and only if one of the following cases occurs:*

(i) *$1 < q < t+1$, $\pi_q = h+1$, $\pi_{q-1} = h$ for some $h \in \{0, 1, \dots, t-2\}$, and $a(h) = a(h+1)$ in the case where $h \geq 1$, and $a(0) + kd = a(1)$ in the case where $h = 0$;*

(ii) *$q = t+1$, $\pi_t = t-1$, and $a(t-1) = 0$;*

(iii) *$q = 1$, $\pi_1 = 0$, and $a(0) = d - 1$;*

(iv) *$q = t+1$, $\pi_t = 0$, and $a(0) = 0$.*

Suppose the algorithm generates the simplex $\sigma(a, \pi)$ as given in Lemma 4.5 and λ_q becomes zero after making an LP pivot step in (4.1). Then the facet τ of σ opposite the vertex y^q is I-complete. In case (iii) the facet τ lies in the face $F(k+1, I)$ of $A(2^{-k-1})$ and the algorithm reaches a 2^{-k-1}-robust stationary point $\bar{x} = \sum_{i=2}^{t+1} \lambda_i^* y^i$ of $\bar{f}$ lying in $F(k+1, I)$. If k is large enough, then $\bar{x}$ is an approximate robust stationary point of f. Otherwise, the algorithm proceeds with $\bar{\sigma}$ by making an LP pivot step in (4.1) with $(f^\top(\bar{y}), 1)^\top$, where $\bar{y}$ is the vertex of $\bar{\sigma}$ opposite the facet τ and $\bar{\sigma}$ in $F(k+1, k+2; I, \gamma(I, I(n-1)))$ is obtained according to Table 1.

In case (iv) the facet τ lies in the face $F(k, I)$ of $A(2^{-k})$ and the algorithm continues with $\bar{\sigma}$ by making an LP pivot step in (4.1) with $(f^\top(\bar{y}), 1)^\top$, where $\bar{y}$ is the vertex of $\bar{\sigma}$ opposite the facet τ and $\bar{\sigma}$ in $F(k-1, k; I, \gamma(I, I(n-1)))$ is also obtained from Table 1.

In case (i) and if $h \geq 1$, the facet τ is a facet of the t-simplex $\bar{\sigma} = \sigma(a, \pi)$ in $F(k, k+1; I)$ lying in the subset $F(k, k+1; I, \bar{\gamma}(I, I(n-1)))$ with

$$\bar{\gamma}(I, I(n-1)) = (J_1, \ldots, J_h, \bar{J}_{h+1}, J_{h+2}, \ldots, J_t),$$

where $\bar{J}_{h+1} \in \mathcal{I}$, $\bar{J}_{h+1} \neq J_{h+1}$, is uniquely determined by the properties that $\bar{J}_{h+1}$ conforms to J_h, has one component less than J_h, and is conformed by J_{h+2}. In case (i) and if $h = 0$, then τ is a facet of the t-simplex $\bar{\sigma} = \sigma(a, \pi)$ in $F(k, k+1; I, \bar{\gamma}(I, \bar{I}(n-1)))$ with $\bar{I}(n-1)$ and $\bar{\gamma}$ defined as follows. Let $J_1 = I(n-1) = (I_1, \ldots, I_{n-1})$. When $J_2 = (I_1, \ldots, I_{n-2})$, we have $\bar{I}(n-1) = (I_1, \ldots, I_{n-2}, \bar{I}_{n-1})$ with $\bar{I}_{n-1} = I_{n-2} \bigcup (N \backslash I_{n-1})$. When $J_2 = (I_2, \ldots, I_{n-1})$, let $\bar{I}(n-1) = (\bar{I}_1, I_2, \ldots, I_{n-1})$ with $\bar{I}_1 = I_2 \backslash I_1$. Finally if $J_2 = (I_1, \ldots, I_k, I_{k+2}, \ldots, I_{n-1})$ for some $k \in \{ 1, \ldots, n-3 \}$, we have $\bar{I}(n-1) = (I_1, \ldots, I_k, \bar{I}_{k+1}, I_{k+2}, \ldots, I_{n-1})$ with $\bar{I}_{k+1} = I_k \bigcup (I_{k+2} \backslash I_{k+1})$. Then $\bar{\gamma}(I, \bar{I}(n-1)) = (\bar{I}(n-1), J_2, \ldots, J_t)$. In both subcases of case (i) the algorithm continues by making a pivot step in (4.1) with $(f^\top(\bar{y}), 1)^\top$, where $\bar{y}$ is the vertex of the new t-simplex $\bar{\sigma}$ opposite the facet τ.

In case (ii) the facet lies in the set $F(k, k+1; J_{t-1})$ of $\mathcal{F}(I)$. More precisely, τ is the $(t-1)$-simplex $\sigma(a, \bar{\pi})$ in $F(k, k+1; \bar{I}, \bar{\gamma}(\bar{I}, I(n-1)))$, where $\bar{I} = J_{t-1}$, $\bar{\gamma}(\bar{I}, I(n-1)) = (J_1, \ldots, J_{t-1})$, and $\bar{\pi} = (\pi_1, \ldots, \pi_{t-1})$. The algorithm now proceeds by making a pivot step in (4.1) with $(-b^\top(I_h), 0)^\top$, where I_h is the unique component of J_{t-1} but not of J_t.

Finally, if through an LP pivot step in (4.1), the variable μ_h becomes 0 for some $h \in \{ 1, \ldots, m \}$, then the algorithm terminates with the approximate robust stationary point $\bar{x} = \sum_i \lambda_i^* y^i$ of f if $m = 1$ and restarts at the point $\bar{x}$ with a smaller grid size in case the accuracy is not satisfactory. Otherwise, the simplex $\sigma(a, \pi)$ is a facet of a unique $(t+1)$-simplex σ in $\mathcal{F}(\bar{I})$ with $\bar{I} = (I_1, \ldots, I_{h-1}, I_{h+1}, \ldots, I_m)$. More precisely, $\bar{\sigma} = \sigma(a, \bar{\pi})$ lies in $F(k, k+1; \bar{I}, \bar{\gamma}(\bar{I}, I(n-1)))$, where $\bar{\gamma}(\bar{I}, I(n-1)) = (\gamma, \bar{I})$, and $\bar{\pi} = (\pi_1, \ldots, \pi_t, t)$. The algorithm continues by making a pivot step in (4.1) with $(f^\top(\bar{y}), 1)^\top$, where $\bar{y}$ is the vertex of $\bar{\sigma}$ opposite the facet σ. This concludes the description of how the algorithm works in the P-triangulation of S^{n-1}.

5. Examples. Now we give some examples to show the power of the robust stationary point concept and the algorithm as well. Let us briefly review the standard model of a pure exchange economy. For details, we refer to Varian [24]. In such an economy there are, say, n commodities and a finite number of consumers, each having a vector of initial endowments. Exchange of commodities is based on relative prices. All consumers exchange goods in order to maximize their utility under their initial wealth constraints. This economy can be characterized by an excess demand function $z : R_+^n \backslash \{0\} \longrightarrow R^n$ which satifies the following standard conditions:

(i) z is a continuous function,

(ii) $z(\lambda p) = z(p)$ for any $\lambda > 0$ and $p \in R_+^n \backslash \{0\}$ (homogeneity),

(iii) $p^\top z(p) = 0$ for $p \in R_+^n \backslash \{0\}$ (Walras' law).

The element $p^* \in R_+^n \backslash \{0\}$ is an equilibrium price vector if $z(p^*) \leq 0$ (see Varian [24, p. 321]). Note that homogeneity permits us to normalize the price vectors to the $(n-1)$-dimensional unit simplex S^{n-1}. Now it is not hard to show that this problem is equivalent to the stationary point problem on S^{n-1}. We first present two examples of such an economic equilibrium model. To keep things simple and interesting, we shall focus on excess demand functions.

Example 1. There are two goods. The excess demand function is given by $z(p) = (p_1 p_2^2(1-p_1^2), -p_1^2 p_2(1-p_1^2))^\top$ for $p \in S^1$. There are two equilibria (i.e., stationary points) $x = (1,0)^\top$, $y = (0,1)^\top$. However only x is a robust stationary point. Further,

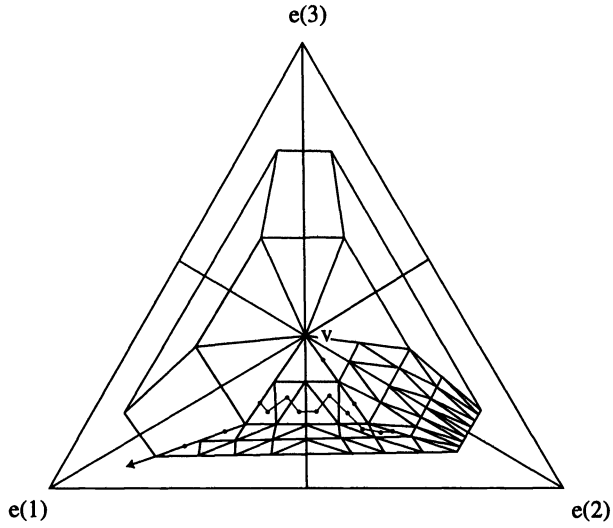


FIG. 3. *The path of the algorithm on* S^2 *for* $d = 2$ *and* $v = (1/3, 1/3, 1/3)^\top$.

x is more sensible than y in economic terms. We need to give some explanation. The basic idea of the Walrasian tatonnement process is as follows. Suppose that an economy is in disequilibrium. Then the Walrasian auctioneer would increase the price of a commodity if the excess demand of that commodity were positive, and decrease the price of a commodity if the excess demand of that commodity were negative. A sensible equilibrium should be stable against some data perturbation. In this example, suppose that the economy is slightly perturbed away from the equilibrium $(0, 1)^\top$. Then the tatonnement process would lead to another equilibrium $(1, 0)^\top$.

Example 2. There are three goods. The excess demand function is given by $z(p) = (p_2p_3, p_1p_3^2, -p_1p_2(1 + p_3))^\top$ for $p \in S^2$. The set of stationary points is $\{p \in S^2 | p_3 = 0\}$. But z only has one robust stationary point: $p^* = (1, 0, 0)^\top$. This fact is quite surprising. Moreover, the equilibrium price vector p^* is also most desirable from an economic point of view.

Finally, we conclude with two more examples.

Example 3. The function is defined by $f(x) = (x_1 + x_2, x_2 + x_3, x_3 + x_1)^\top$ for $x \in S^2$. The set of stationary points is

$$\{(1/3, 1/3, 1/3)^\top, (1, 0, 0)^\top, (0, 1, 0)^\top, (0, 0, 1)^\top\}.$$

However, f just has one robust stationary point: $(1/3, 1/3, 1/3)^\top$.

Example 4. The function is given by $f(x) = (x_1 - 9x_3, -7x_3, -9x_1 - 7x_2 - 7x_3)^\top$ for $x \in S^2$. This example is due to Myerson [16] and is often used in game theory literature. There are three stationary points: $(1, 0, 0)^\top$, $(0, 1, 0)^\top$, and $(0, 0, 1)^\top$. But f just has one robust stationary point: $(1, 0, 0)^\top$. The path followed by the algorithm is illustrated in Fig. 3 where the starting point is $(1/3, 1/3, 1/3)^\top$. The algorithm converges to the robust stationary point $(1, 0, 0)^\top$. At the $13 + 2(k - 1)$ step for $k \in Z_+$, we get the following approximate robust stationary point:

$$x^k = \left(\frac{1}{1 + 2^{-k} + 2^{-2k}}, \frac{2^{-k}}{1 + 2^{-k} + 2^{-2k}}, \frac{2^{-2k}}{1 + 2^{-k} + 2^{-2k}}\right)^\top.$$

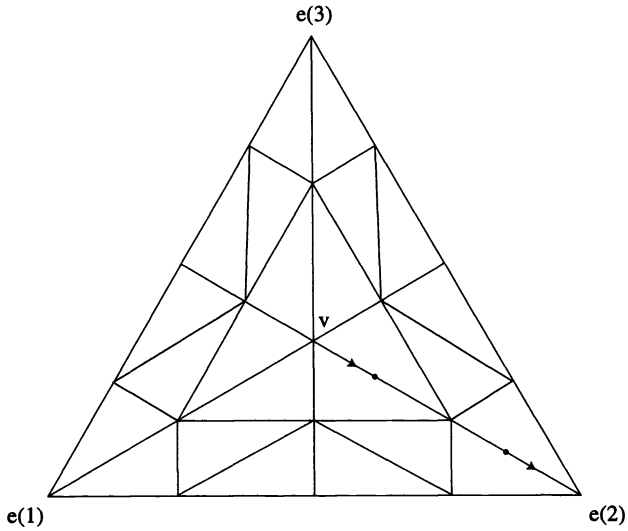


FIG. 4. *The path of the algorithm of van der Laan and Talman on* S^2 *for* $d = 2$ *and* $v = (1/3, 1/3, 1/3)^\top$.

It is easy to see that for each $k \in Z_+$, x^k is in fact a 2^{-k}-robust stationary point of f. By using this example and the same starting point, we also implement the well-known algorithm of van der Laan and Talman [13] (see also Doup and Talman [4]) based on the vector labelling and the V-triangulation [4]. No matter how fine triangulation is employed, their algorithm converges to the stationary point $(0, 1, 0)^\top$, which is not a robust stationary point. The path is shown in Fig. 4.

Acknowledgments. I would like to thank Gerard van der Laan, Dolf Talman, and Yoshi Yamamoto for their stimulating discussions, valuable comments, and constructive suggestions. The referees' helpful comments are gratefully acknowledged.

REFERENCES

[1] E. L. ALLGOWER AND K. GEORG, *Simplicial and continuation methods for approximating fixed points and solutions to systems of equations*, SIAM Rev., 22 (1980), pp. 28–85.
[2] E. VAN DAMME, *Stability and Perfection of Nash Equilibria*, Springer-Verlag, Berlin, 1987.
[3] T. M. DOUP, *Simplicial Algorithms on the Simplotope*, Lecture Notes in Econom. and Math. Systems 318, Springer-Verlag, Berlin, 1988.
[4] T. M. DOUP AND A. J. J. TALMAN, *A new variable dimension simplicial algorithm to find equilibria on the product space of unit simplices*, Math. Programming, 37 (1987), pp. 319–355.
[5] T. M. DOUP, G. VAN DER LAAN, AND A. J. J. TALMAN, *The* $(2^{n+1} - 2)$*-ray algorithm: A new simplicial algorithm to compute economic equilibria*, Math. Programming, 39 (1987), pp. 241–252.
[6] B. C. EAVES, *On the basic theory of complementarity*, Math. Programming, 1 (1972), pp. 68–75.
[7] ———, *Homotopies for computation of fixed points*, Math. Programming, 3 (1972), pp. 1–22.
[8] B. C. EAVES AND R. SAIGAL, *Homotopies for the computation of fixed points on unbounded region*, Math. Programming, 3 (1972), pp. 225–237.
[9] R. W. FREUND, *Variable dimension complexes, Part* I: *Basic theory*, Math. Oper. Res., 9 (1984), pp. 479–497.
[10] M. KOJIMA AND Y. YAMAMOTO, *Variable dimension algorithms: Basic theory, interpretation, and extensions of existing methods*, Math. Programming, 24 (1982), pp. 177–215.
[11] H. W. KUHN, *Approximate search for fixed points*, Comput. Meth. Optim. Problems, 2 (1969), pp. 199–211.

[12] H. W. KUHN AND J. G. MACKINNON, *The Sandwish method for finding fixed points*, J. Optim. Theory Appl., 17 (1975), pp. 189–204.
[13] G. VAN DER LAAN AND A. J. J. TALMAN, *A restart algorithm for computing fixed points without an extra dimension*, Math. Programming, 20 (1979), pp. 33–48.
[14] ———, *An improvement of fixed point algorithms by using a good triangulation*, Math. Programming, 18 (1980), pp. 274–285.
[15] C. E. LEMKE AND J. T. HOWSON, *Equilibrium points of bimatrix games*, SIAM Rev., 12 (1964), pp. 413–423.
[16] R. B. MYERSON, *Refinements of Nash equilibrium concepts*, Internat. J. Game Theory, 8 (1978), pp. 73–80.
[17] H. SCARF, *The approximation of fixed points of a continuous mapping*, SIAM J. Applied Math., 15 (1967), pp. 157–172.
[18] ———, *The Computation of Economic Equilibria*, Yale University Press, New Haven, CT, 1973.
[19] A. SCHRIJVER, *Theory of Linear and Integer Programming*, John Wiley & Sons, New York, 1986.
[20] L. S. SHAPLEY, *On balanced games without side payments*, in Mathematical Programming, T. C. Hu and S. M. Robison, eds., Academic Press, New York, 1973, pp. 261–290.
[21] S. SMALE, *A convergent process of price adjustment and global Newton methods*, J. Math. Econ., 3 (1976), pp. 107–120.
[22] A. J. J. TALMAN AND Y. YAMAMOTO, *A simplicial algorithm for stationary point problems on polytopes*, Math. Oper. Res., 14 (1989), pp. 383–399.
[23] M. J. TODD, *The Computation of Fixed Points and Applications*, Lecture Notes in Econom. and Math. Systems 124, Springer-Verlag, Berlin, 1976.
[24] H. VARIAN, *Microeconomic Analysis*, 3rd Ed., W.W. Norton and Company, New York, 1992.
[25] Y. YAMAMOTO, *A path-following procedure to find a proper equilibrium of finite games*, Internat. J. Game Theory, 22 (1993), pp. 49–59.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 507–520, March 1996

006

RECIPROCAL REALIZATIONS ON THE CIRCLE*

JAN-ÅKE SAND†

Abstract. Reciprocal realizations on a circle are defined. The concepts of minimality, interior observability, and exterior observability are introduced and related to each other, using geometric methods. In particular, the concept of a splitting subspace plays a central role.

Key words. reciprocal process, stochastic realization, splitting subspace

AMS subject classifications. 93E03, 94A12

1. Introduction.

Stochastic models of random phenomena that are spatially distributed are useful in applications such as image processing and computer vision; see, e.g., the paper by Levy [3] and references therein. It is often assumed then that the given object, a random process y, satisfies some strong assumptions about its dependency structure, e.g., that y is a Markov field. In order to relax these assumptions we may look for a model where y is the *output* of some underlying process x satisfying these stronger assumptions, i.e., a stochastic realization of y.

This is a natural generalization of the idea of state-space modeling of time series and leads naturally to the question whether it is possible to develop some stochastic realization theory for spatially indexed stochastic processes in the spirit of the geometric stochastic realization theory of Lindquist, Pavon, Picci and Ruckebusch [5]–[8].

Moreover, let us mention that the idea of realizing y as the output of a Markov field is present in the literaure on so-called hidden Markov models.

In this paper, as a prototype problem, we shall study stochastic processes defined on a discrete circle $\mathbb{T}$. The circle provides a parameter set that exhibits spatial properties but still has the advantage, from an analytical point of view, of being one dimensional.

More specifically, we shall study processes which are outputs of *reciprocal* state processes and define *reciprocal realizations.* Moreover, we shall define stochastic minimality, interior observability, and exterior observability for a reciprocal realization and relate these concepts to each other. Our approach will be geometric in the spirit of Lindquist and Picci, and the concept of a splitting subspace will play an important role when discussing minimality. For reciprocal realizations it turns out that the concept of minimality is quite subtle, since minimality of a reciprocal realization and minimality of its corresponding splitting subspaces do not coincide, as happens in stochastic realization theory for time series.

Reciprocal processes indexed by a discrete set were studied by Levy, Frezza, and Krener in [4], where they obtained finite-dimensional models for representing a certain class of Gaussian reciprocal processes. Such a model will be the basic ingredient in a reciprocal realization. Since our approach is geometric, we shall reformulate in a geometric language some of the results in [4] as well as add some new results on reciprocal processes.

* Received by the editors April 8, 1994. Accepted for publication (in revised form) November 10, 1994.

† Division of Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, 100 44 Stockholm, Sweden.

We shall briefly discuss the construction of reciprocal realizations. For example, it turns out that when using the models of [4] for representing the state process of a reciprocal realization, then the realization must be *external*; i.e., it cannot be constructed out of the given process y only. However, as will be pointed out, the theory on the construction of reciprocal realizations is far from complete.

Let us comment on the choice of a reciprocal state process for a realization on the circle. The reason that the state process x is reciprocal rather than Markovian is simply that the past and future of x with respect to a point in $\mathbb{T}$ coincide. Therefore, the Markov property is not very interesting on $\mathbb{T}$. A reciprocal process has the property that, for any interval, the values of the process in the interval are conditionally independent of those in the exterior, given the values of the process on the boundary of the interval. Hence, the reciprocal property is meaningful on $\mathbb{T}$, and it seems natural that a state process of a stochastic realization on $\mathbb{T}$ should be a reciprocal process.

The paper is organized as follows. In §2 we recall and reformulate in a geometric setting some of the results on reciprocal processes derived by Levy, Frezza, and Krener [4]. Moreover, we shall add some new results on reciprocal processes. In §3 we introduce reciprocal realizations and analyze minimality and observability with geometric tools. In §4 we discuss the construction of reciprocal realizations.

2. Reciprocal processes.

2.1. Preliminaries. In this paper we shall investigate random phenomena defined on a discrete circle $\mathbb{T}$. $\mathbb{T}$ has T elements and is indexed from 0 to $T-1$. All arithmetics on $\mathbb{T}$ is to be interpreted modulo T; e.g., we shall identify -1 with $T-1$. The closed interval $[s,t] \subseteq \mathbb{T}$ will denote the set $\{s, s+1, \dots, t-1, t\}$, and the open interval $(s,t) \subseteq \mathbb{T}$ will denote the set $\{s+1, s+2, \dots, t-1\}$.

Since we are interested in stationary processes on $\mathbb{T}$ which are outputs of systems with reciprocal state processes, we need first to review some of the theory of reciprocal processes on $\mathbb{T}$. We shall follow Levy, Frezza, and Krener [4] but also add some results. Moreover, the approach taken is geometric and in the spirit of Lindquist, Pavon, and Picci [6, 8].

Let H be a Hilbert space. If $M \subseteq H$ is a closed subspace, we write the orthogonal projection onto M as E^M. If η is an n-dimensional column vector with elements $\eta_i \in H$, we shall say that the vector itself belongs to H, i.e., $\eta \in H$. We write $E^M\eta$ for the vector $[E^M\eta_1, \dots, E^M\eta_n]'$ and say that the vectors $x = [\eta_1, \dots, \eta_n]'$ and $y = [\lambda_1, \dots, \lambda_m]'$ of elements in H are orthogonal if $(\eta_i, \lambda_j) = 0$ for all i, j.

If A and B are closed subspaces of H, $E^A B$ denotes the closure of the set $\{E^A b;\, b \in B\}$ and $A \vee B$ denotes the smallest closed vector space containing A and B. The following lemma will prove useful [5, p. 813].

LEMMA 2.1. *Let A and B be closed subspaces of H. Then $A = E^A B \oplus (A \cap B^\perp)$.*

The concept of conditionally orthogonal subspaces is of fundamental importance in stochastic systems theory. We say that the subspaces A and B are conditionally orthogonal given the space X, which we shall write as $A \perp B \,|\, X$, if for all $a \in A$ and $b \in B$ it holds that $(a - E^X a, b - E^X b) = 0$, which is equivalent to $E^{A \vee X} b = E^X b$ for $b \in B$ [5, p. 813]. The following lemma can be found in [5, p. 813].

LEMMA 2.2. *If $A \perp B \,|\, X$, then $A \cap B \subseteq X$.*

2.2. Reciprocal families of subspaces. In this paper all processes are zero-mean vector-valued Gaussian processes. Due to the Gaussian property, conditioning operations are linear projections onto relevant subspaces. Therefore, it is natural to analyze processes by means of the subspaces they generate. This way of handling

processes has the convenient property that any possible linear dependence between the components of $x(t)$ for a given vector process x is factored out of the analysis. To this end, we make the following definitions. Let x be an n-dimensional vector process on $\mathbb{T}$. The space X_t generated by x at t is defined as $X_t := \operatorname{span}\{x_i(t)\,; i = 1, \dots, n\}$. Moreover, let $[t_0, t_1]$ be an interval on $\mathbb{T}$. The spaces $X^i[t_0, t_1]$ and $X^e(t_0, t_1)$ are defined as $X^i[t_0, t_1] := X_{t_0} \vee X_{t_0+1} \vee \cdots \vee X_{t_1}$ and $X^e(t_0, t_1) := X_{t_1} \vee X_{t_1+1} \vee \cdots \vee X_{t_0-1} \vee X_{t_0}$. Finally, define $X_t^i := X^i[t-1, t+1]$ and $X_t^e := X^e(t-1, t+1)$.

Moreover, we notice that a family $\{X_t; t \in \mathbb{T}\}$ of subspaces generated by a stationary process with unitary shift U will have the property that $U^t X_s = X_{t+s}$ for all $t, s \in \mathbb{T}$. A family of subspaces having this shift property with unitary operator U is called a *stationary* family of subspaces.

A process x on $\mathbb{T}$ is *reciprocal* if, given an arbitrary interval $[t_0, t_1] \in \mathbb{T}$, the values of x in the interior and exterior of this interval are conditionally independent given $x(t_0)$ and $x(t_1)$; see [4, p. 1013]. In the Gaussian case it suffices to deal with conditional orthogonality, rather than conditional independence.

The family of subspaces generated by a Gaussian reciprocal process will possess a certain geometric structure.

DEFINITION 2.3. *A family* $\{X_t; t \in \mathbb{T}\}$ *of subspaces is* reciprocal *if for any given interval* $[t_0, t_1]$ *it holds that* $X^i[t_0, t_1]$ *and* $X^e(t_0, t_1)$ *are conditionally orthogonal given* $X_{t_0} \vee X_{t_1}$, *i.e.,*

$$X^i[t_0, t_1] \perp X^e(t_0, t_1) \mid X_{t_0} \vee X_{t_1}. \tag{2.1}$$

REMARK 2.4. *If specialized to the interval* $[t-1, t+1]$, *the condition* (2.1) *reduces to* $X_t^i \perp X_t^e \mid X_{t-1} \vee X_{t+1}$ *and especially for any* $\lambda \in X_t$ *it holds that* $E^{X_t^e}\lambda = E^{X_{t-1} \vee X_{t+1}}\lambda$.

Given a stationary reciprocal family of n-dimensional subspaces $\{X_t\}$ we can construct a vector process $\{x(t)\}$ such that the components of $x(t)$ form a basis for X_t in the following way. Let the set $\{\eta_1, \dots, \eta_n\}$ be a basis for X_0 and set $x(0) := [\eta_1, \dots, \eta_n]'$. Then the process $\{x(t)\}$ defined as $x(t) := [U^t\eta_1, \dots, U^t\eta_n]'$ is a basis process for $\{X_t\}$ in the sense that for every $\eta \in X_t$ there is a unique $a \in \mathbb{R}^n$ such that $\eta = a'x(t)$.

The process $\{x(t)\}$ is clearly reciprocal. In [4] it is shown that a stationary reciprocal process $\{x(t)\}$ satisfies the system

$$x(t) = F_- x(t-1) + F_+ x(t+1) + d(t), \tag{2.2}$$

where d is a certain noise process of which the covariance structure is entirely specified by the matrices F_- and F_+. The matrices F_- and F_+ are determined by the normal equations

$$\begin{bmatrix} \Gamma_1' & \Gamma_1 \end{bmatrix} = \begin{bmatrix} F_- & F_+ \end{bmatrix} \begin{bmatrix} \Gamma_0 & \Gamma_2 \\ \Gamma_2' & \Gamma_0 \end{bmatrix}, \tag{2.3}$$

where Γ_k is the covariance matrix $\Gamma_k := Ex(t)x(t+k)'$.

The system (2.3) always has a solution, but the solution need not be unique. The question of uniqueness is settled by the following lemma.

LEMMA 2.5. *The system* (2.3) *has a unique solution if and only if the vector sum* $X_{t-1} \vee X_{t+1}$ *is direct.*

Proof. Since the second matrix in the right-hand side of (2.3) is the Gram matrix of the set $\{x_1(t-1), \dots, x_n(t-1), x_1(t+1), \dots, x_n(t+1)\}$, the conclusion follows. □

The process d is constructed as $d(t) := E^{(X_t^e)^\perp} x(t)$ and therefore has the property that $d(t)$ is orthogonal to $x(s)$ when $t \neq s$. The process d is called the *two-sided innovation process.* The two-sided innovation process of a stationary reciprocal process is stationary. Moreover, in [4] it is shown that d is locally correlated, i.e., $d(t) \perp d(t+k)$ if $|k| > 1$.

The following lemma gives a necessary and sufficient condition for the two-sided innovation process of a reciprocal process to be identically zero.

LEMMA 2.6. *The two-sided innovation process d of a reciprocal process x is identically zero if and only if $X_t \subseteq (X_{t-1} \vee X_{t+1})$.*

Proof. The result follows from $d(t) = E^{(X_t^e)^\perp} x(t) = x(t) - E^{X_t^e} x(t) = x(t) - E^{X_{t-1} \vee X_{t+1}} x(t)$, where the last equality follows from reciprocity. □

If we identify the process x with a vector constructed by the random vectors $\{x(0), \dots, x(T-1)\}$, i.e., $x' = [x(0)', \dots, x(T-1)]'$, and make the corresponding identification for d, the system (2.2) can be written in matrix form as

$$Fx = d. \tag{2.4}$$

The matrix F is the block-circulant matrix $F = \operatorname{circ}(I, -F_+, 0, \dots, 0, -F_-)$. We say that the model (2.4) is *well posed* if F is invertible.

We shall now introduce a class of reciprocal processes having well-posed models. This is the class of *nonsingular* processes, essentially the class studied in [4]. The family $\{X_t\}$ of subspaces is *nonsingular* if the vector sum $X_0 \vee X_1 \vee \cdots \vee X_{T-1}$ is direct. We say that the vector process $\{x(t)\}$ is nonsingular if it generates a nonsingular family of subspaces. Nonsingularity means that for all $t \in \mathbb{T}$, X_t is linearly independent of X_t^e. Another way of characterizing nonsingularity is that for all $t \in \mathbb{T}$ the components of $d(t)$ form a basis for $(X_t^e)^\perp$, as stated in the following lemma. The proof is simple and omitted.

LEMMA 2.7. *A stationary reciprocal process x is nonsingular if and only if its two-sided innovation process satisfies $E\, d(t)d(t)' > 0$. Moreover, if x is nonsingular then x is uniquely determined by d.*

For reciprocal families nonsingularity is a "local" property, as shown in the following lemma.

LEMMA 2.8. *The reciprocal family is nonsingular if and only if $X_t \cap (X_{t-1} \vee X_{t+1}) = 0$ for all $t \in \mathbb{T}$.*

Proof. If the family is nonsingular it clearly holds that the intersection is trivial. Conversely, suppose that the intersection is trivial. Suppose, to get a contradiction, that the family is not nonsingular. Then there is a t such that $X_t \cap X_t^e \neq 0$; hence let $\lambda \in X_t \cap X_t^e$ and $\lambda \neq 0$. It now follows that $\lambda = E^{X_t^e} \lambda = E^{X_{t-1} \vee X_{t+1}} \lambda \neq \lambda$, which is a contradiction. □

EXAMPLE 2.9. *Let $\{x(t)\}$ be a nonsingular stationary reciprocal process on $\mathbb{T}$ such that $\dim X_t = n$ for all $t \in \mathbb{T}$. Since $\{X_t\}$ is nonsingular, it especially holds that $\dim(X_{t-1} \vee X_{t+1}) = 2n$. Hence, by Lemma 2.5 there is a unique model $x(t) = F_- x(t-1) + F_+ x(t+1) + d(t)$ satisfied by x. Moreover, x is uniquely determined from its two-sided innovation process d* [4].

Form another process $\{\tilde{x}(t)\}$ as $\tilde{x}(t) := [x(t)', x(t+1)']'$. It is easily verified that $\tilde{x}$ is reciprocal. Since $\dim(\tilde{X}_{t-1} \vee \tilde{X}_{t+1}) = 4n$, there is by Lemma 2.5 a unique model with matrices $\tilde{F}_-$ and $\tilde{F}_+$ However, the model is not well posed, since the inclusion

$$\tilde{X}_t = X_t \vee X_{t+1} \subseteq X_{t-1} \vee X_t \vee X_{t+1} \vee X_{t+2} = \tilde{X}_{t-1} \vee \tilde{X}_{t+1}$$

implies that $\tilde{d}(t) = 0$. *Hence, the process* $\tilde{x}$ *is not nonsingular and has a degenerated but unique model* $\tilde{F}\tilde{x} = 0$.

Necessary and sufficient conditions on $\{X_t\}$ for its corresponding model (2.2) to be well posed are not known.

From now on in this paper, we shall focus on *nonsingular* reciprocal processes. The reason for this is that this is a class of reciprocal processes for which we know that the model $Fx = d$ is well posed. Following Levy, Frezza, and Krener [4], we normalize the three-term difference equation (2.2) to obtain the symmetric descriptor model

$$Mx(t) = N'x(t-1) + Nx(t+1) + e(t), \tag{2.5}$$

where $M := D_0^{-1}$, $N := D_0^{-1}F_+$, and $e(t) := D_0^{-1}d(t)$.

Letting Λ be the block circulant $\Lambda := \operatorname{circ}(M, -N, 0, \ldots, -N')$, the normalized model can be written

$$\Lambda x = e. \tag{2.6}$$

It follows that the one-step covariance of the noise e is $Ee(t)e(t+1)' = -N$. Hence, the covariance matrix of e is $Eee' = \Lambda$, and multiplying (2.6) with e' from the right yields

$$Exe' = I. \tag{2.7}$$

From (2.7) it follows that e is the *conjugate process* of x.

In [4] it is shown that the solution x to the system $\Lambda x = e$ is a reciprocal process if Λ is positive definite and e is a stationary locally correlated process with $Ee(t)e(t)' = M$, $Ee(t)e(t+1)' = N$ and $Ee(t)e(t+k)' = 0$ if $|k-t| > 1$. The solution of the well-posed system $\Lambda x = e$ is also a stationary process. Thus we have the following theorem.

THEOREM 2.10. *The solution* x *of the well-posed system* $\Lambda x = e$ *on* $\mathbb{T}$ *is a stationary reciprocal process.*

Proof. The fact that x is a reciprocal process follows from [4]. To prove stationarity, it is enough to show that the covariance matrix of x is a block circulant. Since the model is well posed, we have that $x = \Lambda^{-1}e$ and the covariance of x can be written $Exx' = \Lambda^{-1}Edd'(\Lambda^{-1})' = \Lambda^{-1}$. The inverse of a block circulant is a block circulant [1, p. 181]. Hence, Λ^{-1} is a block circulant. □

We now summarize the results on reciprocal families of subspaces.

THEOREM 2.11. *Let* $\{X_t\}$ *be a family of finite-dimensional subspaces, with basis process* $\{x(t)\}$, *such that* $X_t \cap (X_{t-1} \vee X_{t+1}) = 0$. *Then* $\{X_t\}$ *is a stationary reciprocal family if and only the basis process* $\{x(t)\}$ *satisfies the system*

$$Mx(t) = N'x(t-1) + Nx(t+1) + e(t),$$

where e *is a stationary locally correlated process satisfying the necessary condition* $Eee' = \Lambda > 0$, *and* $\Lambda = \operatorname{circ}(M, -N, 0, \ldots, -N')$.

For a nonsingular stationary family of subspaces the reciprocal property has the following "local" characterization.

COROLLARY 2.12. *A stationary family of nonsingular subspaces* $\{X_t\}$ *is reciprocal if and only if* $X_0^i \perp X_0^e \mid X_{-1} \vee X_1$.

2.3. An abstract description of F_- and F_+. In this subsection we shall present abstract counterparts of the matrices F_- and F_+.

Let $\{X_t\}$ be a given nonsingular stationary reciprocal family and $\{x(t)\}$ be its basis process. For every $\lambda \in X_{-1}$ there is an $a \in \mathbb{R}^n$ such that $\lambda = a'x(-1)$.

We shall now show that F_- is the matrix representation of an operator acting on X_{-1} in this basis. By shifting, $U\lambda = a'x(0)$. Employing (2.2), we get $U\lambda = a'F_-x(-1) + a'F_+x(1) + a'd(0)$. The next step is to project orthogonally onto the subspace $X_{-1} \vee X_1$, which yields

$$E^{X_{-1}\vee X_1}U\lambda = a'F_-x(-1) + F_+a'x(1), \tag{2.8}$$

since $d(0) \perp X_0^e$. Finally, let Π_t be the projection onto X_t *along* X_t^e. By applying Π_{-1} to (2.8) we get $\Pi_{-1}E^{X_{-1}\vee X_1}U\lambda = a'F_-x(-1)$.

Motivated by the previous discussion we define the operator $A_- : X_{-1} \to X_{-1}$ as $A_-\lambda := \Pi_{-1}E^{X_{-1}\vee X_1}U\lambda$ for $\lambda \in X_{-1}$.

In particular, if we let $\lambda = x_k(-1)$, which corresponds to a being the kth unit vector in $\mathbb{R}^n$, we get that $A_-x_k(-1) = \sum_{j=1}^n (F_-)_{k,j}x_j(-1)$. Hence, $F_{-1} : X_- \to X_{-1}$ is the matrix representation of the operator A_- in the basis corresponding to $x(-1)$.

Analogously, we define $A_+ : X_1 \to X_1$ as $A_+\lambda := \Pi_1 E^{X_{-1}\vee X_1}U\lambda$ for $\lambda \in X_1$, and $F_+ : X_1 \to X_1$ is the matrix representation of A_+.

3. Reciprocal realizations. We shall now introduce the main objects of this paper, namely, reciprocal realizations on $\mathbb{T}$.

3.1. Geometric description of a realization. Given a nonsingular stationary reciprocal process x on $\mathbb{T}$ we can define a process y as $y(t) := Cx(t)$. This definition makes y a stationary, but not reciprocal, process on $\mathbb{T}$. If C is of full rank, then y will be nonsingular. However, the stationary process y has a very detailed structure as the output of a reciprocal process, and it makes sense to say that we have a reciprocal realization of y. The purpose of this chapter is to derive and discuss properties of reciprocal realizations. In the next chapter we shall discuss the construction of a realization, starting from the given process y.

DEFINITION 3.1. *The well-posed system*

$$\begin{cases} Mx(t) &= N'x(t-1) + Nx(t+1) + e(t), \\ y(t) &= Cx(t), \end{cases} \tag{3.1}$$

where M, N, and e satisfy the conditions of the preceding section, is a reciprocal realization *of y.*

We say that the process x is the *state process* of the realization and that X_t is the state space. As we will see, though, the relevant splitting subspaces are of the form $X_s \vee X_t$. This differs from Markovian realization theory, where the state space is the splitting subspace.

We shall now point out that the requirement that the state process x be nonsingular has advantages, as well as shortcomings. As advantages we regard the fact that nonsingular reciprocal processes are well studied and enjoy well-posed models of simple structure and that the condition for a stationary family of subspaces $\{X_t\}$ to be reciprocal reduces to the simple condition

$$X_0^i \perp X_0^e \,|\, X_{-1} \vee X_1.$$

As a shortcoming, we regard the result shown in the next section, that a reciprocal realization is necessarily external, which follows from the required nonsingularity of

x. This insight, gained by formulating the realization problem on $\mathbb{T}$ has motivated us to further investigate singular reciprocal processes, an investigation currently being done.

In order to analyze reciprocal realizations with geometric tools, we shall introduce some relevant subspaces. The space $H(y)$ is defined as the space spanned by y, i.e., $H(y) := \operatorname{span}\{y_i(t); \quad t \in \mathbb{T}, i = 1, \dots, m\}$.

The essence of a realization can now be expressed in a geometric and coordinate-free form. The state process x of (3.1) generates a stationary reciprocal family of subspaces $\{X_t\}$, and since the family is nonsingular, reciprocity is equivalent to

$$X_0^e \perp X_0^i \mid X_{-1} \vee X_1. \tag{3.2}$$

On the other hand, given a nonsingular family of subspaces satisfying (3.2) we can introduce a basis and get a three-term difference equation as

$$Mx(t) = N'x(t-1) + Nx(t+1) + e(t).$$

Furthermore, by stationarity, it is easily seen that $y(t) = Cx(t)$ for some matrix C is equivalent to $y(0) \in X_0$. Hence, an equivalent description of (3.1) is that we have a nonsingular stationary family of subspaces $\{X_t\}$ such that

1. $X_0^e \perp X_0^i \mid X_{-1} \vee X_1$,
2. $y(0) \in X_0$.

We shall now give an alternative equivalent geometric description of a reciprocal realization. This description employs the concept of splitting subspaces and opens the door to the geometric stochastic realization theory as developed in, e.g., [5]–[7].

Recall that, if H_1, H_2, and X are subspaces of some Hilbert space H, then X is a *splitting subspace* with respect to H_1 and H_2 if $H_1 \perp H_2 \mid X$. Moreover, a splitting subspace X is *minimal* if there is no proper subspace of X that is a splitting subspace with respect to H_1 and H_2.

We introduce the following subspaces of $H(y)$. Let $[s,t] \subseteq \mathbb{T}$. The subspaces $H^i[s,t]$ and $H^e(s,t)$ of $H(y)$ are defined as $H^i[s,t] := \operatorname{span}\{y_i(k); k \in [s,t], i = 1, \dots, m\}$ and $H^e(s,t) := \operatorname{span}\{y_i(k); k \in (s,t)^c, i = 1, \dots, m\}$. Note that the relation $H^i[s,t] = H^e(t,s)$ holds by symmetry properties of $\mathbb{T}$.

Note that we can write $H(y)$ as $H(y) = H^i[s,t] \vee H^e(s,t)$. For a given realization we see that the condition $y(t) \in X_t$ implies that $H^i[s,t] \subseteq X^i[s,t]$ and $H^e(s,t) \subseteq X^e(s,t)$. From this we get that

$$H^i[s,t] \perp H^e(s,t) \mid X_s \vee X_t. \tag{3.3}$$

Hence, the space $X_s \vee X_t$ is a splitting subspace with respect to $H^i[s,t]$ and $H^e(s,t)$. In contrast to Markovian realization theory on the real line or the integers, we notice that here the splitting subspace is not a space generated by the state process at a single point but instead the splitting subspace is the vector sum of the spaces generated by the state process at two different points in the index set $\mathbb{T}$.

We shall now show that the splitting relation (3.3) completely captures our idea of a realization.

THEOREM 3.2. *For a nonsingular reciprocal stationary family of subspaces $\{X_t\}$ the following are equivalent:*

1. $H_0^i \perp H_0^e \mid X_{-1} \vee X_1$,
2. $y(0) \in X_0$.

Proof. We have already shown that (2) $\Rightarrow$ (1). Suppose that (1) holds. Then by Lemma 2.2 it follows that $H_0^i \cap H_0^e \subseteq X_{-1} \vee X_1$, i.e.,

$$y(-1), y(1) \in X_{-1} \vee X_1. \tag{3.4}$$

By shifting (3.4) with U^2, we get $y(1), y(3) \in X_1 \vee X_3$. Hence, $y(1) \in (X_{-1} \vee X_1) \cap (X_1 \vee X_3)$. But since the family of subspaces is nonsingular, every sum of the type $X_s \vee X_t$ is *direct.* This implies that $(X_{-1} \vee X_1) \cap (X_1 \vee X_3) = X_1$, and we conclude that $y(1) \in X_1$, which, shifted, yields that $y(0) \in X_0$. □

In view of Theorem 3.2 we make the following equivalent definition of a reciprocal realization on $\mathbb{T}$.

DEFINITION 3.3. *Let* $\{y(t);\, t \in \mathbb{T}\}$ *be a stationary process. A nonsingular stationary family of finite-dimensional subspaces* $\{X_t\}$ *is a* reciprocal realization *of* $\{y(t);\, t \in \mathbb{T}\}$ *if*

1. $X_0^e \perp X_0^i \,|\, X_{-1} \vee X_1$,
2. $H_0^i \perp H_0^e \,|\, X_{-1} \vee X_1$.

3.2. Minimality and observability. The general goal of state-space modeling of some phenomena is to achieve data reduction for the behavior of the phenomena. In order to achieve *maximal* data reduction, the state space should be *minimal.*

DEFINITION 3.4. *A reciprocal realization of* $\{y(t);\, t \in \mathbb{T}\}$ *is* minimal *if the corresponding reciprocal family of subspaces* $\{X_k\}$ *has the property that there is no proper subspace* $\hat{X}_0 \subset X_0$ *such that* $\{\hat{X}_t\}$ *is a reciprocal realization of* $\{y(t);\, t \in \mathbb{T}\}$.

Due to stationarity, the inclusion $\hat{X}_0 \subset X_0$ holds if and only if $\hat{X}_t \subset X_t$ for all $t \in \mathbb{T}$.

Note that from this definition of minimality it does *not* follow that all minimal realizations of a given process y have the same dimension.

As previously seen, in realization theory on $\mathbb{T}$ the relevant splitting subspaces are of the form $X_s \vee X_t$, whereas the definition of minimality is given in terms of the state space X_t. Clearly, it is natural to try to relate these two concepts of minimality to each other. In the following theorem we give a sufficient condition for a reciprocal realization to be minimal.

THEOREM 3.5. *Let* $\{X_t\}$ *be a reciprocal realization. If there is a* τ *such that the splitting subspace* $X_0 \vee X_\tau$ *is a minimal splitting subspace, then the realization is minimal.*

Proof. Suppose there is a reciprocal realization $\{\hat{X}_t\}$ such that $\hat{X}_0 \subseteq X_0$. We have to show that $\hat{X}_0 = X_0$. Since $X_0 \vee X_\tau$ is a minimal splitting subspace, it follows that $\hat{X}_0 \vee \hat{X}_\tau = X_0 \vee X_\tau$. By directness of the vector sums it follows that $\hat{X}_\tau = X_\tau$ and stationarity implies that $\hat{X}_0 = X_0$. □

In order to analyze minimality further we shall introduce the concepts of interior and exterior observability. Fix any interval $[0, t]$ and suppose that y is given on $[0, t]$. Now let $\{X_t\}$ be a realization. Clearly, an element in the subspace $(X_0 \vee X_t) \cap H^i[0, t]^\perp$ cannot be distinguished from zero by observing y on $[0, t]$ and is therefore called *unobservable.* Note that this definition is a special case of a general definition of observability given in [7].

DEFINITION 3.6. *Let* $[0, t]$ *be an interval. The realization* $\{X_t\}$ *is* interiorly observable on $[0, t]$ *if* $(X_0 \vee X_t) \cap H^i[0, t]^\perp = 0$.

REMARK 3.7. *Due to stationarity, we can exploit the translation invariance and restrict the analysis to intervals of the type* $[0, t]$, *because on* $\mathbb{T}$ *any interval of the type* $[s, t]$ *can be shifted onto an interval of the type* $[0, t]$.

By Lemma 2.1 the space $X_0 \vee X_t$ can be decomposed as

$$X_0 \vee X_t = E^{X_0 \vee X_t} H^i[0,t] \oplus (X_0 \vee X_t) \cap H^i[0,t]^\perp. \tag{3.5}$$

From this decomposition we read that interior observability is equivalent to the equality

$$X_0 \vee X_t = E^{X_0 \vee X_t} H^i[0,t]. \tag{3.6}$$

DEFINITION 3.8. *Let $[0,t]$ be an interval. The realization $\{X_t\}$ is* exteriorly observable on $[0,t]$ *if* $(X_0 \vee X_t) \cap H^e(0,t)^\perp = 0$.

REMARK 3.9. *The situation here is analogous to that of Markovian realization theory, where it is necessary to introduce the two concepts observability and constructibility.*

The next theorem states that a splitting subspace $X_0 \vee X_t$ is minimal if and only if the realization is interiorly and exteriorly observable on the interval $[0,t]$. The theorem is a modification of Proposition 1 in [7, p. 273].

THEOREM 3.10. *Let $\{X_t\}$ be a reciprocal realization. The splitting subspace $X_0 \vee X_t$ is minimal if and only if the realization is interiorly and exteriorly observable on $[0,t]$.*

The next theorem states that for a given interval $[s,t]$, all minimal splitting subspaces $X_s \vee X_t$ with respect to $H^i[s,t]$ and $H^e(s,t)$ have the same dimension. The theorem follows directly from a general theorem by Lindquist and Picci [5, p. 822].

THEOREM 3.11. *Let $[s,t] \subseteq \mathbb{T}$. All minimal splitting subspaces of the form $X_0 \vee X_t$, with respect to $H^i[0,t]$ and $H^e(0,t)$, have the same dimension.*

3.3. An observability matrix. We shall now give an algebraic criterion for a realization defined by a triplet (M,N,C) and a noise process e to be interiorly observable on $[0,t]$. Recall that the block-circulant Λ is defined as $\Lambda = \operatorname{circ}(M, -N, 0, \dots, 0, -N')$. Since $x = \Lambda^{-1}e$ and $Eee' = \Lambda$, it follows that $Exx' = \Lambda^{-1}$. Let $\Gamma := \Lambda^{-1}$. In the case that T is even we have that

$$\Gamma = \operatorname{circ}(\Gamma_0, \Gamma_1, \dots, \Gamma_{\frac{T}{2}-1}, \Gamma_{\frac{T}{2}}, \Gamma'_{\frac{T}{2}-1}, \dots, \Gamma'_1).$$

The case of T odd is analogous.

Consider the relation (3.6); since the space on the right-hand side is included in the space on the left-hand side, the realization is interiorly observable on $[0,t]$ if and only if the inclusion $X_0 \vee X_t \subseteq E^{X_0 \vee X_t} H^i[0,t]$ holds. Equivalently, we require that

$$x_i(0), x_i(t) \in \operatorname{span}\{\hat{y}_j(s);\ s \in [0,t], j = 1, \dots, m\} \text{ for } i = 1, \dots, n, \tag{3.7}$$

where $\hat{y}_j(s) := E^{X_0 \vee X_t} y_j(s)$.

The condition (3.7) can be expressed as a rank condition on a certain observability matrix. To this end, let $s \in [0,t]$; since $y(s) = Cx(s)$, we get that $\hat{y}(s) = C\,E^{X_0 \vee X_t} x(s)$. The projection $E^{X_0 \vee X_t} x(s)$ can, with suitable matrices H_0 and H_t, be written as $E^{X_0 \vee X_t} x(s) = H_0 x(0) + H_t x(t)$. The normal equations for determining H_0 and H_t are $E\,[x(s) - H_0 x(0) - H_t x(t)]x(0)' = 0$ and $E\,[x(s) - H_0 x(0) - H_t x(t)]x(t)' = 0$. Taken together we get the following system of equations:

$$[E\,x(s)x(0)' \quad E\,x(s)x(t)'] = [H_0 \quad H_t] \begin{bmatrix} E\,x(0)x(0)' & E\,x(0)x(t)' \\ E\,x(t)x(0)' & E\,x(t)x(t)' \end{bmatrix}.$$

Recall that $Ex(s)x(s+k)' = \Gamma_k$, which yields $E\,x(s)x(0)' = \Gamma_s'$, $E\,x(s)x(t)' = \Gamma_{t-s}$ and $E\,x(0)x(t)' = \Gamma_t$. Since the process x is nonsingular, the normal equations have a unique solution for all $t \in \mathbb{T}$, and $\hat{y}(s)$ can be written as

$$\hat{y}(s) = C\begin{bmatrix}\Gamma_s' & \Gamma_{t-s}\end{bmatrix}\begin{bmatrix}\Gamma_0 & \Gamma_t \\ \Gamma_t' & \Gamma_0\end{bmatrix}^{-1}\begin{bmatrix}x(0) \\ x(t)\end{bmatrix}.$$

Stacking all $\hat{y}(s)$'s yields

$$\begin{bmatrix}\hat{y}(0) \\ \hat{y}(1) \\ \hat{y}(2) \\ \cdot \\ \cdot \\ \hat{y}(t)\end{bmatrix} = \begin{bmatrix}C\Gamma_0 & C\Gamma_t \\ C\Gamma_1' & C\Gamma_{t-1} \\ C\Gamma_2' & C\Gamma_{t-2} \\ \cdot & \cdot \\ \cdot & \cdot \\ C\Gamma_t' & C\Gamma_0\end{bmatrix}\begin{bmatrix}\Gamma_0 & \Gamma_t \\ \Gamma_t' & \Gamma_0\end{bmatrix}^{-1}\begin{bmatrix}x(0) \\ x(t)\end{bmatrix}.$$

By this calculation we are lead to define an observability matrix, which as shown can be calculated from the data (M, N, C), for which a rank criterion can be stated.

DEFINITION 3.12. *The observability matrix $\mathcal{O}(t)$ is defined as*

$$\mathcal{O}(t) := \begin{bmatrix}C\Gamma_0 & C\Gamma_t \\ C\Gamma_1' & C\Gamma_{t-1} \\ C\Gamma_2' & C\Gamma_{t-2} \\ \cdot & \cdot \\ \cdot & \cdot \\ C\Gamma_t' & C\Gamma_0\end{bmatrix}.$$

The condition (3.7) can now be expressed as $\ker \mathcal{O}(t) = 0$, and we get the following theorem on interior observability.

THEOREM 3.13. *The realization (M, N, C) is interiorly observable on $[0, t]$ if and only if* $\ker \mathcal{O}(t) = 0$, *i.e., if* $\operatorname{rank} \mathcal{O}(t) = 2n$.

We conclude this section with a numerical example.

EXAMPLE 3.14. *Let $T = 10$ be the length of $\mathbb{T}$ and let the matrices*

$$M = \begin{bmatrix}4 & 0 \\ 0 & 5\end{bmatrix}, N = \begin{bmatrix}-1 & 2 \\ -1 & 2\end{bmatrix}, \text{ and } C = \begin{bmatrix}1 & 0\end{bmatrix}$$

define a reciprocal realization of a process $\{y(t)\}$ as in Definition 3.1.

The covariance matrix $R = E\,yy'$ can be computed as

$$R = circ(0.41, -0.08, -0.02, -0.01, -0.002, -0.001, -0.002, -0.01, -0.02, -0.08),$$

and we conclude that $\{y(t)\}$ is stationary. It is easy to verify that R^{-1} is not a tridiagonal circulant, and consequently, $\{y(t)\}$ is not a reciprocal process. Hence, the dimension of a minimal realization is obviously 2.

The matrix $\mathcal{O}(2)$ is a 3×4 dimensional matrix and its rank cannot be 4, *as required for interior observability on $[0, 2]$. Hence, the splitting subspace $X_0 \vee X_2$ is not a minimal splitting subspace.*

The observability matrix $\mathcal{O}(3)$ is

$$\mathcal{O}(3) = \begin{bmatrix}0.4136 & 0.0322 & -0.0067 & 0.0182 \\ -0.0751 & -0.0894 & -0.0223 & 0.0615 \\ -0.0223 & -0.0265 & -0.0751 & 0.2074 \\ -0.0067 & -0.0077 & 0.4136 & 0.0322\end{bmatrix}$$

and has rank 4. *Consequently, the realization is interiorly observable on* $[0,3]$. *If the realization is also* exteriorly *observable on* $[0,3]$, *we can conclude by Theorem* 3.10 *that* $X_0 \vee X_3$ *is a minimal splitting subspace and by Theorem* 3.5 *that the realization is minimal. Since exterior observability on* $[0,3]$ *is equivalent to interior observability on* $[0,7]$, *we need only check that* rank $\mathcal{O}(7) = 2$, *which happens to be the case, as computations will show.*

4. Construction of reciprocal realizations. The realization problem amounts to the construction of a stochastic realization of a given process. In this paper the given process—y, say—is a stationary process defined on $\mathbb{T}$, and the objective is to produce a reciprocal realization of y, i.e., to find a triplet (M,N,C) and a locally correlated driving noise process e such that

$$\begin{cases} Mx(t) &= N'x(t-1) + Nx(t+1) + e(t), \\ y(t) &= Cx(t). \end{cases} \tag{4.1}$$

The reciprocal realization problem is equivalent to the following geometric problem. Given the stationary process y, find a minimal stationary nonsingular reciprocal family of subspaces $\{X_t\}$ such that $H_0^i \perp H_0^e \,|X_{-1} \vee X_1$.

In general, stochastic realization is a two-step procedure. First, we try to find a model for the process to be realized. In our setting this amounts to finding the matrices (M,N,C). Unfortunately, this problem, which we shall discuss in the next subsection, is still unsolved. Second, we must construct the state process x of the realization. In the following we show that the realization must be *external* and how to construct an external realization.

4.1. The covariance factorization problem. A realization (4.1) can be regarded as a mapping from the input noise e to the output y, which we can write as

$$y = We. \tag{4.2}$$

Since $y(t) = Cx(t)$, we get that $y = (I_T \otimes C)x$, where $\otimes$ denotes the Kronecker product. Furthermore, x satisfies $\Lambda x = e$, or equivalently, $x = \Lambda^{-1}e$. Recall that $Exe' = I$ and $Eee' = \Lambda$. Hence, $y = (I_T \otimes C)\Lambda^{-1}e$ and we define W as

$$W := (I_T \otimes C)\Lambda^{-1}.$$

Since the process y is given, its second-order properties are at hand and we let R_k be defined as $R_k := Ey(t)y(t+k)'$. Moreover, let R denote the covariance matrix

$$R := Eyy';$$

in the case of T even it follows that

$$R = \operatorname{circ}(R_0, R_1, \dots, R_{\frac{T}{2}-1}, R_{\frac{T}{2}}, R'_{\frac{T}{2}-1}, \dots, R'_1).$$

The case of T odd is analogous.

By straightforward calculations we get

$$R = Eyy' = WEee'W' = W\Lambda W'.$$

Inserting the expression for W, we get

$$R = (I_T \otimes C)\Lambda^{-1}(I_T \otimes C').$$

The procedure of going from the given matrix R to the triplet (M, N, C) is clearly a matrix-factorization problem. Moreover, we see that (M, N, C) are determined by the covariance data of y. Note that the smallest possible dimension of M is the minimal dimension of the realization. Necessary and sufficient conditions for a block-circulant R to admit such a factorization are not known.

Although covariance factorization is an inevitable step in constructing a realization we shall not investigate the factorization problem further in this paper but merely assume that the given process y is such that its corresponding R is factorizable and that the triplet (M, N, C) is at hand. Let us just remark that the problem, as it stands, does not seem elementary.

4.2. Nonexistence of internal realizations. If the state process x can be constructed using only the given process y we say that the realization is *internal*, which can be expressed as $X_t \subseteq H(y)$ for all $t \in \mathbb{T}$. Otherwise, the introduction of external random quantities is necessary in order to construct a realization and the state process will live in a larger space than $H(y)$, and we say that the realization is *external*. This larger space containing $H(y) \vee X_0 \vee X_1 \vee \cdots \vee X_{T-1}$ is called the *ambient space* of the realization.

For the case of nonsingular reciprocal realizations it turns out that internal realizations can never occur, except for the trivial case where y is already a reciprocal process itself. This is a consequence of the requirement that the state process be nonsingular and we give the following theorem.

THEOREM 4.1. *Suppose that is y is not reciprocal. Then there are no internal reciprocal realizations of y.*

Proof. To have an internal realization of y we must have a nonsingular reciprocal family of subspaces $\{X_t\}$ such that $X_t \subseteq H(y)$ and consequently, $X_0 \vee X_1 \vee \cdots \vee X_{T-1} \subseteq H(y)$. Moreover, the vector sum $X_0 \vee X_1 \vee \cdots \vee X_{T-1}$ must be direct.

Now if the given process y is an m-dimensional process, then $\dim H(y) \leq mT$. Since y is not reciprocal itself, for a realization it must hold that $\dim X_t > m$ and by the directness of the sum it necessarily follows that $\dim(X_0 \vee X_1 \vee \cdots \vee X_{T-1}) > mT$.

Hence, the existence of a nonsingular internal reciprocal realization would imply the contradiction

$$mT \geq \dim H(y) \geq \dim(X_0 \vee X_1 \vee \cdots \vee X_{T-1}) > mT. \qquad \square$$

4.3. External realizations. In this subsection we shall show how to construct *external* realizations by introducing random quantities that are orthogonal to $H(y)$, but that will help to span the state process in a realization.

Suppose that we are given the process y, such that its covariance matrix R can be factorized to give a minimal triplet (M, N, C). The problem is now to construct a driving noise e, which, fed into the system, gives exactly the process y. Since internal realizations are impossible, we cannot construct e out of y only but have to introduce some external random quantities. The explicit construction will now be given.

To make the calculations more transparent we will work with a white driving noise sequence u instead of e. Hence, let K be a square matrix solution to the equation $K'K = \Lambda^{-1}$. Observe that K is invertible. Define u as $u := Ke$; then it is easily seen that $Euu' = I$. The relation $y = We$ is now transformed to $y = WK^{-1}u$. By letting $S := WK^{-1}$ we have the relation

$$y = Su. \tag{4.3}$$

If we let $S^\sharp$ denote the pseudoinverse of S [9, Chap. 6.11] we can define an $nT \times nT$ matrix Π as

$$\Pi := I - S^\sharp S. \tag{4.4}$$

LEMMA 4.2. *Π is a projection matrix.*

Proof. We show that Π is symmetric and idempotent. $S^\sharp S$ is symmetric, which implies that Π is symmetric. Furthermore, $\Pi^2 = (I - S^\sharp S)(I - S^\sharp S) = I - 2S^\sharp S + S^\sharp S S^\sharp S = I - S^\sharp S = \Pi$, since $S^\sharp S S^\sharp = S^\sharp$. □

We notice that since Π is a projection matrix it is positive semidefinite and thus admissible as a covariance matrix. We can now characterize all possible white-noise solutions u to the equation $y = Su$, with y given. Suppose that u is a solution. Then we have $S^\sharp y = S^\sharp S u = (I - \Pi)u$, which gives that $u = S^\sharp y + \Pi u$. If we define z as $z := \Pi u$, we can write u as $u = S^\sharp y + z$. The covariance of z is Π and z is orthogonal to y, since $E\, yz' = SEuu'\Pi' = S\Pi = S - SS^\sharp S = 0$. This decomposition suggests how to construct the driving noise of an external realization.

THEOREM 4.3. *Let y be given as in* (4.3) *and let z be a process on $\mathbb{T}$ such that $H(z) \perp H(y)$ with covariance $Ezz' = \Pi$. Then the noise $\tilde{u}$ defined as*

$$\tilde{u} := S^\sharp y + \Pi z \tag{4.5}$$

will be white and satisfy (4.3) *and consequently, $\tilde{e}$ defined as $\tilde{e} := K^{-1}\tilde{u}$ will satisfy* (4.2).

Proof. Since y is in the range of S and $S\Pi = 0$, we have that $S\tilde{u} = SS^\sharp y + S\Pi z = y$. Furthermore, $E\tilde{u}\tilde{u}' = S^\sharp E yy' (S^\sharp)' + \Pi = S^\sharp S (S^\sharp S)' + \Pi = S^\sharp S S^\sharp S + \Pi = S^\sharp S + \Pi = I$. Hence, $\tilde{u}$ is white. □

Thus given any process z with the properties of Theorem 4.3, we may choose the basic ambient Hilbert space to be $H = H(y) \oplus H(z)$.

5. Conclusions. Results on reciprocal processes, originally obtained by Levy, Frezza, and Krener [4], are reformulated in a geometric framework. Some new results on reciprocal processes are added.

A nonsingular reciprocal realization is defined as a system of the following type:

$$\begin{cases} Mx(t) &= Nx(t-1) + N'x(t+1) + e(t), \\ y(t) &= Cx(t), \end{cases}$$

where the covariance structure of the driving noise $\{e(t)\}$ is entirely specified by the matrices M and N.

The nonsingularity of the state process, expressed as the vector sum $X_0 \vee X_1 \vee \cdots \vee X_{T-1}$ being direct, enables the characterization of a reciprocal realization in a coordinate-free form. With this geometric characterization at hand we can apply results from stochastic realization theory, as developed by Lindquist, Picci, and others, to define and analyze minimality and observability of a reciprocal realization.

As it turns out, the requirement that the state process be nonsingular implies that a reciprocal realization is necessarily external.

However, under the assumption that the model (M, N, C) is known, we show how to construct the driving noise of an external reciprocal realization.

The nonexistence of internal realizations, caused by the requirement that the state process be nonsingular, has initiated research on models for representing reciprocal processes that are not nonsingular. For example, consider the model

$$\begin{cases} x(t+1) &= Ax(t) + w(t), \\ x(T) &= x(0), \end{cases} \tag{5.1}$$

where w is a white noise defined on $\mathbb{T}$. It is straightforward to show that this model admits a well-posed solution if and only if A avoids certain eigenvalues on the unit circle. Moreover, if $Q := E\, w(t)w(t)' > 0$, it follows, using techniques from [4, pp. 1018–1019], that x is a nonsingular reciprocal process. Moreover, the two models (A, Q) and (M, N) are related by a certain algebraic Riccati equation. This algebraic Riccati equation is equivalent to a certain factorization of Λ in which the factors are circulants themselves. However, even if Q is only positive semidefinite, the solution of (5.1) is still a reciprocal process; this was shown in the continuous-time case by Krener [2]. Hence, this may be a way to bypass the problem of nonsingularity in reciprocal realization theory on $\mathbb{T}$.

It should be pointed out, though, that some results in the geometric analysis seem to depend on the directness of the sum $X_0 \vee X_1 \vee X_1 \vee \cdots \vee X_{T-1}$. Hence the geometric theory also may have to be extended to deal with the nonsingular case.

Acknowledgments. I would like to thank Professor Anders Lindquist, Royal Institute of Technology, for reading a preliminary version of this paper and giving many valuable suggestions for its improvement; Dr. Anders Rantzer, now at Lund Institute of Technology: during the time I worked on this paper, he was an invaluable discussion partner and provided many useful pieces of advice; Dr. Ruggero Frezza, University of Padova, for inviting me to Italy in February 1991 and for numerous discussions on reciprocal processes; and two anonymous referees for many helpful suggestions.

REFERENCES

[1] P. J. Davis, *Circulant Matrices*, Wiley, New York, 1979.

[2] A. J. Krener, *Reciprocal processes and the stochastic realization problem for acausal systems*, in Modeling, Identification and Control, C.I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 197–211.

[3] B. C. Levy, *Noncausal estimation for discrete Gauss-Markov random fields*, in Proceedings of the International Symposium MTNS-89, M.A. Kaashoek, J.H. van Schuppen, and A.C.M. Ran, eds., Birkhäuser Boston, Cambridge, 1989, pp. 13–21.

[4] B. C. Levy, R. Frezza, and A. J. Krener, *Modeling and estimation of discrete time Gaussian reciprocal processes*, IEEE Trans. Automat. Control, 35 (1990), pp. 1013–1023.

[5] A. Lindquist and G. Picci, *Realization theory for multivariate stationary Gaussian processes*, SIAM J. Control Optim., 23 (1985), pp. 809–857.

[6] ———, *A geometric approach to modeling and estimation of linear stochastic systems*, J. Math. Systems Estim. Control, 1 (1991), pp. 241–333.

[7] A. Lindquist, G. Picci, and G. Ruckebusch, *On minimal splitting subspaces and Markovian representations*, Math. Systems Theory, 12 (1979), pp. 271–279.

[8] A. Lindquist and M. Pavon, *On the structure of state-space models for discrete-time stochastic vector processes*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 418–432.

[9] D. G. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 521–541, March 1996

007

ON SPECTRUM AND RIESZ BASIS ASSIGNMENT OF INFINITE-DIMENSIONAL LINEAR SYSTEMS BY BOUNDED LINEAR FEEDBACKS*

CHENG-ZHONG XU† AND GAUTHIER SALLET†

Abstract. This paper deals with spectrum and Riesz basis assignability of infinite-dimensional linear systems via bounded linear feedbacks. The necessary and sufficient condition of Sun [*SIAM J. Control Optim.*, 19 (1981), pp. 730–743] is generalized to a large class of boundary control systems. Two typical examples are presented to illustrate the application of our results. The results obtained in this paper may have potential applications in nondissipative spectral systems.

Key words. distributed parameter systems, bounded linear feedback control perturbation, spectral determination, stability, flexible structures

AMS subject classifications. 93C20, 93D15, 93B60, 93B55

1. Introduction.

In this paper, we consider directly the following linear evolution systems on a separable Hilbert space H (the inner product and induced norm in H are denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, respectively):

$$(\Sigma_O)\ :\ \dot{X}(t) = AX(t) + bu(t),$$

where A is the infinitesimal generator of a C_0-semigroup on H and the input element b is not necessarily admissible in the sense of [7]. Throughout the paper, A and b are assumed to satisfy the conditions H1, H2, and H3.

Hypothesis H1. The operator A has compact resolvent. We suppose that the spectrum $\sigma(A) = \{\lambda_n, n \in \mathbb{N}\}$ is simple.

Hypothesis H2. The domain $\mathcal{D}(A^*)$ of the adjoint operator A^* is a Hilbert space with the graph norm. $\mathcal{D}'(A^*)$ is the topological dual of $\mathcal{D}(A^*)$. We suppose that b belongs to $\mathcal{D}'(A^*)$.

Hypothesis H3. The eigenvectors $\{\phi_k; k \in \mathbb{N}\}$ of A form a Riesz basis in H. The biorthogonal sequence corresponding to the eigenvectors of A^* is denoted by $\{\psi_k; k \in \mathbb{N}\}$ and is also a Riesz basis of H [5, p. 310]. We set $b_k = (\psi_k, b)$, where $(\cdot,\cdot)$ is the classical duality product on $\mathcal{D}(A^*) \times \mathcal{D}'(A^*)$. Here, it is defined as the continuous extension of the inner product on H (H is dense in $\mathcal{D}'(A^*)$). We suppose that $b_k \neq 0$ for all $k \in \mathbb{N}$. Let d_n be the distance of $\{\lambda_n\}$ to the rest of the spectrum $\sigma(A)$. We consider the set of disks $D_n, n \in \mathbb{N}$ centered at $\{\lambda_n\}$ with radius $\frac{1}{3}d_n$. Now suppose that there exists a positive constant M such that for all $\lambda \notin \bigcup_{j\in\mathbb{N}} D_j$ and all $m \in \mathbb{N}$,

$$\sum_{n=1}^{+\infty} \left| \frac{b_n}{\lambda - \lambda_n} \right|^2 \leq M < +\infty \tag{1}$$

and

$$\sum_{n=1, n\neq m}^{+\infty} \left| \frac{b_n}{\lambda_m - \lambda_n} \right|^2 \leq M < +\infty. \tag{2}$$

* Received by the editors June 23, 1993; accepted for publication (in revised form) November 15, 1994.

† Institut National de Recherche en Informatique et en Automatique–Lorraine (Projet CONGE), Centre National de la Recherche Scientifique Unité de Recherche Associée 399, CESCOM, 4 rue Marconi, 57070 Metz, France (xu@ilm.loria.fr, sallet@ilm.loria.fr).

The assumptions H2 and H3 allow us to take into account the cases of boundary control because many linear distributed parameter systems can be formulated in the form of (Σ_O) (see [7] and also [19], [15], [16]). In particular, the cases of [21], the cantilever beam equation with lateral force control [15], [24], or moment force control [15], heat conduction equations [7], and the wave equation [18] enter the class of the systems considered here. For each fixed $\lambda \in \rho(A^*)$, the resolvent set of the adjoint operator A^*, and all $r \in H$, it follows from the hypothesis H3 that

$$\left|\left((\lambda - A^*)^{-1} r, b\right)\right| \leq K_\lambda \|r\|_H.$$

The condition (1) forces the constant K_λ to be bounded uniformly with respect to $\lambda \notin \bigcup_{j\in\mathbb{N}} D_j$ (see §3). In certain cases, the condition (2) of H3 implies the condition (1). See the examples in §2.

A closed linear operator $A : \mathcal{D}(A) \to H$ is called regular spectral if its resolvent is compact and its eigenvectors form a Riesz basis of H [20]. Sun has proved in [21] that under the hypothesis H1, with $b \in H$, and a stronger hypothesis than H3, the following condition is necessary and sufficient for the operator $A + b\langle \cdot, h\rangle$ $(h \in H)$ to be regular spectral and to have the spectrum $\{v_k; k \in \mathbb{N}\}$ assigned:

$$\sum_{k\in\mathbb{N}} \left| \frac{v_k - \lambda_k}{b_k} \right|^2 < +\infty. \tag{3}$$

More results on spectrum assignment via linear feedback at the boundary have been obtained in [6], [12], [11], [15], and [16]. Notably, Rebarber has shown that for some cases, it is possible to assign uniformly an infinite number of eigenvalues by unbounded but admissible linear feedback at the boundary [16]. In [11], Lasiecka and Triggiani have given fine sufficient conditions on $\sigma(A)$ and b such that the operator $A + b\langle \cdot, h\rangle_H$ is regular spectral. In [12], Liu has generalized Sun's condition to the class of systems for which the hypotheses H1–H3 are satisfied, with the following condition replacing those of (1) and (2):

$$\inf_{n\neq m} |\lambda_n - \lambda_m| \geq \delta |\lambda_n|^\alpha \text{ and } \sum_{n=1}^{+\infty} \left(\frac{|b_n|}{|\lambda_n|^\alpha} \right)^2 < +\infty \tag{4}$$

for some $\alpha \in \mathbb{R}$ and $\delta > 0$. However the latter condition is restrictive in the sense that it singles out the one-dimensional wave equation and cantilever beam equation with moment control. The aim of this paper is to expand the result of Sun to a more general class of systems satisfying our conditions H1–H3. It is clear that the condition (4) implies the conditions (1) and (2) of the assumption H3.

In this paper, we do not restrict our study to the case of admissible input elements [6], but consider input elements in $\mathcal{D}'(A^*)$. However, we do restrict our study to the case of bounded linear feedbacks (BLF): $u(t) = \langle x(t), h\rangle$ with $h \in H$. We prove that under the hypotheses H1–H3, the condition (3) is also a necessary and sufficient condition for spectrum and Riesz basis assignment. On the one hand, after each BLF, the controlled operator $A + b\langle \cdot, h\rangle$ is still regular spectral. On the other hand, given a set of points satisfying the condition (3), we can compute explicitly the BLF which realizes the spectrum assignment. The main difference between the work of this paper and that of [6], [12], and [16] is that our condition (Theorem 1) is not only sufficient, but also necessary. The necessary part of the condition may find applications in controller design for infinite-dimensional linear systems (cf., [14] and [9]). Our results

also allow us to explain why the system (Σ_O) cannot be exponentially stabilized by the BLF laws when it has an infinite number of eigenvalues in $\Re e(s) \geq 0$ and its input vector b is admissible in the sense of [7]. However we prove that in some cases, the uniform assignment of the spectrum can be achieved by BLFs. We should point out that the construction of BLF laws is simple and systematic as illustrated by our examples.

The paper is organized as follows. In §2, we present our main result and two typical examples. Section 3 is devoted to the proof of our main theorem. The last section contains our conclusions.

2. Main results. As only BLF laws are considered in the paper, the closed-loop system is governed by the evolution equation (Σ_C) in the phase space H :

$$(\Sigma_C): \ \dot{X}(t) = AX(t) + b\langle X(t), h\rangle.$$

The linear operator $A : \mathcal{D}(A) \longrightarrow H \subset \mathcal{D}'(A^*)$ admits the unique extension

$$\hat{A} \in \mathcal{L}(H, \mathcal{D}'(A^*))$$

by continuity because $\mathcal{D}(A)$ is dense in H. Accordingly, the linear operator

$$A_h = A + b\langle \cdot, h\rangle : \mathcal{D}(A) \to \mathcal{D}'(A^*)$$

admits a unique extension from H to $\mathcal{D}'(A^*)$, still denoted by A_h, and for all $x \in H$, $A_h x = \hat{A}x + b\langle x, h\rangle$.

Define now $\mathcal{D}(A_h) = \{x \in H; \hat{A}x + b\langle x, h\rangle \in H\}$. We use here the same definition for the unbounded linear operator $A_h : \mathcal{D}(A_h) \longrightarrow H$ as that of [16]. In the following, instead of directly dealing with A_h, we study the unbounded linear operator $L_h = A^* + h(\cdot, b)$ because the infinitesimal generation property is equivalent between A_h and L_h if they are adjoint w.r.t. each other. It is easy to see that $\mathcal{D}(L_h) = \mathcal{D}(A^*)$ from the hypothesis H2 and that L_h is closed because $h(\cdot, b)$ is A^*-compact [8, p. 194].

LEMMA 1. *The unbounded linear operator $A_h : \mathcal{D}(A_h) \longrightarrow H$ is the adjoint operator of L_h with the inner product $\langle \cdot, \cdot \rangle$ on H.*

Proof. First, let us prove that for all $x \in H$ and $y \in \mathcal{D}(A^{*^2})$,

$$(y, \hat{A}x) = (A^*y, x). \tag{5}$$

Given all $x \in \mathcal{D}(A)$ and $y \in \mathcal{D}(A^{*^2})$, we have

$$(y, \hat{A}x) = (y, Ax) = \langle y, Ax\rangle = \langle A^*y, x\rangle = (A^*y, x).$$

Since the domain $\mathcal{D}(A)$ is dense in H and both the injection $H \to \mathcal{D}'(A^*)$ and the operator $\hat{A}$ are continuous from H to $\mathcal{D}'(A^*)$, the equality (5) is true for all $x \in H$ and $y \in \mathcal{D}(A^{*^2})$.

Now, for all $x \in \mathcal{D}(L_h^*)$ and $y \in \mathcal{D}(A^{*^2})$,

$$\begin{aligned}(y, A_h x) &= (y, \hat{A}x + b\langle x, h\rangle) = (y, \hat{A}x) + \langle h, x\rangle(y, b)\\ &= (y, \hat{A}x) + \langle h(y, b), x\rangle = (A^*y + h(y, b), x)\\ &= (L_h y, x) = \langle L_h y, x\rangle = \langle y, L_h^* x\rangle = (y, L_h^* x).\end{aligned} \tag{6}$$

With $\mathcal{D}(A^{*^2})$ dense in $\mathcal{D}(A^*)$, $A_h x = L_h^* x$. This means that $A_h \supset L_h^*$.

On the other hand, it follows from the equality (5) that for all $x \in \mathcal{D}(A_h)$ and all $y \in \mathcal{D}(A^{*^2})$, $\langle y, A_h x\rangle = (y, A_h x) = \langle L_h y, x\rangle$. Since A^* is the generator of a C_0-semigroup on H, for each $y \in \mathcal{D}(A^*)$, there is a sequence $y_n \in \mathcal{D}(A^{*^2})$ such that $y_n \to y$ and $L_h y_n \to L_h y$ in H as $n \to +\infty$. This means exactly that $x \in \mathcal{D}(L_h^*)$ and $A_h x = L_h^* x$. Hence $A_h \subset L_h^*$. Therefore $A_h = L_h^*$. □

THEOREM 1. *Assume that the hypotheses* H1–H3 *are satisfied. Then,*

1. *for every $h \in H$, the feedback controlled operator A_h is regular spectral and the spectrum $\sigma(A_h) = \{v_k, k \in \mathbb{N}\}$ satisfies the condition* (3);
2. *given a set $\Lambda = \{v_k, k \in \mathbb{N}\}$ such that $v_j \neq v_k$ for $j \neq k$, there exists an $h \in H$ for the operator A_h to have $\sigma(A_h) = \Lambda$ if and only if the set satisfies the condition* (3). *Moreover the feedback is given by*

$$h = \sum_{j=1}^{+\infty} h_j \psi_j \quad \text{and} \quad \overline{h}_j = \frac{v_j - \lambda_j}{\overline{b}_j} \prod_{n=1, n\neq j}^{+\infty} \frac{\lambda_j - v_n}{\lambda_j - \lambda_n}, \tag{7}$$

where $\overline{h}_j$ denotes the complex conjugate of h_j for $j \in \mathbb{N}$.

We should understand that the infinite product in the theorem is the limit of the sequence in l^2

$$h^N = \left\{ \frac{v_j - \lambda_j}{\overline{b}_j} \prod_{n=1, n\neq j}^{N} \frac{\lambda_j - v_n}{\lambda_j - \lambda_n} \right\}_{j=1,2,\ldots},$$

that is,

$$\lim_{N \to +\infty} \sum_{j=1}^{+\infty} |h_j^N - \overline{h}_j|^2 = 0.$$

We will remark that for any set $\Lambda = \{v_k, k \in \mathbb{N}\}$ assignable by bounded linear feedback, there exists necessarily some integer N such that for all k, $j > N$ and $k \neq j$, $v_k \neq v_j$. The detailed proof and discussion of this result will be given in the next section. The main idea is to prove that L_h, considered as perturbation of A^*, is regular spectral, and that from some rank, the eigenvalues of L_h can be located in the disks centered at the eigenvalues of A^* with radius $6|b_n h_n|$. Moreover we show that the corresponding eigenvectors of L_h form a Riesz basis in H. Then the same result is true for the adjoint operator A_h. It follows from [3] that the controlled operators A_h and L_h are the generators of C_0-semigroups on H. Here we give only two typical examples to illustrate the application of Theorem 1.

Example 1. The wave equation

$$u_{tt}(x,t) = u_{xx}(x,t), \quad u(0,t) = 0, \quad u_x(1,t) = \Gamma(t)$$

with the boundary control $\Gamma(t)$. Define the Hilbert spaces $H = W \times L^2[0,1]$ and $W = \{f; f, f_x \in L^2[0,1], f(0) = 0\}$ with the inner product

$$\left\langle \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \right\rangle_H = \int_0^1 \left[f_{1x}(x)\overline{g_{1x}(x)} + f_2(x)\overline{g_2(x)} \right] dx.$$

Using the techniques of [7], we can formally write the control system on the Hilbert space H

$$\dot{\phi}(t) = A\phi(t) + b\Gamma(t).$$

The semigroup generator $A : \mathcal{D}(A) \to H$ is skew-adjoint with

$$A \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \frac{\partial^2}{\partial x^2} & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$$

and

$$\mathcal{D}(A) = \{(f_1, f_2) \in L^2[0,1] \times L^2[0,1], f_1, f_{1x}, f_{1xx} \in L^2[0,1], f_2 \in W, \\ f_1(0) = f_{1x}(1) = 0\}$$

and $b(x) = [0, \delta(1-x)]'$. By direct computation one can find the following results :

- $L_h = A_h^* = -A + h(\cdot, b)$.
- The spectrum $\sigma(-A) = \{\lambda_{\pm k} = \pm i(k + 1/2)\pi, k = 0, 1, \ldots\}$ and the corresponding eigenvectors

$$\psi_{\pm k}(x) = \begin{bmatrix} 1 \\ -\lambda_{\pm k} \end{bmatrix} \frac{i \sin[(k+1/2)\pi x]}{2(k+1/2)\pi}, \quad k = 0, 1, \ldots.$$

Since the operator A is skew-adjoint and its resolvents are compact, the elements $\{\psi_{\pm k}\}$ form an orthogonal basis of H. It is evident that $b_{\pm k} = \pm(-1)^k/2$, $k = 0, 1, \ldots$. For all $j \geq 0$, we have

$$\sum_{n=0, n\neq j}^{+\infty} \left| \frac{1}{\lambda_j - \lambda_n} \right|^2 + \sum_{n=1}^{+\infty} \left| \frac{1}{\lambda_j - \lambda_{-n}} \right|^2 = \sum_{n=0, n\neq j}^{+\infty} \frac{1}{(j-n)^2\pi^2} + \sum_{n=1}^{+\infty} \frac{1}{(j+n+1)^2\pi^2}$$

$$= \sum_{n=j+1}^{+\infty} \frac{1}{(j-n)^2\pi^2} + \sum_{n=0}^{j-1} \frac{1}{(j-n)^2\pi^2} + \sum_{n=1}^{+\infty} \frac{1}{(j+n+1)^2\pi^2} \leq 3 \sum_{k=1}^{+\infty} \frac{1}{k^2\pi^2} = \frac{1}{2}.$$

The same result is true for $j \leq 0$. Hence the condition (2) is satisfied. We show that in this case the condition (2) does imply the condition (1). Indeed for each $\lambda \notin \bigcup_{j=0}^{+\infty} D_{\pm j}$, there is an integer m_0 such that

$$\Im m(\lambda) \in [\Im m(\lambda_{m_0}), \Im m(\lambda_{m_0+1})]$$

because the distance $d_j = |\Im m(\lambda_j) - \Im m(\lambda_{j+1})|$ is greater than some constant. Then for $j \geq m_0 + 2$,

$$|\lambda - \lambda_j|^2 = \Re e^2(\lambda) + [\Im m(\lambda) - (m_0 + 1 + 1/2)\pi + (m_0 + 1 + 1/2)\pi - (j + 1/2)\pi]^2 \\ \geq \Re e^2(\lambda) + (j - m_0 - 1)^2\pi^2 = \Re e^2(\lambda) + |\lambda_j - \lambda_{m_0+1}|^2.$$

For $j \leq m_0 - 1$,

$$|\lambda - \lambda_j|^2 = \Re e^2(\lambda) + [\Im m(\lambda) - (m_0 + 1/2)\pi + (m_0 + 1/2)\pi - (j + 1/2)\pi]^2 \\ \geq \Re e^2(\lambda) + (m_0 - j)^2\pi^2 = \Re e^2(\lambda) + |\lambda_{m_0} - \lambda_j|^2.$$

Moreover we have

$$|\lambda - \lambda_{m_0}|^2, \ |\lambda - \lambda_{m_0+1}|^2 \geq \pi^2/9.$$

Therefore for each $\lambda \notin \bigcup_{j=0}^{+\infty} D_{\pm j}$,

$$\sum_{j=0}^{+\infty} \left| \frac{1}{\lambda - \lambda_j} \right|^2 + \sum_{j=0}^{+\infty} \left| \frac{1}{\lambda - \lambda_{-j}} \right|^2 \leq \left| \frac{1}{\lambda - \lambda_{m_0}} \right|^2 + \left| \frac{1}{\lambda - \lambda_{m_0+1}} \right|^2 \\ + \sum_{j \geq m_0+2}^{+\infty} \left| \frac{1}{\lambda_j - \lambda_{m_0+1}} \right|^2 + \sum_{-\infty}^{m_0-1} \left| \frac{1}{\lambda_j - \lambda_{m_0}} \right|^2 \leq 1 + \frac{18}{\pi^2}.$$

According to Theorem 1, the spectrum assignable by continuous feedback must satisfy the condition

$$\sum_{n=0}^{+\infty} |v_n - i(n+1/2)\pi|^2 + \sum_{n=1}^{+\infty} |v_{-n} + i(n+1/2)\pi|^2 < +\infty.$$

Therefore the best stability result achievable by continuous feedback is strong stability. However we know that the unbounded feedback $\Gamma(t) = -\frac{1}{2}u_t(1,t)$ exponentially stabilizes the system [10], [18]. Moreover the resulting operator is still regular spectral [18].

Example 2. Consider the Hilbert space $H = W_0^2[0,1] \times L^2[0,1]$ with

$$W_0^n[0,1] = \left\{ f; f, f_x, \ldots, \frac{\partial^n}{\partial x^n} f \in L^2[0,1], f(0) = f_x(0) = 0 \right\}$$

and the inner product

$$\left\langle \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \right\rangle_H = \int_0^1 \left[f_{1xx}(x)\overline{g_{1xx}(x)} + f_2(x)\overline{g_2(x)} \right] dx.$$

The cantilever beam equation with the moment force control can be formally written as follows [15] :

$$\dot{\phi}(t) = A\phi(t) + b\Gamma(t),$$

where the operator $A : \mathcal{D}(A) \to H$ is skew-adjoint with

$$A \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{\partial^4}{\partial x^4} & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$$

and

$$\mathcal{D}(A) = \left\{ (f_1, f_2); f_1 \in W_0^4[0,1], f_2 \in W_0^2[0,1], f_{1xxx}(1) = f_{1xx}(1) = 0 \right\},$$

and $b(x) = [0, \delta'(1-x)]'$.

The spectrum of $-A$ is

$$\sigma(-A) = \left\{ \lambda_{\pm k} = \pm i[k\pi + \pi/2 + O(e^{-k\pi})]^2, k = 1, \ldots \right\}.$$

One may find two positive constants M_1 and M_2 such that $b_{\pm n} = \beta_{\pm n} n$ with $M_1 \leq |\beta_{\pm n}| \leq M_2$ for $n \in \mathbb{N}$ (see [15], [19], and also [24]). The eigenvectors of A form a Riesz basis of H. Without loss of generality, we consider only the case where $j \geq 1$. Then we have

$$|\lambda_j - \lambda_n|^2 = [(j-n)\pi + O(e^{-j\pi}) + O(e^{-n\pi})]^2 \times [(j+n)\pi + O(e^{-j\pi}) + O(e^{-n\pi})]^2$$

and

$$|\lambda_j - \lambda_{-n}|^2 = \left\{ [j\pi + \pi/2 + O(e^{-j\pi})]^2 + [n\pi + \pi/2 + O(e^{-n\pi})]^2 \right\}^2.$$

Then we get the following inequality for some fixed number $\tilde{M}_2$:

$$\sum_{n=1, n\neq j}^{+\infty} \left| \frac{b_n}{\lambda_j - \lambda_n} \right|^2 + \sum_{n=1}^{+\infty} \left| \frac{b_{-n}}{\lambda_j - \lambda_{-n}} \right|^2$$

$$\leq \tilde{M}_2 \left\{ \sum_{n=j+1}^{+\infty} \frac{n^2}{(j-n)^2(j+n)^2} + \sum_{n=1}^{j-1} \frac{n^2}{(j-n)^2(n+j)^2} + \sum_{n=1}^{+\infty} \frac{1}{n^2} \right\} \leq \frac{\tilde{M}_2 \pi^2}{2}.$$

As in Example 1, the reader can verify that in this case the condition (2) also implies the condition (1). Following Theorem 1, all set Λ satisfying

$$\sum_{n=-\infty,n\neq 0}^{+\infty}\left|\frac{v_n-\lambda_n}{n}\right|^2<+\infty$$

can be assigned for the spectrum of the operator A_h by continuous feedback. This is why it is possible to assign the spectrum uniformly by continuous feedback [15]. Here the feedback is simple as given in Theorem 1. For instance, the point set $\Lambda=\{v_{\pm n}=-n^p+\lambda_{\pm n}, p<1/2, n=1,2,\ldots\}$ can be assigned for the spectrum of the controlled operator A_h via the continuous feedback of Theorem 1. The resulting semigroup e^{tA_h} is exponentially stable. In particular, taking $\Lambda=\{v_{\pm n}=-\alpha^2+\lambda_{\pm n}, 0\neq\alpha\in\mathbb{R}, n=1,2,\ldots\}$, we get the controlled semigroup e^{tA_h} satisfying

$$\|e^{tA_h}\|_{\mathcal{L}(H)}\leq M_\alpha e^{-t\alpha^2}$$

for some positive constant M_α (depending on the constant α), where the decay rate is arbitrarily fast by increasing the number α^2. The feedback that realizes the spectrum assignment is

$$h=\sum_{j=-\infty,j\neq 0}^{+\infty}\frac{-\alpha^2}{b_j}\prod_{n=1,n\neq j}^{+\infty}\left(\frac{\overline{\lambda}_j-\overline{v}_n}{\overline{\lambda}_j-\overline{\lambda}_n}\right)\prod_{m=1,m\neq -j}^{+\infty}\left(\frac{\overline{\lambda}_j-\overline{v}_{-m}}{\overline{\lambda}_j-\overline{\lambda}_{-m}}\right)\psi_j.$$

3. Proof of Theorem 1. To simplify the presentation, we introduce the following notation. Define the bounded linear functional $\mathcal{F}\in\mathcal{D}'(A^*)$ such that for all $g\in\mathcal{D}(A^*)$, $\mathcal{F}(g)=(g,b)$. For all $\lambda\in\rho(A^*)$, the resolvent set of the operator A^*, define the characteristic function $F_h(\lambda)=1-\mathcal{F}(R(\lambda,A^*)h)$, where $h\in H$ and $R(\lambda,A^*)=(\lambda-A^*)^{-1}$. For each $\lambda_0\in\rho(A^*)$, the linear functional $\mathcal{F}\circ R(\lambda_0,A^*)\in\mathcal{L}(H,\mathbb{C})$. The complex function $F_h(\lambda)$ is analytic on the resolvent set $\rho(A^*)$ (see [6] for a proof). We set the perturbation operator $T=h\mathcal{F}$ which is A^* compact [8, p. 194]. Then the operator A^*+T is closed with $\mathcal{D}(A^*+T)=\mathcal{D}(A^*)$. The following result can be proved by direct computation.

LEMMA 2. *For all $\lambda\in\rho(A^*)$ such that $F_h(\lambda)\neq 0$, we have*

$$R(\lambda,A^*+T)=R(\lambda,A^*)+R(\lambda,A^*)TR(\lambda,A^*)/F_h(\lambda)$$

and $\lambda\in\rho(A^*+T)$. Moreover the perturbed operator A^*+T has compact resolvents.

From this lemma, we know that the spectrum $\sigma(A^*+T)$ consists entirely of isolated eigenvalues with finite multiplicity [8, p. 187]. Now consider the set of disks $\tilde{D}_j$, $j\in\mathbb{N}$, centered at $\{\overline{\lambda}_j\}$ with radius $\frac{1}{3}d_j$. It is evident that $\tilde{D}_j\bigcap\tilde{D}_l=\emptyset$ for $j\neq l$. *Technically we suppose that $\lambda\notin\bigcup_{j\in\mathbb{N}}\tilde{D}_j$ implies that its complex conjugate $\overline{\lambda}\notin\bigcup_{j\in\mathbb{N}}\tilde{D}_j$.* This assumption is minor because in applications the spectrum $\sigma(A)$ is usually symmetric with respect to the real axis in the complex plane. Otherwise it is sufficient to rewrite the condition (1).

LEMMA 3. *For each $h=\sum_{j=1}^{+\infty}h_j\psi_j$, there exists a positive number $R_1>0$ such that the subset in the complex plane*

$$S=\{\lambda;|\lambda|\leq R_1\}\bigcup\bigcup_{j=1}^{+\infty}\tilde{D}_j$$

contains all the spectrum points $\sigma(A^* + T)$.

Proof. To prove this lemma, it is sufficient to verify that $F_h(\lambda) \neq 0$ for all $\lambda \notin S$ (see Lemma 2). From the hypotheses H1 and H2, for each $h \in H$ and all $\lambda \notin \bigcup_{j=1}^{+\infty} \tilde{D}_j$,

$$\mathcal{F}(R(\lambda, A^*)h) = \sum_{n=1}^{+\infty} \frac{h_n b_n}{\lambda - \overline{\lambda}_n}. \tag{8}$$

Since $\{\psi_n\}_{n=1}^{+\infty}$ is a Riesz basis of H, the following is true for some numbers M_1, $M_2 > 0$:

$$M_1^2 \sum_{j \in \mathbb{N}} |h_j|^2 \leq \|h\|_H^2 \leq M_2^2 \sum_{j \in \mathbb{N}} |h_j|^2. \tag{9}$$

It follows directly from (8) that for all $\lambda \notin \bigcup_{j=1}^{+\infty} \tilde{D}_j$,

$$|\mathcal{F}(R(\lambda, A^*)h)| \leq \sum_{n=1}^{N_1} \frac{|h_n b_n|}{|\overline{\lambda} - \lambda_n|} + \left[\sum_{n \geq N_1 + 1} |h_n|^2 \sum_{n \geq N_1 + 1} \left| \frac{b_n}{\overline{\lambda} - \lambda_n} \right|^2 \right]^{\frac{1}{2}}. \tag{10}$$

Using the condition (1) of the hypothesis H3, we can choose a large integer N_1 such that

$$\sum_{n \geq N_1 + 1} |h_n|^2 \sum_{n \geq N_1 + 1} \left| \frac{b_n}{\overline{\lambda} - \lambda_n} \right|^2 \leq \sum_{n \geq N_1 + 1} |h_n|^2 M \leq \frac{1}{36}, \tag{11}$$

(where we have used the fact that $\lambda \notin \bigcup_{j \in \mathbb{N}} \tilde{D}_j$ implies that $\overline{\lambda} \notin \bigcup_{j \in \mathbb{N}} \tilde{D}_j$) and then choose a positive number R_1 large enough so that for all $|\lambda| \geq R_1$,

$$\sum_{n=1}^{N_1} \frac{|h_n b_n|}{|\overline{\lambda} - \lambda_n|} \leq \frac{1}{6}. \tag{12}$$

It follows from the conditions (10)–(12) that for all $\lambda \notin S$,

$$|\mathcal{F}(R(\lambda, A^*)h)| \leq \frac{1}{3},$$

that is, $|F_h(\lambda)| \geq \frac{2}{3}$. This proves that $\sigma(A^* + T) \subset S$. □

We let $\nu(\lambda, A^*)$ and $\nu(\lambda, A^* + T)$ denote the algebraic multiplicities of λ as eigenvalue of A^* and $A^* + T$, respectively, and $n(\lambda)$ denote the order of λ as zero of the characteristic function $F_h(\lambda)$. (The order of λ as pole of the characteristic function $F_h(\lambda)$ counts as negative and $\nu(\lambda, A^*)$ counts as zero if $\lambda \in \rho(A^*)$.) In [12], Liu has proved the following result.

PROPOSITION 1. *For all* λ *in the complex plane,*

$$\nu(\lambda, A^* + T) = \nu(\lambda, A^*) + n(\lambda). \tag{13}$$

LEMMA 4. *For each* $h \in H$, *there exists an integer* N_1 *such that the infinite part of the spectrum points* $\{\overline{v}_n, n \geq N_1\}$ *of* $A^* + T$ *are simple and the corresponding eigenvectors are given by*

$$\tilde{\psi}_n = \begin{cases} \psi_n + \dfrac{\overline{v}_n - \overline{\lambda}_n}{h_n} \displaystyle\sum_{j \neq n} \frac{h_j \psi_j}{\overline{v}_n - \overline{\lambda}_j}, & \text{if } h_n \neq 0, \\ \psi_n + \dfrac{b_n}{F_h(\overline{\lambda}_n)} \displaystyle\sum_{j \neq n} \frac{h_j \psi_j}{\overline{v}_n - \overline{\lambda}_j}, & \text{if } h_n = 0. \end{cases} \tag{14}$$

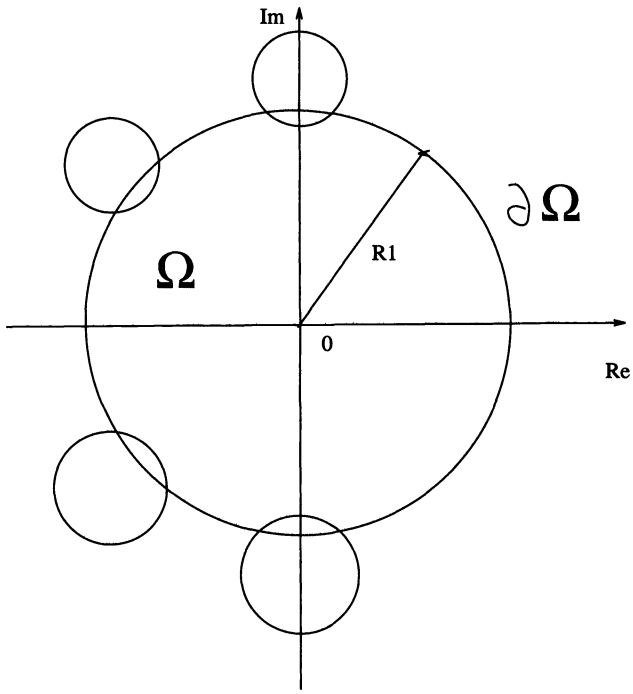


FIG. 1. *Illustrative distribution of the spectrum.*

Moreover the whole spectrum of $A^* + T$ *satisfies the condition*

$$\sum_{n=1}^{+\infty} \left| \frac{v_n - \lambda_n}{b_n} \right|^2 < +\infty.$$

Proof. Consider the disk $D(0, R_1)$ centered at zero with the radius R_1 defined in Lemma 3 such that $\overline{\lambda}_1 \in D(0, R_1)$. For all $\lambda \in \tilde{D}_j$,

$$|\lambda - \overline{\lambda}_1| \geq |\overline{\lambda}_j - \overline{\lambda}_1| - |\lambda - \overline{\lambda}_j| \geq \frac{2}{3}|\overline{\lambda}_j - \overline{\lambda}_1|$$

and

$$|\lambda| \geq |\lambda - \overline{\lambda}_1| - |\overline{\lambda}_1| \geq \frac{2}{3}|\overline{\lambda}_j - \overline{\lambda}_1| - |\overline{\lambda}_1| \geq \frac{2}{3}|\overline{\lambda}_j| - \frac{5}{3}|\overline{\lambda}_1|.$$

Since $\lim_{j \to +\infty} |\lambda_j| = +\infty$, only a finite number N_0 of disks $\tilde{D}_j$ intersect the disk $D(0, R_1)$. Take the boundary $\tilde{C}_j = \{\lambda; |\lambda - \overline{\lambda}_j| = d_j/3\}$ of the disk $\tilde{D}_j$. Define the closed curve $\partial\Omega$ by the boundary of the union

$$\Omega = D(0, R_1) \bigcup_{j=1}^{N_0} \tilde{D}_j$$

(as indicated in Fig. 1). From the proof of Lemma 3, we know that

$$\sup_{\lambda \in \partial\Omega} |\mathcal{F}(R(\lambda, A^*)h)| \leq \frac{1}{3}.$$

From the hypothesis H1, the function $F_h(\lambda)$ has at most N_0 poles in the domain Ω. Applying the Rouché theorem on the functions $g(\lambda) = 1$ and $F_h(\lambda)$ allows us to say that $F_h(\lambda)$ has the same number of zeros as poles in Ω, for $1 - F_h(\lambda) = \mathcal{F}(R(\lambda, A^*)h)$. Then from the identity (13),

$$\sum_{\lambda \in \Omega} \nu(\lambda, A^* + T) = N_0.$$

That means that the operator $A^* + T$ has N_0 eigenvalues (multiplicity counted) in the region Ω. By construction, it is also true that for all $j > N_0$,

$$\sup_{\lambda \in \tilde{C}_j} |\mathcal{F}(R(\lambda, A^*)h)| \leq \frac{1}{3}.$$

The function $F_h(\lambda)$ has either one pole $\overline{\lambda}_j$ or no pole in $\tilde{D}_j$. Applying the Rouché theorem on the functions $g(\lambda) = 1$ and $F_h(\lambda)$ allows us to say that $F_h(\lambda)$ has the same number of zeros as poles in the disk $\tilde{D}_j$. Then from the identity of Liu (13), the operator $A^* + T$ has a simple eigenvalue in $\tilde{D}_j$ for $j > N_0$. Moreover this eigenvalue is either $\{\overline{\lambda}_j\}$ or the unique zero $\overline{v}_j \neq \overline{\lambda}_j$ of $F_h(\lambda)$ in $\tilde{D}_j$.

Actually the simple eigenvalue $\overline{v}_j$ is situated in a smaller disk contained in $\tilde{D}_j$. Because

$$\lim_{\substack{j \to +\infty \\ j \neq n}} \frac{|h_j b_j|}{|\lambda_j - \lambda_n|} = 0,$$

we can take some $N_1 > N_0$ sufficiently large such that for any λ with $|\lambda - \overline{\lambda}_j| \leq 6|h_j b_j|$ and any integer n,

$$\sup_{j \geq N_1} \sup_{n \neq j} \left| \frac{\overline{\lambda}_n - \overline{\lambda}_j}{\lambda - \overline{\lambda}_n} \right| = \sup_{j \geq N_1} \sup_{n \neq j} \frac{1}{\left| 1 + \frac{\lambda - \overline{\lambda}_j}{\overline{\lambda}_j - \overline{\lambda}_n} \right|} \leq \sup_{j \geq N_1} \sup_{n \neq j} \frac{1}{1 - \frac{6|h_j b_j|}{|\lambda_j - \lambda_n|}} \leq \frac{3}{2}. \tag{15}$$

It follows from the definition (8) that for all $j \geq N_1$ and all $|\lambda - \overline{\lambda}_j| = 6|h_j b_j|$,

$$\begin{aligned}
|\mathcal{F}(R(\lambda, A^*)h)| &\leq \sum_{n=1, n \neq j}^{N_2} \frac{|h_n b_n|}{|\overline{\lambda} - \lambda_n|} + \sum_{n > N_2, n \neq j} \frac{|h_n b_n|}{|\overline{\lambda} - \lambda_n|} + \frac{1}{6} \\
&= \sum_{n=1, n \neq j}^{N_2} \frac{|h_n b_n|}{|\overline{\lambda}_j - \overline{\lambda}_n|} \left| \frac{\overline{\lambda}_n - \overline{\lambda}_j}{\lambda - \overline{\lambda}_n} \right| + \sum_{n > N_2, n \neq j} \frac{|h_n b_n|}{|\overline{\lambda}_n - \overline{\lambda}_j|} \left| \frac{\overline{\lambda}_n - \overline{\lambda}_j}{\lambda - \overline{\lambda}_n} \right| + \frac{1}{6} \\
&\leq \sum_{n=1, n \neq j}^{N_2} \frac{3|h_n b_n|}{2|\overline{\lambda}_j - \overline{\lambda}_n|} + \sum_{n > N_2, n \neq j} \frac{3|h_n b_n|}{2|\lambda_n - \lambda_j|} + \frac{1}{6}.
\end{aligned}$$

Take an integer N_2 so large that

$$\sum_{n > N_2, n \neq j} \frac{3|h_n b_n|}{2|\lambda_n - \lambda_j|} \leq \frac{1}{6}.$$

Then we can always choose the integer $N_1 > N_0$ so large that for all $j \geq N_1$,

$$\sum_{n=1, n \neq j}^{N_2} \frac{3|h_n b_n|}{2|\lambda_j - \lambda_n|} \leq \frac{1}{3}.$$

Therefore there is some integer N_1 such that for all $j \geq N_1$ and all $|\lambda - \overline{\lambda}_j| = 6|h_j b_j|$,

$$|\mathcal{F}(R(\lambda, A^*)h)| \leq \frac{2}{3}.$$

Applying again the Rouché theorem on the functions $g(\lambda) = 1$ and $F_h(\lambda)$ and reasoning with the identity (13) as above allows us to prove that $|v_j - \lambda_j| \leq 6|h_j b_j|$ for all $j \geq N_1$. It follows from the above that

$$\sum_{n=1}^{+\infty} \left| \frac{v_n - \lambda_n}{b_n} \right|^2 \leq \sum_{n=1}^{N_1 - 1} \left| \frac{v_n - \lambda_n}{b_n} \right|^2 + \sum_{n \geq N_1} 36|h_j|^2 < +\infty.$$

Now, we compute the corresponding eigenvectors of the operator $A^* + T$:

$$A^* \tilde{\psi}_j + h\mathcal{F}(\tilde{\psi}_j) = \overline{v}_j \tilde{\psi}_j. \tag{16}$$

Observe that for any eigenvector $\tilde{\psi}_j$, $\mathcal{F}(\tilde{\psi}_j) \neq 0$. Suppose that $\mathcal{F}(\tilde{\psi}_j) = 0$ for some j. The only solution of the above eigenvalue equation is $v_j = \lambda_j$ and $\tilde{\psi}_j = \psi_j$. This implies that $\mathcal{F}(\psi_j) = 0$, which contradicts the hypothesis H3. Setting

$$\begin{aligned} \tilde{\psi}_j &= \sum_{m=1}^{+\infty} \alpha_m \psi_m, \\ h &= \sum_{m=1}^{+\infty} h_m \psi_m, \end{aligned} \tag{17}$$

we prove that $v_j = \lambda_j$ in the eigenvalue equation, that is, $\overline{\lambda}_j \in \sigma(A^* + T)$ if and only if $h_j = 0$. Substituting the expression (17) into the equation (16) allows us to obtain the following:

$$(\overline{v}_j - \overline{\lambda}_m)\alpha_m = h_m \mathcal{F}(\tilde{\psi}_j), \quad m = 1, 2, \ldots. \tag{18}$$

It is evident that $h_j = 0$ if $v_j = \lambda_j$. Suppose that $h_j = 0$. Then the characteristic function

$$F_h(\lambda) = 1 - \sum_{n=1, n \neq j}^{+\infty} \frac{h_n b_n}{\lambda - \overline{\lambda}_n}$$

is analytic at the point $\lambda = \overline{\lambda}_j$. This implies that the order $n(\overline{\lambda}_j)$ of the point $\lambda = \overline{\lambda}_j$ as zero of the function $F_h(\lambda)$ is greater than or equal to zero. It follows from Proposition 1 that $\nu(\overline{\lambda}_j, A^* + T) \geq 1$, or $\overline{v}_j = \overline{\lambda}_j$. In particular, for all $j \geq N_1$, $\nu(\overline{\lambda}_j, A^* + T) = 1$. Now we are interested in the eigenvalue equation only for $j \geq N_1$. For $h_j \neq 0$, we know from the above that $v_j \neq \lambda_j$. Then direct computation from (18) leads to

$$\tilde{\psi}_j = \psi_j + \frac{\overline{v}_j - \overline{\lambda}_j}{h_j} \sum_{m \neq j} \frac{h_m \psi_m}{\overline{v}_j - \overline{\lambda}_m}.$$

For $h_j = 0$ and $j \geq N_1$, $v_j = \lambda_j$ and $F_h(\overline{\lambda}_j) \neq 0$ because $\nu(\overline{\lambda}_j, A^* + T) = 1$. One can find that

$$\tilde{\psi}_j = \sum_{m \neq j} \frac{h_m \mathcal{F}(\tilde{\psi}_j)}{\overline{v}_j - \overline{\lambda}_m} \psi_m + \frac{F_h(\overline{\lambda}_j)\mathcal{F}(\tilde{\psi}_j)}{b_j} \psi_j.$$

This finishes the proof of Lemma 4. □

LEMMA 5. *For some integer $N_2 \geq N_1$, the sequence $\{\psi_1, \ldots, \psi_{N_2}, \tilde{\psi}_{N_2+1}, \ldots\}$ forms a Riesz basis of H.*

Proof. Define the linear application $\Lambda : H \to H$ by $\Lambda(\psi_j) = \psi_j$ for $1 \leq j \leq N_2$ and $\Lambda(\psi_j) = \tilde{\psi}_j$ for $j \geq N_2 + 1$. We will prove that the application Λ as well as its inverse Λ^{-1} are bounded. Then the above sequence is also a Riesz basis because it is equivalent to the Riesz basis $\{\psi_j, j \in \mathbb{N}\}$ (see [5, p. 309]).

For all $g = \sum_{j=1}^{+\infty} \alpha_j \psi_j \in H$, using Lemma 4, we get

$$\begin{aligned}\Lambda(g) &= \sum_{j=1}^{+\infty} \alpha_j \psi_j + \sum_{j=N_2+1}^{+\infty} \alpha_j (\tilde{\psi}_j - \psi_j) = \sum_{j=1}^{+\infty} \alpha_j \psi_j + \sum_{j=N_2+1}^{+\infty} \alpha_j \beta_j \sum_{\substack{m \in \mathbb{N} \\ m \neq j}} \frac{h_m \psi_m}{\overline{v}_j - \overline{\lambda}_m} \\ &= g + \Delta\Lambda(g),\end{aligned}$$

where

$$\beta_j = \begin{cases} \dfrac{v_j - \lambda_j}{h_j} & \text{if } h_j \neq 0, \\ \dfrac{b_j}{F_h(\lambda_j)} & \text{if } h_j = 0, \end{cases} \tag{19}$$

and

$$\Delta\Lambda(g) = \sum_{j=N_2+1}^{+\infty} \alpha_j \beta_j \sum_{\substack{m \in \mathbb{N} \\ m \neq j}} \frac{h_m \psi_m}{\overline{v}_j - \overline{\lambda}_m}.$$

For $h_j = 0$, the function $\mathcal{F}(R(\lambda, A^*)h)$ is analytic in $\tilde{D}_j$. Then

$$\sup_{\lambda \in \tilde{D}_j} |\mathcal{F}(R(\lambda, A^*)h)| \leq \sup_{\lambda \in \tilde{C}_j} |\mathcal{F}(R(\lambda, A^*)h)| \leq \frac{1}{3}.$$

This implies that $|F_h(\lambda)| \geq 2/3$ for all $\lambda \in \tilde{D}_j$. In particular, $|F_h(\overline{\lambda}_j)| \geq 2/3$. From (19), for all $j \geq N_2$,

$$|\beta_j| \leq 6|b_j|. \tag{20}$$

Using the fact (9) that $\{\psi_j\}$ is a Riesz basis, we can obtain the following estimates:

$$\begin{aligned}\|\Delta\Lambda(g)\| &\leq \sum_{j=N_2+1}^{+\infty} |\alpha_j \beta_j| \left\| \sum_{m \neq j} \frac{h_m \psi_m}{\overline{v}_j - \overline{\lambda}_m} \right\| \leq M_2 \sum_{j=N_2+1}^{+\infty} |\alpha_j \beta_j| \left[\sum_{m \neq j} \left| \frac{h_m}{v_j - \lambda_m} \right|^2 \right]^{\frac{1}{2}} \\ &\leq M_2 \left[\sum_{j=N_2+1}^{+\infty} |\alpha_j|^2 \right]^{\frac{1}{2}} \left[\sum_{j=N_2+1}^{+\infty} |\beta_j|^2 \sum_{m \neq j} \left| \frac{h_m}{v_j - \lambda_m} \right|^2 \right]^{\frac{1}{2}} \\ &\leq \|g\| \frac{M_2}{M_1} \left[\sum_{j=N_2+1}^{+\infty} |\beta_j|^2 \sum_{m \neq j} \left| \frac{h_m}{v_j - \lambda_m} \right|^2 \right]^{\frac{1}{2}}. \end{aligned} \tag{21}$$

Let us prove that for some integer $N_2 \geq N_1$,

$$\frac{M_2^2}{M_1^2} \sum_{j=N_2+1}^{+\infty} |\beta_j|^2 \sum_{m \neq j} \left| \frac{h_m}{v_j - \lambda_m} \right|^2 \leq \left(\frac{2}{3}\right)^2.$$

From the conditions (20) and (15), choose an N_3 with $N_2 \geq N_1$ such that

$$\frac{M_2^2}{M_1^2} \sum_{j=N_2+1}^{+\infty} |\beta_j|^2 \sum_{m \neq j, m > N_3} \left| \frac{h_m}{v_j - \lambda_m} \right|^2 = \frac{M_2^2}{M_1^2} \sum_{m>N_3}^{+\infty} |h_m|^2 \sum_{j>N_2, j\neq m} \left| \frac{\beta_j}{\lambda_j - \lambda_m} \right|^2 \left| \frac{\lambda_j - \lambda_m}{v_j - \lambda_m} \right|^2$$

$$(22) \qquad \leq \frac{M_2^2}{M_1^2} \sum_{m \geq N_3+1}^{+\infty} |h_m|^2 \sum_{j>N_2, j\neq m} 36 \left| \frac{b_j}{\lambda_j - \lambda_m} \right|^2 \times \left(\frac{3}{2}\right)^2 \leq \left(\frac{1}{3}\right)^2.$$

Since from the hypothesis (2)

$$\lim_{N_2 \to +\infty} \sum_{m=1}^{N_3} |h_m|^2 \sum_{j>N_2, j\neq m} \left| \frac{b_j}{\lambda_j - \lambda_m} \right|^2 = 0,$$

we can always choose the $N_2 \geq N_1$ so large that

$$\frac{M_2^2}{M_1^2} \sum_{j=N_2+1}^{+\infty} |\beta_j|^2 \sum_{m \neq j, m=1}^{N_3} \left| \frac{h_m}{v_j - \lambda_m} \right|^2 = \frac{M_2^2}{M_1^2} \sum_{m=1}^{N_3} |h_m|^2 \sum_{j>N_2, j\neq m} \left| \frac{\beta_j}{\lambda_j - \lambda_m} \right|^2 \left| \frac{\lambda_j - \lambda_m}{v_j - \lambda_m} \right|^2$$

$$(23) \qquad \leq \frac{M_2^2}{M_1^2} \sum_{m=1}^{N_3} |h_m|^2 \sum_{j>N_2, j\neq m} 36 \left| \frac{b_j}{\lambda_j - \lambda_m} \right|^2 \times \left(\frac{3}{2}\right)^2 \leq \left(\frac{1}{3}\right)^2.$$

Substituting (22) and (23) into (21) proves that for some integer $N_2 \geq N_1$, $\|\Delta\Lambda\| \leq 2/3$. Therefore the linear operator Λ and its inverse are bounded. It follows from [5, p. 309] that the sequence $\{\psi_1, \ldots, \psi_{N_2}, \tilde{\psi}_j, j \geq N_2 + 1\}$ is a Riesz basis of H. □

Now let us prove Theorem 1.

Proof of Theorem 1. We take the linearly independent elements $\{\tilde{\psi}_1, \ldots, \tilde{\psi}_{N_2}\}$, which are the generalized eigenvectors of the operator $A^* + T$ corresponding to the finite set $\{\overline{v}_1, \ldots, \overline{v}_{N_2}\}$ of the spectrum $\sigma(A^* + T)$ in the following sense. Without loss of generality, we suppose that in the set $\{\overline{v}_1, \ldots, \overline{v}_{N_2}\}$ there are s distinct eigenvalues $\{\overline{v}_1, \ldots, \overline{v}_s\}$ with the respective algebraic multiplicities $\{\nu_1, \ldots, \nu_s\}$ such that $\sum_{j=1}^{s} \nu_j = N_2$. Then

$$\mathrm{Ker}(\overline{v}_1 - A^* - T)^{\nu_1} = \mathrm{Span}\{\tilde{\psi}_j, j = 1, 2, \ldots, \nu_1\}$$

$$\mathrm{Ker}(\overline{v}_2 - A^* - T)^{\nu_2} = \mathrm{Span}\{\tilde{\psi}_j, j = \nu_1 + 1, \ldots, \nu_1 + \nu_2\},$$

and so on. We want to prove that the sequence $\{\tilde{\psi}_j, j \in \mathbb{N}\}$ is still a Riesz basis of H. Two sequences of vectors $\{g_j, j \in \mathbb{N}\}$ and $\{f_j, j \in \mathbb{N}\}$ are said to be quadratically close if $\sum_{j \in \mathbb{N}} \|g_j - f_j\|^2 < +\infty$. It is obvious that the two sequences $\{\psi_1, \ldots, \psi_{N_2}, \tilde{\psi}_j, j >$

$N_2\}$ and $\{\tilde{\psi}_j, j \in \mathbb{N}\}$ are quadratically close to each other. A sequence of almost normalized vectors $\{g_j\}$ is said to be ω-linearily independent if the equality

$$\sum_{j=1}^{+\infty} c_j g_j = 0 \text{ for } \sum_{j=1}^{+\infty} |c_j|^2 < +\infty$$

implies that $c_j = 0$ for all $j \geq 1$. The Bari theorem [5, Thm. 2.3, p. 317] says that a sequence of ω-linearily independent vectors quadratically close to a Riesz basis is also a Riesz basis. Since we have already shown that the sequence $\{\psi_1, \ldots, \psi_{N_2}, \tilde{\psi}_j, j \geq N_2+1\}$ is a Riesz basis (Lemma 5), we now need only prove the ω-linear independence.

The Laurent series of the resolvent $R(\lambda, A^* + T)$ at $\{\overline{v}_1, \ldots, \overline{v}_s\}$ takes the form [8, p. 181]

$$R(\lambda, A^* + T) = \sum_{m=1}^{s} \left[\frac{P_m}{\lambda - \overline{v}_m} + \sum_{n=1}^{\nu_m - 1} \frac{D_m^n}{(\lambda - \overline{v}_m)^{n+1}} \right] + R_0(\lambda, A^* + T), \tag{24}$$

where the operator P_m is the projector on the subspace $\operatorname{Ker}(\overline{v}_m - A^* - T)^{\nu_m}$ and D_m is the nilpotent commuting with P_m for $m = 1, 2, \ldots, s$ and $R_0(\lambda, A^* + T)$ is analytic at $\overline{v}_j$, $j = 1, \ldots, s$ (see [4, p. 2292]). Let $\{\alpha_j\}_{j=1}^{+\infty} \in l^2$ and $\sum_{j=1}^{+\infty} \alpha_j \tilde{\psi}_j = 0$. Write

$$\sum_{j=1}^{N_2} \alpha_j \tilde{\psi}_j = - \sum_{j=N_2+1}^{+\infty} \alpha_j \tilde{\psi}_j. \tag{25}$$

Applying the resolvent operator on the two sides of the last identity, we obtain

$$R(\lambda, A^* + T) \sum_{j=1}^{N_2} \alpha_j \tilde{\psi}_j = - \sum_{j=N_2+1}^{+\infty} \frac{\alpha_j}{\lambda - \overline{v}_j} \tilde{\psi}_j. \tag{26}$$

Since $P_m P_n = \delta_{n,m} P_n$, $P_n D_n = D_n P_n = D_n$ and $R_0(\lambda, A^* + T) P_n = 0$ for all $\lambda \in \rho(A^* + T)$ [8, p. 181],

$$R(\lambda, A^* + T) \sum_{j=1}^{N_2} \alpha_j \tilde{\psi}_j = \sum_{m=1}^{s} \left\{ \left[\frac{P_m}{\lambda - \overline{v}_m} + \sum_{n=1}^{\nu_m - 1} \frac{D_m^n}{(\lambda - \overline{v}_m)^{n+1}} \right] \sum_{j=1}^{\nu_m} \alpha_j \tilde{\psi}_j \right\}. \tag{27}$$

Substituting (27) into (26) leads to

$$\begin{aligned} \left[\frac{P_1}{\lambda - \overline{v}_1} + \sum_{n=1}^{\nu_1 - 1} \frac{D_1^n}{(\lambda - \overline{v}_1)^{n+1}} \right] \sum_{j=1}^{\nu_1} \alpha_j \tilde{\psi}_j &= - \sum_{j=N_2+1}^{+\infty} \frac{\alpha_j}{\lambda - \overline{v}_j} \tilde{\psi}_j \\ &\quad - \sum_{m=2}^{s} \left\{ \left[\frac{P_m}{\lambda - \overline{v}_m} + \sum_{n=1}^{\nu_m - 1} \frac{D_m^n}{(\lambda - \overline{v}_m)^{n+1}} \right] \sum_{j=1}^{\nu_m} \alpha_j \tilde{\psi}_j \right\}. \end{aligned} \tag{28}$$

Since the controlled operator $A^* + T$ has compact resolvents, its spectrum has no accumulation point different from ∞, that is to say, $\lim_{j \to +\infty} |v_j| = +\infty$. So there is a positive integer δ such that for all $1 \leq k \leq s$ and all $j > N_2$,

$$|v_k - v_j| \geq \delta.$$

On the other hand, the sequence $\{\tilde{\psi}_j, j > N_2\}$ is a part of the Riesz basis in Lemma 5. This implies that for all $1 \leq k \leq s$,

$$\left\| \sum_{j>N_2}^{+\infty} \frac{\alpha_j}{\overline{v}_k - \overline{v}_j} \tilde{\psi}_j \right\|_H^2 \leq M_2^2 \sum_{j>N_2}^{+\infty} \left| \frac{\alpha_j}{\overline{v}_k - \overline{v}_j} \right|^2 \leq \frac{M_2^2}{\delta^2} \sum_{j>N_2}^{+\infty} |\alpha_j|^2.$$

Thus the right side of (28) is bounded on some domain containing $\overline{v}_1$. By multiplying the two sides by $(\lambda - \overline{v}_1)^j$, $j = \nu_1, \nu_1 - 1, \ldots, 1$ and taking the limit for $\lambda \to \overline{v}_1$, we get, successively,

$$D_1^m \sum_{j=1}^{\nu_1} \alpha_j \tilde{\psi}_j = 0, \quad m = \nu_1 - 1, \ldots, 1 \tag{29}$$

and

$$P_1 \left(\sum_{j=1}^{\nu_1} \alpha_j \tilde{\psi}_j \right) = \sum_{j=1}^{\nu_1} \alpha_j \tilde{\psi}_j = 0. \tag{30}$$

Since the elements $\{\tilde{\psi}_1, \ldots, \tilde{\psi}_{\nu_1}\}$ are linearly independent, the relation (30) implies that $\alpha_j = 0$, $j = 1, \ldots, \nu_1$. The relations (29) and (30) imply also that for all $\lambda \in \rho(A^* + T)$, the right side of (28) is identically zero. Repeating the same procedure for the other eigenvalues $\{\overline{v}_2, \overline{v}_3, \ldots, \overline{v}_s\}$ we can prove that $\alpha_j = 0$, $j = 1, \ldots, N_2$. Since the sequence $\{\tilde{\psi}_j, j \geq N_2 + 1\}$ is part of the Riesz basis, the relation (25) implies that $\alpha_j = 0$, $j \geq N_2 + 1$. Thus we have proved ω-linear independence of the sequence $\{\tilde{\psi}_j, j \in \mathbb{N}\}$. Therefore, it is a Riesz basis of H. So we have proved that for every $h \in H$, the contolled operator A_h is regular spectral. This result with Lemma 4 proves the assertion (1) of Theorem 1 and also the necessary part of the assertion (2).

Now let us prove the sufficient part of the assertion (2). Suppose that the given set satisfies the condition (3). Consider the Hilbert space l^2. We show that the element $\{\overline{h}_j\}_{j=1,2,\ldots}$ belongs to l^2, where the $\overline{h}_j$'s are given by the infinite products in (7). For this purpose, we consider the following sequence $h^N \in l^2$:

$$h_j^N = \frac{v_j - \lambda_j}{\overline{b}_j} \prod_{n \neq j}^{N} \left(\frac{\lambda_j - v_n}{\lambda_j - \lambda_n} \right).$$

This is a Cauchy sequence in l^2. In fact,

$$h^N - h^L = \left\{ \frac{v_j - \lambda_j}{\overline{b}_j} \left[\prod_{n \neq j}^{N} \left(\frac{\lambda_j - v_n}{\lambda_j - \lambda_n} \right) - \prod_{n \neq j}^{L} \left(\frac{\lambda_j - v_n}{\lambda_j - \lambda_n} \right) \right] \right\}_{j=1,2,\ldots}.$$

Since

$$\begin{aligned} \sum_{n \neq j} \left| 1 - \frac{\overline{\lambda}_j - \overline{v}_n}{\overline{\lambda}_j - \overline{\lambda}_n} \right| &= \sum_{n \neq j} \left| \frac{\lambda_n - v_n}{\lambda_j - \lambda_n} \right| = \sum_{n \neq j} \left| \frac{\lambda_n - v_n}{b_n} \right| \left| \frac{b_n}{\lambda_j - \lambda_n} \right| \\ &\leq \left[\sum_{n \neq j} \left| \frac{\lambda_n - v_n}{b_n} \right|^2 \sum_{n \neq j} \left| \frac{b_n}{\lambda_j - \lambda_n} \right|^2 \right]^{\frac{1}{2}} \leq M^2, \end{aligned} \tag{31}$$

we have (see [17, p. 291])

$$\prod_{n\neq j}\left|\frac{\lambda_j - v_n}{\lambda_j - \lambda_n}\right| \leq e^{M^2}.$$

Given an $\epsilon > 0$, there is an integer $\tilde{N}$ such that

$$\sum_{j>\tilde{N}}^{+\infty} |h_j^N - h_j^L|^2 < \frac{\epsilon}{2}.$$

Since the condition (31) is true, the infinite sum

$$\sum_{n\neq j}\left|1 - \frac{\overline{\lambda}_j - \overline{v}_n}{\overline{\lambda}_j - \overline{\lambda}_n}\right|$$

converges uniformally with respect to $j \leq \tilde{N}$. This implies that the infinite product

$$\prod_{n\neq j}^{N}\left(\frac{\lambda_j - v_n}{\lambda_j - \lambda_n}\right)$$

converges uniformally with respect to $j \leq \tilde{N}$. Hence for sufficiently large N and L, we have

$$\sum_{j=1}^{\tilde{N}} |h_j^N - h_j^L|^2 = \sum_{j=1}^{\tilde{N}}\left|\frac{v_j - \lambda_j}{\bar{b}_j}\right|^2\left|\prod_{n\neq j}^{N}\left(\frac{\lambda_j - v_n}{\lambda_j - \lambda_n}\right) - \prod_{n\neq j}^{L}\left(\frac{\lambda_j - v_n}{\lambda_j - \lambda_n}\right)\right|^2 \leq \frac{\epsilon}{2}.$$

Thus for sufficiently large N and L,

$$\|h^N - h^L\|_{l^2} < \epsilon.$$

This proves that the limit h, which is the feedback, belongs to the Hilbert space H. We must still prove that the controlled operator has the spectrum assigned $\sigma(A^* + T) = \{\overline{v}_j\}_{j=1}^{+\infty}$. With the feedback element h given in (7) and from the proof of Lemma 4, we know that the controlled operator A^*+T has N_0 eigenvalues (multiplicity counted) in the region Ω and the other eigenvalues are simple and each of them is situated in the corresponding disk. It is sufficient to verify that the finite part of the spectrum contained in Ω is equal to the subset $\{\overline{v}_j; j = 1, 2, \ldots, N_0\}$ and the only simple eigenvalue in the disk $\tilde{D}_j$ is equal to $\overline{v}_j$ for $j > N_0$. Reorder the elements $\overline{v}_1, \overline{v}_2, \ldots, \overline{v}_{N_0}$ such that $v_j = \lambda_j$ if some $\overline{v}_k \in \sigma(A^*)$ for $k \leq N_0$. From the expression of the feedback element, $h_j = 0$ if $v_j = \lambda_j$. So $\overline{\lambda}_j \in \sigma(A^* + T)$ with simple algebraic multiplicity (see Proposition 1). The rest of the spectrum $\sigma(A^* + T)$ is equal to the zero set of the function $F_h(\lambda)$. From Proposition 1 and the hypothesis that $v_j \neq v_n$ for $j \neq n$, it is easy to see that each eigenvalue $\{v_j\}$ is of simple algebraic multiplicity.

We claim that imposing the function $F_h(\lambda)$ to be zero on the point $\overline{v}_j \notin \sigma(A^*)$ gives the following unique solution (7) in l^2 :

$$F_h(\overline{v}_m) = 1 - \sum_{n=1}^{+\infty}\frac{b_n h_n}{\overline{v}_m - \overline{\lambda}_n} = 0.$$

Without loss of generality, the above equation is equivalent to

$$h_m + \sum_{n\neq m}^{+\infty} \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \frac{b_n h_n}{\overline{v}_m - \overline{\lambda}_n} = \frac{\overline{v}_m - \overline{\lambda}_m}{b_m}, \quad m \in \mathbb{N}. \tag{32}$$

Define the linear operator Θ, $T : l^2 \to l^2$ such that

$$\Theta h = \left\{ \sum_{n\neq m}^{+\infty} \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \frac{b_n h_n}{\overline{v}_m - \overline{\lambda}_n} \right\}_{m=1}^{+\infty},$$

and

$$T = I + \Theta.$$

We shall prove that the operator Θ is compact and that the operator T is one-to-one. Then the operator T has a bounded inverse. Set $g = \{(\overline{v}_m - \overline{\lambda}_m)/b_m\}_{m=1}^{+\infty}$. Then the above equation has a unique solution : $h = T^{-1}g$. Define also the sequence of operators $T_n : l^2 \to l^2$ by

$$T_n r = \begin{cases} r_m + \sum_{j\neq m}^{n} \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \frac{b_j r_j}{\overline{v}_m - \overline{\lambda}_j} & \text{for } m \leq n, \\ r_m & \text{for } m \geq n+1. \end{cases}$$

Using the same argument as that used to prove that Θ is compact, we can see that this sequence of operators is bounded. Direct matrix computations (tedious but elementary) allow us to show that T_n is invertible and that

$$T_n^{-1} g = \begin{cases} g_m \prod_{j\neq m}^{n} \frac{\overline{\lambda}_m - \overline{v}_j}{\overline{\lambda}_m - \overline{\lambda}_j} & \text{for } m \leq n, \\ g_m & \text{for } m \geq n+1. \end{cases}$$

As in the above we can show that

$$\lim_{n\to+\infty} T_n^{-1} g = \left\{ g_m \prod_{j\neq m}^{+\infty} \frac{\overline{\lambda}_m - \overline{v}_j}{\overline{\lambda}_m - \overline{\lambda}_j} \right\}_{m=1,2,\ldots}.$$

By direct matrix computations (which are also tedious) we can prove that the sequence of operators T_n^{-1} is bounded. Let us prove that

$$h = T^{-1}g = \lim_{n\to+\infty} T_n^{-1} g.$$

Since we have the identity

$$T_n^{-1} g - T^{-1} g = T_n^{-1}(T - T_n) T^{-1} g,$$

it is sufficient to prove that for any $r \in l^2$,

$$\lim_{n\to+\infty} (T - T_n) r = 0, \tag{33}$$

which will be evident in the following.

Now return to proving the compactness of the operator Θ and the one-to-one property of the operator T. Take any weakly convergent sequence $g^k \in l^2$. Then $\|g^k\|_{l^2} \leq M$ and

$$\Theta g^k = \left\{ \sum_{n \neq m}^{+\infty} \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \frac{b_n g_n^k}{\overline{v}_m - \overline{\lambda}_n} \right\}_{m=1}^{+\infty}.$$

We prove that for every $\epsilon > 0$, there is an $N > 0$ such that for all $k \geq N$,

$$\|\Theta g^k\|_{l^2}^2 < \epsilon. \tag{34}$$

In fact,

$$\|\Theta g^k\|_{l^2}^2 = \sum_{m=1}^{+\infty} \left| \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \right|^2 \left| \sum_{n \neq m}^{+\infty} \frac{b_n g_n^k}{\overline{v}_m - \overline{\lambda}_n} \right|^2.$$

Note that by hypothesis

$$\lim_{N_1 \to +\infty} \sum_{m \geq N_1, m \neq n} \left| \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \right| \left| \frac{b_m}{\lambda_m - \lambda_n} \right| = 0.$$

Therefore there is an integer N_1 such that for all $m \geq N_1$ and $n \neq m$,

$$|v_m - \lambda_m| \leq \frac{1}{3} |\lambda_m - \lambda_n|,$$

$$|v_m - \lambda_n| \geq |\lambda_m - \lambda_n| - |v_m - \lambda_m| \geq \frac{2}{3} |\lambda_m - \lambda_n|.$$

Then for all $m \geq N_1$ and $n \neq m$,

$$\frac{|\lambda_m - \lambda_n|}{|v_m - \lambda_n|} \leq \frac{3}{2}.$$

Since the conditions (3) and (1) are satisfied, we can choose an $\tilde{N} > N_1$ such that for all $m > \tilde{N}$,

$$\begin{aligned}
&\sum_{m=\tilde{N}}^{+\infty} \left| \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \right|^2 \left| \sum_{n \neq m}^{+\infty} \frac{b_n g_n^k}{\overline{v}_m - \overline{\lambda}_n} \right|^2 \\
&\quad \leq \sum_{m=\tilde{N}}^{+\infty} \left| \frac{v_m - \lambda_m}{b_m} \right|^2 \sum_{n \neq m}^{+\infty} \left| \frac{b_n}{v_m - \lambda_n} \right|^2 \sum_{l \neq m}^{+\infty} |g_l^k|^2 \\
&\quad \leq \sum_{m=\tilde{N}}^{+\infty} \left| \frac{v_m - \lambda_m}{b_m} \right|^2 \sum_{n \neq m}^{+\infty} \left| \frac{b_n}{\lambda_m - \lambda_n} \right|^2 \left| \frac{\lambda_m - \lambda_n}{v_m - \lambda_n} \right|^2 \sum_{l \neq m}^{+\infty} |g_l^k|^2 \\
&\quad \leq \sum_{m=\tilde{N}}^{+\infty} \left| \frac{v_m - \lambda_m}{b_m} \right|^2 \sum_{n \neq m}^{+\infty} \left| \frac{b_n}{\lambda_m - \lambda_n} \right|^2 \left(\frac{3}{2} \right)^2 \sum_{l \neq m}^{+\infty} |g_l^k|^2 \leq \frac{\epsilon}{2}.
\end{aligned} \tag{35}$$

Since the point $\overline{v}_j \in \rho(A^*)$, $\mathcal{F} \circ R(\overline{v}_j, A^*) \in \mathcal{L}(H, \mathbb{C})$. It is implied that the elements $r^m = \{b_n/(\overline{v}_m - \overline{\lambda}_n)\}_{n=1}^{+\infty}$ belong to the Hilbert space l^2. The weak convergence of the sequence g^k implies that there is an $N > 0$ such that for all $k > N$

$$\sum_{m=1}^{\tilde{N}-1} \left| \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \right|^2 \left| \sum_{n \neq m}^{+\infty} \frac{b_n g_n^k}{\overline{v}_m - \overline{\lambda}_n} \right|^2 < \frac{\epsilon}{2}. \tag{36}$$

The addition of the inequalities (35) and (36) implies that of (34). Therefore we have proved that

$$\lim_{k \to +\infty} \|\Theta g^k\|_{l^2}^2 = 0$$

and, as a result, the compactness of the operator Θ. Note that for all $r \in l^2$,

$$(T - T_n)r = \begin{cases} \dfrac{\overline{v}_m - \overline{\lambda}_m}{b_m} \displaystyle\sum_{j>n}^{+\infty} \frac{b_j r_j}{\overline{v}_m - \overline{\lambda}_j} & \text{for } m \leq n, \\ \dfrac{\overline{v}_m - \overline{\lambda}_m}{b_m} \displaystyle\sum_{j \neq m}^{+\infty} \frac{b_j r_j}{\overline{v}_m - \overline{\lambda}_j} & \text{for } m \geq n+1. \end{cases}$$

In fact, we have

$$\|(T - T_n)r\|_{l^2}^2 = \sum_{m=1}^{n} \left| \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \right|^2 \left| \sum_{j>n}^{+\infty} \frac{b_j r_j}{\overline{v}_m - \overline{\lambda}_j} \right|^2 + \sum_{m \geq n+1}^{+\infty} \left| \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \right|^2 \left| \sum_{j \neq m}^{+\infty} \frac{b_j r_j}{\overline{v}_m - \overline{\lambda}_j} \right|^2$$

$$\leq \sum_{m=1}^{n} \left| \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \right|^2 \sum_{j>n}^{+\infty} \left| \frac{b_j}{\overline{v}_m - \overline{\lambda}_j} \right|^2 \sum_{l>n}^{+\infty} |r_l|^2 + \sum_{m \geq n+1}^{+\infty} \left| \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \right|^2 \sum_{j \neq m}^{+\infty} \left| \frac{b_j}{\overline{v}_m - \overline{\lambda}_j} \right|^2 \sum_{l \neq m}^{+\infty} |r_l|^2,$$

which tends to zero for $n \to +\infty$, for the two terms

$$\sum_{l>n}^{+\infty} |r_l|^2, \quad \sum_{m \geq n+1}^{+\infty} \left| \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \right|^2$$

go to zero for $n \to +\infty$. Suppose that $Tr = 0$. Then $r = T_n^{-1}(T_n - T)r$. Taking the limit for $n \to +\infty$, we prove that $r = 0$. So the operator T is one-to-one. Finally the unique solution of (32) is

$$h = T^{-1} \left\{ \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \right\}_{m=1}^{+\infty} = \lim_{n \to +\infty} T_n^{-1} \left\{ \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \right\}_{m=1}^{+\infty}$$

$$= \left\{ \frac{\overline{v}_m - \overline{\lambda}_m}{b_m} \prod_{j \neq m}^{+\infty} \frac{\overline{\lambda}_m - \overline{v}_j}{\overline{\lambda}_m - \overline{\lambda}_j} \right\}_{m=1}^{+\infty}.$$

So we have finished the proof of Theorem 1. □

4. Conclusions. In this paper, the necessary and sufficient condition of Sun [21] has been generalized to a large class of distributed parameter systems with boundary controls, which allows us to exploit the limitations imposed by BLF. For example,

the input vector being admissible implies that $\{b_k\} \in l^\infty$ [22], [25]. In this case, BLF cannot uniformly assign the spectrum of the systems. We have proved that it is possible to achieve exponential stabilization of some systems by means of BLF only (Example 2). For an assignable spectrum set, we have given an explicit feedback law which realizes the spectrum assignment with the resulting controlled operator being regular spectral. The paper has also given a systematic method to assign a finite number of spectrum points. This method could find potential applications in damped flexible systems as illustrated by [24] (see [19] and [1] for other models). We should mention that the assumption that the eigenvectors of the operator A constitute a Riesz basis practically reduces the applications to evolution systems in space-dimension one.

Acknowledgments. The authors would like to thank the referees for their valuable suggestions and constructive comments on the paper.

REFERENCES

[1] J. Baillieul and M. Levi, *Rotational elastic dynamics*, Physica, 27D (1987), pp. 43–62.
[2] G. Chen, M. C. Delfour, A. M. Krall, and G. Payre, *Modeling, stabilization, and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.
[3] R. F. Curtain, *Spectral systems*, Internat. J. Control, 39 (1984), pp. 657–666.
[4] N. Dunford and J. Schwartz, *Linear Operators Part* III: *Spectral Operators*, Wiley-Interscience, New York, 1971.
[5] I. C. Gohberg and M. G. Kreĭn, *Introduction to the Theory of Linear Nonselfadjoint Operators*, American Mathematical Society, Providence, RI, 1969.
[6] L. F. Ho, *Spectral assignability of systems with scalar control and application to a degenerate hyperbolic system*, SIAM J. Control Optim., 24 (1986), pp. 1212–1231.
[7] L. F. Ho and L. Russell, *Admissible elements for systems in Hilbert space and a Carleson measure criterion*, SIAM J. Control Optim., 21 (1983), pp. 614–639.
[8] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1976.
[9] T. Kobayashi, *A digital PI-controller for distributed parameter systems*, SIAM J. Control Optim., 26 (1988), pp. 1399–1414.
[10] V. Komornik, *Rapid boundary stabilization of the wave equation*, SIAM J. Control Optim., 29 (1991), pp. 197–208.
[11] I. Lasiecka and R. Triggiani, *Finite rank, relatively bounded perturbations of semigroup generators. Part* II : *spectrum and Riesz basis assignment with application to feedback systems*, Ann. Mat. Pura Appl. (4), 143 (1986), pp. 47–100.
[12] J. Q. Liu, *Perturbation of one rank and the pole assignment*, J. Systems Sci. Math. Sci., no.2(2) (1982), pp. 81–94. (In Chinese with English abstract.)
[13] A. Pazy, *Semigroup of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[14] S. A. Pohjolainen, *Robust multivariable PI-controller for infinite-dimensional systems*, IEEE Trans. Automat. Control, 27 (1982), pp. 17–30.
[15] R. L. Rebarber, *Spectral determination for a cantilever beam*, IEEE Trans. Automat. Control, 34 (1989), pp. 502–510.
[16] ———, *Spectral assignability for distributed parameter systems with unbounded scalar control*, SIAM J. Control Optim., 27 (1989), pp. 148–169.
[17] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
[18] P. Rideau, *Contrôle d'un assemblage de poutres flexibles par des capteurs-actionneurs ponctuels: étude du spectre du système*, Thèse, Ecole Nationale Supérieure des Mines de Paris, Sophia-Antipolis, France, 1985.
[19] D. L. Russell, *Mathematical models for the elastic beam and their control-theoretic implications*, in Autumn College on Semigroups and Applications, International Center for Theoretical Physics, Italy, 1984.
[20] J. Schwartz, *Perturbation of spectral operators and applications* : I. *Bounded perturbation*, Pacific J. Math., 4 (1954), pp. 415–458.
[21] S. H. Sun, *On spectrum distribution of completely controllable linear systems*, SIAM J. Control Optim., 19 (1981), pp. 730–743.
[22] G. Weiss, *Admissibility of input elements for diagonal semigroup on* l^2, Systems Control Lett., 10 (1988), pp. 79–82.

[23] C. Z. Xu and J. Baillieul, *Stabilizability and stabilization of a rotating body-beam system with torque control*, IEEE Trans. Automat. Control, 38 (1993), pp. 1754–1765.

[24] C. Z. Xu and G. Sallet, *Boundary stabilization of rotating flexible systems*, in Lecture Notes in Control and Information Sciences 185, R.F. Curtain, A. Bensoussan, and J.L. Lions, eds., Springer-Verlag, Berlin, New York, pp. 347–365.

[25] ———, *On spectrum assignment of infinite-dimensional linear systems by bounded linear feedback*, Rapport de Recherche, No. 1705, INRIA, 1992.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 542–553, March 1996

008

SUPERIOR INFORMATION IS INSUFFICIENT TO WIN IN GAMES BETWEEN FINITE AUTOMATA*

VLADIMIR CHERNORUTSKII†, RAUF IZMAILOV‡, AND ALEXEI POKROVSKII§

Abstract. A game between two computers is considered: the first computer generates a binary sequence while the second one tries to predict the next element of this sequence using the previous elements. Both computers operate with the same pool of strategies, which is the set of all boolean functions of N arguments. Notwithstanding the asymmetry of the game, it turns out that the value of the game is zero. An algorithm for choosing an optimal superstrategy for the first computer is proposed, and several generalizations of the game are considered.

Key words. repeated games, finite automata, information, games on graphs

AMS subject classifications. 90D20, 90D43

1. Introduction.

Games with finite automata have been investigated for more than 30 years starting with the classical works by Tsetlin [8, 9]. The approach considered here is related to [1, 2, 4]. To describe the model, we start with the description of the motivating experiment.

In 1960s a group of researchers in Voronezh University (Russia) was carrying out the following experiments. Participating undergraduate students (later joined by postgraduate students and professors) were asked to generate sufficiently long (about several hundred digits) binary sequences. The only restriction was the request to generate these sequences without using coins, dice, or other devices—that was considered cheating. The elements of each sequence were transmitted one by one to the input of a computer program specially designed to predict the next element of the sequence using the already accumulated data on the previous ones. It turned out that humans could not generate "good" random sequences and a "smart" computer program predicted from 55% to 80% of elements of any human sequence.

So human intelligence proved to be unable to win against the computer in this game. But what can happen if two computers play against each other in such a game? Obviously, the answer depends on the relative "strength" or "complexity" of the computers and programs involved. But to keep matters simple, let us assume the game is fair and that both computers are of the same class—there is no advantage or disadvantage in complexity from either side. What would be the outcome of the game in this case, and what sort of strategies would be used?

To formalize the problem, consider two players **P** and **S** playing the matrix game with the payoff matrix, known to both players,

$$\mathbf{V} = \begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}.$$

* Received by the editors November 9, 1992; accepted for publication (in revised form) November 29, 1994.

† Institute for Information Transmission Problems, 19 Ermolovoi Street, 101447 Moscow GSP-7, Russia (vvchern@ippi.msk.su). The research of this author was supported in part by Russian Foundation for Fundamental Researches grant 93012884.

‡ NEC USA, Inc. C and C Research Laboratories, 4 Independence Way, Princeton, NJ 08540 (rauf@ccrl.nj.nec.com).

§ Mathematical Department, University of Queensland, Brisbane, Queensland 4072, Australia (ap@maths.uq.oz.au). The research of this author was supported by Australian Research Council grant A 89132609.

Here **P**, being the minimizer, plays rows, whereas the maximizer **S** plays columns. In other words, **S** wins 1 if he guesses **P**'s decision correctly; otherwise **S** loses 1. For each player we are interested in the expected value of his gain during an infinite time. The key features of this game are the description of information accessible to the players and the extraction of a class of rules available to the players from which to choose their moves. Such situations are usually called *supergames* [11], and the permissible rules are called *superstrategies.*

Suppose in addition that **P** has no information about the previous performance of **S**, although **P** knows several of his own previous decisions. Moreover, we assume that exactly the same previous decisions of **P** are available to **S**, so both players make their choices based on the same amount of information. Similar assumptions about players' memory are typical for so-called games with inertia [10]. Thus every pure superstrategy of player **P** consists of choosing a rule for processing a sequence of his preceding decisions. We assume that all realizable rules are deterministic, while randomization is possible only before the beginning of a supergame, when each player chooses his superstrategy.

We consider only the case when every pure superstrategy of **S** consists of processing the available information about the previous moves of **P**, but not about **S**. In other words, **P** generates binary bits, where every bit is determined by N previous ones, and **S** tries to predict these bits, knowing N previous bits in **P**'s sequence. Both players can use the same set of programs for processing of previous bits.

Intuitively, the second player should not lose. After all, he has the advantage of knowing the history of **P**'s moves, whereas **P** has no information about the performance of **S**. On the other hand, **S** has the same complexity as **P**. Therefore the question is whether **S** can use his information edge without using more complex programs than **P**? In other words, *can an advantage in information be exploited without an accompanying advantage in intelligence?*

In the following sections we address this question and present several approaches. First, we describe the exact formulation of the game and present the basic result of this paper, Theorem 1, which gives a negative answer to the question posed above. Next, we present a geometric interpretation of the game which is used to describe an optimal algorithm for **P**, and for other generalizations of the game. Finally, we briefly outline several related problems which will be probably addressed in future research. All proofs are collected in the appendix.

2. Model and main result. To describe the game **G** formally, we fix a natural number N characterizing the "memory depth" of both **P** and **S**: **P** chooses his next move as a function of his N previous moves, while **S** predicts the move of **P** using N previous moves of **P**. The number N is called the *dimension* of the supergame. Let $\mathbb{B}$ be the set $\{0,1\}$; the elements of the set $\mathbb{B}^N$ are the binary representations of the numbers $0, 1, \ldots, 2^N - 1$. Denote by $\mathcal{F}$ the set of all pure superstrategies of the players. Then $\mathcal{F}$ is a set of boolean functions $F : \mathbb{B}^N \to \mathbb{B}$ such that if $F \in \mathcal{F}$, so also is the negation $1 - F \in \mathcal{F}$. Fix $\alpha_0 \in \mathbb{B}^N$ and superstrategies $P, S \in \mathcal{F}$ of players **P** and **S**, respectively. A realization of the supergame is a pair of binary sequences,

$$\{P(\alpha_0), P(\alpha_1), \ldots\} \quad \text{and} \quad \{S(\alpha_0), S(\alpha_1), \ldots\}, \tag{1}$$

of the players' moves, where

$$\alpha_k = P(\alpha_{k-1}) + (2\alpha_{k-1} \bmod 2^N), \quad k = 1, 2, \ldots. \tag{2}$$

Consider, for example, $N = 1$: both players "remember" only the last move. In this case $\mathbf{P}$ can use any of three functions P_1, P_2, and P_3 as a generator of his binary sequence:

$$\begin{cases} P_1(0) = 0, & P_1(1) = 0, \\ P_2(0) = 1, & P_2(1) = 0, \\ P_3(0) = 1, & P_3(1) = 1. \end{cases} \tag{3}$$

Actually, $\mathbf{P}$ can also use the function P_4, defined by

$$P_4(0) = 0, \ P_4(1) = 1,$$

but P_4 may be disregarded since the output sequence it generates is the same as one of the functions P_1 or P_3. The player $\mathbf{S}$ may use four functions $S_1, \ldots, S_4$ for his predictions:

$$\begin{cases} S_1(0) = 0, & S_1(1) = 1, \\ S_2(0) = 1, & S_2(1) = 0, \\ S_3(0) = 0, & S_3(1) = 0, \\ S_4(0) = 1, & S_4(1) = 1; \end{cases} \tag{4}$$

here function S_1 predicts that the next element of the sequence of $\mathbf{P}$ will be the same as the previous one, while the function S_2 predicts exactly the opposite. Each of the other functions S_3 and S_4 steadily predicts 0 or 1 without referring to the history of the game.

Given the superstrategies P and S, the mean payoff $V(P, S)$ of the supergame is determined by

$$V(P, S) = \lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} \eta(P(\alpha_i), S(\alpha_i)), \tag{5}$$

where the payoff function $\eta : \mathbb{B}^2 \to \mathbb{B}$ is defined as $\eta(a, b) = 1 - 2|a - b|$. Since the sequences (1) are periodic, the limit (5) exists.

Considering again the same example $N = 1$, the payoff V may be written in the matrix form

$$(V(P_i, S_j)) = \begin{pmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & 0 & 0 \\ 1 & -1 & -1 & 1 \end{pmatrix}. \tag{6}$$

For any fixed pure superstrategy $S_* \in \mathcal{F}$ of $\mathbf{S}$ the equality

$$\min_{P \in \mathcal{F}} \{V(P, S_*)\} = -1$$

holds (if $P = S_*$); also

$$\max_{S \in \mathcal{F}} \{V(P_*, S)\} = 1$$

holds (if $S = P_*$) for any fixed pure superstrategy P_* of $\mathbf{P}$. These two relations mean that the supergame $\mathbf{G}$ has no solutions in the class of pure superstrategies.

Therefore, mixed superstrategies (probability distributions $p(\cdot)$ on the finite set $\mathcal{F}$ of pure superstrategies) have to be considered. The choice of mixed superstrategy

is the choice of a random rule for using pure superstrategies. If mixed superstrategies $p_1(\cdot)$ and $p_2(\cdot)$ are chosen, the payoff $\overline{V}(p_1(\cdot), p_2(\cdot))$ of the supergame is determined as

$$\overline{V}(p_1(\cdot), p_2(\cdot)) = \sum_{P,S \in \mathcal{F}} p_1(P) p_2(S) V(P, S).$$

So the payoff of the supergame coincides with the payoff of the zero-sum two-person game [6] with the matrix $\mathbf{V} = \{\mathbf{v}_{ij}\}$, where $\mathbf{v}_{ij} = \overline{V}(F_i, F_j)$. The dimension of the matrix $\mathbf{V}$ is equal to the cardinality of the finite set $\mathcal{F}$. Therefore, we use the terms "games" and "strategy" instead of "supergame" and "superstrategy."

The payoff of the game $\mathbf{G}$ is nonnegative since the player $\mathbf{S}$ can guarantee the zero result using the mixed strategy of arbitrary pair of functions S and $1 - S$ with the weights $1/2$. For $\mathbf{P}$ there is no such simple strategy-provided nonnegative result. Nevertheless, the following theorem holds.

THEOREM 1. *The value of the game* $\mathbf{G}$ *is zero.*

Theorem 1 is the main result of this paper. As we can see from (5), the value $V(P, S)$ depends on the initial vector α_0. However, it will shown further that the price of the game does not depend on α_0. Theorem 1 provides the answer to the question formulated in the introduction: *an advantage in information cannot be exploited without an accompanying advantage in intelligence.*

3. Geometric interpretation. The following geometrical interpretation of the game $\mathbf{G}$ is convenient. We will follow the terminology in [3]. Denote by $\Gamma = \Gamma(N)$ the (directed) graph with 2^N vertices enumerated from 0 to $2^N - 1$ where each vertex α is the initial endpoint of exactly two arcs with the terminal endpoints $(2\alpha \bmod 2^N)$ and $(1+(2\alpha \bmod 2^N))$ (compare with the realization (2)).

Denote by $U = U(\Gamma)$ the set of all arcs of the graph Γ. The arc passing from the vertex α to the vertex β is denoted by $e_{\alpha\beta}$ (Figure 1). A subset $\mathcal{U} \subset U$ is called *proper* if for any vertex α the set $\mathcal{U}$ contains exactly 1 arc with the initial endpoint α.

We establish a one-to-one correspondence between the set $\mathcal{F}$ of strategies and the family of proper subsets of arcs. Let a strategy F belong to $\mathcal{F}$. Then for any vertex α choose $\beta = F(\alpha) + (2\alpha \bmod 2^N)$ and include the arc $e_{\alpha\beta}$ in $\mathcal{U}$. Conversely, if $e_{\alpha\beta} \in \mathcal{U}$, then put $F(\alpha) = \beta \bmod 2$. Now it is evident that a choice of the strategy F is equivalent to the choice of the proper subset $\mathcal{U} \in 2^{U(\Gamma)}$. If $\mathcal{U}$ is proper, write

$$\overline{\mathcal{U}} = \{e_{\alpha\gamma} : \gamma = \beta + 1 - 2(\beta \bmod 2),\ e_{\alpha\beta} \in \mathcal{U}\}.$$

Then the set $\mathcal{F}$ is that family of proper subsets $\mathcal{U}$ such that both $\mathcal{U}$ and $\overline{\mathcal{U}}$ belong to the family $\mathcal{F}$.

Fix strategies P, S of both players, and fix an initial point $\alpha_0 \in \mathbb{B}^N$. After an initial period the sequences (1) become periodic. The period of the first sequence corresponds to a elementary circuit in Γ. Denote this cycle by $C(P, \alpha_0)$ and the number of vertices in the cycle by $q(P, \alpha_0)$. Then (5) can be rewritten as

$$V(P, S) = [q(P, \alpha_0)]^{-1} \sum_{\alpha \in C(P, \alpha_0)} \eta(P(\alpha), S(\alpha)). \tag{7}$$

Note that $V(P, S)$ in (7) depends on α_0, but the value of the game does not: it follows from Theorem 1.

Consider an elementary q-cycle C. That is, a cycle with length q. All the arcs generating this cycle form the proper subset $\mathcal{U}(C) \in \mathcal{F}$. This defines the strategy $F(C)$

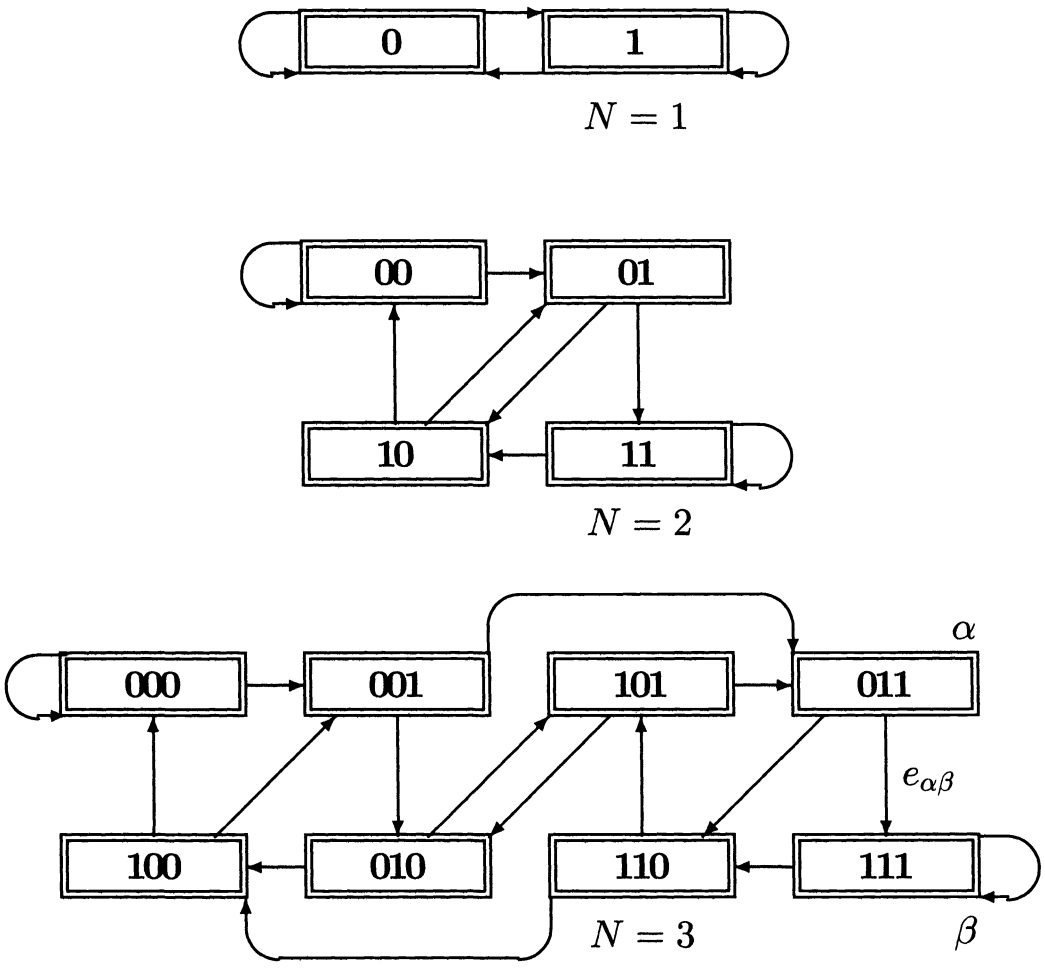


FIG. 1. *Graphs* $\Gamma(N)$.

that generates the cycle C. Denote by $\mathcal{C}(q,N)$ the set of all different elementary q-cycles in the graph $\Gamma(N)$, and let $\xi(q,N)$ be the number of elements in $\mathcal{C}(q,N)$. For example, Figure 2 shows all six elementary cycles for $N=2$. In this case $\xi(1,2) = \xi(3,2) = 2$ and $\xi(2,2) = \xi(4,2) = 1$.

4. Optimal strategies. Let the set $\mathcal{F}$ be the set of all boolean functions. Then an optimal strategy for **P** may be constructed effectively.

Denote by $\nu(N)$ the minimal number of active strategies in an optimal mixed strategy of the player **P**. As mentioned above, the minimal number of active strategies for **S** is 2.

THEOREM 2. *The minimum number* $\nu(N)$ *of active strategies in an optimal strategy of player* **P** *satisfies the upper bound*

$$\nu(N) \leq \sum_{q|(N+1)} \xi(q,N).$$

One of the optimal strategies of **P** *is the mixed strategy consisting of a weighted sum*

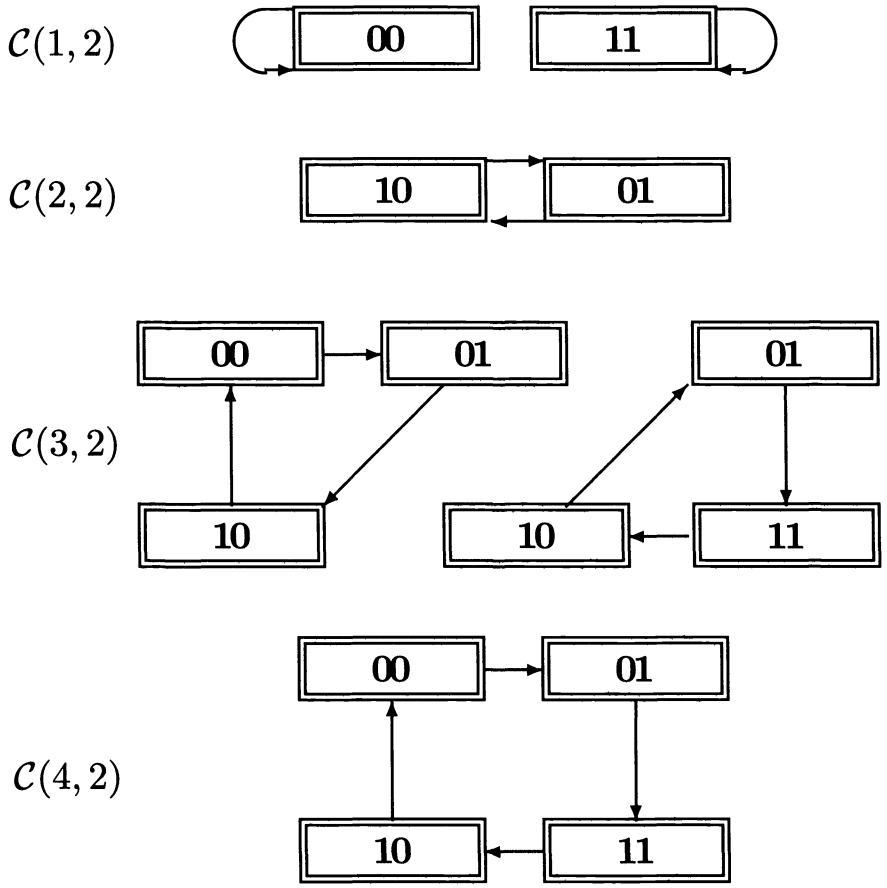


FIG. 2. *Elementary cycles in* $\Gamma(2)$.

of the functions $F(C)$ *generating all the cycles* $C \in \mathcal{C}(q,N)$, *where* $q|(N+1)$. *Every such function* $F(C)$ *has the weight* $2^{-(N+1)}q$ *in the mixed strategy.*

Therefore, the optimal mixture of strategies suggested in Theorem 2 prescribes the weights of the pure strategies to be chosen proportional to the periods of the binary sequences they generate.

For example, if $N = 3$, the optimal strategy may be constructed with 1-cycles, 2-cycles, and 4-cycles ($d = 1, 2, 4$ are the divisors of $N+1$). According to Figure 1, elementary cycles with the corresponding weights shown in Table 1 may be chosen.

For certain cases, an optimal strategy for **P** may be even simpler, as shown in the following statement.

COROLLARY 1. *If* $N+1$ *is prime, then* $\nu(N) \leq 2 + (2^{N+1} - 2)/(N+1)$.

For example, if $N = 2$, **P** may use only two kinds of cycles: 1-cycles and 3-cycles (see Figure 2). This is because 1 and 3 are the only divisors of $N+1$. Choosing them with the weights 1/8 and 3/8, respectively, we obtain the results in Table 2.

If $\mathcal{F}$ does not coincide with the set of all boolean functions, Theorem 2 is not applicable. However, **P** can always choose a "simple" optimal strategy (see [5] for

TABLE 1

Period	Weight	Sequence
1	$\frac{1}{16}$	000000000...
1	$\frac{1}{16}$	111111111...
2	$\frac{1}{8}$	010101010...
4	$\frac{1}{4}$	000100010...
4	$\frac{1}{4}$	001100110...
4	$\frac{1}{4}$	011101110...

TABLE 2

Period	Weight	Sequence
1	$\frac{1}{8}$	000000000...
1	$\frac{1}{8}$	111111111...
2	$\frac{3}{8}$	011011011...
4	$\frac{3}{8}$	001001001...

details).

LEMMA 1. *The set of optimal strategies of* $\mathbf{P}$ *contains a mixed strategy, all components of which have equal weight.*

5. Generalizations. The game $\mathbf{G}$ may be further generalized to the game $\mathbf{G}^*$. Maps $\Phi : I\!B^k \to I\!B^k$ will be called *filters* and positive functions $\Psi : I\!B^k \to I\!R_+$ will be called *factors*, where $I\!R_+$ denotes the set of nonnegative real numbers. The vectors of $I\!B^k$ will be called *strategies*.

Let the set of strategies $\mathcal{F}$ be a subset of $I\!B^k$, and suppose also that it contains along with any element $F = (f_1, \ldots, f_k) \in I\!B^k$ its boolean negation

$$1 - F = (1 - f_1, \ldots, 1 - f_k).$$

The number k is referred to as the *dimension* of the game $\mathbf{G}^*$. The game $\mathbf{G}^*$ is determined by its dimension k, filter Φ, and factor Ψ and the weight vector $M = (\mu_1, \ldots, \mu_k) \in I\!R^k_+$. Given the strategies $P = (p_1, \ldots, p_k)$ and $S = (s_1, \ldots, s_k)$, the payoff is determined by

$$W(P, S) = \Psi(P) \sum_{i=1}^{k} \mu_i \phi_i(P)(1 - 2|p_i - s_i|), \tag{8}$$

where $(\phi_1(P), \ldots, \phi_k(P))$ are the coordinates of the vector $\Phi(P)$. The values (8) form the matrix $\mathbf{W}$ of the game $\mathbf{G}^*$: $\mathbf{W} = (W(P_i, S_j))$.

The game $\mathbf{G}^*$ has the following interpretation. Let k cards be given, where each card i has its own value μ_i. Player **P** can write 0 or 1 on every card: that is, player **P** chooses a binary vector P. We assume that P can be chosen from the set $\mathcal{F}$ of admissible strategies. Then this vector P is filtered by the rule Φ. Only the cards with $\phi_i(P) = 1$ remain under consideration. Player **S** knows the rule Φ, although he does not know the vector P. He can also write 0 or 1 on every card trying to guess the vector P. Player **S** wins $\Psi(P)\mu_i$ if the ith card remains after filtration, $\phi_i(P) = 1$, and his prediction s_i coincides with p_i. Otherwise, if $p_i \neq s_i$, **S** loses $\Psi(P)\mu_i$. The game is repeated infinitely.

THEOREM 3. *For any filter and any factor the value of the game* $\mathbf{G}^*$ *is zero.*

Yet another generalization of the game **G** may be described in terms of the graphical interpretation of §3. Let Γ be an directed graph with exactly two arcs outgoing from every vertex. Denote by $\mathcal{C}(\Gamma)$ the set of all elementary cycles in Γ and by $\mathcal{D}(\Gamma)$ the set of all proper subsets of arcs. For every cycle $C \in \mathcal{C}(\Gamma)$ and for any set $D \in \mathcal{D}(\Gamma)$ define the number $V(C, D)$, which equals the fraction d/c, where d is the number of arcs of D belonging to cycle C and c is the length of this cycle. Enumerating the sets $\mathcal{C}(\Gamma)$ and $\mathcal{D}(\Gamma)$, one can obtain the matrix $\mathbf{V}$ with elements $\mathbf{v}_{ij} = V(C_i, D_j)$. Theorem 1 implies the following theorem.

THEOREM 4. *The value of the game with the matrix* $\mathbf{V}$ *equals* $1/2$.

6. Summary and conclusions. We considered a supergame between two computers, where the first computer generates a binary sequence while its opponent tries to predict the next element of this sequence using the previous elements. Both computers operate with the same pool of strategies, which is the set of boolean functions of N arguments. Despite asymmetry of the game, it is shown that the value of the game is zero. An algorithm for choosing an optimal superstrategy to generate the binary sequences is given.

We have addressed only several issues of this problem. It is interesting and important to address other open issues here. For example, suppose the players **P** and **S** have *different* memory capacities. What is the value of the game **G** in this case, and what would be the optimal strategies?

Another question is to gain an insight into the main reason for the superiority of the computer in the experiments described in the introduction: was it human predictability or the asymmetry of the game (the students were not aware of the computer predictions)? See [7] for other issues related to these experiments.

7. Appendix.

7.1. Proof of Theorem 3. We start with some notations and definitions. Nonzero vectors with nonnegative coordinates will be called *positive.*

Vectors $(1, 1, \ldots, 1)$ (of corresponding length) will be denoted by I. Denote by E_n^i the $2^n \times 2^n$-matrix with its ith row being equal to I and all other elements being zero. Let n-vector A (belonging to $\mathbb{B}^n$ or $\mathbb{R}^n$) be given, and $l < m \leq n$. We use the notations $A|_m$ for the mth coordinate of A and $A|_{l\,..\,m}$ for the vector $(A|_l, \ldots, A|_m)$. Given m-vector B, denote by $\langle A, B\rangle$ the concatenation of the A and B: the $(n+m)$-vector C such that $C|_{1,\ldots,n} = A$ and $C|_{n+1,\ldots,n+m} = B$.

It is sufficient to show (see, for example, [11]) that there exist positive vectors A and B (weights in the mixed strategies for the players **S** and **P**, correspondingly) such that

$$\mathbf{W}A = 0, \quad \mathbf{W}^T B = 0. \tag{9}$$

Moreover, further it will be proven that we may assume $A = I$.

Suppose that $\mathcal{F} = I\!B^k$ (the maximal set of functions). The relation (9) will be proved with induction on k. Let $k = 1$. Then the weight vector is $M = \mu_1$, the number of strategies equals to 2. Let the factor Ψ be determined by the relations

$$\Psi(0) = \psi_1, \ \Psi(1) = \psi_2.$$

Then four different filters Φ have to be considered:

$$\Phi\begin{pmatrix}0\\1\end{pmatrix} \in \left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}\right\}.$$

Hence only four different matrices $\mathbf{W}$ may occur in this case:

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ -\psi_2\mu_1 & \psi_2\mu_1 \end{pmatrix}, \begin{pmatrix} \psi_1\mu_1 & -\psi_1\mu_1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} \psi_1\mu_1 & -\psi_1\mu_1 \\ -\psi_2\mu_1 & \psi_2\mu_1 \end{pmatrix}.$$

All these matrices satisfy (9) for appropriate pairs A and B. The sum of columns is zero, so we may assume $A = I$.

Suppose that (9) holds for all dimensions up to k inclusively. Fix a filter Φ in the game with the dimension $k+1$. Let $(\mu_1, \dots, \mu_{k+1})$ be the weight vector, $K = 2^{k+1}$, and let $(\psi_1, \dots, \psi_K)$ be the vector consisting of all values of factor Ψ, i.e., $\psi_i = \Psi(\alpha_i)$ (where $\alpha_i \in I\!B^{k+1}$). Split the strategy set into four nonintersecting subsets:

$$\begin{aligned} J_1 &= \{S : \ S|_{k+1} = 0, \ \Phi(S)|_{k+1} = 1\}, \\ J_2 &= \{S : \ S|_{k+1} = 0, \ \Phi(S)|_{k+1} = 0\}, \\ J_3 &= \{S : \ S|_{k+1} = 1, \ \Phi(S)|_{k+1} = 1\}, \\ J_4 &= \{S : \ S|_{k+1} = 1, \ \Phi(S)|_{k+1} = 0\}. \end{aligned}$$

Note that the set $J^0 = (J_1 \cup J_2)$ consists of all numbers from 0 to $K/2-1$ (in the sense of the one-to-one correspondence between natural numbers and elements of $I\!B^k$ stated above). In the same way, the set $J^1 = (J_3 \cup J_4)$ consists of all numbers from $K/2$ to $K-1$. Therefore, the sets J^0, J^1 can be considered as $I\!B^k$. Define two filters Φ^1 and Φ^2 as restrictions of Φ on J^0 and J^1, respectively:

$$\Phi^1(S) = \Phi(\langle S, 0\rangle)|_{1,\dots,k}, \quad \Phi^2(S) = \Phi(\langle S, 1\rangle)|_{1,\dots,k},$$

where $S \in I\!B^k$. Consider two games $\mathbf{G}^*$ of dimension k: the first game is determined by the filter Φ^1 and the factor $(\psi_1, \dots, \psi_{K/2})$, and the second game is determined by the filter Φ^2 and the factor $(\psi_{K/2+1}, \dots, \psi_K)$; the weight vector $(\mu_1, \dots, \mu_k)$ is the same for both games. Denote by $\mathbf{W}_1$ and $\mathbf{W}_2$ the matrices of these games. Denote

$$\mathbf{W}_1^{\pm} = \mathbf{W}_1 \pm \mu_{k+1} \sum_{i \in J_1} \psi_i E_k^i, \quad \mathbf{W}_2^{\pm} = \mathbf{W}_2 \pm \mu_{k+1} \sum_{i \in J_3} \psi_i E_k^i. \tag{10}$$

By definition, the matrix $\mathbf{W}$ has the form

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1^+ & \mathbf{W}_1^- \\ \mathbf{W}_2^- & \mathbf{W}_2^+ \end{pmatrix}. \tag{11}$$

The induction assumption implies that the sum of all the columns of $\mathbf{W}$ is zero. That proves the first equality in (9) for $A = I$. Using the induction assumption

once again we conclude that there exist positive (2^k)-vectors $B^1 = (b_1^1, b_2^1, \ldots)$ and $B^2 = (b_1^2, b_2^2, \ldots)$ such that $\mathbf{W}_1^T B^1 = \mathbf{W}_2^T B^2 = 0$.

Suppose at least one of the numbers

$$p = \sum_{i \in J_1} \psi_i b_i^1, \quad q = \sum_{i \in J_3} \psi_i b_i^2 \tag{12}$$

is not equal to zero. Then (2^{k+1})-vector $B = \langle qB^1, pB^2 \rangle$ is positive. Therefore (10)–(12) imply

$$\mathbf{W}^T B = 0. \tag{13}$$

If, on the other hand, both numbers $p = q = 0$, then all the terms in (12) equal zero and (assuming $B = \langle B^1, B^2 \rangle$) we obtain (13). This proves the theorem for $\mathcal{F} = \mathbb{B}^k$.

Now let $\mathcal{F} \neq \mathbb{B}^k$. Since any element F belongs to $\mathcal{F}$ together with $I - F$, this is also true for the set $\mathbb{B}^k \setminus \mathcal{F}$. The numbers i and j are referred to as symmetrical if $i + j = 2^k$. The game matrix $\mathbf{W}$ can be obtained from the above construction by deleting some rows and columns with symmetrical numbers. It follows that $\mathbf{W}$ has the form of (10)–(11). Thus Theorem 3 remains valid.

7.2. Proof of Theorem 1. We will show that the game $\mathbf{G}$ is a special case of the game $\mathbf{G}^*$.

Let $k = 2^N$. (N and k are the dimensions of the games $\mathbf{G}$ and $\mathbf{G}^*$, respectively.) Consider the graph $\Gamma(\mathbf{G})$ corresponding to the game $\mathbf{G}$, and enumerate all its vertices with the numbers $0, \ldots, k-1$. Fix an initial vertex α_0. Now we establish a one-to-one correspondence between strategy sets $\mathcal{F}(\mathbf{G})$ and $\mathcal{F}(\mathbf{G}^*)$. As stated in §3, any strategy $P \in \mathcal{F}(\mathbf{G})$ may be considered as a proper subset of $U(\Gamma)$. Let the arc e_{ij} belong to P and denote $P^* = (p_0^*, \ldots, p_{k-1}^*) \in \mathbb{B}^k$, where $p_i^* = j \bmod 2$. Similarly, let $P^* \in \mathcal{F}(\mathbf{G}^*)$ and assume

$$P = \{e_{ij} : j = (2i \bmod k) + p_i^*;\ 0 \leq i \leq k-1\}.$$

The one-to-one correspondence is established.

As shown in §3, any strategy $P \in \mathcal{F}(\mathbf{G})$ defines the elementary cycle $C(P, \alpha_0)$ of length $q(P, \alpha_0)$. Choose the weight vector $M = I$, the factor $\Psi(P^*) = q^{-1}(P, \alpha_0)$, and the filter $\Phi(P^*) = (\phi_1(P^*), \ldots, \phi_k(P^*))$, where

$$\phi_i(P^*) = \begin{cases} 1 & \text{if the vertex } (i-1) \text{ belongs to the cycle } C(P, \alpha_0), \\ 0 & \text{otherwise.} \end{cases}$$

These definitions with (7) and (8) imply $\mathbf{V} = \mathbf{W}$; i.e., the games matrices are equal. Generally, different initial vertices α_0 define different games $\mathbf{G}^*$. Nevertheless, Theorem 3 implies that the price of any of these games is zero. This completes the proof of Theorem 1.

7.3. Proof of Theorem 2. The following *frequency* vectors are used further. Consider 2^{N+1}-dimensional vectors and enumerate their coordinates with the corresponding binary $(N+1)$-vectors. For example, if $N = 1$, then $2^{N+1} = 4$ and the components of 4-dimensional vectors are enumerated by (00), (01), (10), and (11), correspondingly. Each periodic binary sequence π defines the following relative frequencies of $(N+1)$-words in π. The relative frequency of a word $b = (b_1, \ldots, b_{N+1})$ is the limit

$$\lim_{M \to \infty} \left(\frac{1}{M} F(b, \pi_M) \right),$$

where π_M is the initial segment of π of length M and $F(b, \pi_M)$ is the number of the words b in π_M.

All the sequences π generated by $\mathbf{P}$ may be enumerated by corresponding 2^{N+1}-dimensional frequency vectors p. For example, for $N = 1$ the functions P_1, P_2, P_3 (see (3)) generate three different sequences π_i described by their frequency vectors p_i:

$$
\begin{array}{lll}
 & (00)\quad (01)\quad (10)\quad (11) \\
\pi_1 = 000000\ldots \Longrightarrow & p_1 = (1, \quad 0, \quad 0, \quad 0), \\
\pi_2 = 010101\ldots \Longrightarrow & p_2 = \left(0, \quad \frac{1}{2}, \quad \frac{1}{2}, \quad 0\right), \\
\pi_3 = 111111\ldots \Longrightarrow & p_3 = (0, \quad 0, \quad 0, \quad 1).
\end{array} \tag{14}
$$

The strategies of the player $\mathbf{S}$ are described by 2^{N+1}-dimensional payoff vectors s: if a strategy s predicts the $(N+1)$-th element d_{N+1} after a series $(d_1, \ldots, d_N)$, then the $(d_1, \ldots, d_N, d_{N+1})$-th coordinate of the vector s is equal to $+1$ and its $(d_1, \ldots, d_N, 1 - d_{N+1})$-th coordinate is equal to -1. For example, if $N = 1$, then the strategy S_1 repeating the last symbol generated by $\mathbf{P}$ is described by the vector $(+1, -1, -1, +1)$. Another available strategy S_2 in the case $N = 1$ is described by the vector $s_2 = (-1, +1, +1, -1)$ (see (4)). Other two functions s_3 and s_4 are described by the vectors $(+1, -1, +1, -1)$ and $(-1, +1, -1, +1)$.

In these notations the matrix of the game is the matrix of all pairwise vector products (p_i, s_j). In the case $N = 1$ this matrix has the same form as (6):

	s_1	s_2	s_3	s_4
π_1	1	−1	1	−1
π_2	−1	1	0	0
π_3	1	−1	−1	1.

The 2^{N+1}-dimensional vector I is orthogonal to any vector s_j (this fact follows from the definition of vectors s_j). If such nonnegative weights ζ_i are chosen that

$$
\sum_i \zeta_i p_i = I, \tag{15}
$$

then the mixed strategy of p_i with weights ζ_i (where $\zeta_1 + \zeta_2 + \cdots = 1$) has the zero price for player $\mathbf{P}$. For example, for $N = 1$ the coefficients

$$
\zeta_1 = \frac{1}{4}, \quad \zeta_2 = \frac{1}{2}, \quad \zeta_3 = \frac{1}{4}
$$

may be chosen; then the formulas (14) imply (15) for $N = 1$. Therefore it is sufficient to indicate an algorithm of choosing the base vectors p_i and the weights ζ_i so that (15) holds: in fact, this algorithm provides the way of choosing an optimal mixed strategy. This is what is to be done next, in the final part of the proof.

Consider a strategy F of $\mathbf{P}$ which generates a q-periodic binary sequence π. Then its frequency vector has $2^{N+1} - q$ zero components, and all its q nonzero components are equal to q^{-1}.

Let $q = N+1$. Consider a N-word b and both possible $(N+1)$-periodic extensions of this word:

$$
b0b0b0b0\ldots \quad \text{or} \quad b1b1b1b1\ldots.
$$

The next step is to demonstrate that each of these $(N+1)$-periodic extensions corresponds to some sequence generated by a boolean function of N arguments. The definition of such function F is straightforward: for any N-word $B = (b_k, \ldots, b_{k+N-1})$ in the infinite $(N+1)$-periodic sequence (b_i) we define $F(B) = b_{k+N}$. This way cannot lead to contradictions: if B_1 and B_2 are two $(N+1)$-words in the infinite $(N+1)$-periodic sequence (b_i) and the first N elements of both words B_1 and B_2 coincide, then their $(N+1)$-th elements also coincide (since the period of the sequence has fixed number of 0's and 1's).

Let $d|(N+1)$. Then *any* d-periodic sequence may be generated by some boolean function of N arguments. Denote by P_N the set of all boolean functions generating all d-periodic sequences. Consider two arbitrary strategies $F_1, F_2 \in P_N$ and denote the frequency vectors of these functions by p_1 and p_2. Then there are two possibilities:

1. The vectors p_1 and p_2 coincide.
2. The vectors p_1 and p_2 have no common nonzero components.

To prove it, suppose that there is a common nonzero component $(N+1)^{-1}$ of the vectors p_1 and p_2. Therefore two $(N+1)$-periodic sequences π_1 and π_2 share a $(N+1)$-word, which means they coincide. Now (15) holds for all *distinct* sequences p_i with periods d_i (where $d_i|(N+1)$) with the weights $\zeta_i = d_i^{-1}$.

7.4. Proof of Corollary 1. Using the approach of the proof of Theorem 2 and the fact $N+1$ is prime, the period d of a d-periodic subsequence of an $(N+1)$-periodic output sequence may be equal to either 1 or $(N+1)$. Hence there are two 1-periodical sequences (0000... and 1111...) which must be chosen with the weights $2^{-(N+1)}$, and other $(2^{N+1}-2)$ elements are covered by $(2^{N+1}-2)/(N+1)$ distinct $(N+1)$-periodic sequences; each of these $(N+1)$-periodic sequences has to be chosen with the weights $(N+1)2^{-(N+1)}$.

That gives the estimate of Corollary 1.

Acknowledgment. The authors would like to thank Phil Diamond for reading a draft of the manuscript and making many useful suggestions. We are also grateful to the anonymous reviewers for their helpful comments.

REFERENCES

[1] D. Abreu and A. Rubinstein, *The structure of Nash equilibrium in repeated games with finite automata*, Econometrica, 56 (1988), pp. 1259–1281.

[2] S. Alpern, *Games with repeated decisions*, SIAM J. Control Optim., 26 (1988), pp. 468–477.

[3] C. Berge, *Graphs*, North-Holland Mathematical Library 6, Part 1, North-Holland, Amsterdam, 1991.

[4] E. Kalai and W. Stanford, *Finite rationality and interpersonal complexity in repeated games*, Econometrica, 56 (1988), pp. 397–410.

[5] V. I. Opoytsev and V. V. Chernorutskii, *On the new principle of solution of combinatorial problems*, Automat. Remote Control, 7 (1993), pp. 201–204.

[6] G. Owen, *Game Theory*, Academic Press, New York, 1982.

[7] A. V. Pokrovskii, *Measures of unpredictability of binary sequences*, Soviet Phys. Dokl., 34 (1989), pp. 594–596.

[8] M. L. Tsetlin, *A note on a game of finite automata versus a player using a mixed strategy*, Soviet Math. Dokl., 149 (1963), pp. 52–53. (In Russian.)

[9] M. L. Tsetlin and V. Yu. Krylov, *Examples of games of automata*, Soviet Math. Dokl., 149 (1963), pp. 284–287. (In Russian.)

[10] A. A. Vladimirov, I. V. Emelin, M. A. Krasnosel'skii, and A. V. Pokrovskii, *Infinitely repeating matrix games with inertia*, Transactions VNIISI Game Dynamic Systems, 4 (1982), pp. 84–98. (In Russian.)

[11] S. Zamir, *Topics in Noncooperative Game Theory*, Lecture Notes in Math. 1330, Springer-Verlag, New York, 1988, pp. 72–128.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 554–571, March 1996

009

A NEW FORMULATION OF STATE CONSTRAINT PROBLEMS FOR FIRST-ORDER PDES *

HITOSHI ISHII† AND SHIGEAKI KOIKE‡

Abstract. The first-order Hamilton–Jacobi–Bellman equation associated with the state constraint problem for optimal control is studied. Instead of the boundary condition which Soner introduced, a new and appropriate boundary condition for the PDE is proposed. The uniqueness and Lipschitz continuity of viscosity solutions for the boundary value problem are obtained.

Key words. state constraint problem, viscosity solution, comparison principle

AMS subject classifications. 35B37, 35F30, 49L25

1. Introduction. In this paper we study state constraint (SC) problems for first-order PDEs. The name, state constraint problems, or state-space constraint problems, comes from optimal control. In [10] Soner first considered the problem of characterizing the value functions of state constraint problems in optimal control as the unique viscosity solutions of the associated Hamilton–Jacobi–Bellman equations.

Let us recall Soner's formulation. Consider the first-order PDE

$$H(x,u,Du)=0 \quad \text{in } \Omega, \tag{1}$$

where $H:\overline{\Omega}\times\mathbf{R}\times\mathbf{R}^n\to\mathbf{R}$ is given by

$$H(x,r,p)=\max_{a\in A}\{\lambda r-\langle g(x,a),p\rangle-f(x,a)\}. \tag{2}$$

Here Ω is a bounded open subset of $\mathbf{R}^n$, A is an index set, $\lambda>0$ is a constant, and $g:\overline{\Omega}\times A\to\mathbf{R}^n$ and $f:\overline{\Omega}\times A\to\mathbf{R}$ are given functions.

Equation (1) is the Hamilton–Jacobi–Bellman equation associated with the optimal control problem, where the sets $\overline{\Omega}$ and A, and the constant λ are the state space, the control set, and the discount factor, respectively, and where the functions g and f describe the dynamics and the running cost, respectively.

In Soner's formulation, a function $u\in C(\overline{\Omega})$ is a viscosity solution of the SC problem for (1) if

$$\begin{cases} H(x,u,Du)\le 0 \text{ in } \Omega, \\ H(x,u,Du)\ge 0 \text{ in } \overline{\Omega} \end{cases} \tag{3}$$

in the viscosity sense. We refer to [3] for the definition of viscosity solutions and for a general scope of the theory of viscosity solutions.

The above formulation (3) does not take into account the boundary behavior of u as a subsolution of (1), which reflects the fact that uniqueness results for (3) require

* Received by the editors June 15, 1993; accepted for publication (in revised form) December 5, 1994.

† Department of Mathematics, Chuo University, Kasuga, Bunkyo-ku, Tokyo 112, Japan. The research of this author was supported in part by Ministry of Education, Science, and Culture Grants-in-Aid for Scientific Research 04640189 and 02640150.

‡ Department of Mathematics, Saitama University, Shimo-Ohkubo, Urawa, Saitama 338, Japan. The research of this author was supported in part by Ministry of Education, Science, and Culture Grant-in-Aid for Scientific Research 04740093.

the continuity of the solution u of (3) near the boundary $\partial\Omega$. See for this [10], [2], and [6].

The primary purpose of this paper is to point out that the value function V of the SC problem in optimal control corresponding to the Hamiltonian H satisfies a condition on $\partial\Omega$ in addition to (3). If we take this boundary condition into account in the formulation of the SC problem for (1), then this new SC problem for (1) characterizes the value function as a unique solution among (not necessarily continuous) bounded functions on $\overline{\Omega}$ under natural hypotheses on Ω, g, and f. This will be done in §§2 and 3. It will turn out that the additional boundary condition behaves like an oblique boundary condition. In fact, to show our comparison theorem, we will employ some ideas from [4], [5], [7], and [8] that were useful to oblique boundary value problems. We refer to [1] and [11] for other ideas to show comparison theorems for oblique problems. We also refer to [11] for a general framework to boundary value problems for first-order PDEs.

In §5 we also take up the same subject for SC problems with state-space Ω (replacing $\overline{\Omega}$).

The secondary purpose is to show the Lipschitz continuity of the solution of the SC problem provided that λ is large enough. This result has been proved in [9] via a direct estimation of the value function and the method here gives a new approach to the result. This will be dealt with in §4.

Section 6 is devoted to the proof of certain lemmas.

2. New formulation. In this section we introduce and explain our new formulation of the SC problem for (1).

We assume throughout that A is a compact metric space and that

$$\text{(A1)} \qquad \begin{cases} g : \overline{\Omega} \times A \to \mathbf{R}^n,\ f : \overline{\Omega} \times A \to \mathbf{R} \text{ are continuous,} \\ \sup\limits_{a \in A} \|g(\cdot, a)\|_{C^{0,1}(\overline{\Omega})} < \infty. \end{cases}$$

Note that the function $f : \overline{\Omega} \times A \to \mathbf{R}$ is bounded and uniformly continuous. We may assume that g and f are defined on $\mathbf{R}^n \times A$ and moreover that $\sup_{a \in A} \|g(\cdot, a)\|_{C^{0,1}(\mathbf{R}^n)} < \infty$ and $f : \mathbf{R}^n \times A \to \mathbf{R}$ is bounded and uniformly continuous. We use this convention throughout this paper.

For a given Lipschitz function ξ on $\mathbf{R}^n$ and $x \in \mathbf{R}^n$, the unique solution $Y(t)$ of the initial value problem

$$\frac{dY}{dt}(t) = \xi(Y(t)) \quad \text{for } t > 0 \quad \text{and} \quad Y(0) = x$$

will be denoted by $Y(t; x, \xi)$. For any $z \in \overline{\Omega}$, let $A(z)$ denote the set

$$\{a \in A \mid \exists r > 0 \text{ such that } Y(t; x, g(\cdot, a)) \in \overline{\Omega} \text{ for } x \in \overline{\Omega} \cap B(z, r),\ t \in [0, r]\}.$$

Here and henceforth $B(z, r)$ denotes the closed ball with center z and radius r. It is clear that the multifunction $\overline{\Omega} \ni x \mapsto A(x)$ is lower semicontinuous and that $A(x) = A$ for all $x \in \Omega$.

In what follows we assume that

$$\text{(A2)} \qquad A(z) \neq \emptyset \quad \text{for all} \quad z \in \partial\Omega.$$

We define the inward Hamiltonian $H_{in} : \overline{\Omega} \times \mathbf{R} \times \mathbf{R}^n \to \mathbf{R}$ by

$$H_{in}(x,r,p) = \sup_{a \in A(x)} \{\lambda r - \langle g(x,a), p\rangle - f(x,a)\}.$$

Of course, $H_{in}(x,r,p) = H(x,r,p)$ for $(x,r,p) \in \Omega \times \mathbf{R} \times \mathbf{R}^n$. Moreover, by the semicontinuity of $x \mapsto A(x)$ and the uniform continuity of g, f we see that H_{in} is lower semicontinuous.

Our definition of solutions of the SC problem for (1) is as follows.

DEFINITION 2.1. *We call a bounded function $u : \overline{\Omega} \to \mathbf{R}$ a* viscosity subsolution (*respectively, a* viscosity supersolution) of the SC problem for (1) *if*

$$H_{in}(x, u^*, Du^*) \le 0 \quad \text{in } \overline{\Omega} \qquad (\text{respectively, } H(x, u_*, Du_*) \ge 0 \quad \text{in } \overline{\Omega})$$

in the viscosity sense. Here u^* and u_* denote the upper semicontinuous and the lower semicontinuous envelopes of u, respectively, i.e.,

$$u^*(x) = \lim_{r \downarrow 0} \sup\{u(y) \mid y \in \overline{\Omega} \cap B(x,r)\} \quad \text{for } x \in \overline{\Omega}$$

and $u_* = -(-u)^*$ on $\overline{\Omega}$. We call a bounded function $u : \overline{\Omega} \to \mathbf{R}$ a viscosity solution of the SC problem for (1) *if u is both a viscosity subsolution and a viscosity supersolution of the SC problem for* (1).

Now we consider the SC problem in optimal control associated with the Hamiltonian H given by (2). The value function $V : \overline{\Omega} \to \mathbf{R}$ of the problem is defined as follows: For $x \in \overline{\Omega}$ and a measurable function $\alpha : [0,\infty) \to A$ we consider the state equation

$$\frac{dX}{dt}(t) = g(X(t), \alpha(t)) \quad \text{for} \quad t > 0 \quad \text{and} \quad X(0) = x. \tag{4}$$

We write $X(t; x, \alpha)$ for the solution of (4) to indicate the dependence on x, α. For each $x \in \overline{\Omega}$ define the set $\mathcal{A}(x)$ as the set of all measurable functions $\alpha : [0,\infty) \to A$ such that the solution $X(t)$ of (4) stays in $\overline{\Omega}$ for all $t \in [0,\infty)$. It is easily seen that under assumption (A2), $\mathcal{A}(x) \neq \emptyset$.

Define the value function V on $\overline{\Omega}$ by

$$V(x) = \inf_{\alpha \in \mathcal{A}(x)} \int_0^\infty e^{-\lambda t} f(X(t; x, \alpha), \alpha(t)) dt. \tag{5}$$

THEOREM 2.2. *Under assumptions* (A1) *and* (A2), *V is a viscosity solution of the SC problem for* (1).

Proof. First of all, we observe that V is bounded on $\overline{\Omega}$. As in [10] we see that V satisfies (3) in the viscosity sense. (In [10] it is assumed that V is continuous on $\overline{\Omega}$ but this is not assumed here. Accordingly, we have to modify the argument in [10] a little.)

It remains to show that if $\varphi \in C^1(\overline{\Omega})$ and if $z \in \partial\Omega$ is a maximum point of $V^* - \varphi$, then

$$H_{in}(z, V^*(z), D\varphi(z)) \le 0. \tag{6}$$

To do this fix any $\varphi \in C^1(\overline{\Omega})$ and let $z \in \partial\Omega$ be a maximum point of $V^* - \varphi$. We may assume that $V^*(z) - \varphi(z) = 0$.

Fix any $a \in A(z)$. Set $\alpha(t) \equiv a$. By the definition of $A(z)$, there is $r > 0$ such that

$$X(t; x, \alpha) \in \overline{\Omega} \quad \text{for } x \in \overline{\Omega} \cap B(z, r) \text{ and } t \in [0, r].$$

It is obvious that if the choice of r is appropriate and if $x \in \overline{\Omega} \cap B(z, r)$, then the restriction $\alpha|_{[0,r]}$ can be extended to $[0, \infty)$ so that the resulting function which we denote by α_x belongs to $\mathcal{A}(x)$.

Fix $\varepsilon \in (0, r]$ and choose $x_\varepsilon \in \overline{\Omega} \cap B(z, \varepsilon)$ such that

$$V(x_\varepsilon) > \varphi(x_\varepsilon) - \varepsilon \int_0^\varepsilon e^{-\lambda t} dt. \tag{7}$$

The dynamic programming principle yields

$$V(x_\varepsilon) \leq \int_0^\varepsilon e^{-\lambda t} f(X(t; x_\varepsilon, \alpha_{x_\varepsilon}), a) dt + e^{-\lambda \varepsilon} V(X(\varepsilon; x_\varepsilon, \alpha_{x_\varepsilon})).$$

Hence, using (7) and observing that $V \leq \varphi$ on $\overline{\Omega}$, we get

$$\begin{aligned} 0 &< \int_0^\varepsilon e^{-\lambda t} \{ f(X(t; x_\varepsilon, \alpha_{x_\varepsilon}), a) + \varepsilon \} dt + e^{-\lambda \varepsilon} \varphi(X(\varepsilon; x_\varepsilon, \alpha_{x_\varepsilon})) - \varphi(x_\varepsilon) \\ &= \int_0^\varepsilon e^{-\lambda t} \{ f(X(t; x_\varepsilon, \alpha_{x_\varepsilon}), a) - \lambda \varphi(X(t; x_\varepsilon, \alpha_{x_\varepsilon})) \\ &\qquad + \langle g(X(t; x_\varepsilon, \alpha_{x_\varepsilon}), a), D\varphi(X(t; x_\varepsilon, \alpha_{x_\varepsilon})) \rangle + \varepsilon \} dt. \end{aligned}$$

Thus, there is $y_\varepsilon = X(t_\varepsilon; x_\varepsilon, \alpha_{x_\varepsilon})$ with $t_\varepsilon \in (0, \varepsilon)$ such that

$$-\varepsilon < f(y_\varepsilon, a) - \lambda \varphi(y_\varepsilon) + \langle g(y_\varepsilon, a), D\varphi(y_\varepsilon) \rangle.$$

Now, sending $\varepsilon \downarrow 0$, we conclude that (6) holds. □

3. Comparison of solutions. In what follows we use the condition introduced by Soner [10] which guarantees the continuity of value functions of SC problems in optimal control. Let $G(x)$ denote the set $\{g(x, a) \mid a \in A(x)\}$ for $x \in \overline{\Omega}$. For $B \subset \mathbf{R}^m$ let $\operatorname{co} B$ denote the convex hull of B. Soner's condition is stated as follows:

(A3) For each $z \in \partial\Omega$ there are $r > 0$ and $\xi \in \operatorname{co} G(z)$ such that

$$B(x + t\xi, rt) \subset \overline{\Omega} \quad \text{for all } x \in \overline{\Omega} \cap B(z, r) \text{ and } 0 \leq t \leq r. \tag{8}$$

Let us give a few remarks concerning condition (A3). First, it is obvious that if (A3) holds, then (A2) is satisfied. Second, if (8) holds for some $z \in \partial\Omega$, $r > 0$, and $\xi \in \mathbf{R}^n$, then we have

$$B(x - t\xi, rt) \subset \Omega^c \quad \text{for all } x \in \Omega^c \cap B(z, r) \text{ and } 0 \leq t \leq r.$$

(See the proof of Lemma 6.3 in §6.) That is, if (A3) holds, then the condition (A3) with Ω^c and $-G(z)$ replacing $\overline{\Omega}$ and $G(z)$, respectively, is satisfied. Condition (A3) thus implies that Ω is a Lipschitz domain.

Our main result in this section can now be stated.

THEOREM 3.1. *Let* (A1) *and* (A3) *hold. Let u and v be a viscosity subsolution and a viscosity supersolution of the SC problem for* (1), *respectively. Then $u \leq v$ on $\overline{\Omega}$.*

To prove this theorem, we need a few lemmas.

LEMMA 3.2. *Let* (A1) *and* (A3) *hold. Then there are* $\xi_0 \in C^{0,1}(\mathbf{R}^n, \mathbf{R}^n)$, $\eta_0 \in C(\mathbf{R}^n)$, *and* $r_0 > 0$ *such that*

$$(\xi_0(x), \eta_0(x)) \in \operatorname{co}\{(g(x,a), f(x,a)) \mid a \in A(x)\} \quad \text{for } x \in \partial\Omega$$

and

$$B(x + t\xi_0(x), r_0 t) \subset \overline{\Omega} \quad \text{for all } x \in \partial\Omega \text{ and } 0 \le t \le r_0.$$

Henceforth we assume that (A1) and (A3) are satisfied and fix ξ_0, η_0, and r_0 so that the above conditions are satisfied.

LEMMA 3.3. *There is a function* $\psi \in C^\infty(\overline{\Omega})$ *such that*

$$\langle \xi_0(x), D\psi(x) \rangle \ge 1 \quad \text{on } \partial\Omega.$$

LEMMA 3.4. *There are* $w \in C^1(\overline{\Omega} \times \overline{\Omega})$, *constants* $C_i > 0$, $i = 1, 2, 3$, *and* $r_1 > 0$ *such that*

$$\langle \xi_0(x), D_x w(x,y) \rangle \le 0 \quad \text{for all } x \in \partial\Omega \text{ and } y \in \overline{\Omega} \cap B(x, r_1),$$

and for all $x, y \in \overline{\Omega}$,

$$\begin{gathered} |x-y|^2 \le w(x,y) \le C_1 |x-y|^2, \\ \max\{|D_x w(x,y)|, |D_y w(x,y)|\} \le C_2 |x-y|, \\ |D_x w(x,y) + D_y w(x,y)| \le C_3 |x-y|^2. \end{gathered}$$

We postpone the proof of these lemmas until §6 and here complete the proof of Theorem 3.1, assuming the validity of these lemmas.

Proof. Let ψ be as in Lemma 3.3 and let μ be a constant to be fixed later. Define functions $\tilde{u}$ and $\tilde{v}$ on $\overline{\Omega}$ by

$$\tilde{u}(x) = u^*(x) + \mu\psi(x) \quad \text{and} \quad \tilde{v}(x) = v_*(x) + \mu\psi(x),$$

so that

$$\widetilde{H}_{in}(x, \tilde{u}, D\tilde{u}) \le 0 \quad \text{on } \overline{\Omega}$$

and

$$\widetilde{H}(x, \tilde{v}, D\tilde{v}) \ge 0 \quad \text{on } \overline{\Omega} \tag{9}$$

in the viscosity sense. Here $\widetilde{H}_{in}$ and $\widetilde{H}$ are defined by

$$\begin{aligned} &\widetilde{H}_{in}(x, r, p) \\ &\quad = \sup_{a \in A(x)} \{\lambda r - \langle g(x,a), p \rangle - f(x,a) - \lambda\mu\psi(x) + \mu\langle g(x,a), D\psi(x) \rangle\} \end{aligned}$$

and

$$\begin{aligned} &\widetilde{H}(x, r, p) \\ &\quad = \max_{a \in A}\{\lambda r - \langle g(x,a), p \rangle - f(x,a) - \lambda\mu\psi(x) + \mu\langle g(x,a), D\psi(x) \rangle\}. \end{aligned}$$

In particular, $\tilde{u}$ satisfies

$$\begin{cases} \widetilde{H}(x, \tilde{u}, D\tilde{u}) \leq 0 \ \text{ in } \Omega, \\ -\langle \xi_0(x), D\tilde{u} \rangle \leq -\lambda u^*(x) + \eta_0(x) - \mu \langle \xi_0(x), D\psi(x) \rangle \ \text{ on } \partial\Omega \end{cases}$$

in the viscosity sense (see Definition 7.4 in [3]). Now fix μ large enough so that

$$-\lambda u^*(x) + \eta_0(x) - \mu \langle \xi_0(x), D\psi(x) \rangle \leq -1 \ \text{ on } \partial\Omega.$$

Then $\tilde{u}$ satisfies

$$\begin{cases} \widetilde{H}(x, \tilde{u}, D\tilde{u}) \leq 0 \ \text{ in } \Omega, \\ -\langle \xi_0(x), D\tilde{u} \rangle \leq -1 \ \text{ on } \partial\Omega \end{cases} \tag{10}$$

in the viscosity sense.

It is enough to show that $\tilde{u} \leq \tilde{v}$ on $\overline{\Omega}$. To do this we argue by contradiction. Suppose that $\max_{\overline{\Omega}}(\tilde{u} - \tilde{v}) > 0$. Let $w \in C^1(\overline{\Omega} \times \overline{\Omega})$ be a function as in Lemma 3.4. Let $\varepsilon > 0$ and $(x_\varepsilon, y_\varepsilon)$ be a maximum point of the function

$$(x, y) \mapsto \tilde{u}(x) - \tilde{v}(y) - \frac{1}{\varepsilon} w(x, y) \ \text{ on } \overline{\Omega} \times \overline{\Omega}.$$

Standard arguments show that $(1/\varepsilon) w(x_\varepsilon, y_\varepsilon) \to 0$ as $\varepsilon \downarrow 0$. Choosing $\varepsilon > 0$ small enough, we may assume that $|x_\varepsilon - y_\varepsilon| \leq r_1$, where $r_1 > 0$ is from Lemma 3.4. If $x_\varepsilon \in \partial\Omega$, then in view of (10) we have either

$$-\frac{1}{\varepsilon} \langle \xi_0(x_\varepsilon), D_x w(x_\varepsilon, y_\varepsilon) \rangle \leq -1 \quad \text{or} \quad \widetilde{H}\left(x_\varepsilon, \tilde{u}(x_\varepsilon), \frac{1}{\varepsilon} D_x w(x_\varepsilon, y_\varepsilon) \right).$$

By our choice of w we have

$$\langle \xi_0(x_\varepsilon), D_x w(x_\varepsilon, y_\varepsilon) \rangle \leq 0.$$

This means that we always have

$$\widetilde{H}\left(x_\varepsilon, \tilde{u}(x_\varepsilon), \frac{1}{\varepsilon} D_x w(x_\varepsilon, y_\varepsilon) \right) \leq 0.$$

By (9) we have

$$\widetilde{H}\left(y_\varepsilon, \tilde{v}(y_\varepsilon), -\frac{1}{\varepsilon} D_y w(x_\varepsilon, y_\varepsilon) \right) \geq 0.$$

We proceed as in the standard comparison argument and get a contradiction. □

4. Lipschitz continuity. In this section we assume

$$\sup_{a \in A} \| f(\cdot, a) \|_{C^{0,1}(\overline{\Omega})} < \infty. \tag{A4}$$

In view of Theorems 2.2 and 3.1 there is a unique viscosity solution $u \in C(\overline{\Omega})$ of the SC problem for (1) under the hypotheses (A1) and (A3).

THEOREM 4.1. *Let* (A1), (A3), *and* (A4) *hold. Set*

$$L_1 = \sup_{a\in A} \|g(\cdot,a)\|_{C^{0,1}(\overline{\Omega})} \quad \text{and} \quad \lambda_0 = L_1(C_2+C_3)/2,$$

where C_2 and C_3 are constants from Lemma 3.4. *Let $\lambda > \lambda_0$. Then the viscosity solution $u \in C(\overline{\Omega})$ of the SC problem for* (1) *is Lipschitz continuous on $\overline{\Omega}$.*

Remark. Loreti and Tessitore [9] have already obtained the above result under a slightly stronger hypothesis on Ω. Their choice of λ_0 may differ from ours.

Proof. Let ξ_0, $\tilde{u}$, $\widetilde{H}$, w, r_1, μ, and ψ be as in the proof of Theorem 3.1. Let $C_i > 0$, $i = 2,3$, be constants from Lemma 3.4. We set

$$\tilde{f}(x,a) = f(x,a) + \lambda\mu\psi(x) - \mu\langle g(x,a), D\psi(x)\rangle.$$

Let $\lambda > \lambda_0$. We will prove that

$$\tilde{u}(x) - \tilde{u}(y) \le \alpha w(x,y)^{1/2} \quad \text{for all } x,y \in \overline{\Omega} \tag{11}$$

and for some $\alpha > 0$. Observe that the Lipschitz continuity of u is a direct consequence of (11).

To see that (11) holds, let $\alpha > 0$ be a constant to be fixed later and (x_α, y_α) be a maximum point of the function

$$\Phi(x,y) = \tilde{u}(x) - \tilde{u}(y) - \alpha w(x,y)^{1/2} \quad \text{on } \overline{\Omega}\times\overline{\Omega}.$$

Let $\alpha_0 = 2\|u\|_{C(\overline{\Omega})}/r_1$. Noting that $\Phi(x_\alpha, y_\alpha) \ge 0$, we see that if $\alpha \ge \alpha_0$, then $|x_\alpha - y_\alpha| \le r_1$. We assume henceforth that $\alpha \ge \alpha_0$. If $x_\alpha \ne y_\alpha$, then, as in the proof of Theorem 3.1, we have

$$\widetilde{H}\left(x_\alpha, \tilde{u}(x_\alpha), \frac{\alpha}{2} w(x_\alpha,y_\alpha)^{-1/2} D_x w(x_\alpha,y_\alpha)\right) \le 0,$$

and

$$\widetilde{H}\left(y_\alpha, \tilde{u}(y_\alpha), -\frac{\alpha}{2} w(x_\alpha,y_\alpha)^{-1/2} D_y w(x_\alpha,y_\alpha)\right) \ge 0.$$

Using these and still assuming that $x_\alpha \ne y_\alpha$, we compute that

$$\begin{aligned}
0 &\ge \min_{a\in A}\Big\{\lambda(\tilde{u}(x_\alpha) - \tilde{u}(y_\alpha)) - \frac{\alpha}{2} w(x_\alpha,y_\alpha)^{-1/2}(\langle g(x_\alpha,a) - g(y_\alpha,a), D_x w(x_\alpha,y_\alpha)\rangle \\
&\qquad + \langle g(y_\alpha,a), D_x w(x_\alpha,y_\alpha) + D_y w(x_\alpha,y_\alpha)\rangle) - \tilde{f}(x_\alpha,a) + \tilde{f}(y_\alpha,a)\Big\} \\
&\ge \alpha\lambda w(x_\alpha,y_\alpha)^{1/2} - \frac{\alpha}{2} w(x_\alpha,y_\alpha)^{-1/2} \\
&\quad \times \left(L_1|x_\alpha - y_\alpha| C_2 |x_\alpha - y_\alpha| + L_1 C_3 |x_\alpha - y_\alpha|^2\right) - L_2|x_\alpha - y_\alpha| \\
&\ge w(x_\alpha,y_\alpha)^{1/2}(\alpha(\lambda - \lambda_0) - L_2),
\end{aligned}$$

where

$$L_2 = \sup_{a\in A} \|\tilde{f}(\cdot,a)\|_{C^{0,1}(\overline{\Omega})}.$$

We now fix

$$\alpha = \max\{\alpha_0, (L_2+1)/(\lambda - \lambda_0)\}.$$

We see from the above computation that if $x_\alpha \ne y_\alpha$ then $w(x_\alpha, y_\alpha) \le 0$, which is impossible by our choice of w, and hence we conclude that $x_\alpha = y_\alpha$. This implies (11), and the proof is completed. □

5. The SC problem with state-space Ω. In this section we briefly discuss the SC problem with Ω as its state-space.

We define $\widehat{\mathcal{A}}(x)$ for each $x \in \Omega$ as the set of all measurable functions $\alpha : [0,\infty) \to A$ such that $X(t;x,\alpha) \in \Omega$ for all $t \geq 0$, and the value function $\widehat{V}$ on Ω by

$$\widehat{V}(x) = \inf_{\alpha \in \widehat{\mathcal{A}}(x)} \int_0^\infty e^{-\lambda t} f(X(t;x,\alpha),\alpha(t))dt.$$

Note that $\widehat{\mathcal{A}}(x) \subset \mathcal{A}(x)$ and $V(x) \leq \widehat{V}(x)$ for $x \in \Omega$. For $m = 1,2,\ldots$, we set

$$\Lambda_m = \left\{ (\gamma_1,\ldots,\gamma_m,a_1,\ldots,a_m) \mid \gamma_i \geq 0,\ a_i \in A,\ \sum_{i=1}^m \gamma_i = 1 \right\},$$

and define $\Lambda_m(z)$ for $z \in \overline{\Omega}$ as the set of those $(\gamma_1,\ldots,\gamma_m,a_1,\ldots,a_m) \in \Lambda_m$ which satisfy the following condition: there is $r > 0$ such that if $x,y \in \Omega \cap B(z,r)$ and $0 \leq t \leq r$, then $x + t\sum_{i=1}^m \gamma_i g(y,a_i) \in \Omega$. We set

$$\Lambda = \bigcup_{m\geq 1} \Lambda_m \quad \text{and} \quad \Lambda(z) = \bigcup_{m\geq 1} \Lambda_m(z) \quad \text{for } z \in \overline{\Omega}.$$

It is clear that $\Lambda(x) = \Lambda$ for $x \in \Omega$ under the assumption (A1). We define $\widehat{H}_{in} : \overline{\Omega} \times \mathbf{R} \times \mathbf{R}^n \to \mathbf{R}$ by

$$\widehat{H}_{in}(x,r,p) = \sup \sum_{i=1}^m \gamma_i \{\lambda r - \langle g(x,a_i),p\rangle - f(x,a_i)\},$$

where the supremum is taken over all $(\gamma_1,\ldots,\gamma_m,a_1,\ldots,a_m) \in \Lambda(x)$. We note that $\widehat{H}_{in}(x,r,p) = H(x,r,p)$ if $x \in \Omega$.

We have to modify the definition of viscosity solutions. To this end, we change the definition of upper and lower semicontinuous envelopes. For a bounded function $u : \Omega \to \mathbf{R}$ we define $u^*, u_* : \overline{\Omega} \to \mathbf{R}$ by

$$u^*(x) = \limsup_{r\downarrow 0} \{u(y) \mid y \in \Omega \cap B(x,r)\} \quad \text{and} \quad u_* = -(-u)^*.$$

DEFINITION 5.1. *A bounded function $u : \Omega \to \mathbf{R}$ is a* viscosity subsolution (*respectively, a* viscosity supersolution) of the SC problem for (1) with state-space Ω *if*

$$\widehat{H}_{in}(x,u^*,Du^*) \leq 0 \quad \text{in } \overline{\Omega} \qquad (\text{respectively, } H(x,u_*,Du_*) \geq 0 \quad \text{in } \overline{\Omega})$$

in the viscosity sense. Moreover, a bounded function $u : \Omega \to \mathbf{R}$ is a viscosity solution of the SC problem for (1) with state-space Ω *if it is both a viscosity subsolution and a viscosity supersolution of the SC problem for* (1) *with state-space* Ω.

The following conditions (A5) and (A6) play the roles of (A2) and (A3), respectively, in the current problem.

(A5) $$\Lambda(z) \neq \emptyset \quad \text{for all } z \in \partial\Omega.$$

(A6) For each $z \in \partial\Omega$ there are $r > 0$ and $(\gamma_1,\ldots,\gamma_m,a_1,\ldots,a_m) \in \Lambda$ such that if we set $\xi = \sum_{i=1}^m \gamma_i g(z,a_i)$, then

$$B(x + t\xi, tr) \subset \overline{\Omega} \quad \text{for all } x \in \overline{\Omega} \cap B(z,r) \text{ and } 0 \leq t \leq r.$$

We can now state the following theorems.

THEOREM 5.2. *Let* (A1) *and* (A5) *hold. Then the value function* $\widehat{V}$ *is a viscosity solution of the SC problem for* (1) *with state-space* Ω.

THEOREM 5.3. *Let* (A1) *and* (A6) *hold. Let* u *and* v *be a viscosity subsolution and a viscosity supersolution of the SC problem for* (1) *with state-space* Ω, *respectively. Then* $u \le v$ *on* Ω.

THEOREM 5.4. *Let* (A1), (A4), *and* (A6) *hold. Then there is* $\lambda_0 > 0$ *such that if* $\lambda > \lambda_0$ *and if* $u : \Omega \to \mathbf{R}$ *is a (unique) viscosity solution of the SC problem for* (1) *with state-space* Ω, *then* u *is Lipschitz continuous in* Ω.

We remark that conditions (A1) and (A6) together imply (A5).

LEMMA 5.5. *There are* $r_0 > 0$, $\zeta_i \in C^\infty(\mathbf{R}^n)$, *and* $a_i \in A$, *with* $i = 1, \dots, m$, *such that*

$$(\zeta_1(x), \dots, \zeta_m(x), a_1, \dots, a_m) \in \Lambda_m(x) \quad \text{for all } x \in \partial\Omega$$

and such that if we set $\xi_0(x) = \sum_{i=1}^m \zeta_i(x) g(x, a_i)$, *then*

$$B(x + t\xi_0(x), r_0 t) \subset \overline{\Omega} \quad \text{for } x \in \partial\Omega \text{ and } 0 \le t \le r_0.$$

The proof of this lemma parallels that of Lemma 6.2, so we omit giving it here.

Let $\zeta_i \in C^\infty(\mathbf{R}^n)$, $a_i \in A$, with $i = 1, \dots, m$, and ξ_0 be as in Lemma 5.5. Set

$$\eta_0(x) = \sum_{i=1}^m \zeta_i(x) f(x, a_i).$$

If u is a viscosity subsolution of the SC problem for (1) with state-space Ω, then it satisfies

$$\begin{cases} H(x, u^*, Du^*) \le 0 & \text{in } \Omega, \\ -\langle \xi_0(x), Du^* \rangle \le -\lambda u^*(x) + \eta_0(x) & \text{on } \partial\Omega \end{cases}$$

in the viscosity sense. Using this and proceeding as in the proof of Theorems 3.1 and 4.1, we can prove Theorems 5.3 and 5.4 without difficulty. We omit the details.

The proof of Theorem 5.2 is a little harder than that of Theorem 2.2. We may assume that $g \in C(\mathbf{R}^n \times A, \mathbf{R}^n)$ and $\sup_{a \in A} \|g(\cdot, a)\|_{C^{0,1}(\mathbf{R}^n)} < \infty$. The key observation is stated as Lemma 5.6.

LEMMA 5.6. *Let* $(\gamma_1, \dots, \gamma_m, a_1, \dots, a_m) \in \Lambda$. *Define* $\xi \in C^{0,1}(\mathbf{R}^n, \mathbf{R}^n)$ *by*

$$\xi(x) = \sum_{i=1}^m \gamma_i g(x, a_i).$$

(1) *Then there is a sequence of measurable* $\alpha_k : [0, \infty) \to \{a_1, \dots, a_m\}$ *such that for each* $T > 0$ *and* $h \in C([0, T] \times A)$,

$$\int_0^T h(t, \alpha_k(t)) dt \to \sum_{i=1}^m \gamma_i \int_0^T h(t, a_i) dt \quad \text{as } k \to \infty. \tag{12}$$

(2) *Fix such a sequence* $\{\alpha_k\}$. *Then*

$$X(t; x, \alpha_k) \to Y(t; x, \xi) \quad \text{as } k \to \infty,$$

uniformly on compact subsets of $[0,\infty)$.

Proof. We begin with (1). Set

$$I_1 = [0,\gamma_1),\ I_2 = [\gamma_1,\gamma_1+\gamma_2),\dots,I_m = [\gamma_1+\cdots+\gamma_{m-1},1).$$

Then, the interval $[0,1)$ is the direct sum of $I_1,\dots,I_m$. Define $\alpha_0 : [0,\infty) \to \{a_1,\dots,a_m\}$ by setting $\alpha_0(t) = a_i$ if $t \in I_i$, with $i = 1,\dots,m$, and $\alpha_0(t) = \alpha_0(t-k)$ if $t \in [k,k+1)$, with $k = 1,2,\dots$. Now, define $\alpha_k : [0,\infty) \to \{a_1,\dots,a_m\}$ for $k = 1,2,\dots$ by $\alpha_k(t) = \alpha_0(kt)$. It is not hard to see that for each $T > 0$ and $h \in C([0,T]\times A)$,

$$\int_0^T h(t,\alpha_k(t))dt \to \sum_{i=1}^m \gamma_i \int_0^T h(t,a_i)dt \ \text{ as } k\to\infty.$$

Next, we turn to (2). Fix $x \in \mathbf{R}^n$. By virtue of the Ascoli–Arzela theorem, we can extract a subsequence of $\{k\}$ along which $\{X(t;x,\alpha_k)\}$ converges to a function $Y \in C([0,\infty),\mathbf{R}^n)$ uniformly on compact subsets of $[0,\infty)$. Now, the sequence $\{\alpha_k\}$ satisfies (12). Therefore,

$$\int_0^T g(X(t;x,\alpha_k),\alpha_k(t))dt \to \int_0^T \xi(Y(t))dt \ \text{ as } k\to\infty$$

for each $T > 0$ along the subsequence. From this it is easily seen that $Y(t)$ is a solution of

$$\frac{dY}{dt}(t) = \xi(Y(t)) \ \text{ for } t>0 \ \text{ and } \ Y(0) = x.$$

That is, $Y(t;x,\xi) \equiv Y(t)$. This implies that

$$X(t;x,\alpha_k) \to Y(t;x,\xi) \ \text{ as } k\to\infty,$$

uniformly on compact subsets of $[0,\infty)$. □

As a consequence of Lemma 5.6, we deduce that under assumption (A5), $\widehat{\mathcal{A}}(x) \neq \emptyset$ for all $x \in \Omega$.

Outline of proof of Theorem 5.2. We write $u = \widehat{V}$. We shall prove only that if $\varphi \in C^1(\overline{\Omega})$ and if $z \in \partial\Omega$ is a maximum point of $u^* - \varphi$, then

$$\widehat{H}_{in}(z,u^*(z),D\varphi(z)) \le 0.$$

We may assume that $u^*(z) = \varphi(z)$.

Fix any $(\gamma_1,\dots,\gamma_m,a_1,\dots,a_m) \in \Lambda(z)$. Define ξ as in Lemma 5.6 and $\eta \in C(\overline{\Omega})$ by

$$\eta(x) = \sum_{i=1}^m \gamma_i f(x,a_i).$$

We write $Y(t;x)$ for $Y(t;x,\xi)$ for notational simplicity. We fix $r > 0$ so that $Y(t;x) \in \Omega$ for all $x \in \Omega\cap B(z,r)$ and $t \in [0,r]$.

Fix $\varepsilon \in (0,r]$ and choose $x_\varepsilon \in \Omega\cap B(z,\varepsilon)$ such that

$$u(x_\varepsilon) > \varphi(x_\varepsilon) - \varepsilon\int_0^\varepsilon e^{-\lambda t}dt.$$

By virtue of Lemma 5.6, there is $\alpha_\varepsilon \in \widehat{\mathcal{A}}(x_\varepsilon)$ such that

$$\left|\int_0^\varepsilon e^{-\lambda t}\{f(X(t;x_\varepsilon,\alpha_\varepsilon),\alpha_\varepsilon(t)) - \eta(Y(t;x_\varepsilon))\}dt\right| < \varepsilon\int_0^\varepsilon e^{-\lambda t}dt$$

and such that

$$\left|\int_0^\varepsilon e^{-\lambda t}\{\langle g(X(t;x_\varepsilon,\alpha_\varepsilon),\alpha_\varepsilon(t))D\varphi(X(t;x_\varepsilon,\alpha_\varepsilon))\rangle\right.$$

$$\left.-\langle \xi(Y(t;x_\varepsilon)), D\varphi(Y(t;x_\varepsilon))\rangle\}dt\right| < \varepsilon\int_0^\varepsilon e^{-\lambda t}dt.$$

By the dynamic programming principle, we have

$$u(x_\varepsilon) \le \int_0^\varepsilon e^{-\lambda t} f(X(t;x_\varepsilon,\alpha_\varepsilon),\alpha_\varepsilon(t))dt + e^{-\lambda\varepsilon}u(X(\varepsilon;x_\varepsilon,\alpha_\varepsilon)).$$

Combining these, we get

$$\begin{aligned} 0 &< \int_0^\varepsilon e^{-\lambda t}\{f(X(t;x_\varepsilon,\alpha_\varepsilon),\alpha_\varepsilon(t)) + \varepsilon\}dt + e^{-\lambda\varepsilon}\varphi(X(\varepsilon;x_\varepsilon,\alpha_\varepsilon)) - \varphi(x_\varepsilon) \\ &= \int_0^\varepsilon e^{-\lambda t}\{\eta(Y(t;x_\varepsilon)) - \lambda\varphi(X(t;x_\varepsilon,\alpha_\varepsilon)) + \langle\xi(Y(t;x_\varepsilon)), D\varphi(Y(t;x_\varepsilon))\rangle + 3\varepsilon\}dt. \end{aligned}$$

Thus, there is $y_\varepsilon = Y(t_\varepsilon;x_\varepsilon)$ and $z_\varepsilon = X(t_\varepsilon;x_\varepsilon,\alpha_\varepsilon)$ with $t_\varepsilon \in (0,\varepsilon)$ such that

$$-3\varepsilon < \eta(y_\varepsilon) - \lambda\varphi(z_\varepsilon) + \langle\xi(y_\varepsilon), D\varphi(y_\varepsilon)\rangle.$$

Now, sending $\varepsilon \downarrow 0$, we conclude that

$$\widehat{H}_{in}(z,u^*(z),D\varphi(z)) \le 0. \qquad \square$$

6. Proof of lemmas. Throughout this section we assume that (A1) and (A3) hold. We also assume that g is defined on $\mathbf{R}^n \times A$ and $\sup_{a\in A}\|g(\cdot,a)\|_{C^{0,1}(\mathbf{R}^n)} < \infty$.

The following constructions of ψ and w in Lemmas 3.3 and 3.4, respectively, are similar to those in Dupuis and Ishii [4]. We begin with the following lemma.

LEMMA 6.1. *Let $z \in \mathbf{R}^n$. Let $\xi,\eta \in \mathbf{R}^n$ and $r > 0$ satisfy*

$$B(x+t\xi,rt) \subset \overline{\Omega} \quad \text{for all } x \in \overline{\Omega}\cap B(z,r) \text{ and } 0 \le t \le r, \tag{13}$$

$$B(x+t\eta,rt) \subset \overline{\Omega} \quad \text{for all } x \in \overline{\Omega}\cap B(z,r) \text{ and } 0 \le t \le r. \tag{14}$$

Let $\zeta = \gamma\xi + (1-\gamma)\eta$ for some $\gamma \in [0,1]$. Then there is $s > 0$ such that

$$B(x+t\zeta,st) \subset \overline{\Omega} \quad \text{for all } x \in \overline{\Omega}\cap B(z,s) \text{ and } 0 \le t \le s. \tag{15}$$

Proof. We set

$$s = \min\left\{\frac{r}{2}, \frac{r}{2(r+|\xi|)}\right\},$$

and prove (15) for this s.

Let $0 < t \le s$ and $y \in B(x + t\zeta, st)$. For some $p \in B(0, s)$ we have

$$\begin{aligned} y &= x + t\zeta + tp \\ &= (x + t\gamma\xi + t\gamma p) + t(1-\gamma)\eta + t(1-\gamma)p. \end{aligned}$$

In view of (13) we see that $x + t\gamma\xi + t\gamma p \in \overline{\Omega}$. By our choice of s, it is easily checked that $x + t\gamma\xi + t\gamma p \in B(z, r)$. Hence, by using (14) we conclude that $y \in \overline{\Omega}$. □

LEMMA 6.2. *There are $r > 0$, $\zeta_i \in C^\infty(\mathbf{R}^n)$, and $a_i \in A$, with $i = 1, \dots, m$, satisfying $\zeta_i \ge 0$ and $\sum_{i=1}^m \zeta_i = 1$ on $\partial\Omega$ such that if we set $\xi(x) = \sum_{i=1}^m \zeta_i(x) g(x, a_i)$, then*

$$\xi(x) \in \operatorname{co} G(x) \quad \textit{for } x \in \partial\Omega,$$

and

$$B(x + t\xi(x), rt) \subset \overline{\Omega} \quad \textit{for } x \in \partial\Omega \textit{ and } 0 \le t \le r. \tag{16}$$

Proof. By the compactness of $\partial\Omega$, we deduce from (A3) that there are $r > 0$, $a_i \in A$, $z_j \in \partial\Omega$, and $\gamma_{ij} \ge 0$, with $i = 1, \dots, m$ and $j = 1, \dots, l$, such that

$$\sum_{i=1}^m \gamma_{ij} = 1 \text{ for all } j \quad \text{and} \quad \partial\Omega \subset \bigcup_{j=1}^l B(z_j, r/2),$$

and such that if we set $\xi_j(x) = \sum_{i=1}^m \gamma_{ij} g(x, a_i)$, then for all j,

$$\xi_j(x) \in \operatorname{co} G(x) \text{ for all } x \in B(z_j, r) \cap \overline{\Omega},$$

and

$$B(x + t\xi_j(x), rt) \subset \overline{\Omega} \text{ for all } x \in \overline{\Omega} \cap B(z_j, r) \text{ and } 0 \le t \le r.$$

By a standard argument, we find that there are $\zeta_j \in C_0^\infty(\mathbf{R}^n)$, with $j = 1, \dots, l$, such that $\zeta_j \ge 0$ and $\operatorname{supp} \zeta_j \subset \operatorname{Int} B(z_j, r)$ for all j and $\sum_{j=1}^l \zeta_j = 1$ on $\partial\Omega$. Using Lemma 6.1 and setting

$$\xi(x) = \sum_{i=1}^m \sum_{j=1}^l \zeta_j(x) \gamma_{ij} g(x, a_i),$$

we see that $\xi(x) \in \operatorname{co} G(x)$ for all $x \in \partial\Omega$ and that there is $s \in (0, r]$ such that (16) holds with s in place of r. □

Proof of Lemma 3.2. Lemma 3.2 follows immediately from Lemma 6.2. □

LEMMA 6.3. *Let $\xi \in C(\mathbf{R}^n, \mathbf{R}^n)$ and $r > 0$. Assume that* (16) *holds with these ξ and r. Then, for all $z \in \partial\Omega$, $x \in B(z, r)$ and $0 \le t \le r$,*

$$B(x + t\xi(x), rt) \subset \overline{\Omega} \quad \textit{if } x \in \overline{\Omega},$$

and

$$B(x - t\xi(x), rt) \subset \Omega^c \quad \textit{if } x \in \Omega^c. \tag{17}$$

Proof. Let $z \in \partial\Omega$, $x \in B(z, r)$, and $0 < t \le r$. First we suppose $x \in \overline{\Omega}$ and prove that

$$B(x + t\xi(x), rt) \cap B(z, r) \subset \overline{\Omega}. \tag{18}$$

Suppose that this is not true, and choose $p \in B(0,r)$ so that

$$x + t\xi(x) + tp \in B(z,r) \cap (\overline{\Omega})^c.$$

Then there is $\tau \in (0,t)$ such that $x+\tau\xi(x)+\tau p \in \partial\Omega$. It is obvious that $x+\tau\xi(x)+\tau p \in B(z,r)$. Hence, noting that

$$x + t\xi(x) + tp = (x + \tau\xi(x) + \tau p) + (t-\tau)\xi(x) + (t-\tau)p,$$

and using (16) we see that $x + t\xi(x) + tp \in \overline{\Omega}$. This is a contradiction, which implies that (18) holds.

Next we show that

$$B(x - t\xi(x), rt) \cap B(z,r) \subset \Omega^c \quad \text{if } x \in \Omega^c. \tag{19}$$

It is enough to prove (19) for the case when $x \in (\overline{\Omega})^c$. We thus assume that $x \in (\overline{\Omega})^c$. Again, we suppose that (19) is not true, and get a contradiction. Now, we have

$$x - t\xi(x) + tp \in B(z,r) \cap \Omega$$

for some $p \in B(0,r)$, and moreover,

$$x - \tau\xi(x) + \tau p \in \partial\Omega$$

for some $0 < \tau < t$. Since $x - \tau\xi(x) + \tau p \in B(z,r)$, noting that

$$x = (x - \tau\xi(x) + \tau p) + \tau\xi(x) + \tau(-p),$$

we see that $x \in \overline{\Omega}$, which is a contradiction. □

Proof of Lemma 3.3. Let $\xi_0 \in C^{0,1}(\mathbf{R}^n, \mathbf{R}^n)$ be from Lemma 3.2. By Lemma 6.3, there is $r > 0$ such that for all $z \in \partial\Omega$, $x \in B(z,r)$, and $0 \le t \le r$,

$$B(x + t\xi_0(x), 2rt) \subset \overline{\Omega} \quad \text{if } x \in \overline{\Omega}, \tag{20}$$

and

$$B(x - t\xi(x), 2rt) \subset \Omega^c \quad \text{if } x \in \Omega^c. \tag{21}$$

The following arguments are based on the idea of value functions in optimal control or the method of characteristics.

For $s > 0$ we write Γ_s for the open s-neighborhood of $\partial\Omega$, i.e.,

$$\Gamma_s = \{x \in \mathbf{R}^n \mid \operatorname{dist}(x, \partial\Omega) < s\}.$$

We may assume that $|\xi_0(x)| \le 1$ for $x \in \mathbf{R}^n$. We fix $\varepsilon > 0$ so that $\varepsilon(1/r + 1) \le r$ and so that if $x \in \Gamma_r$ and $y \in B(x, \varepsilon/r)$, then $|\xi_0(y) - \xi_0(x)| \le r$.

We write $Y(t;x)$ for $Y(t;x,-\xi_0)$ for notational simplicity. Now, we claim that

$$Y(t;x) \in (\Gamma_\varepsilon)^c \quad \text{for all } x \in \Gamma_\varepsilon \text{ and } t \ge 2\varepsilon/r. \tag{22}$$

To show this, we will prove that

$$x \in \Omega \cap \Gamma_\varepsilon \Longrightarrow \exists t \in (0, \varepsilon/r] \text{ such that } Y(t;x) \in \Omega^c, \tag{23}$$

$$x \in \Omega^c \cap \Gamma_\varepsilon \Longrightarrow \exists t \in (0, \varepsilon/r] \text{ such that } Y(t;x) \in (\Gamma_\varepsilon)^c \cap \Omega^c, \tag{24}$$

and

$$x \in \Omega^c \cap (\Gamma_\varepsilon)^c \Longrightarrow Y(t;x) \in (\Gamma_\varepsilon)^c \ \forall t > 0. \tag{25}$$

Once we have proved (23)–(25), we can easily conclude that (22) holds.

To prove (23), fix $x \in \Omega \cap \Gamma_\varepsilon$ and suppose that $Y(t;x) \in \Omega$ for all $t \in (0, \varepsilon/r]$. Fix $z \in \partial\Omega$ so that $|x - z| < \varepsilon$. Then, for all $t, \tau \in [0, \varepsilon/r]$,

$$|Y(t;x) - Y(\tau;x)| \leq \int_0^{\varepsilon/r} |\xi_0(Y(s;x))| ds \leq \varepsilon/r,$$

and hence,

$$|Y(t;x) - z| < \varepsilon/r + \varepsilon \leq r \quad \text{and} \quad |\xi_0(Y(t;x)) - \xi_0(Y(\tau,x))| \leq r.$$

Thus, setting $y = Y(\varepsilon/r;x)$, we have

$$\begin{aligned} x &= y + (\varepsilon/r)\xi_0(y) + \int_0^{\varepsilon/r} (\xi_0(Y(s;x)) - \xi_0(y)) ds \\ &\in B(y + (\varepsilon/r)\xi_0(y), \varepsilon). \end{aligned}$$

Hence, in view of (20) we see that $B(x,\varepsilon) \subset \overline{\Omega}$ and therefore that $\operatorname{dist}(x, \partial\Omega) \geq \varepsilon$. This is a contradiction, which proves (23).

To prove (24), let $x \in \Omega^c \cap \Gamma_\varepsilon$. As above, we find that for some $z \in \partial\Omega$, we have $x \in B(z,\varepsilon)$ and also $Y(t;x) \in B(z,r)$ and $|\xi_0(Y(t;x)) - \xi_0(x)| \leq r$ for all $0 \leq t \leq \varepsilon/r$. We have

$$\begin{aligned} Y(\varepsilon/r;x) &= x - (\varepsilon/r)\xi_0(x) - \int_0^{\varepsilon/r} (\xi_0(Y(s;x)) - \xi_0(x)) ds \\ &\in B(x - (\varepsilon/r)\xi_0(x), \varepsilon), \end{aligned}$$

and hence $B(Y(\varepsilon/r;x),\varepsilon) \subset \Omega^c$ by (21). Therefore we have $Y(\varepsilon/r;x) \in \Omega^c \cap (\Gamma_\varepsilon)^c$.

To prove (25), we let $x \in \Omega^c \cap (\Gamma_\varepsilon)^c$. Suppose that $Y(\tau;x) \in \Gamma_\varepsilon$ for some $\tau > 0$. We may assume that $Y(t;x) \in \Omega^c$ for all $0 \leq t \leq \tau$. Then there is $\sigma \in (0,\tau)$ such that $Y(\sigma;x) \in \partial(\Gamma_\varepsilon)$ and $Y(t;x) \in \Gamma_\varepsilon$ for $\sigma < \forall t < \tau$. We may assume that $\tau \leq \sigma + \varepsilon/r$. Then, for all $t \in [\sigma,\tau]$, $|Y(t;x) - Y(\sigma;x)| \leq \varepsilon/r$ and hence, $|\xi_0(Y(t;x)) - \xi_0(Y(\sigma;x) + p)| \leq r$ for all $p \in B(0,\varepsilon)$. Therefore, noting that for any $p \in \mathbf{R}^n$,

$$\begin{aligned} Y(\tau;x) + p = Y(\sigma;x) + p &- (\tau - \sigma)\xi_0(Y(\sigma;x) + p) \\ &+ \int_\sigma^\tau \{\xi_0(Y(\sigma;x) + p) - \xi_0(Y(t;x))\} dt, \end{aligned}$$

we find that for all $p \in B(0,\varepsilon)$,

$$Y(\tau;x) + p \in B(Y(\sigma;x) + p - (\tau - \sigma)\xi_0(Y(\sigma;x) + p), r(\tau - \sigma)).$$

Since $B(Y(\sigma;x),\varepsilon) \subset \Omega^c \cap \Gamma_r$, in view of (21) we have

$$B(Y(\tau;x) + p, r(\tau - \sigma)) \subset \Omega^c \quad \text{for all } p \in B(0,\varepsilon),$$

and, in particular, $B(Y(\tau;x),\varepsilon) \subset \Omega^c$. That is, $Y(\tau;x) \in (\Gamma_\varepsilon)^c$. This is a contradiction, from which (25) follows.

Now we choose $h \in C^\infty(\mathbf{R}^n)$ so that $h \geq 0$ on $\mathbf{R}^n$, $\operatorname{supp} h \subset \Gamma_\varepsilon$, and $h(x) \geq 2$ for $x \in \Gamma_{\varepsilon/2}$. Define $v : \Gamma_\varepsilon \to \mathbf{R}$ by

$$v(x) = \int_0^\infty h(Y(t;x))dt.$$

By virtue of (22) we see that

$$v(x) = \int_0^{2\varepsilon/r} h(Y(t;x))dt.$$

It is now easy to check that $v \in C^{0,1}(\Gamma_\varepsilon)$ and that v satisfies

$$\langle \xi_0(x), Dv \rangle = h(x) \quad \text{in } \Gamma_\varepsilon$$

in the viscosity sense, and as a result

$$\langle \xi_0(x), Dv \rangle = h(x) \quad \text{a.e. in } \Gamma_\varepsilon.$$

By mollifying both sides of this, we conclude that there is a function $\psi \in C^\infty(\mathbf{R}^n)$ such that $\langle \xi_0(x), D\psi(x) \rangle \geq 1$ on $\Gamma_{\varepsilon/2}$. □

We shall work in $\mathbf{R}^2$ for a while. Let e_1 denote the unit vector $(1,0) \in \mathbf{R}^2$. Fix $0 < \delta < \rho < 1$, and set

$$K = \bigcup_{t \geq 0} B(te_1, \delta t) \quad \text{and} \quad L = \bigcup_{t \geq 0} B(te_1, \rho t). \tag{26}$$

LEMMA 6.4. *There is a function $v \in C(\mathbf{R}^2) \cap C^{1,1}(\mathbf{R}^2 \setminus \{0\})$ such that v is convex and symmetric with respect to the x_1-axis,*

$$v(tx) = tv(x) \quad \textit{for } x \in \mathbf{R}^2 \textit{ and } t \geq 0,$$
$$v(x) > 0 \quad \textit{if } x \neq 0,$$

and

$$\langle q, Dv(x) \rangle \leq 0 \quad \textit{for all } q \in K \textit{ and } x \in L^c. \tag{27}$$

Proof. Set

$$\varepsilon = (\rho - \delta)\delta, \quad K_\varepsilon = \bigcup_{0 \leq t \leq \delta} B(te_1, \delta t + \varepsilon),$$

and

$$v(x) = \inf\{t > 0 \mid x \in tK_\varepsilon\} \quad \text{for } x \in \mathbf{R}^2.$$

We observe that K_ε contains the origin as its interior point; is convex, bounded, and symmetric with respect to the x_1-axis; and is the closed ε-neighborhood of the convex set

$$\bigcup_{0 \leq t \leq \delta} B(te_1, \delta t).$$

Then we have that v is convex and symmetric with respect to the axis $x_2 = 0$,

$$v(x) > 0 \text{ if } x \neq 0, \quad v(tx) = tv(x) \text{ for } x \in \mathbf{R}^2 \text{ and } t \geq 0,$$
$$v \in C(\mathbf{R}^2) \cap C^{1,1}(\mathbf{R}^2 \setminus \{0\}),$$
$$v(x) = 1 \iff x \in \partial(K_\varepsilon),$$
$$v(x) \leq 1 \iff x \in K_\varepsilon.$$

Next we must verify that (27) holds. Fix $x \in L^c$. We may assume without loss of generality that $x \in \partial(K_\varepsilon)$. Then $x = te_1 + p$ for some $0 \leq t \leq \delta$ and $p \in \partial B(0, \delta t + \varepsilon)$. Suppose for the moment that $t = \delta$. Then, $|p| \leq \delta^2 + \varepsilon = \rho\delta = \rho t$, and this implies that $x \in L$. This means that $t < \delta$. Therefore, there is $\eta > 0$ such that

$$x + K \cap B(0, \eta) \subset K_\varepsilon.$$

Fix any $q \in K$. If $s > 0$ is small enough, then we have

$$x + sq \in K_\varepsilon \text{ and hence } v(x + sq) \leq 1 = v(x).$$

Thus, differentiating $v(x + sq)$ with respect to s, we conclude that $\langle q, Dv(x) \rangle \leq 0$. □

Proof of Lemma 3.4. Let $\xi_0 \in C^{0,1}(\mathbf{R}^n, \mathbf{R}^n)$ and $0 < r < 1$ satisfy

$$B(x - t\xi_0(x), rt) \subset \Omega^c \text{ for all } x \in \partial\Omega \text{ and } 0 \leq t \leq r.$$

We shall build a function $w \in C^1(\overline{\Omega} \times \overline{\Omega})$ which satisfies the conditions of Lemma 3.4 concerning w with this ξ_0. To this end we may assume that $|\xi_0(x)| = 1$ for $x \in \Gamma_r \equiv \{y \mid \operatorname{dist}(y, \partial\Omega) < r\}$. Moreover, replacing r by a smaller number if necessary, we may assume that for some $\varepsilon > 0$,

$$B(x - t\xi_0(x), (r + \varepsilon)t) \subset \Omega^c \text{ for all } x \in \partial\Omega \text{ and } 0 \leq t \leq r.$$

We set $\rho = r/\sqrt{1 - r^2}$ and $\delta = \rho/2$, so that $r = \rho/\sqrt{1 + \rho^2}$ and $0 < \delta < \rho < 1$. In the above inclusion, we may assume that $\varepsilon/(1 - \varepsilon) \leq \delta/\sqrt{1 + \delta^2}$. Let K and L be the sets defined by (26) with the above δ and ρ. Observe that

$$K = \left\{(q_1, q_2) \mid |q_2| \leq \frac{\delta}{\sqrt{1 - \delta^2}} q_1\right\} \text{ and } L = \left\{(z_1, z_2) \mid |z_2| \leq \frac{\rho}{\sqrt{1 - \rho^2}} z_1\right\}.$$

By approximating ξ_0, we can choose a function $\xi \in C^\infty(\mathbf{R}^n, \mathbf{R}^n)$ and $s \in (0, r]$ so that $|\xi(y) - \xi_0(x)| \leq \varepsilon$ for all $x \in \partial\Omega$ and $y \in B(x, s)$ and $|\xi(x)| = 1$ for all $x \in \Gamma_r$. It follows that

$$B(x - t\xi(y), rt) \subset \Omega^c \text{ for all } x \in \partial\Omega, \ y \in B(x, s) \text{ and } 0 \leq t \leq r. \tag{28}$$

Let $v \in C(\mathbf{R}^2) \cap C^{1,1}(\mathbf{R}^2 \setminus \{0\})$ be from Lemma 6.4. We define $w \in C(\mathbf{R}^{2n})$ by

$$w(x, y) = v(\langle x - y, \xi(y) \rangle, |x - y - \langle x - y, \xi(y) \rangle \xi(y)|)^2.$$

It is convenient to introduce the following notation:

$$Q(y) = I - \xi(y) \otimes \xi(y) \equiv \Big(\delta_{ij} - \xi_i(y)\xi_j(y)\Big)_{1 \leq i,j \leq n}.$$

Note that

$$w(x,y) = v(\langle x-y, \xi(y)\rangle, |Q(y)(x-y)|)^2,$$

that if $|\xi(y)| = 1$ then $Q(y) : \mathbf{R}^n \to \mathbf{R}^n$ is the orthogonal projection onto the orthogonal complement to the space generated by $\xi(y)$, and that

$$x - y = \langle x-y, \xi(y)\rangle \xi(y) + Q(y)(x-y).$$

Noting that the function $v(z_1, z_2)$ is symmetric in the variable z_2, we see that $w \in C^1(\mathbf{R}^{2n})$.

For given $x, y \in \mathbf{R}^n$ we write

$$z_1 = \langle x-y, \xi(y)\rangle, \quad z_2 = |Q(y)(x-y)|, \quad \text{and} \quad z = (z_1, z_2).$$

Differentiation gives

$$D_x w(x,y) = 2v(z)\left\{ D_1 v(z)\xi(y) + D_2 v(z)\frac{Q(y)(x-y)}{|Q(y)(x-y)|}\right\},$$

and

$$D_y w(x,y) = 2v(z)\Big\{ D_1 v(z)(-\xi(y) + R_1(x,y)) + D_2 v(z)(-I + R_2(x,y))\frac{Q(y)(x-y)}{|Q(y)(x-y)|}\Big\},$$

provided $Q(y)(x-y) \neq 0$, where $R_1(x,y)$ and $R_2(x,y)$ are C^∞ functions on $\mathbf{R}^{2n}$ with values in $\mathbf{R}^n$ and in the space of real $n \times n$ matrices, respectively, and moreover, the estimate

$$|R_1(x,y)| + \|R_2(x,y)\| \leq C|x-y|$$

holds for all $x, y \in \mathbf{R}^n$ and for some $C > 0$. Here and later $D_1 v$ and $D_2 v$ denote the first and the second components of Dv, respectively. If, on the other hand, $Q(y)(x-y) = 0$, then we have

$$D_x w(x,y) = 2v(z)D_1 v(z)\xi(y), \quad D_y w(x,y) = 2v(z)D_1 v(z)(-\xi(y) + R_1(x,y)).$$

Now, we fix $x \in \partial\Omega$ and $y \in \overline{\Omega}$ so that $|x-y| \leq s$. We will show that $\langle \xi_0(x), D_x w(x,y)\rangle \leq 0$. If $x - y = 0$, then, trivially, the inequality holds. We shall assume that $x - y \neq 0$. We consider the case when $z_2 = |Q(y)(x-y)| = 0$. Observe that if $z_1 = \langle x-y, \xi(y)\rangle > 0$, then by (28) we have

$$B(x - z_1\xi(y), rz_1) = B(y, rz_1) \subset \Omega^c, \quad \text{which is a contradiction.}$$

Therefore, we have $z_1 \leq 0$ and hence $D_1 v(z) \leq 0$. We observe that

$$1 \geq \langle \xi_0(x), \xi(y)\rangle = 1 + \langle \xi_0(x), \xi(y) - \xi_0(x)\rangle \geq 1 - \varepsilon > 0.$$

Now, we see that $\langle \xi_0(x),\ D_x w(x,y)\rangle \leq 0$.

Next we consider the case when $z_2 = |Q(y)(x-y)| \neq 0$. We set

$$q_1 = \langle \xi_0(x), \xi(y)\rangle, \quad q_2 = \left\langle \xi_0(x), \frac{Q(y)(x-y)}{|Q(y)(x-y)|}\right\rangle \quad \text{and} \quad q = (q_1, q_2),$$

and note that

$$\langle \xi_0(x), D_x w(x,y)\rangle = 2v(z)\big(q_1 D_1 v(z) + q_2 D_2 v(z)\big) = 2v(z)\langle q, Dv(z)\rangle.$$

We now check that $q \in K$. To do this, we recall that $1-\varepsilon \le q_1 \le 1$, and observe that

$$\frac{|q_2|}{q_1} \le \frac{1}{1-\varepsilon}\left|\left\langle \xi_0(x) - \xi(y), \frac{Q(y)(x-y)}{|Q(y)(x-y)|}\right\rangle\right| \le \frac{\varepsilon}{1-\varepsilon}.$$

Thus, $|q_2| \le (\delta/\sqrt{1-\delta^2})q_1$ and hence $q \in K$. Next, we want to show that z is in the closure of L^c. If $z_1 \le 0$, we have immediately that $z \in L^c$. We thus assume that $z_1 > 0$. Observing that if $z_2/z_1 < r$, then

$$y = x - z_1\xi(y) + (Q(y)(y-x)) \in \operatorname{Int} B(x - z_1\xi(y), rz_1) \subset \Omega^c$$

by (28), we see that $z_2/z_1 \ge r = \rho/\sqrt{1+\rho^2}$ and hence that z is in the closure of L^c. Now we conclude in view of (27) that $\langle \xi_0(x), D_x w(x,y)\rangle \le 0$.

Multiplying w by an appropriate positive constant, we may assume that $w(x,y) \ge |x-y|^2$ for all $x, y \in \mathbf{R}^n$. It is easy to check that w satisfies the other requirements. □

Note Added in Proof. The authors recently learned that results similar to Theorem 4.1 are obtained in [8, Thm. X.2].

REFERENCES

[1] G. Barles and P.-L. Lions, *Fully nonlinear Neumann type boundary conditions for first-order Hamilton-Jacobi equations*, Nonlinear Anal., 16 (1991), pp. 143–153.

[2] I. Capuzzo Dolcetta and P.-L. Lions, *Hamilton-Jacobi equations with state constraint*, Trans. Amer. Math. Soc., 318 (1990), pp. 643–683.

[3] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[4] P. Dupuis and H. Ishii, *On oblique derivative problems for fully nonlinear second-order elliptic partial differential equations on nonsmooth domains*, Nonlinear Anal., 15 (1990), pp. 1123–1138.

[5] ———, *On oblique derivative problems for fully nonlinear second-order elliptic PDE's on domains with corners*, Hokkaido Math. J., 20 (1991), pp. 135–164.

[6] H. Ishii, *A boundary value problem of the Dirichlet type for Hamilton-Jacobi equations*, Ann. Scoula Norm. Sup. Pisa Cl. Sci. (4), 16 (1989), pp. 14–45.

[7] ———, *Fully nonlinear oblique derivative problems for nonlinear second-order elliptic PDE's*, Duke Math. J., 62 (1991), pp. 633–661.

[8] P.-L. Lions, *Neumann type boundary conditions for Hamilton-Jacobi equations*, Duke Math. J., 52 (1985), pp. 793–820.

[9] P. Loreti and M. E. Tessitore, *Approximation and regularity results on constrained viscosity solutions of Hamilton-Jacobi-Bellman equations*, J. Math. Systems Estim. Control, 4 (1994), pp. 467–483.

[10] M. Soner, *Optimal control with state-space constraint I*, SIAM J. Control Optim., 24 (1986), pp. 552–562.

[11] D. Tataru, *Boundary value problems for first order Hamilton-Jacobi equations*, Nonlinear Anal., 19 (1992), pp. 1091–1110.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 572–600, March 1996

010

CONDITIONS FOR ROBUSTNESS AND NONROBUSTNESS OF THE STABILITY OF FEEDBACK SYSTEMS WITH RESPECT TO SMALL DELAYS IN THE FEEDBACK LOOP*

HARTMUT LOGEMANN†, RICHARD REBARBER‡, AND GEORGE WEISS§

Abstract. It has been observed that for many stable feedback control systems, the introduction of arbitrarily small time delays into the loop causes instability. In this paper we present a systematic frequency domain treatment of this phenomenon for distributed parameter systems. We consider the class of all matrix-valued transfer functions which are bounded on some right half-plane and which have a limit at $+\infty$ along the real axis. Such transfer functions are called regular. Under the assumption that a regular transfer function is stabilized by unity output feedback, we give sufficient conditions for the robustness and for the nonrobustness of the stability with respect to small time delays in the loop. These conditions are given in terms of the high-frequency behavior of the open-loop system. Moreover, we discuss robustness of stability with respect to small delays for feedback systems with dynamic compensators. In particular, we show that if a plant with infinitely many poles in the closed right half-plane is stabilized by a controller, then the stability is not robust with respect to delays. We show that the instability created by small delays is itself robust to small delays. Three examples are given to illustrate these results.

Key words. small time delays, robust stabilization, linear distributed parameter systems, regular transfer functions, dynamic stabilization

AMS subject classifications. 93C20, 93C25, 93D09, 93D15, 93D25

1. The main results.

Consider the linear feedback system shown in Fig. 1, where u is the input function and y is the output function, both $\mathbb{C}^m$-valued. $\mathbf{H}$ is the open-loop transfer function, with values in $\mathbb{C}^{m\times m}$, which we assume to be regular and in particular well posed. *Wellposedness* means that $\mathbf{H}$ is bounded on some right half-plane, and *regularity* means that, in addition, $\mathbf{H}$ has a limit at $+\infty$ along the real axis (see §2 for more detail on these concepts). The block with transfer function $e^{-\varepsilon s}$ represents a delay by ε, where $\varepsilon \geq 0$. The transfer function of the closed-loop system is given by

$$\mathbf{G}^\varepsilon(s) \;=\; \mathbf{H}(s)\left(I + e^{-\varepsilon s}\mathbf{H}(s)\right)^{-1}. \tag{1.1}$$

$\mathbf{G}^\varepsilon$ can be obtained from $\mathbf{G}^0$ by

$$\mathbf{G}^\varepsilon(s) \;=\; \mathbf{G}^0(s)\left[I - (1 - e^{-\varepsilon s})\mathbf{G}^0(s)\right]^{-1}. \tag{1.2}$$

To avoid possible complications with domains of transfer functions, we make the following convention: If a meromorphic function is defined on some right half-plane and can be extended meromorphically to a greater right half-plane, we will not make any distinction between the initial function and its extension. This will not lead to confusions.

* Received by the editors June 23, 1993; accepted for publication (in revised form) December 7, 1994.

† School of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom (hl@maths.bath.ac.uk).

‡ Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, NE 68588-0323 (rrebarbe@math.unl.edu). This research was supported in part by National Science Foundation grant DMS-9206986.

§ Department of Electrical Engineering, Ben-Gurion University, 84105 Beer Sheva, Israel (weiss@bguvm.bgv.ac.il).

We say that $\mathbf{G}^\varepsilon$ is *L^2-stable* if $\mathbf{G}^\varepsilon \in H^\infty(\mathbb{C}^{m\times m})$; i.e., $\mathbf{G}^\varepsilon$ is a bounded analytic function on the open right half-plane $\mathbb{C}_0 = \{s \in \mathbb{C} \mid \operatorname{Re} s > 0\}$. Indeed, as is well known, this property is equivalent to the one that $u \in L^2([0,\infty),\mathbb{C}^m)$ implies $y \in L^2([0,\infty),\mathbb{C}^m)$.

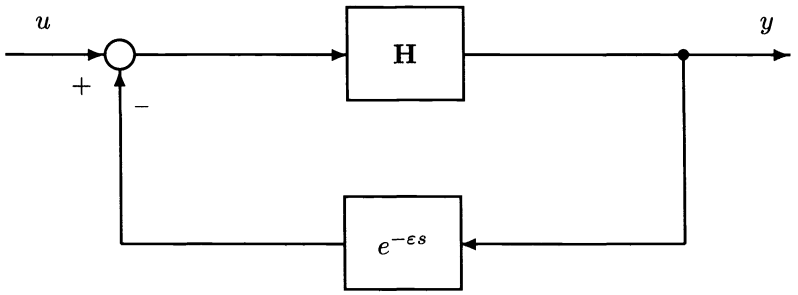


FIG. 1. *Feedback system with delay.*

In many engineering applications the aim is to stabilize a plant by a feedback controller. Here, stability may have various meanings—for example, exponential stability in the state space. We may think of $\mathbf{H}$ as the transfer function of the plant and the controller connected in cascade, and the stability of the corresponding closed-loop system implies that $\mathbf{G}^0$ is L^2-stable. However, stability might be lost if tiny (and often inevitable) delays are present in the feedback loop, leading to the feedback system shown in Fig. 1. Indeed, it might be that for arbitrarily small $\varepsilon > 0$, $\mathbf{G}^\varepsilon$ has poles in $\mathbb{C}_0$, which implies that the system cannot be stable in any reasonable state space sense either. Our aim in this paper is to find conditions on $\mathbf{H}$ (necessary and/or sufficient) for this phenomenon (observed by many authors) to happen.

We say that $\mathbf{G}^0$ is *robustly stable with respect to delays* if there is an $\varepsilon_0 > 0$ such that for any $\varepsilon \in [0,\varepsilon_0]$, $\mathbf{G}^\varepsilon$ is L^2-stable. The absence of this property means that arbitrarily small destabilizing delays can be found for $\mathbf{G}^0$.

If the transfer function $\mathbf{H}$ is meromorphic on the half-plane $\mathbb{C}_0$, then we denote by $\mathfrak{P}_\mathbf{H}$ the (discrete) set of its poles in $\mathbb{C}_0$. (We say that p is a pole of $\mathbf{H}$ if p is a pole of at least one entry of $\mathbf{H}$.) We define

$$\gamma = \limsup_{\substack{|s|\to\infty \\ s\in\mathbb{C}_0\setminus\mathfrak{P}_\mathbf{H}}} r(\mathbf{H}(s)), \tag{1.3}$$

where $r(\mathbf{H}(s))$ denotes the spectral radius of the matrix $\mathbf{H}(s)$. It might happen that $\gamma = \infty$—for example, if $\mathbf{H}$ is scalar and has an unbounded sequence of poles on the imaginary axis. If $\mathbf{G}^0$ is L^2-stable, then from the formula $\mathbf{H} = \mathbf{G}^0(I - \mathbf{G}^0)^{-1}$ it is not difficult to see that $\mathbf{H}$ is meromorphic on $\mathbb{C}_0$. Indeed, for s in the half-plane where $\mathbf{H}(s)$ was originally defined, $I - \mathbf{G}^0(s) = (I + \mathbf{H}(s))^{-1}$ so that $\det(I - \mathbf{G}^0(s))$ is not identically 0. Hence, if $\mathbf{G}^0$ is L^2-stable, then (1.3) makes sense. This fact is used implicitly in the statement of our main result, which is the following theorem.

THEOREM 1.1. *Let $\mathbf{H}$ be a $\mathbb{C}^{m\times m}$-valued regular transfer function and suppose that $\mathbf{G}^0 = \mathbf{H}(I + \mathbf{H})^{-1}$ is L^2-stable. Let γ be defined by* (1.3).

(i) *If $\gamma < 1$, then $\mathbf{G}^0$ is robustly stable with respect to delays.*

(ii) *If $\gamma > 1$, then $\mathbf{G}^0$ is not robustly stable with respect to delays.*

The proof of (i) is much easier than the proof of (ii) and is in §6. It is shown in the same section that (i) cannot be extended to multidelay perturbations. The proof of (ii) is very involved, so in order to present the ideas clearly, without multivariable

technicalities, we first give the proof for $m = 1$ in §§3 and 4. The multivariable case is treated in §5. We were not able to prove a general result for the case $\gamma = 1$. However, trivial examples (e.g., $\mathbf{H}(s) \equiv I$) show that $\mathbf{G}^0$ will in general not be robustly stable with respect to delays if $\gamma = 1$.

In §7 we show that the instability created by a small delay in the closed loop is itself robust to small delays.

In §8 we discuss destabilization and robustness with respect to delays for systems with dynamic feedback. Let $\mathbf{P}$ and $\mathbf{K}$ be meromorphic functions on $\mathbb{C}_0$ of appropriate dimensions such that the products $\mathbf{PK}$ and $\mathbf{KP}$ exist. We say that $\mathbf{K}$ *stabilizes* $\mathbf{P}$ if

$$\begin{bmatrix} I & \mathbf{P} \\ -\mathbf{K} & I \end{bmatrix}^{-1}$$

is L^2-stable. Intuitively, this means that if we connect the plant $\mathbf{P}$ and the controller $\mathbf{K}$ in a feedback loop with two external inputs, then all the possible transfer functions in the loop are L^2-stable (see §8 for details).

Let us denote $\mathbf{K}_\varepsilon(s) = e^{-\varepsilon s}\mathbf{K}(s)$. We say that $\mathbf{K}$ *stabilizes* $\mathbf{P}$ *robustly with respect to delays* if there is an $\varepsilon_0 > 0$ such that for any $\varepsilon \in [0, \varepsilon_0]$ $\mathbf{K}_\varepsilon$ stabilizes $\mathbf{P}$. Intuitively, this means that the introduction of sufficiently small delays into the feedback loop mentioned earlier does not destroy its stability. By a corollary of Theorem 1.1 stated in §8, if $\mathbf{H} = \mathbf{PK}$ is regular and γ is defined by (1.3), then $\gamma < 1$ implies that $\mathbf{K}$ stabilizes $\mathbf{P}$ robustly, while $\gamma > 1$ implies that the opposite is true. Let $\mathbb{C}_0^{\mathrm{cl}}$ denote the closure of $\mathbb{C}_0$. Using the above mentioned corollary and a lemma of independent interest, we prove (still in §8) the following theorem.

THEOREM 1.2. *Let* $\mathbf{P}$ *and* $\mathbf{K}$ *be matrix-valued meromorphic functions on a right open half-plane containing* $\mathbb{C}_0^{\mathrm{cl}}$. *Assume that* $\mathbf{PK}$ *is regular and* $\mathbf{K}$ *stabilizes* $\mathbf{P}$. *If* $\mathbf{P}$ *has infinitely many poles in* $\mathbb{C}_0^{\mathrm{cl}}$, *then* $\mathbf{K}$ *does not stabilize* $\mathbf{P}$ *robustly with respect to delays.*

Thus, roughly speaking, if a plant with infinitely many poles in $\mathbb{C}_0^{\mathrm{cl}}$ is given, we cannot find a controller such that the resulting feedback system is robustly stable with respect to small delays in the loop.

In §9 we give three simple examples.

2. Preliminaries and discussion of earlier results. There are many examples in the literature of systems described by partial differential equations which are exponentially stabilized by a feedback but are destabilized by arbitrarily small time delays in the feedback loop. The first example of this sort appeared in Datko, Lagnese, and Polis [9], where a one-dimensional wave equation with boundary feedback was studied. The same phenomenon has been subsequently described in many other examples; see Datko [10], Desch, Hannsgen, Renardy, and Wheeler [13], Hannsgen, Renardy, and Wheeler [20], Bontsema and deVries [3] and Grimmer, Lenczewski, and Schappacher [19]. In a more abstract framework, this destabilization by small delays was demonstrated for classes of distributed parameter feedback systems in Datko [10], [11] and Desch and Wheeler [12]. In these classes of systems the open loop semigroup is unitary, and only one stabilizing feedback is considered for each given plant, typically a type of co-located control.

In contrast with the works referred to above, we will take a frequency domain approach here, which is not tied to a specific form for the stabilizing feedback. Our approach is similar in spirit to that in the considerably older paper of Barman, Callier, and Desoer [1]. In that paper, necessary and sufficient conditions were given for a

class of single-input single-output (SISO) systems to be robustly stable with respect to delays. The results in [1] are limited by several restrictions, including the requirement that the open-loop transfer function has at most finitely many poles in the closed right half-plane. These results were applied to some systems described by partial differential equations in [3].

A more general class of perturbations of feedback systems, including small delays in the loop, were considered by Georgiou and Smith [17], [18] (see also Curtain [8]). Their concept of w-stability is considerably stronger than robust stability with respect to delays; it covers a large class of perturbations which represent high-frequency modelling uncertainties. The necessary and sufficient criterion for w-stability in [18] resembles our Corollaries 8.2 and 8.4, especially in the SISO case. For multiple-input multiple-output (MIMO) systems, there is a curious difference: the result for w-stability is in terms of the norms of the transfer functions, while our result for robustness with respect to delays is in terms of their spectral radius. The proof of destabilization results for w-stability is considerably easier than for robustness with respect to delays.

We will now recall some concepts and results needed in this paper. We will work with finite-dimensional input and output spaces, but we mention that these concepts and results have natural counterparts for Hilbert space-valued input/output functions, which means operator-valued transfer functions.

Let $\alpha \in \mathbb{R}$. We will use the notation

$$\mathbb{C}_\alpha = \{ s \in \mathbb{C} \mid \operatorname{Re} s > \alpha \},$$

and $H_\alpha^\infty(\mathbb{C}^{p\times m})$ will denote the space of all bounded analytic $\mathbb{C}^{p\times m}$-valued functions on $\mathbb{C}_\alpha$. We write H^∞ for H_0^∞. The norm $\|\mathbf{G}\|_\infty$ of a function $\mathbf{G} \in H_\alpha^\infty$ is the supremum of $\|\mathbf{G}(s)\|$ over $\mathbb{C}_\alpha$, the matrix norm being defined as the greatest singular value. After the identification of functions with their meromorphic extensions, which was mentioned in §1, we have that

$$H_\alpha^\infty \subset H_\beta^\infty \quad \text{if} \quad \alpha \le \beta.$$

DEFINITION 2.1. *A* well-posed $\mathbb{C}^{p\times m}$-valued transfer function *is an element of one of the spaces* $H_\alpha^\infty(\mathbb{C}^{p\times m})$.

Well-posed transfer functions correspond to shift invariant operators on $L^2_{loc}[0,\infty)$ with finite growth bound; see Weiss [32, §3]. In particular, the transfer function of any abstract linear system is well posed, as follows from [32, Prop. 4.1]. Conversely, for any well-posed transfer function $\mathbf{H}$ we can find an abstract linear system whose transfer function is $\mathbf{H}$, as follows from results in Salamon [29].

DEFINITION 2.2. *A well-posed matrix-valued transfer function* $\mathbf{H}$ *is* regular *if the limit* $\lim_{\lambda\to+\infty} \mathbf{H}(\lambda) = D$ *exists (where* λ *is real). Then* D *is the* feedthrough matrix *of* $\mathbf{H}$.

Practically all well-posed transfer functions of interest are regular. (In fact, it is a nontrivial exercise in complex analysis to construct an example of a well-posed and nonregular transfer function.) If the transfer function of an abstract linear system is regular (such systems are called regular), then the system has a simple and convenient state space representation, like finite-dimensional systems; see [32, §2].

For SISO systems, we will use the term *feedthrough value* instead of feedthrough matrix. We introduce a notation for angular domains in $\mathbb{C}$: for any number $\psi \in (0,\pi]$,

$$\mathcal{W}(\psi) = \left\{ re^{i\phi} \mid r \in (0,\infty),\ \phi \in (-\psi,\psi) \right\}.$$

We will need the following simple fact about regular transfer functions.

PROPOSITION 2.3. *Let* $\mathbf{H}$ *be a regular matrix-valued transfer function, with feedthrough matrix* D. *Then for any* $\psi \in (0, \frac{\pi}{2})$,

$$\lim_{\substack{|s|\to\infty \\ s\in\mathcal{W}(\psi)}} \mathbf{H}(s) = D.$$

This follows from Duren [14, Thm. 1.3], after mapping the half-plane onto the unit disk. It follows also from the results in [32, §5], where Laplace transform techniques are used.

Remark 2.4. If $\mathbf{H}$ and $\mathbf{G}^\varepsilon$ are related as in (1.1) and $\varepsilon > 0$, then it is easy to see that $\mathbf{H}$ is well posed (regular) if and only if $\mathbf{G}^\varepsilon$ is well posed (regular). If $\mathbf{H}$ and $\mathbf{G}^0$ are both well posed, then one of them is regular if and only if the other is. Similar statements are true for operator-valued transfer functions but are more difficult to prove; see Weiss [33].

3. Nonrobustness: The SISO case with $|D| \leq 1$. In this section we prove (ii) of Theorem 1.1 for SISO systems and under the additional assumption that D, the feedthrough value of $\mathbf{H}$, satisfies $|D| \leq 1$. This situation is fairly typical for transfer functions of unstable vibrating systems. Since we can say slightly more than what is written in (ii) of Theorem 1.1, we restate the result.

THEOREM 3.1. *Let* $\mathbf{H}$ *be a regular SISO transfer function and, for any* $\varepsilon \geq 0$, *the function* $\mathbf{G}^\varepsilon$ *be defined by* (1.1). *Let* D *denote the feedthrough value of* $\mathbf{H}$. *Assume that*

(1) $\mathbf{G}^0 \in H^\infty$, *so that* γ *can be defined by* (1.3),
(2) $\gamma > 1$,
(3) $|D| \leq 1$.

Then there exist sequences (ε_n) *and* (p_n) *with*

$$\varepsilon_n > 0, \quad \varepsilon_n \to 0, \quad p_n \in \mathbb{C}_0, \quad |\mathrm{Im}\, p_n| \to \infty$$

and such that for any $n \in \mathbb{N}$, p_n *is a pole of* $\mathbf{G}^{\varepsilon_n}$.

Proof. From condition (1), using (1.1) we can see that the point -1 has a neighborhood which does not intersect the range of $\mathbf{H}$ (regarded as a meromorphic function on $\mathbb{C}_0$). This implies that there exist $\eta > 0$ and $\gamma_1 > 1$ such that the set

$$S_1 = \left\{ re^{i\phi} \;\middle|\; r \in [1, \gamma_1],\ |\phi - \pi| < \eta \right\}$$

does not intersect the range of $\mathbf{H}$:

$$\mathbf{H}(s) \notin S_1 \qquad \forall\, s \in \mathbb{C}_0. \tag{3.1}$$

Since we may choose γ_1 arbitrarily close to 1, by condition (2) we may assume

$$1 < \gamma_1 < \gamma. \tag{3.2}$$

The definition of γ and (3.2) enable us to show that there exists a sequence (z_n) in $\mathbb{C}_0$ with the following properties:

(a) $|z_n| \to \infty$ as $n \to \infty$;
(b) for any $n \in \mathbb{N}$,

$$|\mathbf{H}(z_n)| \geq \gamma_1;$$

(c) for any $n \in \mathbb{N}$, $\mathbf{H}$ is analytic on the ray

$$\Gamma_n = \{z_n + a \mid a \in [0, \infty)\}$$

(i.e., there are no poles on these rays).

By Proposition 2.3, for any $\psi \in (0, \frac{\pi}{2})$ there exists an $r_\psi > 0$ such that for any $s \in \mathcal{W}(\psi)$ with $|s| > r_\psi$ we have $|\mathbf{H}(s) - D| < \gamma_1 - 1$. Using that $|\mathbf{H}(s)| \le |D| + |\mathbf{H}(s) - D|$ and condition (3), we get that, for s as above, $|\mathbf{H}(s)| < \gamma_1$. By property (b) it follows that the sequence (z_n) lies in the set $\{s \in \mathbb{C}_0 \mid |s| \le r_\psi \text{ or } |\arg s| \ge \psi\}$. Here and in the rest of this proof, the argument of a (nonzero) complex number s is defined such that $\arg s \in (-\pi, \pi]$. By property (a), for n sufficiently large, $|z_n| \le r_\psi$ is not possible, so that $|\arg z_n| \ge \psi$. Since the choice of $\psi \in (0, \frac{\pi}{2})$ was arbitrary, we conclude that

$$\lim_{n \to \infty} |\arg z_n| = \frac{\pi}{2}. \tag{3.3}$$

Together with property (a) this implies that

$$\lim_{n \to \infty} |\operatorname{Im} z_n| = \infty. \tag{3.4}$$

We may assume without loss of generality that for all $n \in \mathbb{N}$, $\operatorname{Im} z_n > 0$. Indeed, if such a subsequence does not exist, then a similar argument can be made assuming that $\operatorname{Im} z_n < 0$ for all $n \in \mathbb{N}$.

By property (c), $\mathbf{H}$ is continuous on each ray Γ_n, and by Proposition 2.3, $\mathbf{H}(s) \to D$ as $s \to \infty$ on Γ_n. Since $|D| \le 1$ (by condition (3)), the numbers

$$a_n' = \max\{a \in [0, \infty) \mid |\mathbf{H}(z_n + a)| \ge \gamma_1\},$$
$$a_n'' = \min\{a \in [a_n', \infty] \mid |\mathbf{H}(z_n + a)| \le 1\}$$

are well defined. (If $|D| = 1$, then it might happen that $a_n'' = \infty$.) Put

$$z_n' = z_n + a_n', \qquad z_n'' = z_n + a_n''.$$

(It might happen that $z_n'' = \infty$.) We will be looking for poles of $\mathbf{G}^\varepsilon$ in the open horizontal segments $(z_n', z_n'') \subset \Gamma_n$.

From the definition of z_n' and z_n'', using property (c) we see that the image of $[z_n', z_n'']$ through $\mathbf{H}$ is a curve Π_n contained in $\{s \in \mathbb{C} \mid 1 \le |s| \le \gamma_1\}$. The possibility that $z_n'' = \infty$ is not disturbing since (by Proposition 2.3) $\mathbf{H}$ is continuous at infinity along the ray Γ_n. By (3.1) Π_n cannot enter S_1, so it is confined to the set

$$S_2 = \{re^{i\phi} \mid r \in [1, \gamma_1],\ |\phi| \le \pi - \eta\};$$

see Fig. 2. Thus we have

$$\Pi_n \subset S_2 \qquad \forall\, n \in \mathbb{N}. \tag{3.5}$$

Using (1.1) we see that p is a pole of $\mathbf{G}^\varepsilon$ if and only if

$$e^{-\varepsilon p}\mathbf{H}(p) = -1.$$

A sufficient condition for this is

$$\log \mathbf{H}(p) - \varepsilon p = -i\pi, \tag{3.6}$$

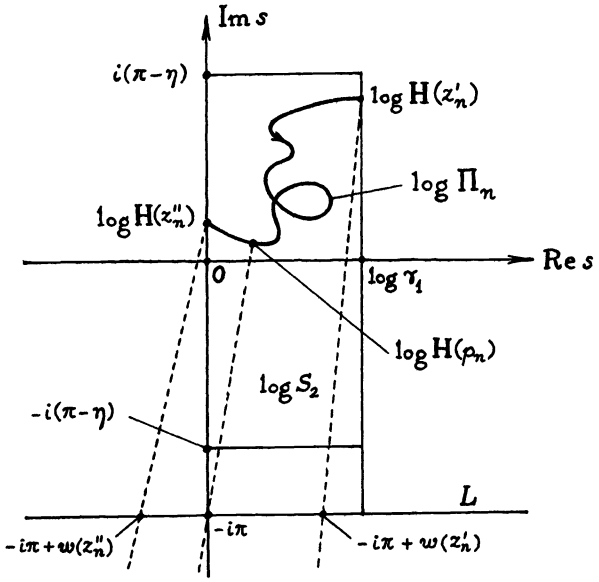


FIG. 2. *The sets S_1, S_2 and the curve Π_n.*

where we choose the branch of the logarithm to be $\log z = \log|z| + i \arg z$, with $\arg z \in (-\pi, \pi]$, as agreed earlier, and $\log|z| \in \mathbb{R}$.

For each $s \in \mathbb{C}_0$ with $\operatorname{Im} s > 0$, the ray $R(s) = \{\log \mathbf{H}(s) - \varepsilon s \mid \varepsilon \in [0, \infty)\}$ intersects the horizontal line $L = \{s \in \mathbb{C} \mid \operatorname{Im} s = -\pi\}$ in a point $w(s) - i\pi$. Indeed, for $\varepsilon = 0$ the corresponding point of $R(s)$ is above L, while for large $\varepsilon > 0$ the corresponding point of $R(s)$ is below L. Thus we can define the real-valued functions $w(s)$, $e(s)$ for each $s \in \mathbb{C}_0$ with $\operatorname{Im} s > 0$ such that $e(s) > 0$ and

$$\log \mathbf{H}(s) - e(s)s = w(s) - i\pi. \tag{3.7}$$

A simple computation shows that

$$e(s) = \frac{\arg \mathbf{H}(s) + \pi}{\operatorname{Im} s}, \tag{3.8}$$

$$w(s) = \log|\mathbf{H}(s)| - \frac{(\arg \mathbf{H}(s) + \pi)\operatorname{Re} s}{\operatorname{Im} s}. \tag{3.9}$$

CLAIM. *For all $n \in \mathbb{N}$ sufficiently large, there is a point $p_n \in (z'_n, z''_n)$ such that $w(p_n) = 0$.*

Figure 3 is intended to give an intuitive picture of this claim. In this figure we see the curve $\log \Pi_n$ which, according to (3.5), is contained in the rectangle $\log S_2$. It is assumed in the picture that z''_n is finite. The rays $R(z'_n)$, $R(z''_n)$, and $R(p_n)$ (dotted lines) and the horizontal line L are marked.

Proof of the claim. We define

$$w(\infty) = -\infty.$$

Then, as a $[-\infty, \infty)$-valued function, w is continuous on each segment $[z'_n, z''_n]$. Indeed, by (3.5) $\arg \mathbf{H}(s)$ has no jumps on such a segment. If z''_n is finite, then the continuity of w is clear from (3.9). If $z''_n = \infty$, then it is easy to see from (3.9) that $\lim_{s \to \infty} w(s) = -\infty$, where $s \in (z'_n, z''_n)$.

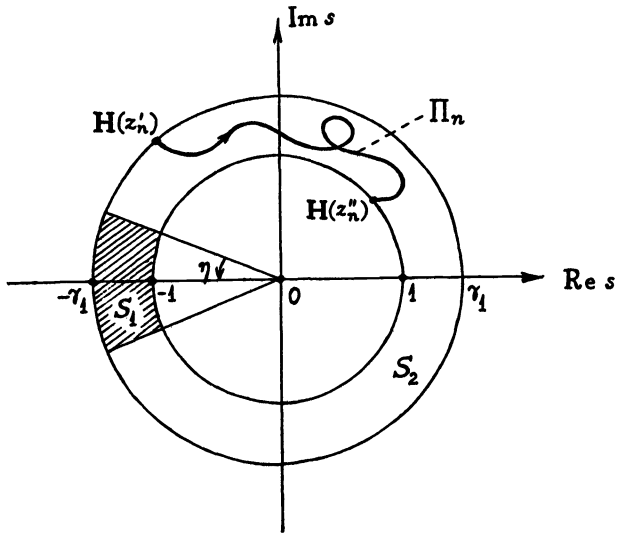


FIG. 3. *The claim (the existence of* p_n*).*

Next we show that for all $n \in \mathbb{N}$ sufficiently large,

$$w(z_n') > 0 \quad \text{and} \quad w(z_n'') < 0. \tag{3.10}$$

The sequence (z_n') shares with (z_n) properties (a) and (b) (also (c), but this is not needed now). By the exact same argument used to prove (3.3), and by the assumption $\operatorname{Im} z_n > 0$, we get that

$$\lim_{n\to\infty} \arg z_n' = \frac{\pi}{2}.$$

Hence

$$\lim_{n\to\infty} \frac{(\arg \mathbf{H}(z_n') + \pi)\operatorname{Re} z_n'}{\operatorname{Im} z_n'} = 0. \tag{3.11}$$

On the other hand, it is clear that

$$\frac{(\arg \mathbf{H}(z_n'') + \pi)\operatorname{Re} z_n''}{\operatorname{Im} z_n''} > 0. \tag{3.12}$$

By the definition of z_n' and z_n'', $|\mathbf{H}(z_n')| = \gamma_1 > 1$ and $|\mathbf{H}(z_n'')| = 1$. Therefore $\log|\mathbf{H}(z_n')| = \log\gamma_1 > 0$ and $\log|\mathbf{H}(z_n'')| = 0$. Combining this with (3.9), (3.11), and (3.12) we see that for all n sufficiently large, (3.10) holds.

Since $w(s)$ is continuous on $[z_n', z_n'']$, (3.10) implies that for all $n \in \mathbb{N}$ sufficiently large there exists $p_n \in (z_n', z_n'')$ such that $w(p_n) = 0$. (This is indicated in Fig. 3 by having the ray $R(p_n)$ go through $-i\pi$.) This completes the proof of the claim.

Returning to the main proof, we may now assume without loss of generality that for each $n \in \mathbb{N}$ there is a $p_n \in (z_n', z_n'')$ such that $w(p_n) = 0$. (If not, select an appropriate subsequence.) We denote $\varepsilon_n = e(p_n)$. By (3.7) (with $s = p_n$) we get that p_n and ε_n satisfy (3.6), so p_n is a pole of $\mathbf{G}^{\varepsilon_n}$. By (3.8), $\varepsilon_n > 0$. Since $\operatorname{Im} p_n = \operatorname{Im} z_n$, by (3.4) we have $|\operatorname{Im} p_n| \to \infty$. By (3.8) and (3.4) we have $\varepsilon_n \to 0$. □

4. Nonrobustness: The SISO case with $|D| > 1$. In this section we prove (ii) of Theorem 1.1 for SISO systems and under the additional assumption that D, the feedthrough value of $\mathbf{H}$, satisfies $|D| > 1$. Then obviously $\gamma > 1$, so that this does not have to be assumed. In fact, we will prove a stronger result, in which the assumption that $\mathbf{G}^0$ is L^2-stable is not needed. We do not even need that $\mathbf{H}$ should be meromorphic on $\mathbb{C}_0$.

THEOREM 4.1. *Let $\mathbf{H}$ be a regular SISO transfer function and, for any $\varepsilon \geq 0$, the function $\mathbf{G}^\varepsilon$ be defined by* (1.1). *Let D denote the feedthrough value of $\mathbf{H}$ and assume that*

$$|D| > 1.$$

Then there exist sequences (ε_n) and (p_n) with

$$\varepsilon_n > 0, \quad \varepsilon_n \to 0, \quad \operatorname{Re} p_n \to \infty, \quad \operatorname{Im} p_n \to \infty$$

and such that, for any $n \in \mathbb{N}$, p_n is a pole of $\mathbf{G}^{\varepsilon_n}$.

Proof. If D is not a negative real number, then we define the argument of any (nonzero) complex number s such that $\arg s \in (-\pi, \pi]$, as in the proof of Theorem 3.1. If D is negative then it is more convenient to change the definition such that $\arg s \in \left(-\frac{\pi}{2}, \frac{3\pi}{2}\right]$ (to avoid a jump discontinuity at D). The function log is defined by $\log s = \log|s| + i \arg s$ with $\log|s| \in \mathbb{R}$.

For any $r > 0$, B_r will denote the closed disk of radius r with center in D. Due to $|D| > 1$ and the way in which we have defined the function log, we can find a $\rho > 0$ such that (1) $|z| > 1$ for all $z \in B_\rho$ and (2) arg (and hence log) is continuous on B_ρ.

The simple inequalities

$$\min_{z \in B_\rho} (\arg z + \pi) > 0, \qquad \min_{z \in B_\rho} \log|z| > 0$$

enable us to find numbers $0 < \alpha < \beta$ such that for any $z \in B_\rho$

$$\log|z| - \frac{\arg z + \pi}{\alpha} < 0, \qquad \log|z| - \frac{\arg z + \pi}{\beta} > 0. \tag{4.1}$$

Let $\psi \in (0, \frac{\pi}{2})$ be such that $\beta < \operatorname{tg} \psi$. By Proposition 2.3, there exists a $\sigma > 0$ such that

$$\mathbf{H}(s) \in B_\rho \qquad \forall\, s \in \mathcal{W}(\psi) \cap \mathbb{C}_\sigma. \tag{4.2}$$

(The notation $\mathbb{C}_\sigma$ was introduced in §2.) Let (x_n) be a sequence of real numbers with $x_n > \sigma$ and such that $x_n \to \infty$. Define

$$z_n' = (1 + i\alpha)x_n, \qquad z_n'' = (1 + i\beta)x_n.$$

We will be looking for poles of $\mathbf{G}^\varepsilon$ in the open vertical segments (z_n', z_n'').

The remainder of this proof resembles the part of the proof of Theorem 3.1 which starts after (3.5), so we will be brief. Using (1.1) we see that p is a pole of $\mathbf{G}^\varepsilon$ if and only if $e^{-\varepsilon p}\mathbf{H}(p) = -1$. A sufficient condition for this is that (3.6) holds.

For each $s \in \mathbb{C}_0$ with $\operatorname{Im} s > 0$, the ray $R(s) = \{ \log \mathbf{H}(s) - \varepsilon s \mid \varepsilon \in [0, \infty) \}$ intersects the horizontal line $L = \{ s \in \mathbb{C} \mid \operatorname{Im} s = -\pi \}$ in a point $w(s) - i\pi$, as explained in

the previous proof. Thus we can define the functions $w(s)$, $e(s)$ for each $s \in \mathbb{C}_0$ with $\operatorname{Im} s > 0$ such that $e(s) > 0$ and (3.7) holds. These functions are given by (3.8) and (3.9). The following claim is almost identical to the one in the proof of Theorem 3.1.

CLAIM. *For all $n \in \mathbb{N}$ there exists $p_n \in (z_n', z_n'')$ such that $w(p_n) = 0$.*

The proof is simpler in this case: By (4.2) and property (2) it is clear that w is continuous on each segment $[z_n', z_n'']$. Moreover, (3.9), (4.1), and (4.2) imply (3.10), from which the claim follows.

We return to the proof of the theorem. We denote $\varepsilon_n = e(p_n)$. By (3.7) (with $s = p_n$) we get that p_n and ε_n satisfy (3.6), so p_n is a pole of $\mathbf{G}^{\varepsilon_n}$. By the definition of the function $e(\cdot)$, $\varepsilon_n > 0$. We have $\operatorname{Re} p_n = x_n$ so that $\operatorname{Re} p_n \to \infty$. Since $\operatorname{Im} p_n > \alpha x_n$, we also have $\operatorname{Im} p_n \to \infty$. Now by (3.8) we have $\varepsilon_n \to 0$. □

5. Nonrobustness: The MIMO case. In this section we show that the results in §§3 and 4 extend to the multivariable case. In particular, we prove part (ii) of Theorem 1.1 for $m > 1$. In order to do this, it is convenient to state some preliminary facts and results. If $V \subset U \subset \mathbb{C}$, we say that V is a *discrete* set in U if V has no accumulation points in U. Let $\mathcal{H}_\alpha$ denote the ring of holomorphic functions defined on $\mathbb{C}_\alpha$ and $\mathcal{M}_\alpha$ denote the field of meromorphic functions on $\mathbb{C}_\alpha$. The vector spaces of $\mathbb{C}^{p\times m}$-valued holomorphic functions and of $\mathbb{C}^{p\times m}$-valued meromorphic functions on $\mathbb{C}_\alpha$ will be denoted by $\mathcal{H}_\alpha(\mathbb{C}^{p\times m})$ and $\mathcal{M}_\alpha(\mathbb{C}^{p\times m})$, respectively. It is clear that $\mathcal{H}_\alpha(\mathbb{C}^{p\times m}) = {\mathcal{H}_\alpha}^{p\times m}$ and $\mathcal{M}_\alpha(\mathbb{C}^{p\times m}) = {\mathcal{M}_\alpha}^{p\times m}$. A complex number $s_0 \in \mathbb{C}_\alpha$ is a *pole* of $\mathbf{H} \in \mathcal{M}_\alpha(\mathbb{C}^{p\times m})$ if and only if s_0 is a pole of at least one of the entries of $\mathbf{H}$.

In the following let $\mathbf{H}$ be in $\mathcal{M}_\alpha(\mathbb{C}^{m\times m})$. The set of all poles of $\mathbf{H}$ is denoted by $\mathfrak{P}_\mathbf{H}$. Moreover, define

$$\psi(\cdot\,, \lambda) := \det(\lambda I - \mathbf{H}(\cdot)) \in \mathcal{M}_\alpha[\lambda]\,,$$

where $\mathcal{M}_\alpha[\lambda]$ denotes the ring of polynomials over $\mathcal{M}_\alpha$. Since $\mathcal{M}_\alpha$ is a field, there exist a unique $\ell \in \mathbb{N}$ and unique monic irreducible polynomials $\psi_i(\cdot\,, \lambda) \in \mathcal{M}_\alpha[\lambda]$ such that

$$\psi(s, \lambda) = \prod_{i=0}^{\ell} \psi_i(s, \lambda)\,.$$

Let $\Delta_i(\cdot) \in \mathcal{M}_\alpha$ denote the discriminant of $\psi_i(\cdot\,, \lambda) \in \mathcal{M}_\alpha[\lambda]$. If $s_0 \notin \mathfrak{P}_\mathbf{H}$, then it is not difficult to show that the coefficients of the polynomials $\psi_i(\cdot\,, \lambda)$ are holomorphic in a sufficiently small neighborhood of s_0 (cf. Baumgärtel [2, p. 397]) and hence $\Delta_i(s_0)$ is the discriminant of $\psi_i(s_0, \lambda) \in \mathbb{C}[\lambda]$. Thus $\psi_i(s_0, \lambda)$ has only simple zeros if and only if $\Delta_i(s_0) \neq 0$; see, for example, Cohn [7, p. 175]. Since $\psi_i(\cdot\,, \lambda)$ is irreducible, it follows that $\Delta_i(s) \not\equiv 0$ and hence the set $\mathfrak{C}_\mathbf{H}$ of *critical points* of $\mathbf{H}$ defined by

$$\mathfrak{C}_\mathbf{H} := \left\{ s \in \mathbb{C}_\alpha \mid \prod_{i=0}^{\ell} \Delta_i(s) = 0 \right\}$$

is a discrete set in $\mathbb{C}_\alpha$. We shall need the following lemma from Forster [15, p. 52].

LEMMA 5.1. *Let $s_0 \in \mathbb{C}$, let $U \subset \mathbb{C}$ be an open neighborhood of s_0, and suppose that $c_1(s), \ldots, c_n(s)$ are holomorphic functions on U. If $\lambda_0 \in \mathbb{C}$ is a simple zero of the polynomial*

$$\lambda^n + c_1(s_0)\lambda^{n-1} \ldots + c_n(s_0) \in \mathbb{C}[\lambda],$$

then there exists an open neighborhood $V \subset U$ of s_0 and a function ξ holomorphic on V such that $\xi(s_0) = \lambda_0$ and

$$\xi^n(s) + c_1(s)\xi^{n-1}(s) \ldots + c_n(s) = 0 \quad \forall\, s \in V.$$

For $s_0 \in \mathbb{C}_\alpha$, $\phi \in [0, 2\pi)$, and $0 < a \leq \infty$ set $\Gamma := \{s_0 + e^{i\phi}t \,|\, 0 \leq t < a\}$. Hence Γ is a half-line ($a = \infty$) or a line segment ($a \neq \infty$) with initial point s_0.

PROPOSITION 5.2. *Suppose that $\Gamma^{\mathrm{cl}} \subset \mathbb{C}_\alpha \setminus (\mathfrak{P}_{\mathbf{H}} \cup \mathfrak{C}_{\mathbf{H}})$. If $\lambda_0 \in \sigma(\mathbf{H}(s_0))$ (the spectrum of $\mathbf{H}(s_0)$), then there exists a region $V \subset \mathbb{C}_\alpha$ satisfying $\Gamma^{\mathrm{cl}} \subset V \subset \mathbb{C}_\alpha \setminus (\mathfrak{P}_{\mathbf{H}} \cup \mathfrak{C}_{\mathbf{H}})$ and a function ξ holomorphic on V such that $\xi(s_0) = \lambda_0$ and*

$$\psi(s, \xi(s)) = \det(\xi(s)I - \mathbf{H}(s)) = 0 \quad \forall\, s \in V.$$

Moreover, if $a = \infty$, then under the extra assumption that the limit

$$\lim_{|s| \to \infty,\, s \in \Gamma} \mathbf{H}(s) = D \in \mathbb{C}^{m \times m}$$

exists, it follows that

$$\lim_{|s| \to \infty,\, s \in \Gamma} \xi(s) =: \xi_\infty \in \sigma(D).$$

The above proposition remains true if Γ is replaced by more general curves. However, for our purposes Proposition 5.2 is sufficient.

Proof. Let $\lambda_0 \in \sigma(\mathbf{H}(s_0))$. After a suitable renumbering we may assume that $\psi_0(s_0, \lambda_0) = 0$. Let us first consider the case when $a \neq \infty$. Define $\gamma(t) := s_0 + e^{i\phi}t$ for $t \in [0, a]$. It follows from the assumption that for any $t \in [0, a]$ the polynomial $\psi_0(\gamma(t), \lambda) \in \mathbb{C}[\lambda]$ has no multiple zeros. Let n denote the degree of $\psi_0(\cdot\,, \lambda)$ and $\lambda_t^1, \ldots, \lambda_t^n$ denote the n different simple zeros of $\psi_0(\gamma(t), \lambda) \in \mathbb{C}[\lambda]$. Moreover, for $t \in [0, a]$ let $\mathfrak{B}_t$ denote the set of all open balls B_t with center in $\gamma(t)$ such that $B_t \subset \mathbb{C}_\alpha \setminus (\mathfrak{P}_{\mathbf{H}} \cup \mathfrak{C}_{\mathbf{H}})$ and such that there exist n functions ξ_t^i holomorphic on B_t satisfying $\xi_t^i(\gamma(t)) = \lambda_t^i$,

$$\xi_t^i(s) \neq \xi_t^j(s) \;\; \forall\, s \in B_t\,, \forall\, i, j \in \underline{n},\; i \neq j, \;\text{ and }\; \psi_0(s, \xi_t^i(s)) = 0 \;\; \forall\, s \in B_t\,, \forall\, i \in \underline{n}\,,$$

where $\underline{n}$ denotes the set $\{1, \ldots, n\}$. By Lemma 5.1 the set $\mathfrak{B}_t$ is nonempty, and by setting $\hat{B}_t := \bigcup_{B_t \in \mathfrak{B}_t} B_t$ we obtain the maximal element of $\mathfrak{B}_t$. Denoting the radius of $\hat{B}_t$ by ϱ_t we claim that

$$\varrho := \inf_{t \in [0,a]} \varrho_t > 0\,. \tag{5.1}$$

Assume that (5.1) is not true. Then there exist numbers $t_j \in [0, a]$ such that

$$\lim_{j \to \infty} \varrho_{t_j} = 0\,. \tag{5.2}$$

Since $a < \infty$, we may assume without loss of generality that $\lim_{j \to \infty} t_j =: t^* \in [0, a]$ exists. By assumption $\gamma(t^*) \in \mathbb{C}_\alpha \setminus (\mathfrak{P}_{\mathbf{H}} \cup \mathfrak{C}_{\mathbf{H}})$ and hence $\varrho_{t^*} > 0$. Using $\lim_{j \to \infty} \gamma(t_j) = \gamma(t^*)$, we conclude that there exist $j_0 \in \mathbb{N}$ and $\delta > 0$ such that $B(\gamma(t_j), \delta) \subset \hat{B}_{t^*}$ for all $j \geq j_0$, where $B(\gamma(t_j), \delta)$ denotes the open ball of radius δ with center in $\gamma(t_j)$. But this implies that $B(\gamma(t_j), \delta) \in \mathfrak{B}_{t_j}$ for all $j \geq j_0$ and therefore $B(\gamma(t_j), \delta) \subset \hat{B}_{t_j}$ for all $j \geq j_0$ (by the maximality of $\hat{B}_{t_j}$). Thus we obtain

that $\varrho_{t_j} \geq \delta$ for all $j \geq j_0$, which contradicts (5.2). It follows that (5.1) holds, and therefore there exists a largest number $m \in \mathbb{N}$ such that $m\varrho \leq a$. Setting $\tau_j := \varrho j$ it is clear that

$$B(\gamma(\tau_j), \varrho) \subset \hat{B}_{\tau_j} \text{ for } j = 0, \ldots, m \text{ and } \Gamma^{cl} \subset \bigcup_{j=0,\ldots,m} B(\gamma(\tau_j), \varrho).$$

Let $j_0 \in \underline{n}$ be such that $\lambda_{\tau_0}^{j_0} = \lambda_0^{j_0} = \lambda_0$, and set

$$\xi_0(s) := \xi_{\tau_0}^{j_0}(s) = \xi_0^{j_0}(s).$$

Now $S_0 := B(\gamma(0), \varrho) \cap B(\gamma(\tau_1), \varrho) \neq \emptyset$ is contained in $\mathbb{C}_\alpha \setminus (\mathfrak{P}_{\mathbf{H}} \cup \mathfrak{C}_{\mathbf{H}})$, and hence, for any $s \in S_0$, the polynomial $\psi_0(s, \lambda) \in \mathbb{C}[\lambda]$ has n different simple zeros, which are given by $\xi_{\tau_1}^1(s), \ldots, \xi_{\tau_1}^n(s)$. On the other hand we have

$$\psi_0(s, \xi_0(s)) = 0 \quad \forall s \in S_0,$$

and hence there must exist $j_1 \in \underline{n}$ such that

$$\xi_0(s) = \xi_{\tau_1}^{j_1}(s) \quad \forall s \in S_0.$$

Setting

$$\xi_1(s) := \xi_{\tau_1}^{j_1}(s) \ \forall s \in B(\gamma(\tau_1), \varrho) \text{ and } S_1 := B(\gamma(\tau_1), \varrho) \cap B(\gamma(\tau_2), \varrho) \neq \emptyset,$$

the same argument can be used to conclude that there exists $j_2 \in \underline{n}$ such that

$$\xi_1(s) = \xi_{\tau_2}^{j_2}(s) \quad \forall s \in S_1.$$

Repeating the above argument shows that there exist $m+1$ holomorphic functions $\xi_j : B(\gamma(\tau_j), \varrho) \to \mathbb{C}$ $(j = 0, \ldots, m)$ such that $\xi_{j+1}(s) = \xi_j(s)$ for all $s \in S_j := B(\gamma(\tau_j), \varrho) \cap B(\gamma(\tau_{j+1}), \varrho) \neq \emptyset$ $(j = 0, \ldots, m-1)$. On the region

$$V := \bigcup_{j=0,\ldots,m} B(\gamma(\tau_j), \varrho)$$

we define a function $\xi(s)$ by setting

$$\xi(s) := \xi_j(s) \text{ if } s \in B(\gamma(\tau_j), \varrho).$$

By construction ξ is a well-defined holomorphic function on V such that $\xi(s_0) = \lambda_0$ and $\det(\xi(s)I - \mathbf{H}(s)) = 0$ for all $s \in V$.

Let us now consider the case when $a = \infty$. For $j \in \mathbb{N}$ define $\Gamma_j := \{s_0 + e^{i\phi}t \,|\, 0 \leq t < j\}$. The above construction shows that there exist regions V_j satisfying $\Gamma_j^{\text{cl}} \subset V_j \subset \mathbb{C}_\alpha \setminus (\mathfrak{P}_{\mathbf{H}} \cup \mathfrak{C}_{\mathbf{H}})$ and $V_j \subset V_{j+1}$ and functions ξ_j holomorphic on V_j such that $\xi_j(s_0) = \lambda_0$, $\psi_0(s, \xi_j(s)) = 0$ for all $s \in V_j$ and $\xi_{j+1}|_{V_j} = \xi_j$. Hence on $V := \bigcup_{j=1}^{\infty} V_i$ we obtain a well-defined holomorphic function $\xi(s)$ with the desired properties by setting

$$\xi(s) := \xi_j(s) \text{ if } s \in V_j.$$

Finally, the last statement in the proposition is a consequence of the fact that the eigenvalues of $\mathbf{H}(\gamma(t))$ are continuous in t; see Kato [21, p. 106]. □

The next result extends Theorem 3.1 to the multivarable case.

THEOREM 5.3. *Let* $\mathbf{H}(s)$ *be a* $\mathbb{C}^{m\times m}$*-valued regular transfer function and define for any* $\varepsilon \geq 0$

$$\mathbf{G}^{\varepsilon}(s) := \mathbf{H}(s)(I + e^{-\varepsilon s}\mathbf{H}(s))^{-1}. \tag{5.3}$$

Let $D \in \mathbb{C}^{m\times m}$ *denote the feedthrough matrix of* $\mathbf{H}$ *and assume the following:*

(1) $\mathbf{G}^0 \in H^\infty(\mathbb{C}^{m\times m})$,

(2) $\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} r(\mathbf{H}(s)) =: \gamma > 1$,

(3) $r(D) \leq 1$.

Then there exist sequences (ε_n) *and* (p_n) *with*

$$\varepsilon_n > 0\,,\ \ \varepsilon_n \to 0\,,\ \ p_n \in \mathbb{C}_0\,,\ \ |\operatorname{Im} p_n| \to \infty$$

and such that for any $n \in \mathbb{N}$, p_n *is a pole of* $\mathbf{G}^{\varepsilon_n}$.

Remark 5.4. (i) Since $r(\mathbf{H}(s))$ is not defined, if s is a pole of $\mathbf{H}$, condition (2) in the previous theorem should be formulated more precisely as

$$\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0\setminus\mathfrak{P}_{\mathbf{H}}} r(\mathbf{H}(s)) =: \gamma > 1.$$

To simplify the notation, we make the convention that if $s \in \mathfrak{P}_{\mathbf{H}}$, then $r(\mathbf{H}(s)) = 0$.

(ii) Suppose that $\lim_{\mu\downarrow 0} \mathbf{H}(\mu + i\omega) =: \mathbf{H}(i\omega)$ exists for almost all $\omega \in \mathbb{R}$. (For example, this is the case if $\mathbf{H} \in H^\infty(\mathbb{C}^{m\times m})$.) Then condition (2) in Theorem 5.3 is satisfied if

$$\lim_{\varrho\to\infty} \operatorname{ess\,sup}\{r(\mathbf{H}(i\omega)) \,|\, |\omega| > \varrho\} > 1. \tag{5.4}$$

Under the extra assumption that $\mathbf{H}(i\omega)$ is almost periodic, (5.4) holds if

$$\operatorname{ess\,sup}\{r(\mathbf{H}(i\omega) \,|\, \omega \in \mathbb{R}\} > 1.$$

For the proof of Theorem 5.3 we need the following simple lemma.

LEMMA 5.5. *Let the set* $A \subset \mathbb{C}_0$ *be discrete in* $\mathbb{C}_0$. *Then for any* $\varepsilon > 0$ *and* $\beta \in \mathbb{R}$, *there exists* $y_0 \in \mathbb{R}$ *such that*

$$|y_0 - \beta| \leq \varepsilon \quad \text{and} \quad \{x + iy_0 \,|\, x > 0\} \cap A = \emptyset\,.$$

Proof. It is easy to see that A is countable, so we can choose $y_0 \in (\beta - \varepsilon, \beta + \varepsilon)$ such that $y_0 \notin \operatorname{Im} A$. □

Proof of Theorem 5.3. As in the SISO case it follows from the assumptions (1) and (2) that there exist constants $\eta > 0$ and $\gamma_1 \in (1, \gamma)$ such that the set

$$S_1 = \{re^{i\phi} \mid r \in [1, \gamma_1]\,,\ |\phi - \pi| < \eta\}$$

does not intersect $\sigma(\mathbf{H}(s))$ for all $s \in \mathbb{C}_0 \setminus \mathfrak{P}_{\mathbf{H}}$:

$$\sigma(\mathbf{H}(s)) \cap S_1 = \emptyset \quad \forall\, s \in \mathbb{C}_0 \setminus \mathfrak{P}_{\mathbf{H}}. \tag{5.5}$$

Since $\gamma_1 < \gamma$, by assumption (2) there exists a sequence (s_n) in $\mathbb{C}_0 \setminus \mathfrak{P}_{\mathbf{H}}$ such that $|s_n| \to \infty$ and $r(\mathbf{H}(s_n)) > \gamma_1$ for all $n \in \mathbb{N}$. It is obvious that we can find numbers $\delta_n \in (0, 1)$ such that the vertical segment $J_n := [s_n - i\delta_n, s_n + i\delta_n]$ is contained in $\mathbb{C}_0 \setminus \mathfrak{P}_{\mathbf{H}}$ and

$$r(\mathbf{H}(s)) \geq \gamma_1 \ \ \forall\, s \in \bigcup_{n\in\mathbb{N}} J_n\,.$$

It follows from Lemma 5.5 that there exists $z_n \in J_n$ such that the ray

$$\Gamma_n = \{z_n + a \,|\, a \in [0, \infty)\}$$

does not intersect the set $\mathfrak{P}_{\mathbf{H}} \cup \mathfrak{C}_{\mathbf{H}}$:

$$\Gamma_n \cap (\mathfrak{P}_{\mathbf{H}} \cup \mathfrak{C}_{\mathbf{H}}) \neq \emptyset \quad \forall n \in \mathbb{N}. \tag{5.6}$$

Using assumption (3) it can be shown as in the SISO case that

$$\lim_{n \to \infty} |\mathrm{Im}\, z_n| = \infty. \tag{5.7}$$

Again, without loss of generality, we may assume that $\mathrm{Im}\, z_n > 0$ for all $n \in \mathbb{N}$. By construction we have that $r(\mathbf{H}(z_n)) \geq \gamma_1$ and hence there exists $\lambda_n \in \sigma(\mathbf{H}(z_n))$ such that $|\lambda_n| \geq \gamma_1 > 1$. Since (5.6) holds, an application of Proposition 5.2 shows that for all $n \in \mathbb{N}$ there exists a region V_n satisfying $\Gamma_n \subset V_n \subset \mathbb{C}_0 \setminus (\mathfrak{P}_{\mathbf{H}} \cup \mathfrak{C}_{\mathbf{H}})$ and a function ξ_n holomorphic on V_n such that $\xi_n(z_n) = \lambda_n$ and

$$\xi_n(s) \in \sigma(\mathbf{H}(s)) \quad \forall\, s \in V_n.$$

Moreover, by Proposition 2.3

$$\lim_{|s| \to \infty,\, s \in \Gamma_n} \mathbf{H}(s) = D,$$

and hence we obtain by Proposition 5.2 that

$$\lim_{|s| \to \infty,\, s \in \Gamma_n} \xi_n(s) =: \xi_n^{\infty} \in \mathbb{C}$$

exists and $\xi_n^{\infty} \in \sigma(D)$. As a consequence of assumption (3) we have that $|\xi_n^{\infty}| \leq 1$. Therefore the extended real numbers

$$\begin{aligned} a_n' &= \max\{a \in [0, \infty) \,|\, |\xi_n(z_n + a)| \geq \gamma_1\}, \\ a_n'' &= \min\{a \in [a_n', \infty] \,|\, |\xi_n(z_n + a)| \leq 1\} \end{aligned}$$

are well defined. (If $r(D) = 1$, it might happen that $a_n'' = \infty$.) Setting

$$z_n' = z_n + a_n' \quad \text{and} \quad z_n'' = z_n + a_n''$$

(where $z_n'' = \infty$ is possible), we will be looking for poles of $\mathbf{G}^{\varepsilon}$ in (z_n', z_n''). Notice that a sufficient condition for $p \in (z_n', z_n'')$ to be a pole of $\mathbf{G}^{\varepsilon}$ is that

$$\log \xi_n(p) - \varepsilon p = -i\pi.$$

By (5.5) it follows that for all $n \in \mathbb{N}$

$$\xi_n([z_n', z_n'']) \subset S_2 := \{re^{i\phi} \,|\, r \in [1, \gamma_1],\ |\phi| \leq \pi - \eta\}. \tag{5.8}$$

It follows that $\log \xi_n$ and $\arg \xi_n$ (where log and arg are defined as in §2) are continuous functions on $[z_n', z_n'']$. For $s \in [z_n', z_n'']$ define

$$R_n(s) := \{\log \xi_n(s) - \varepsilon s \,|\, \varepsilon \in [0, \infty)\}.$$

Then for each $n \in \mathbb{N}$ and $s \in [z_n', z_n'']$ the ray $R_n(s)$ intersects the line $L = \{s \in \mathbb{C} \mid \operatorname{Im} s = -\pi\}$ in a point $w_n(s) - i\pi$. Thus we can define functions $w_n, e_n : [z_n', z_n''] \to \mathbb{R}$ such that $e_n(s) > 0$ for all $s \in [z_n', z_n'']$ and

$$\log \xi_n(s) - e_n(s)s = w_n(s) - i\pi \quad \forall\, s \in [z_n', z_n''],\ \forall\, n \in \mathbb{N}. \tag{5.9}$$

As in the proof of Theorem 3.1 it follows that for all sufficiently large n there exists $p_n \in (z_n', z_n'')$ such that $w_n(p_n) = 0$. Thus, by (5.9)

$$\log \xi_n(p_n) - e_n(p_n)p_n = -i\pi\,.$$

Finally, it is clear that

$$e_n(p_n) = \frac{\arg \xi_n(p_n) + \pi}{\operatorname{Im} p_n}\,. \tag{5.10}$$

The sequence $(\arg \xi_n(p_n))$ is bounded, and by (5.7), we have that $\operatorname{Im} p_n = \operatorname{Im} z_n \to \infty$ as $n \to \infty$. Therefore we obtain from (5.10) that $\lim_{n\to\infty} e_n(p_n) = 0$. Setting $\varepsilon_n = e_n(p_n)$ it follows that $\mathbf{G}^{\varepsilon_n}$ has a pole in $p_n \in \mathbb{C}_0$. □

Along the same lines a multivariable extension of Theorem 4.1 can be obtained. We state this result without proof.

THEOREM 5.6. *Let* $\mathbf{H}$ *be a* $\mathbb{C}^{m\times m}$*-valued regular transfer function, and for any* $\varepsilon \geq 0$ *let* $\mathbf{G}^\varepsilon$ *be defined by* (5.3). *Let* D *denote the feedthrough matrix of* $\mathbf{H}$, *and assume that*

$$r(D) > 1.$$

Then there exist sequences (ε_n) *and* (p_n) *with*

$$\varepsilon_n > 0\,,\ \varepsilon_n \to 0\,,\ \operatorname{Re} p_n \to \infty\,,\ \operatorname{Im} p_n \to \infty$$

and such that for any $n \in \mathbb{N}$, p_n *is a pole of* $\mathbf{G}^{\varepsilon_n}$.

Combining Theorem 5.3 and Theorem 5.6 yields part (ii) of Theorem 1.1.

6. Robustness of stability. So far we have proved only part (ii) of Theorem 1.1. In this section we conclude the proof. In fact, since for part (i) of Theorem 1.1 the wellposedness and regularity assumptions are not needed and we obtain additionally uniform boundedness for the matrices $\mathbf{G}^\varepsilon$, we restate our result.

THEOREM 6.1. *Suppose* $\mathbf{G}^0 \in H^\infty(\mathbb{C}^{m\times m})$ *and denote* $\mathbf{H} = \mathbf{G}^0(I - \mathbf{G}^0)^{-1}$ (*so that* $\mathbf{H} \in \mathcal{M}_0(\mathbb{C}^{m\times m})$). *If*

$$\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} r(\mathbf{H}(s)) < 1, \tag{6.1}$$

then there exist numbers $\varepsilon_0 > 0$ *and* $M > 0$ *such that* $\mathbf{G}^\varepsilon(s) = \mathbf{H}(s)(I + e^{-\varepsilon s}\mathbf{H}(s))^{-1} \in H^\infty(\mathbb{C}^{m\times m})$ *and* $\|\mathbf{G}^\varepsilon\|_\infty \leq M$ *for all* $\varepsilon \in [0, \varepsilon_0]$.

Remark 6.2. Suppose that $\mathbf{H} \in H^\infty(\mathbb{C}^{m\times m})$. Then, as is well known,

$$\lim_{\mu\downarrow 0} \mathbf{H}(\mu + i\omega) =: \mathbf{H}(i\omega)$$

exists for almost all $\omega \in \mathbb{R}$. It is easy to show that $r(\mathbf{H}(s))$ is a subharmonic function on $\mathbb{C}_0$. Using standard results on subharmonic functions (see, for example, Narasimhan [25, p. 227]) it is not difficult to prove that

$$\sup\{r(\mathbf{H}(s)) \mid s \in \mathbb{C}_0\} = \operatorname{ess\,sup}\{r(\mathbf{H}(i\omega)) \mid \omega \in \mathbb{R}\}.$$

As a consequence, (6.1) will be satisfied if $\operatorname{ess\,sup}\{r(\mathbf{H}(i\omega)) \mid \omega \in \mathbb{R}\} < 1$.

The proof of Theorem 6.1 requires some preparation. If $\mathbf{H} \in \mathcal{M}_0(\mathbb{C}^{m\times m})$ and s_0 is a pole of $\mathbf{H}$, then trivial examples show that $r(\mathbf{H}(s))$ does not necessarily blow up as $s \to s_0$. However, the next lemma reveals that this phenomenon cannot occur if $\mathbf{H}(I+\mathbf{H})^{-1}$ is L^2-stable.

LEMMA 6.3. *Let $U \subset \mathbb{C}$, suppose $\mathbf{G}^0$ is bounded and holomorphic on U, and denote $\mathbf{H} = \mathbf{G}^0(I-\mathbf{G}^0)^{-1}$. If $\sup_{s\in U} r(\mathbf{H}(s)) < \infty$, then $\sup_{s\in U} \|\mathbf{H}(s)\| < \infty$.*

Proof. Assume the claim is not true; i.e., there exists a sequence (s_n) in U such that $\lim_{n\to\infty} \|\mathbf{H}(s_n)\| = \infty$. Using Cramer's rule and the boundedness of $\mathbf{G}^0$ on U, it follows that

$$\lim_{n\to\infty} \det(I - \mathbf{G}^0(s_n)) = 0. \tag{6.2}$$

Now $(\mathbf{G}^0(s_n))$ is a bounded sequence in $\mathbb{C}^{m\times m}$, and hence we may assume without loss of generality that $\lim_{n\to\infty} \mathbf{G}^0(s_n) =: D^0 \in \mathbb{C}^{m\times m}$ exists. From (6.2) it follows that $1 \in \sigma(D^0)$. This in turn implies that there exist eigenvalues $\lambda_n \in \sigma(\mathbf{H}(s_n))$ such that $\lim_{n\to\infty}[\lambda_n/(1+\lambda_n)] = 1$. But this leads to a contradiction, since the sequence (λ_n) is bounded by assumption. □

Proof of Theorem 6.1. *Step* 1: For $\varrho > 0$ set

$$\mathbb{C}_0^\varrho := \{s \in \mathbb{C}_0 \mid |s| \geq \varrho\}. \tag{6.3}$$

By (6.1) there exists $R > 0$ and $q \in (0,1)$ such that

$$r(\mathbf{H}(s)) \leq q < 1 \ \ \forall s \in \mathbb{C}_0^R. \tag{6.4}$$

Combining (6.4) and Lemma 6.3 shows that $\mathbf{H}(s)$ is bounded on $\mathbb{C}_0^R$.

Step 2: We claim that there exists a number $L > 0$ such that $\|\mathbf{G}^\varepsilon(s)\| \leq L$ for all $\varepsilon \geq 0$ and for all $s \in \mathbb{C}_0^R$. Suppose the claim is not true. Then, since by Step 1 $\mathbf{H}(s)$ is bounded on $\mathbb{C}_0^R$, it follows from Cramer's rule that there exists a sequence (s_n) in $\mathbb{C}_0^R$ and a sequence (ε_n) of nonnegative numbers such that

$$\lim_{n\to\infty} \det(I + e^{-\varepsilon_n s_n}\mathbf{H}(s_n)) = 0. \tag{6.5}$$

Now $(\mathbf{H}(s_n))$ is a bounded sequence in $\mathbb{C}^{m\times m}$ and $|e^{-\varepsilon_n s_n}| \leq 1$ for all $n \in \mathbb{N}$, and therefore (as in Step 1) we may assume without loss of generality that the limits $\lim_{n\to\infty} e^{-\varepsilon_n s_n} =: d$ and $\lim_{n\to\infty} \mathbf{H}(s_n) =: E$ exist. Using (6.4) and the fact that $|d| \leq 1$ we see that $r(dE) < 1$. On the other hand it follows from (6.5) that $-1 \in \sigma(dE)$, a contradiction.

Step 3: Choose $\varepsilon_0 > 0$ such that for any $s \in \mathbb{C}_0$ with $|s| \leq R$ and any $\varepsilon \in [0, \varepsilon_0]$

$$|1 - e^{-\varepsilon s}| \leq \frac{1}{2\|\mathbf{G}^0\|_\infty}.$$

The identity $\mathbf{G}^\varepsilon(s) = \mathbf{G}^0(s)(I - (1-e^{-\varepsilon s})\mathbf{G}^0(s))^{-1}$ shows that, for all s and ε as above, $\|\mathbf{G}^\varepsilon(s)\| \leq 2\|\mathbf{G}^0\|_\infty$.

Combining Steps 2 and 3, we obtain that $\mathbf{G}^\varepsilon \in H^\infty(\mathbb{C}^{m\times m})$ and $\|\mathbf{G}^\varepsilon\|_\infty \leq M$ for all $\varepsilon \in [0, \varepsilon_0]$, where $M := \max(L, 2\|\mathbf{G}^0\|_\infty)$. □

Theorem 6.1 deals with delay perturbations of the form $e^{-\varepsilon s}$. A natural question to ask is whether it remains true for *multidelay perturbations* of the form $\operatorname{diag}_{1\leq j\leq m}(e^{-\varepsilon^j s})$, where $\varepsilon^j \geq 0$, $j = 1, \ldots, n$. The answer is no, as the following example shows.

Example 6.4. Consider the transfer function $\mathbf{H}(s) \equiv D$, where D is given by

$$D = \begin{pmatrix} -1/2 & 1/4 \\ -1 & 1/2 \end{pmatrix}.$$

The matrix D is nilpotent; i.e., $\sigma(D) = \{0\}$ and hence $\mathbf{G}^0(s) = \mathbf{H}(s)(I + \mathbf{H}(s))^{-1} \equiv D(I + D)^{-1}$ belongs to $H^\infty(\mathbb{C}^{2\times 2})$. Moreover

$$\limsup_{|s|\to\infty, s\in\mathbb{C}_0} r(\mathbf{H}(s)) = r(D) = 0\,,$$

and thus, by Theorem 6.1, $\mathbf{G}^0$ is robustly stable with respect to delays. Setting $\Delta := \operatorname{diag}(1, -1)$, a trivial computation shows that $\sigma(\Delta D) = \{-1, 0\}$. Therefore, defining $\varepsilon_n^1 := 2\pi/n$ and $\varepsilon_n^2 := \pi/n$, it follows that

$$\det(I + \operatorname{diag}(e^{-in\varepsilon_n^1}, e^{-in\varepsilon_n^2})D) = \det(I + \Delta D) = 0 \quad \forall n \in \mathbb{N}\,.$$

As a consequence, for all $n \in \mathbb{N}$, $p_n = in$ is a pole of the closed-loop transfer function $\mathbf{G}^{\varepsilon_n}$ with multidelay $\varepsilon_n = (\varepsilon_n^1, \varepsilon_n^2)$, defined by

$$\mathbf{G}^{\varepsilon_n}(s) := \mathbf{H}(s)(I + \operatorname{diag}(e^{-\varepsilon_n^1 s}, e^{-\varepsilon_n^2 s})\mathbf{H}(s))^{-1}\,.$$

In order to give a sufficient condition for robust stability in the presence of multidelay perturbations, set

$$\mathbf{\Delta} := \{\operatorname{diag}_{1\le j\le m}(s_j) \mid s_j \in \mathbb{C}\} \subset \mathbb{C}^{m\times m}$$

and define the structured singular value $\mu_{\mathbf{\Delta}}(M)$ of $M \in \mathbb{C}^{m\times m}$ with respect to $\mathbf{\Delta}$ by

$$\mu_{\mathbf{\Delta}}(M) := \frac{1}{\min\{\|\Delta\| \mid \Delta \in \mathbf{\Delta},\ \det(I - M\Delta) = 0\}}\,,$$

unless no $\Delta \in \mathbf{\Delta}$ makes $I - M\Delta$ singular, in which case $\mu_{\mathbf{\Delta}}(M) := 0$ (cf. Packard and Doyle [26]).

THEOREM 6.5. *Suppose* $\mathbf{G}^0 \in H^\infty(\mathbb{C}^{m\times m})$ *and denote* $\mathbf{H} = \mathbf{G}^0(I - \mathbf{G}^0)^{-1}$ (*so that* $\mathbf{H} \in \mathcal{M}_0(\mathbb{C}^{m\times m})$). *If*

$$\limsup_{|s|\to\infty,\ s\in\mathbb{C}_0} \mu_{\mathbf{\Delta}}(\mathbf{H}(s)) < 1, \tag{6.6}$$

then there exist numbers $\delta > 0$ *and* $M > 0$ *such that*

$$\mathbf{G}^\varepsilon(s) := \mathbf{H}(s)(I + \operatorname{diag}_{1\le j\le m}(e^{-\varepsilon^j s})\mathbf{H}(s))^{-1} \in H^\infty(\mathbb{C}^{m\times m}) \tag{6.7}$$

and $\|\mathbf{G}^\varepsilon\|_\infty \le M$ *for all* $\varepsilon = (\varepsilon^1, \dots, \varepsilon^m) \in \mathbb{R}_+^m$ *satisfying* $\|\varepsilon\| < \delta$.

Using standard properties of structured singular values [26] the proof of Theorem 6.5 is a straightforward extension of the proof of Theorem 6.1 and is therefore left to the reader.

Condition (6.6) holds if

$$\limsup_{|s|\to\infty,\ s\in\mathbb{C}_0} \|\mathbf{H}(s)\| < 1 \tag{6.8}$$

is satisfied. Equation (6.8) is not necessary for robustness with respect to multidelay perturbations, as the following simple example shows.

Example 6.6. Let h_1, h_2, and h_3 be in H^∞ with $\|h_1\|_\infty < 1$, $\|h_3\|_\infty < 1$ and

$$\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} |h_2(s)| > 1\,.$$

If we define

$$\mathbf{H} = \begin{pmatrix} h_1 & h_2 \\ 0 & h_3 \end{pmatrix},$$

then it is clear that $\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} \|\mathbf{H}(s)\| > 1$, so (6.8) is not satisfied. Since

$$\det(I + \operatorname{diag}(e^{-\varepsilon^1 s}, e^{-\varepsilon^2 s})\mathbf{H}(s)) = (1 + e^{-\varepsilon^1 s}h_1(s))(1 + e^{-\varepsilon^2 s}h_2(s)) \;\forall\, (\varepsilon^1, \varepsilon^2) \in \mathbb{R}^2_+\,,$$

it follows that, denoting $\varepsilon = (\varepsilon^1, \varepsilon^2) \in \mathbb{R}^2_+$,

$$\inf_{\varepsilon\in\mathbb{R}^2_+} \inf_{s\in\mathbb{C}_0} |\det(I + \operatorname{diag}(e^{-\varepsilon^1 s}, e^{-\varepsilon^2 s})\mathbf{H}(s))| > 0\,.$$

Let $\mathbf{G}^\varepsilon$ be defined by (6.7). Using Cramer's rule we obtain that $\mathbf{G}^\varepsilon \in H^\infty(\mathbb{C}^{2\times 2})$ and $\sup_{\varepsilon\in\mathbb{R}^2_+} \|\mathbf{G}^\varepsilon\|_\infty < \infty$; in particular, we have robust stability with respect to multidelay perturbations.

It seems to be a difficult open problem whether the condition

$$\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} \mu_{\mathbf{\Delta}}(\mathbf{H}(s)) > 1$$

implies lack of robustness with respect to small multidelay perturbations.

7. Robustness of instability. Given a transfer function $\mathbf{H}$ of size $m \times m$, we have shown in the previous sections that, under certain conditions, there exists a positive sequence (ε_n) with $\varepsilon_n \to 0$ such that the closed-loop transfer function $\mathbf{G}^{\varepsilon_n}$ has at least one pole in $\mathbb{C}_0$ for all $n \in \mathbb{N}$. In this section we show that this property is robust in the following sense: For any $n \in \mathbb{N}$ there exists $\delta_n \in (0, \varepsilon_n)$ such that for any $\varepsilon \in \mathbb{R}^m_+$ with $\varepsilon = (\varepsilon^1, \ldots, \varepsilon^m) \in \bigcup_{n\in\mathbb{N}} (\varepsilon_n - \delta_n, \varepsilon_n + \delta_n)^m$, $\mathbf{G}^\varepsilon$ (defined by (6.7)) has a pole in $\mathbb{C}_0$.

In the following we shall need the notion of a right-coprime (or left-coprime) factorization of a matrix-valued meromorphic function.

LEMMA 7.1. *Suppose* $\mathbf{H} \in \mathcal{M}_\alpha(\mathbb{C}^{p\times m})$. *Then the following statements hold:*

(i) $\mathbf{H}$ *admits a right-coprime factorization over* $\mathcal{H}_\alpha$; *i.e., there exist matrices* $N \in \mathcal{H}_\alpha(\mathbb{C}^{p\times m})$, $D, Y \in \mathcal{H}_\alpha(\mathbb{C}^{m\times m})$, *and* $X \in \mathcal{H}_\alpha(\mathbb{C}^{m\times p})$ *such that*

$$\mathbf{H} = ND^{-1} \;\text{ and }\; XN + YD = I\,.$$

The matrices N *and* D *are unique up to multiplication from the right by a unimodular factor. A number* $s_0 \in \mathbb{C}_\alpha$ *is a pole of* $\mathbf{H}$ *if and only if* $\det D(s_0) = 0$.

(ii) $\mathbf{H}$ *admits a left-coprime factorization over* $\mathcal{H}_\alpha$; *i.e., there exist matrices* $\tilde{N} \in \mathcal{H}_\alpha(\mathbb{C}^{p\times m})$, $\tilde{D}, \tilde{Y} \in \mathcal{H}_\alpha(\mathbb{C}^{p\times p})$, *and* $\tilde{X} \in \mathcal{H}_\alpha(\mathbb{C}^{m\times p})$ *with*

$$\mathbf{H} = \tilde{D}^{-1}\tilde{N} \;\text{ and }\; \tilde{N}\tilde{X} + \tilde{D}\tilde{Y} = I\,.$$

The matrices $\tilde{N}$ *and* $\tilde{D}$ *are unique up to multiplication from the left by a unimodular factor. A number* $s_0 \in \mathbb{C}_\alpha$ *is a pole of* $\mathbf{H}$ *if and only if* $\det \tilde{D}(s_0) = 0$.

(iii) *If* $\mathbf{H} = ND^{-1}$ *is a right-coprime factorization over* $\mathcal{H}_\alpha$ *and* $\mathbf{H} = \tilde{D}^{-1}\tilde{N}$ *is a left-coprime factorization over* $\mathcal{H}_\alpha$*, then the zeros of* $\det D$ *and* $\det \tilde{D}$ *in* $\mathbb{C}_\alpha$ *coincide (counting multiplicities).*

Proof. It is well known that the ring $\mathcal{H}_\alpha$ is a Bézout domain; i.e., every finitely generated ideal is principal (see, for example, Narasimhan [25, p. 136]). Now $\mathcal{M}_\alpha$ is the quotient field of $\mathcal{H}_\alpha$, and statements (i) and (ii) follow from Vidyasagar [31, p. 330]. Statement (iii) is proved in [31, p. 76] for rational matrices. An inspection of the proof in [31] shows that it only utilizes the fact that the elementary divisor theorem holds for the ring of stable rational functions; i.e., any matrix with stable rational entries is equivalent to its Smith form. Since this is also true for the ring $\mathcal{H}_\alpha$ (see [25, p. 139]), it follows that the proof in [31] carries over to matrices with entries in $\mathcal{M}_\alpha$. □

If $f \in \mathcal{M}_0$ and $V \subset \mathbb{C}_0$ is compact, let $Z(f, V)$ denote the number of zeros of f in V, counting multiplicities. Moreover, if $\gamma : [0,1] \to \mathbb{C}$ is a closed curve and $a \in \mathbb{C} \setminus \gamma([0,1])$, we denote the winding number (index) of γ around a by $\operatorname{ind}(\gamma, a)$.

PROPOSITION 7.2. *Let* $\mathbf{H}$ *be in* $\mathcal{M}_0(\mathbb{C}^{m\times m})$ *and suppose that* $\mathbf{H}(s)(I + \mathbf{H}(s))^{-1}$ *has at least one pole in* $\mathbb{C}_0$*. Then there exists* $\delta > 0$ *such that* $\mathbf{G}^\eta$ *defined by*

$$\mathbf{G}^\eta(s) = \mathbf{H}(s)(I + \operatorname{diag}_{1\le j\le m}(e^{\eta^j s})\mathbf{H}(s))^{-1}$$

has at least one pole in $\mathbb{C}_0$ *for all* $\eta = (\eta^1, \dots, \eta^m) \in \mathbb{C}^m$ *satisfying* $\|\eta\| < \delta$.

Proof. Let $\mathbf{H} = ND^{-1}$ be a right-coprime factorization over $\mathcal{H}_0$ and $s_0 \in \mathbb{C}_0$ be a pole of $\mathbf{H}(I + \mathbf{H})^{-1}$. Set $V := \{s \in \mathbb{C} \mid |s_0 - s| \le \varrho\}$ and choose $\varrho > 0$ such that

$$V \subset \mathbb{C}_0 \quad \text{and} \quad Z(\det(D + N), \partial V) \neq 0 \quad \forall\, s \in \partial V. \tag{7.1}$$

Let $\gamma_V : [0,1] \to \mathbb{C}$ be the continuous parametrization of ∂V given by $t \mapsto s_0 + \varrho e^{2\pi i t}$. For $\eta = (\eta^1, \dots, \eta^m) \in \mathbb{C}^m$ set

$$\begin{aligned} N_\eta(s) &:= \operatorname{diag}_{1\le j\le m}(e^{\eta^j s})N(s)\,, \\ \Gamma_\eta(t) &:= \det[D(\gamma_V(t)) + N_\eta(\gamma_V(t))]\,. \end{aligned}$$

It is clear that

$$\lim_{\eta\to 0}\left(\sup_{t\in[0,1]} |\Gamma_0(t) - \Gamma_\eta(t)|\right) = 0. \tag{7.2}$$

Now it follows from (7.1) that

$$\inf_{t\in[0,1]} |\Gamma_0(t)| > 0,$$

and therefore we may conclude, using (7.2), that there exists $\delta > 0$ such that

$$\inf_{t\in[0,1]} |\Gamma_\eta(t)| > 0 \text{ for all } \eta \in \mathbb{C} \text{ such that } \|\eta\| < \delta\,. \tag{7.3}$$

Choose $\eta \in \mathbb{C}^m$ with $\|\eta\| < \delta$, and define the map

$$\Lambda : [0,1] \times [0,1] \to \mathbb{C}\,,\ (t,\tau) \mapsto \Gamma_{\tau\eta}(t)\,.$$

Then Λ is continuous and, by (7.3), $0 \notin \Lambda([0,1] \times [0,1])$. Trivially, it holds that

$$\begin{aligned} \Lambda(t,0) &= \Gamma_0(t) \ \ \forall\, t \in [0,1]\,, \\ \Lambda(t,1) &= \Gamma_\eta(t) \ \ \forall\, t \in [0,1], \end{aligned}$$

and furthermore we obtain for all $\tau \in [0,1]$ that

$$\Lambda(0,\tau) = \Gamma_{\tau\eta}(0) = \Gamma_{\tau\eta}(1) = \Lambda(1,\tau)\,.$$

Thus we have shown that Γ_0 and Γ_η are homotopic in $\mathbb{C} \setminus \{0\}$, and therefore (cf. Rudin [27, Thm. 10.40]) it follows that

$$\operatorname{ind}(\Gamma_0, 0) = \operatorname{ind}(\Gamma_\eta, 0)\,. \tag{7.4}$$

Using the principle of the argument we obtain

$$Z(\det(D+N_\eta), V) = \operatorname{ind}(\Gamma_\eta, 0) = \operatorname{ind}(\Gamma_0, 0) = Z(\det(D+N), V)\,. \tag{7.5}$$

Now, $s_0 \in V \subset \mathbb{C}_0$ is a pole of $\mathbf{H}(I+\mathbf{H})^{-1}$ or equivalently $\det(D(s_0)+N(s_0)) = 0$, and thus, by (7.5)

$$Z(\det(D+N_\eta), V) = Z(\det(D+N), V) \neq 0\,. \tag{7.6}$$

It is easy to see that $\mathbf{G}^\eta = N(D+N_\eta)^{-1}$ is a right-coprime factorization over $\mathcal{H}_0$, and thus it follows from (7.6) that $\mathbf{G}^\eta$ has a pole in $V \subset \mathbb{C}_0$. □

Combining Proposition 7.2 and Theorem 5.3 we obtain the main result of this section, a "robust" version of Theorem 5.3.

THEOREM 7.3. *Let $\mathbf{H}(s)$ be a $\mathbb{C}^{m\times m}$-valued regular transfer function and, for $\varepsilon = (\varepsilon^1, \dots, \varepsilon^m) \in \mathbb{R}_+^m$, $\mathbf{G}^\varepsilon$ be given by* (6.7). *If the conditions* (1)–(3) *of Theorem* 5.3 *are satisfied, then there exist sequences (ε_n) and (δ_n) with $\varepsilon_n > 0$, $\varepsilon_n \to 0$, $\delta_n \in (0, \varepsilon_n)$ and such that $\mathbf{G}^\varepsilon$ has poles in $\mathbb{C}_0$ for all $\varepsilon \in \bigcup_{n\in\mathbb{N}} (\varepsilon_n - \delta_n, \varepsilon_n + \delta_n)^m$.*

It is clear that Theorem 5.6 can be strengthened in a similar way.

8. Dynamic output feedback. In this section we apply our results to systems with dynamic output feedback. In particular we show that—roughly speaking—for a plant with infinitely many unstable poles there does *not* exist any stabilizing (dynamic) output feedback compensator such that the stability of the closed-loop system is robust with respect to small delays.

DEFINITION 8.1. *If $\mathbf{P} \in \mathcal{M}_\alpha(\mathbb{C}^{p\times m})$ and $\mathbf{K} \in \mathcal{M}_\alpha(\mathbb{C}^{m\times p})$ for some $\alpha \in \mathbb{R}$, we say that* $\mathbf{K}$ stabilizes $\mathbf{P}$ *if* $\det(I + \mathbf{P}(s)\mathbf{K}(s)) \not\equiv 0$ *and*

$$\begin{pmatrix} I & \mathbf{P} \\ -\mathbf{K} & I \end{pmatrix}^{-1} \in H^\infty(\mathbb{C}^{(m+p)\times(m+p)}). \tag{8.1}$$

It follows from a well-known formula of Frobenius (see Gantmacher [16, p. 73]) that $\mathbf{K}$ stabilizes $\mathbf{P}$ if and only if $\det(I + \mathbf{P}(s)\mathbf{K}(s)) \not\equiv 0$ and the transfer function

$$F(\mathbf{P},\mathbf{K}) = \begin{pmatrix} \mathbf{K}(I+\mathbf{P}\mathbf{K})^{-1} & -\mathbf{K}\mathbf{P}(I+\mathbf{K}\mathbf{P})^{-1} \\ \mathbf{P}\mathbf{K}(I+\mathbf{P}\mathbf{K})^{-1} & \mathbf{P}(I+\mathbf{K}\mathbf{P})^{-1} \end{pmatrix} \tag{8.2}$$

is in $H^\infty(\mathbb{C}^{(m+p)\times(m+p)})$. Note that $F(\mathbf{P},\mathbf{K})$ is the transfer function from (u_1,u_2) to (y_1,y_2) of the feedback system shown in Fig. 4, if we take there $\varepsilon = 0$. If $\mathbf{K} \in H^\infty(\mathbb{C}^{m\times p})$, then $\mathbf{K}$ stablizes $\mathbf{P}$ if and only if $\mathbf{P}(I+\mathbf{K}\mathbf{P})^{-1} \in H^\infty(\mathbb{C}^{p\times m})$. The next result follows trivially from Theorems 5.3 and 5.6.

COROLLARY 8.2. *Let $\mathbf{P}$ and $\mathbf{K}$ be matrix-valued transfer functions of size $p \times m$ and $m \times p$, respectively. Suppose that $\mathbf{P}\mathbf{K}$ is regular, and for $\varepsilon \geq 0$ set $\mathbf{K}_\varepsilon(s) := e^{-\varepsilon s}\mathbf{K}(s)$ and define*

$$F^\varepsilon(\mathbf{P},\mathbf{K}) := \begin{pmatrix} \mathbf{K}(I+\mathbf{P}\mathbf{K}_\varepsilon)^{-1} & -\mathbf{K}_\varepsilon\mathbf{P}(I+\mathbf{K}_\varepsilon\mathbf{P})^{-1} \\ \mathbf{P}\mathbf{K}(I+\mathbf{P}\mathbf{K}_\varepsilon)^{-1} & \mathbf{P}(I+\mathbf{K}_\varepsilon\mathbf{P})^{-1} \end{pmatrix}. \tag{8.3}$$

Then, if $\mathbf{K}$ *stabilizes* $\mathbf{P}$ *(i.e.,* $F^0(\mathbf{P},\mathbf{K}) \in H^\infty(\mathbb{C}^{(m+p)\times(m+p)})$*) and*

$$\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} r(\mathbf{P}(s)\mathbf{K}(s)) = \limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} r(\mathbf{K}(s)\mathbf{P}(s)) > 1,$$

there exist sequences (ε_n) *and* (p_n) *with*

$$\varepsilon_n > 0\,,\ \varepsilon_n \to 0\,,\ p_n \in \mathbb{C}_0\,,\ |\mathrm{Im}\, p_n| \to \infty$$

and such that for any $n \in \mathbb{N}$, p_n *is a pole of* $\mathbf{PK}(I+\mathbf{PK}_{\varepsilon_n})^{-1}$ *and hence of the overall closed-loop transfer function* $F^{\varepsilon_n}(\mathbf{P},\mathbf{K})$.

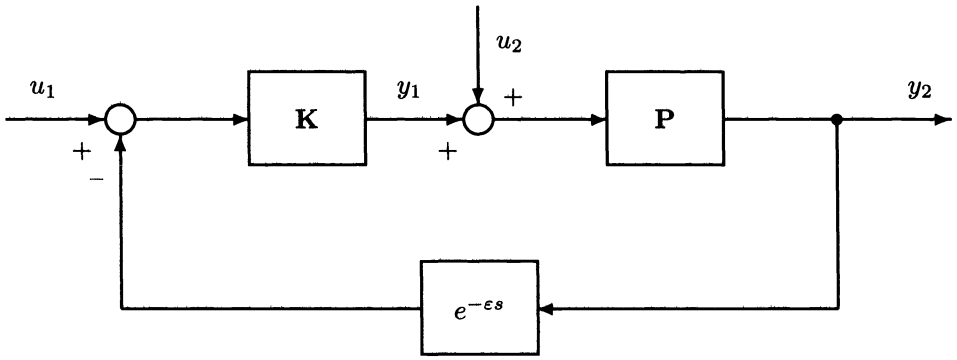


FIG. 4. *Feedback system with plant, compensator, and delay.*

The feedback system corresponding to $F^\varepsilon(\mathbf{P},\mathbf{K})$ is shown in Fig. 4; in particular we have that $(y_1,y_2)^T = F^\varepsilon(\mathbf{P},\mathbf{K})(u_1,u_2)^T$. It is clear that $F(\mathbf{P},\mathbf{K}_\varepsilon)$ is L^2-stable if $F^\varepsilon(\mathbf{P},\mathbf{K})$ is. Conversely, under the assumptions that $\mathbf{P}$ and $\mathbf{K}$ are well posed and $\mathbf{P}$ stabilizes $\mathbf{K}$,[1] it is easy to show that $F^\varepsilon(\mathbf{P},\mathbf{K})$ is L^2-stable if $F(\mathbf{P},\mathbf{K}_\varepsilon)$ is.

In order to apply Theorem 6.1 to systems with dynamic output feedback we need the following lemma.

LEMMA 8.3. *For some* $\alpha \in \mathbb{R}$ *let* $\mathbf{P}$ *and* $\mathbf{K}$ *be in* $\mathcal{M}_\alpha(\mathbb{C}^{p\times m})$ *and* $\mathcal{M}_\alpha(\mathbb{C}^{m\times p})$*, respectively, and suppose that* $\mathbf{K}$ *stabilizes* $\mathbf{P}$. *If* $U \subset \mathbb{C}_0$ *and*

$$\sup_{s\in U} \|\mathbf{P}(s)\| = \infty \ \text{ or } \ \sup_{s\in U} \|\mathbf{K}(s)\| = \infty\,,$$

then it follows that $\sup_{s\in U} \|\mathbf{P}(s)\mathbf{K}(s)\| = \infty$.

Proof. From the assumption that $\mathbf{K}$ stabilizes $\mathbf{P}$ it follows that the entries of $\mathbf{P}$ and $\mathbf{K}$ belong to the the quotient field of H^∞; i.e., they can be written as the fraction of two H^∞-functions. Moreover, it follows from Smith [30] that $\mathbf{P}$ and $\mathbf{K}$ both have right- and left-coprime factorizations over H^∞. This means in particular that there exist matrices $\tilde{N}_\mathbf{P}$, $\tilde{D}_\mathbf{P}$, $\tilde{X}_\mathbf{P}$, $\tilde{Y}_\mathbf{P}$, $N_\mathbf{K}$, $D_\mathbf{K}$, $X_\mathbf{K}$, and $Y_\mathbf{K}$ with entries in H^∞ satisfying

$$\begin{aligned} \mathbf{P} &= \tilde{D}_\mathbf{P}^{-1}\tilde{N}_\mathbf{P}\,, & \tilde{N}_\mathbf{P}\tilde{X}_\mathbf{P} + \tilde{D}_\mathbf{P}\tilde{Y}_\mathbf{P} &= I,\\ \mathbf{K} &= N_\mathbf{K}D_\mathbf{K}^{-1}\,, & X_\mathbf{K}N_\mathbf{K} + Y_\mathbf{K}D_\mathbf{K} &= I. \end{aligned}$$

Moreover, since by assumption the closed-loop system is stable, it follows trivially that I stabilizes $\mathbf{PK}$. Therefore, using again the result in Smith [30], we conclude that

[1] If $\mathbf{PK}$ is well posed and $\mathbf{P}$ stabilizes $\mathbf{K}$, then Lemma 8.3 shows that $\mathbf{P}$ and $\mathbf{K}$ are well posed.

$\mathbf{PK}$ admits a right-coprime factorization over H^∞; i.e., there exist matrices N, D, X, and Y with entries in H^∞ such that

$$\mathbf{PK} = ND^{-1}, \quad XN + YD = I. \tag{8.4}$$

It is well known (see Vidyasagar [31, p. 364]) that closed-loop stability is equivalent to

$$\inf_{s\in\mathbb{C}_0} |\det(\tilde{D}_{\mathbf{P}}(s)D_{\mathbf{K}}(s) + \tilde{N}_{\mathbf{P}}(s)N_{\mathbf{K}}(s))| > 0. \tag{8.5}$$

Let us assume that $\sup_{s\in U} \|\mathbf{P}(s)\| = \infty$. Then there exists a sequence (s_n) in U such that $\lim_{n\to\infty} \|\mathbf{P}(s_n)\| = \infty$, and hence, using the boundedness of $\tilde{D}_{\mathbf{P}}(s)$ and $\tilde{N}_{\mathbf{P}}(s)$, we obtain

$$\lim_{n\to\infty} \det(\tilde{D}_{\mathbf{P}}(s_n)) = 0. \tag{8.6}$$

Realizing that

$$\begin{aligned} \det(\tilde{D}_{\mathbf{P}}D_{\mathbf{K}} + \tilde{N}_{\mathbf{P}}N_{\mathbf{K}}) &= \det(\tilde{D}_{\mathbf{P}})\det(D_{\mathbf{K}})\det(I + \mathbf{PK}) \\ &= \frac{\det(\tilde{D}_{\mathbf{P}})\det(D_{\mathbf{K}})}{\det(D)}\det(D + N) \end{aligned} \tag{8.7}$$

and combining (8.5)–(8.7) we see that

$$\lim_{n\to\infty} \det(D(s_n)) = 0. \tag{8.8}$$

Moreover, using (8.4), it follows that

$$\det(X\mathbf{PK} + Y) = \frac{1}{\det(D)}. \tag{8.9}$$

Finally, (8.8), (8.9), and the boundedness of the matrices $X(s)$ and $Y(s)$ imply that $\lim_{n\to\infty} \|\mathbf{P}(s_n)\mathbf{K}(s_n)\| = \infty$ and thus $\sup_{s\in U} \|\mathbf{P}(s)\mathbf{K}(s)\| = \infty$. With a similar argument we can prove the claim if we assume that $\sup_{s\in U} \|\mathbf{K}(s)\| = \infty$. □

COROLLARY 8.4. *Let* $\mathbf{P} \in \mathcal{M}_\alpha(\mathbb{C}^{p\times m})$ *and* $\mathbf{K} \in \mathcal{M}_\alpha(\mathbb{C}^{m\times p})$ *for some* $\alpha \in \mathbb{R}$. *If* $F^0(\mathbf{P},\mathbf{K}) \in H^\infty(\mathbb{C}^{(m+p)\times(m+p)})$ *and*

$$\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} r(\mathbf{P}(s)\mathbf{K}(s)) < 1, \tag{8.10}$$

then there exist numbers $\varepsilon_0 > 0$ *and* $M > 0$ *such that* $F^\varepsilon(\mathbf{P},\mathbf{K}) \in H^\infty(\mathbb{C}^{(m+p)\times(m+p)})$ *and* $\|F^\varepsilon(\mathbf{P},\mathbf{K})\|_\infty \le M$ *for all* $\varepsilon \in [0,\varepsilon_0]$.

Proof. For $\varrho > 0$ let $\mathbb{C}_0^\varrho$ be defined by (6.3). Combining (8.10), the fact that $\mathbf{PK}(I + \mathbf{PK})^{-1} \in H^\infty(\mathbb{C}^{p\times p})$, and Lemma 6.3, we see that there exists numbers $R_1 > 0$ and $L_1 > 0$ such that

$$\|\mathbf{P}(s)\mathbf{K}(s)\| \le L_1 \ \forall\, s \in \mathbb{C}_0^{R_1}.$$

Hence we obtain using Lemma 8.3 that

$$\|\mathbf{P}(s)\| \le L_2 \text{ and } \|\mathbf{K}(s)\| \le L_2 \ \forall\, s \in \mathbb{C}_0^{R_2}, \tag{8.11}$$

where L_2 and R_2 are suitable positive constants. By Theorem 6.1 there exist numbers $\varepsilon_1 > 0$ and $M_1 > 0$ such that

$$\|\mathbf{PK}(I + \mathbf{PK}_\varepsilon)^{-1}\|_\infty \leq M_1 \quad \forall \varepsilon \in [0, \varepsilon_1] \tag{8.12}$$

and so

$$\|(I + \mathbf{PK}_\varepsilon)^{-1}\|_\infty \leq 1 + M_1 \quad \forall \varepsilon \in [0, \varepsilon_1]\,.$$

Therefore, and by (8.11), there exists $\tilde{M}_2 > 0$ such that

$$\|\mathbf{K}(s)(I + \mathbf{P}(s)\mathbf{K}_\varepsilon(s))^{-1}\| \leq \tilde{M}_2 \quad \forall s \in \mathbb{C}_0^{R_2}\,,\ \forall \varepsilon \in [0, \varepsilon_1]\,. \tag{8.13}$$

Setting $\mathbf{L}_\varepsilon := \mathbf{K}(I + \mathbf{PK}_\varepsilon)^{-1}$, we have $\mathbf{L}_0 \in H^\infty(\mathbb{C}^{m\times p})$ and $\mathbf{PL}_0 \in H^\infty(\mathbb{C}^{p\times p})$. Choosing $\varepsilon_2 \in (0, \varepsilon_1]$ such that for any $s \in \mathbb{C}_0$ with $|s| \leq R_2$ and any $\varepsilon \in [0, \varepsilon_2]$

$$|1 - e^{-\varepsilon s}| \leq \frac{1}{2\|\mathbf{PL}_0\|_\infty}$$

and realizing that

$$\mathbf{L}_\varepsilon(s) = \mathbf{L}_0(s)[I - (1 - e^{-\varepsilon s})\mathbf{P}(s)\mathbf{L}_0(s)]^{-1},$$

we obtain that for all s and ε as above

$$\|\mathbf{L}_\varepsilon(s)\| \leq 2\|\mathbf{L}_0\|_\infty\,. \tag{8.14}$$

Combining (8.13) and (8.14) shows that

$$\|\mathbf{K}(I + \mathbf{PK}_\varepsilon)^{-1}\|_\infty \leq M_2 \quad \forall \varepsilon \in [0, \varepsilon_2]\,, \tag{8.15}$$

where $M_2 := \max(\tilde{M}_2, 2\|\mathbf{L}_0\|_\infty)$. Finally, using similar arguments, it can be shown that

$$\|\mathbf{K}_\varepsilon\mathbf{P}(I + \mathbf{K}_\varepsilon\mathbf{P})^{-1}\|_\infty \leq M_3 \quad \forall \varepsilon \in [0, \varepsilon_3] \tag{8.16}$$

and

$$\|\mathbf{P}(I + \mathbf{K}_\varepsilon\mathbf{P})^{-1}\|_\infty \leq M_4 \quad \forall \varepsilon \in [0, \varepsilon_4]\,, \tag{8.17}$$

where M_3, M_4, ε_3, and ε_4 are suitable positive numbers. The claim now follows from (8.12) and (8.15)–(8.17). □

Using Corollary 8.2 and Lemma 8.3 it is easy to give the proof of Theorem 1.2. More precisely, we prove the following result which is slightly stronger than Theorem 1.2.

THEOREM 8.5. *Let* $\mathbf{P} \in \mathcal{M}_\alpha(\mathbb{C}^{p\times m})$ *and* $\mathbf{K} \in \mathcal{M}_\alpha(\mathbb{C}^{m\times p})$ *for some* $\alpha \in \mathbb{R}$, *and suppose that* $\mathbf{PK}$ *is regular. Then, if* $\mathbf{K}$ *stabilizes* $\mathbf{P}$ *and* $\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} \|\mathbf{P}(s)\| = \infty$, *there exist sequences* (ε_n) *and* (p_n) *with*

$$\varepsilon_n > 0\,,\ \varepsilon_n \to 0\,,\ p_n \in \mathbb{C}_0\,,\ |\mathrm{Im}\, p_n| \to \infty$$

and such that, for any $n \in \mathbb{N}$, p_n *is a pole of* $\mathbf{PK}(I + \mathbf{PK}_{\varepsilon_n})^{-1}$ *and hence of the overall closed-loop transfer function* $F^{\varepsilon_n}(\mathbf{P}, \mathbf{K})$ *given by* (8.3).

Proof. Since $\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} \|\mathbf{P}(s)\| = \infty$ and $\mathbf{K}$ stabilizes $\mathbf{P}$, it follows from Lemma 8.3 that $\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} \|\mathbf{P}(s)\mathbf{K}(s)\| = \infty$. Now, by assumption, $\mathbf{PK}(I +$

$\mathbf{PK})^{-1} \in H^\infty(\mathbb{C}^{p\times p})$, and hence an application of Lemma 6.3 and Corollary 8.2 yields the claim. □

The following remark shows that for a large class of transfer functions which are bounded at high frequencies there always exists a stabilizing compensator such that the stability of the closed loop is robust with respect to small delays.

Remark 8.6. Define $\mathcal{T} := \bigcup_{\alpha<0} H_\alpha^\infty + \mathcal{R}_{spu}$, where $\mathcal{R}_{spu}$ denotes the ring of strictly proper totally unstable rational functions, i.e., $\mathcal{R}_{spu} := \{f \in \mathbb{C}(s) \mid f(\infty) = 0 \text{ and } f(s) \neq \infty \text{ for all } s \in \mathbb{C} \setminus \mathbb{C}_0^{cl}\}$. Note that if $\mathbf{P} \in \mathcal{T}^{p\times m}$, then

$$\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} \|\mathbf{P}(s)\| < \infty,$$

which implies in particular that $\mathbf{P}$ has at most finitely many poles in $\mathbb{C}_0^{cl}$. The ring $\mathcal{T}$ contains the so-called Callier–Desoer ring of transfer functions (cf. Callier and Desoer [4],[5]). It is known that for any $\mathbf{P} \in \mathcal{T}^{p\times m}$ there exists a *strictly proper rational* compensator $\mathbf{K}$ such that $F(\mathbf{P},\mathbf{K}) \in H^\infty(\mathbb{C}^{(m+p)\times(m+p)})$; see Logemann [24] and the references therein. Combining this result with Corollary 8.4, it follows that for any $\mathbf{P} \in \mathcal{T}^{p\times m}$ there exists a compensator $\mathbf{K} \in \mathcal{T}^{m\times p}$ and a number $\varepsilon_0 > 0$ such that $F^\varepsilon(\mathbf{P},\mathbf{K}) \in H^\infty(\mathbb{C}^{(m+p)\times(m+p)})$ for all $\varepsilon \in [0,\varepsilon_0]$.

Remark 8.7. We claim that the conclusions of Theorem 8.5 do not remain true if the assumption $\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} \|\mathbf{P}(s)\| = \infty$ is replaced by the weaker assumption that there exist a sequence of poles of $\mathbf{P}$ in the open left half-plane going to ∞ tangentially along the imaginary axis. To this end let $\mathbf{P}$ be the transfer function of the following neutral system:

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) - \dot{x}_2(t-h) \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 1 & -a \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} u(t)$$
$$y(t) = x_2(t)\,,$$

i.e.,

$$\mathbf{P}(s) = \frac{1}{s+1}\,\frac{1}{s(1-e^{-hs})+a}\,,$$

where $a, h > 0$. It is shown in Logemann [23] that $\mathbf{P} \in H^\infty$. Trivially, for any compensator $\mathbf{K} \in H^\infty$ satisfying $\|\mathbf{PK}\|_\infty < 1$ the closed-loop transfer function $F^\varepsilon(\mathbf{P},\mathbf{K})$ is in $H^\infty(\mathbb{C}^{2\times 2})$ for all $\varepsilon \geq 0$. However, using Rouché's theorem, it is not difficult to show that there exist a sequence of poles $p_n \in \mathbb{C} \setminus \mathbb{C}_0^{cl}$ of $\mathbf{P}$ and numbers $\ell_n \in \mathbb{N}$ with $\ell_n \to \infty$ as $n \to \infty$ such that

$$\lim_{n\to\infty} |p_n - i\frac{2\pi}{h}\ell_n| = 0\,.$$

9. Examples. In this section we illustrate Theorem 1.1 with three examples.

Example 9.1. In this example we analyze the robustness with respect to delays for a damped wave equation. For $x \in (0,1)$ and $t > 0$ we consider the following system:

$$w_{tt}(x,t) - w_{xx}(x,t) + 2aw_t(x,t) + a^2 w(x,t) = 0\,, \tag{9.1}$$

$$w(0,t) = 0\,, \qquad w_x(1,t) = u(t)\,, \tag{9.2}$$

$$y(t) = kw_t(1,t)\,. \tag{9.3}$$

We assume here that the viscous damping parameter a is nonnegative and the boundary damping parameter k is positive. It is known that the feedback control

$$u(t) = -y(t) \tag{9.4}$$

exponentially stabilizes the system (see, for instance, Chen [6]). Hence, if the transfer function of (9.1)–(9.3) is denoted by $\mathbf{H}$, then it follows that $\mathbf{H}(I+\mathbf{H})^{-1} \in H^\infty$. An easy computuation shows that $\mathbf{H}$ is given by

$$\mathbf{H}(s) = \frac{ks}{s+a}\left(\frac{1-e^{-2(s+a)}}{1+e^{-2(s+a)}}\right).$$

In Datko, Lagnese, and Polis [9] the robustness of the closed loop system (9.1)–(9.4) with respect to small delays was analyzed. We will obtain frequency domain versions of their results, using Theorem 1.1. We need to compute γ as defined by (1.3) for this system.

CLAIM.

$$\gamma = \limsup_{|s|\to\infty, s\in\mathbb{C}_0} |\mathbf{H}(s)| = k\frac{e^{2a}+1}{e^{2a}-1}\,.$$

Proof. The following simple estimates are clear for Re $s > 0$:

$$|1-e^{-2(s+a)}| \le 1+e^{-2a}\,, \qquad |1+e^{-2(s+a)}| \ge 1-e^{-2a}\,, \qquad \left|\frac{s}{s+a}\right| \le 1\,.$$

These estimates show that $\gamma \le k(e^{2a}+1)/(e^{2a}-1)$. To obtain the opposite inequality, let $s_n = (1/n) + i(2n+1)\pi/2$ for $n \in \mathbb{N}$. Then

$$\lim_{n\to\infty} \mathbf{H}(s_n) = \lim_{n\to\infty} \frac{ks_n}{s_n+a}\frac{1+e^{-2/n}e^{-2a}}{1-e^{-2/n}e^{-2a}} = k\frac{e^{2a}+1}{e^{2a}-1}\,.$$

This shows that $\gamma \ge k(e^{2a}+1)/(e^{2a}-1)$, completing the proof of the claim. □

Let us apply Theorem 1.1 to this system. We consider two cases.

Case 1: $k \ge 1$. In this case $\gamma > 1$ for any $a \ge 0$, so the transfer function $\mathbf{H}(s)(I+e^{-\varepsilon s}\mathbf{H}(s))^{-1}$ has poles in $\mathbb{C}_0$ for arbitrarily small $\varepsilon > 0$.

Case 2: $k < 1$. In this case $\gamma > 1$ if and only if

$$a < \frac{1}{2}\ln\frac{1+k}{1-k}\,.$$

If a satisfies this estimate, then the same conclusion as in case 1 holds. When

$$a > \frac{1}{2}\ln\frac{1+k}{1-k}\,,$$

the delayed feedback system is L^2-stable for all sufficiently small delays.

Example 9.2. We consider the following first-order neutral system:

$$\dot{x}(t) - a\dot{x}(t-h) = -bx(t) + u(t)\,,$$
$$y(t) = c\dot{x}(t-h)\,.$$

Here $a \geq 1$, $b > 0$, $c \in \mathbb{R}$, and $h > 0$. We consider the feedback $u(t) = -y(t)$, so the free dynamics of the closed loop are described by

$$\dot{x}(t) + (c-a)\dot{x}(t-h) = -bx(t)\,.$$

This system is exponentially stable if $|c-a| < 1$. The open-loop transfer function is

$$\mathbf{H}(s) = \frac{sce^{-hs}}{s(1-ae^{-sh})+b}.$$

$\mathbf{H}$ is clearly well posed and regular with feedthrough 0. If $a > 1$, then the equation $1-ae^{-sh} = 0$ has a zero at $s = \log(a)/h$, which is in $\mathbb{C}_0$. Hence, by a result in Salamon [28, p. 160], the characteristic equation $s(1-ae^{-sh})+b = 0$ has infinitely many zeros in $\mathbb{C}_0$. (This follows also directly from the periodicity of $1-ae^{-hs}$ and an application of Rouché's theorem.) As a consequence $\gamma = \infty > 1$, so the closed-loop stability is destroyed by arbitrarily small delays. If $a = 1$, then the equation $1-e^{-sh} = 0$ has a zero at $s = 0$. It is easy to see that $\mathbf{H}(s)$ has no poles in $\mathbb{C}_0^{\mathrm{cl}}$. However, as shown in Logemann [23], we have that $\limsup_{\omega\to\infty} |\mathbf{H}(i\omega)| = \infty$. Hence $\gamma = \infty > 1$, and so the closed-loop system is not robustly stable with respect to delays.

Example 9.3. In this example the input space and the output space are $\mathbb{R}^2$. We consider two coupled vibrating strings, one with spatial extent $0 \leq x \leq 1$ and the other with spatial extent $1 \leq x \leq 2$. Each string satisfies the damped wave equation

$$w_{tt}(x,t) - w_{xx}(x,t) + 2aw_t(x,t) + a^2 w(x,t) = 0\,, \quad x \in (0,1)\cup(1,2)\,,$$

where the viscous damping parameter $a \geq 0$. At the linkage we assume the displacement is continuous, so

$$w(1^-,t) = w(1^+,t)\,,$$

and we set the discontinuity of the vertical tension force equal to a control variable:

$$w_x(1^-,t) - w_x(1^+,t) = u_1(t)\,.$$

We take the right endpoint fixed, and at the left endpoint we set the vertical tension force equal to another control variable, leading to

$$w(2,t) = 0\,, \qquad w_x(0,t) = u_2(t)\,.$$

We take one observation proportional to the velocity at the linkage, and the other observation negatively proportional to the velocity at the left endpoint, leading to

$$y_1(t) = k_1 w_t(1,t)\,, \qquad y_2(t) = -k_2 w_t(0,t)\,,$$

for $k_1, k_2 \geq 0$. Let $u(t) = [u_1(t), u_2(t)]^T$ and $y(t) = [y_1(t), y_2(t)]^T$. The transfer function $\mathbf{H}$ for this system can be computed to be

$$\mathbf{H}(s) = \frac{s}{s+a}\begin{pmatrix} -\dfrac{k_1}{2}\dfrac{e^{-4(s+a)}-1}{e^{-4(s+a)}+1} & k_1\dfrac{e^{-(s+a)}(e^{-2(s+a)}-1)}{e^{-4(s+a)}+1} \\ k_2\left(\dfrac{e^{-(s+a)}(e^{-2(s+a)}-1)}{e^{-4(s+a)}+1}\right) & k_2\left(\dfrac{1-e^{-4(s+a)}}{e^{-4(s+a)}+1}\right) \end{pmatrix}.$$

TABLE 1
Values of γ for given values of a, k_1, and k_2.

(k_1, k_2)	$a = .1$	$a = .25$	$a = .5$	$a = 3$	$a = 10$
(.1, .1)	.7562	.3146	.1755	.1005	.1
(.1, .5)	2.7790	1.1714	.6881	.5003	.5
(.1, 1)	5.3112	2.2513	1.3426	1.0090	1
(.5, .1)	1.7647	.7348	.4120	.2508	.25
(.5, .5)	3.7810	1.5730	.8777	.5024	.5
(.5, 1)	6.3051	2.6337	1.4992	1.0016	1
(1, .1)	3.0278	1.2673	.7274	.5006	.5
(1, .5)	5.0410	2.0957	1.1633	.5351	.5003
(1, 1)	7.5621	3.1461	1.7553	1.0049	1

Clearly, $\mathbf{H}$ is regular with feedthrough matrix

$$D = \begin{pmatrix} k_1/2 & 0 \\ 0 & k_2 \end{pmatrix}.$$

It is not hard to show that for any values of $a \geq 0$, $k_1 \geq 0$, $k_2 \geq 0$, $a + k_1 + k_2 > 0$ the closed-loop transfer function $\mathbf{H}(I+\mathbf{H})^{-1}$ is in $H^\infty(\mathbb{C}^{2\times 2})$. In the case when $a = 0$ this follows also from the fact that the closed-loop semigroup is shown in Liu, Huang, and Chen [22] to be exponentially stable.

There are some values of k_1, k_2, and a where no further computation needs to be done in order to apply the results in the preceeding sections. If $k_1 > 2$ or $k_2 > 1$, the spectral radius of D is greater than 1, so Theorem 5.6 implies that there exists $\varepsilon_n \downarrow 0$ such that $\mathbf{G}^{\varepsilon_n}(s) = \mathbf{H}(s)(I + e^{-\varepsilon_n s}\mathbf{H}(s))^{-1}$ has poles $p_n \in \mathbb{C}_0$ such that the real and imaginary parts of p_n go to infinity as n goes to infinity. Another simple case is when $a = 0$ and $k_1 + k_2 > 0$. In this case $\mathbf{G}^0$ is stable and $\mathbf{H}$ has poles at $s = \pi i(1 + 2n)/4$ for all integers n. Thus we obtain from Lemma 6.3 that $\gamma = \limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} r(\mathbf{H}(s)) = \infty$, and hence, by Theorem 1.1, $\mathbf{G}^0$ is not robustly stable with respect to delays.

In the case where $a > 0$, $0 < k_1 < 2$, $0 < k_2 < 1$ we need to compute γ. First note that γ is the same for $\tilde{\mathbf{H}}(s) := ((s + a)/s)\mathbf{H}(s)$ as it is for $\mathbf{H}(s)$. To compute the spectral radius of $\tilde{\mathbf{H}}(s)$, we need to compute the eigenvalues of $\tilde{\mathbf{H}}(s)$. These are found to be

$$\eta_\pm(k_1, k_2, s, a) = -\frac{k_1 + 2k_2}{4}\frac{1 - e^{4(s+a)}}{1 + e^{4(s+a)}} \pm \frac{1}{4(1 + e^{4(s+a)})}\sqrt{g(k_1, k_2, s, a)},$$

where

$$g(k_1, k_2, s, a) = (k_1 - 2k_2)^2 + 16k_1k_2e^{2(s+a)} - 2e^{4(s+a)}(k_1^2 + 12k_1k_2 + 4k_2^2) \\ +16k_1k_2e^{6(s+a)} + e^{8(s+a)}(k_1 - 2k_2)^2.$$

Since $\eta_\pm(k_1, k_2, s + \pi i, a) = \eta_\pm(k_1, k_2, s, a)$ and $\tilde{\mathbf{H}}$ is in $H^\infty(\mathbb{C}^{2\times 2})$, we obtain, using Remark 6.2, that

$$\gamma = \limsup_{|s|\to\infty, s\in\mathbb{C}_0} r(\tilde{\mathbf{H}}(s)) = \sup_{s\in\mathbb{C}_0} r(\tilde{\mathbf{H}}(s)) = \sup_{\omega\in\mathbb{R}} r(\tilde{\mathbf{H}}(i\omega)) = \sup_{0\leq\omega\leq\pi} r(\tilde{\mathbf{H}}(i\omega)).$$

Thus, computing γ is a fairly straightfoward numerical problem. Using Mathematica, we obtain Table 1, giving values of γ for some values of k_1, k_2, and a. As we see from the table, the possibility of robustness increases as a increases and decreases as

k_1 and k_2 increase. Note that the last column, with $a = 10$, is almost the same as that obtained by taking the limit of γ as $a \to \infty$, which is easily computed to be $\max\{k_2, k_1/2\}$. Thus, for large values of the viscous damping coefficient a, robustness is determined in a simple way by k_1 and k_2.

Acknowledgments. We would like to thank Bengt Mårtensson (Bremen) for a helpful discussion on Example 6.4 and Fabian Wirth (Bremen) for some useful remarks which led to an improvement of an earlier version of Theorem 6.5.

REFERENCES

[1] J. F. BARMAN, F. M. CALLIER, AND C. A. DESOER, *L^2-stability and L^2-instability of linear time-invariant distributed feedback systems perturbed by a small delay in the loop*, IEEE Trans. Automat. Control, 18 (1973), pp. 479–484.

[2] H. BAUMGÄRTEL, *Analytic Perturbation Theory for Matrices and Operators*, Birkhäuser, Basel, 1985.

[3] J. BONTSEMA AND S. A. DE VRIES, *Robustness of flexible systems against small time delays*, in Proc. 27th Conference on Decision and Control, Austin, Texas, Dec. 1988.

[4] F. M. CALLIER AND C. A. DESOER, *An algebra of transfer functions for distributed linear time-invariant systems*, IEEE Trans. Circuits Systems, 25 (1978), pp. 651–662. (Correction: 26 (1979), p. 360.)

[5] ———, *Simplifications and clarifications on the paper "An algebra of transfer functions for distributed linear time-invariant systems,"* IEEE Trans. Circuits Systems, 27 (1980), pp. 320–323.

[6] G. CHEN, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–274.

[7] P. M. COHN, *Algebra*, Vol. 1, Wiley, New York, 1974.

[8] R. CURTAIN, *A synthesis of time and frequency domain methods for the control of infinite-dimensional systems: A system theoretic approach*, in Control and Estimation in Distributed Parameter Systems, H. T. Banks, ed., Frontiers in Applied Mathematics, Vol. 11, Society for Industrial and Applied Mathematics, Philadelphia, 1992, pp. 171–224.

[9] R. DATKO, J. LAGNESE, AND M. P. POLIS, *An example of the effect of time delays in boundary feedback stabilization of wave equations*, SIAM J. Control Optim., 24 (1986), pp. 152–156.

[10] R. DATKO, *Not all feedback stabilized hyperbolic systems are robust with respect to small time delays in their feedbacks*, SIAM J. Control Optim., 26 (1988), pp. 697–713.

[11] ———, *The destabilizing effect of delays on certain vibrating systems*, in Advances in Computing and Control, W. A. Porter, S. C. Kak, and J. L. Aravena, eds., Lecture Notes in Control and Inform. Sci. 130, Springer-Verlag, New York, 1989.

[12] W. DESCH AND R. L. WHEELER, *Destabilization due to delay in one-dimensional feedback*, in Control and Estimation of Distributed Parameter Systems, Internat. Ser. Numer. Math., Vol. 91, F. Kappel, K. Kunisch and W. Schappacher, eds., Birkhäuser-Verlag, Boston, 1989.

[13] W. DESCH, K. B. HANNSGEN, Y. RENARDY, AND R. L. WHEELER, *Boundary stabilization of an Euler–Bernoulli beam with viscoelastic damping*, in Proc. 26th Conference on Decision and Control, Los Angeles, CA, Dec. 1987.

[14] P. L. DUREN, *Theory of H^p Spaces*, Academic Press, New York, 1970.

[15] O. FORSTER, *Lectures on Riemann Surfaces*, Springer-Verlag, New York, 1981.

[16] F. R. GANTMACHER, *Matrizentheorie*, Springer-Verlag, Berlin, 1986.

[17] T. GEORGIOU AND M. C. SMITH, *w-Stability of feedback systems*, Systems Control Lett., 13 (1989), pp. 271–277.

[18] ———, *Graphs, causality and stabilizability: linear, shift-invariant systems on $\mathcal{L}_2[0,\infty)$*, Math. Control, Signals, Systems, 6 (1993), pp. 195–223.

[19] R. GRIMMER, R. LENCZEWSKI, AND W. SCHAPPACHER, *Well-posedness of hyperbolic equations with delay in the boundary conditions*, in Semigroup Theory and Applications, P. Clement, S. Invernizzi, E. Mitidieri, and I. Vrabie, eds., Lecture Notes in Pure and Appl. Math. 116, Marcel Dekker, New York and Basel, 1989.

[20] K. B. HANNSGEN, Y. RENARDY, AND R. L. WHEELER, *Effectiveness and robustness with respect to time delays of boundary feedback stabilization in one-dimensional viscoelasticity*, SIAM J. Control Optim., 26 (1988), pp. 1200–1234.

[21] T. KATO, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, 1976.

[22] K. S. LIU, F. L. HUANG, AND G. CHEN, *Exponential stability analysis of a long chain of coupled vibrating strings with dissipative linkage*, SIAM J. Appl. Math., 49 (1989), pp. 1694–1707.
[23] H. LOGEMANN, *On the transfer matrix of a neutral system: Characterizations of exponential stability in input-output terms*, Systems Control Lett., 9 (1987), pp. 393–400.
[24] ———, *Stabilization and regulation of infinite-dimensional systems using coprime factorizations*, in Analysis and Optimization of Systems: State and Frequency Domain Approaches for Infinite-Dimensional Systems, R. F. Curtain, A. Bensoussan, and J. L. Lions, eds., Lecture Notes in Control and Inform. Sci. 185, Springer-Verlag, Berlin, 1993.
[25] R. NARASIMHAN, *Complex Analysis in One Variable*, Birkhäuser, Boston, 1985.
[26] A. PACKARD AND J. DOYLE, *The complex structured singular value*, Automatica, 29 (1993), pp. 71–109.
[27] W. RUDIN, *Real and Complex Analysis*, 2nd ed., McGraw-Hill, New York, 1974.
[28] D. SALAMON, *Control and Observation of Neutral Systems*, Pitman, Boston, 1984.
[29] ———, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147–164.
[30] M. C. SMITH, *On stabilization and existence of coprime factorizations*, IEEE Trans. Automat. Control, 34 (1989), pp. 1005–1007.
[31] M. VIDYASAGAR, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
[32] G. WEISS, *Transfer functions of regular linear systems, Part* I: *Characterizations of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.
[33] ———, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 601–619, March 1996

011

GENERALIZED DISCRETE-TIME RICCATI THEORY*

VLAD IONESCU† AND CRISTIAN OARĂ‡

Abstract. A Riccati-like equation, termed the generalized (discrete-time) algebraic Riccati equation, which incorporates as special cases both the standard and the constrained discrete-time algebraic Riccati equations, is introduced and investigated under the weakest possible assumptions imposed on the initial data. A complete characterization of the conditions under which such an equation of general form has a stabilizing solution is presented in terms of the so-called proper deflating subspace of the extended Hamiltonian pencil. An evaluation of an associated quadratic index along constrained stable trajectories is given in terms of the stabilizing solution to the generalized Riccati equation. Possible applications of the developed theory range from nonstandard spectral and inner-outer factorizations to H^2 and H^∞ control in singular cases. The results exposed in the present paper are the discrete-time counterpart of those stated in the authors' previous paper concerning the generalized (singular) continuous-time Riccati theory. The results could be also seen as an extension to singular cases of the usual discrete-time algebraic Riccati equation theory (of indefinite sign).

Key words. generalized Riccati equation, extended Hamiltonian pencil, proper deflating subspace, Kronecker canonical form, constrained dynamics with quadratic cost

AMS subject classifications. 11D09, 15A21, 15A22, 15A24, 47A56, 49N10, 58F05, 93C55

1. Introduction.

In the past decade the topics on the Riccati equation theory have been continuously enlarged. Thus the usual field consisting of standard Riccati equations (control and estimation) encountered in the linear quadratic theory [5], [19] has been considerably increased by considering those Riccati equations which are typically used for game theoretic situations (see, for instance, [3], [5], [25]). The unorthodox cases, such as those regarding the constrained Riccati equation used for nonstandard inner-outer factorization (see [6], [14], [15], [26]), must be also seen as significant factors that prove the progress in the field.

Among the paths that have been explored in this area, a relevant one is that based on the Popov–Yakubovich–type approach in conjunction with the matrix pencil theory (see [11], [12], [13], [16], [17]). The present paper contains the discrete counterpart of the theory developed in [17] and presents a unified approach of the various cases in which discrete-time Riccati-like equations are involved. As we shall see, the results obtained in the continuous case [17] cannot be entirely transferred to the discrete case and there are many situations when significant differences occur.

The main tool used in the paper is based on the notion of the so-called *proper deflating subspace* of the extended Hamiltonian pencil (EHP) (for EHP see, for instance, [16]). In fact the significant differences between the continuous and the discrete cases have their origin in the differences that occur in the intimate structure of the corresponding EHPs.

Subsequently the following notations will be used. The open unit disk and its closure in the complex plane will be denoted by $\mathbf{D}_1(0)$ and $\overline{\mathbf{D}}_1(0)$, respectively. $\mathbf{R}^n$ and $\mathbf{R}^{m\times n}$ will be used to denote the real n-dimensional Euclidean space and the ring of real $m \times n$ matrices, respectively. $\mathbf{R}(\lambda)$ will stand for the field of rational functions over $\mathbf{R}$. Any matrix of full column (row) rank will be called monic (epic).

* Received by the editors January 19, 1994; accepted for publication (in revised form) December 8, 1994.

† Faculty of Automatic Control and Computers, University Politehnica Bucharest, 3 Emile Zola, 71272, Bucharest, Romania.

‡ Faculty of Automatic Control and Computers, University Politehnica Bucharest, 34 Austrului, 73115, Bucharest, Romania.

The spectrum of a square matrix A will be denoted by $\Lambda(A)$. For the transpose and the Moore–Penrose pseudo inverse of a matrix A we shall use the notations A^T and $A^{\#}$, respectively. By A^{-T} we denote $(A^{-1})^T$, if A^{-1} exists. Script capital letters will be used for subspaces in $\mathbf{R}^n$. If $\mathcal{V} \subset \mathbf{R}^n$ and $A\mathcal{V} \subset \mathcal{V}$, write $A|\mathcal{V}$ for the restriction of A to $\mathcal{V}$. The image and the null space of a matrix A will be denoted by $\mathrm{Im}A$ and $\ker A$, respectively. Strict equivalence of two matrix pencils $\lambda M - N$ and $\lambda\tilde{M} - \tilde{N}$ will be denoted by $\lambda M - N \sim \lambda\tilde{M} - \tilde{N}$ and means the existence of two constant nonsingular matrices H and Z of appropriate dimensions such that $H(\lambda M - N)Z = \lambda\tilde{M} - \tilde{N}$. By $l^2(\mathbf{N}; \mathbf{R}^r)$ we shall denote the Hilbert space of norm-square-summable $\mathbf{R}^r$-valued sequences. Here $\mathbf{N}$ stands for the set of positive integers. The inner product in $l^2(\mathbf{N}; \mathbf{R}^r)$ will be denoted by $\langle , \rangle$. Irrelevant block entries of a matrix are denoted by x.

2. Definitions and basic notions. Any triplet $\Sigma = (A, B, P)$, where $A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, and $P \in \mathbf{R}^{(n+m)\times(n+m)}$, with

$$P = \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} = P^T,$$

will be called a *Popov triplet.* Such a triplet expresses synthetically a tandem which consists of a linear (discrete-time) system $\sigma x = Ax + Bu$ and an associated quadratic cost criterion $\sum_{k=0}^{\infty} w_k^T P w_k$, $w_k := [x_k^T u_k^T]^T$. Here σ stands for the unit shift operator, i.e., $(\sigma x)_k = x_{k+1}$.

DEFINITION 1. *Let Σ be a Popov triplet. The algebraic equation*

$$\begin{bmatrix} A^TXA - X + Q & A^TXB + L \\ L^T + B^TXA & R + B^TXB \end{bmatrix} \begin{bmatrix} I \\ F \end{bmatrix} V = 0, \tag{1}$$

where $X \in \mathbf{R}^{n\times n}, V \in \mathbf{R}^{n\times r}$ is monic with $r \leq n$, r unfixed, and F such that $(A+BF)V = VS$ for an appropriate matrix $S \in \mathbf{R}^{r\times r}$ will be termed the generalized discrete-time algebraic Riccati equation (GDTARE).

A quadruple (r, X, V, F) for which (1) holds and $V^TXV = V^TX^TV$ will be called a solution to GDTARE. A solution to GDTARE will be called stabilizing if, in addition, $\Lambda(S) \subset \mathbf{D}_1(0)$.

Remark 1. If (r, X, V, F) is a stabilizing solution to the GDTARE, then $\mathcal{V} = \mathrm{Im}V$ is a stable (A, B) invariant subspace. Indeed, we have directly from Definition 1 that $(A + BF)\mathcal{V} \subset \mathcal{V}$ and $\Lambda((A + BF)|\mathcal{V}) \subset \mathbf{D}_1(0)$. □

Remark 2. If V is invertible, we may assume that $V = I$ and (1) becomes

$$\begin{aligned} A^TXA - X + Q + (A^TXB + L)F = 0, \\ L^T + B^TXA + (R + B^TXB)F = 0. \end{aligned} \tag{2}$$

In this case a stabilizing solution reduces to a pair (X, F) satisfying (2) with $X = X^T$ and $A + BF$ stable. For $X = X^T$ the system (2) can be rewritten

$$\begin{aligned} A^TXA - X - F^T(R + B^TXB)F + Q = 0, \\ L^T + B^TXA + (R + B^TXB)F = 0 \end{aligned} \tag{3}$$

or as

$$\begin{aligned} A^TXA - X - (A^TXB + L)(R + B^TXB)^{\#}(B^TXA + L^T) + Q = 0, \\ \ker(R + B^TXB) \subset \ker(A^TXB + L). \end{aligned} \tag{4}$$

The form (3) is the discrete counterpart of the form encountered in [1] and [20], while the form (4) is known as the *constrained discrete-time algebraic Riccati equation* (CDTARE). (See [14] for the continuous-time case and [15] for the discrete-time case.) If, in addition, $R + B^T X B$ is invertible, (2) reduces to the well-known discrete-time algebraic Riccati equation:

$$A^T X A - X - (A^T X B + L)(R + B^T X B)^{-1}(B^T X A + L^T) + Q = 0. \tag{5}$$

Note that no assumption on the inertia of the symmetric matrix R or on the invertibility of A are made. □

For earlier works on existence, convergence, and numerical solution of Riccati equations in such standard cases as, for example, the control and estimation Riccati equations, refer to [2], [19], [21]. A very good survey of the progress in the field of the Riccati equation, including the indefinite-sign case, can be also found in [5].

With the above definitions we can now state the following proposition.

PROPOSITION 2.1. *If* (r, X, V, F) *and* $(r, \tilde{X}, V, \tilde{F})$ *are two stabilizing solutions to* (1), *then* $V^T X V = V^T \tilde{X} V$.

Proof. We may write $(A + BF)V = VS$ and $(A + B\tilde{F})V = V\tilde{S}$ with both $\Lambda(S)$ and $\Lambda(\tilde{S})$ in $\mathbf{D}_1(0)$. From (1) we have $A^T X A V - XV + QV + LFV + A^T X B F V = A^T X (A + BF)V - XV + QV + LFV = A^T X V S - XV + QV + LFV = 0$ and thus

$$V^T A^T X V S - V^T X V + V^T Q V + V^T L F V = 0. \tag{6}$$

Again using (1) we also have

$$V^T \tilde{F}^T (L^T + B^T X A)V + V^T \tilde{F}^T (R + B^T X B) F V$$

$$= V^T \tilde{F}^T B^T X (A + BF)V + V^T \tilde{F}^T L^T V + V^T \tilde{F}^T R F V = 0,$$

that is,

$$V^T \tilde{F}^T B^T X V S + V^T \tilde{F}^T L^T V + V^T \tilde{F}^T R F V = 0. \tag{7}$$

Adding (6) and (7) yields

$$\tilde{S}^T V^T X V S - V^T X V + V^T (LF + \tilde{F}^T L^T)V + V^T \tilde{F}^T R F V + V^T Q V = 0. \tag{8}$$

A similar procedure applied to (1) for $\tilde{X}$ and $\tilde{F}$ instead of X and F, respectively, yields

$$S^T V^T \tilde{X} V \tilde{S} - V^T \tilde{X} V + V^T (L\tilde{F} + F^T L^T)V + V^T F^T R \tilde{F} V + V^T Q V = 0. \tag{9}$$

As $V^T \tilde{X} V = V^T \tilde{X}^T V$, we get, by transposing (9),

$$\tilde{S}^T V^T \tilde{X} V S - V^T \tilde{X} V + V^T (\tilde{F}^T L^T + LF)V + V^T \tilde{F}^T R F V + V^T Q V = 0, \tag{10}$$

and subtracting (10) from (8) we have

$$\tilde{S}^T (V^T X V - V^T \tilde{X} V) S - (V^T X V - V^T \tilde{X} V) = 0. \tag{11}$$

Hence, as both $\Lambda(\tilde{S})$ and $\Lambda(S)$ are in $\mathbf{D}_1(0)$, the Stein equation (11) has a unique zero solution, and consequently $V^T X V = V^T \tilde{X} V$. □

Let Σ be a Popov triplet and associate with it the discrete-time linear system

$$\begin{aligned} \sigma x &= Ax + Bu, \\ \lambda &= Qx + A^T\sigma\lambda + Lu, \\ v &= L^Tx + B^T\sigma\lambda + Ru. \end{aligned} \tag{12}$$

Let $w = [x^T\ \lambda^T\ u^T]^T \in \mathbf{R}^n \times \mathbf{R}^n \times \mathbf{R}^m$. Then for $v = 0$, (12) can be written in the descriptor form $M\sigma w = Nw$, where

$$M = \begin{bmatrix} I & 0 & 0 \\ 0 & -A^T & 0 \\ 0 & -B^T & 0 \end{bmatrix}, \qquad N = \begin{bmatrix} A & 0 & B \\ Q & -I & L \\ L^T & 0 & R \end{bmatrix}; \tag{13}$$

$M, N \in \mathbf{R}^{(2n+m)\times(2n+m)}$.

DEFINITION 2 (see [13], [15]). *The matrix pencil $\lambda M - N$ with M and N defined via* (13) *is called the EHP associated with Σ.*

As we shall see in §3, the EHP introduced by Definition 2 will play a crucial role in studying the GDTARE. In order to investigate the EHP let us briefly recall some geometric notions intimately related to the general matrix pencil theory, which have been introduced and studied in [13], [17], [22], [26].

DEFINITION 3. *Let $\lambda M - N$ with $M, N \in \mathbf{R}^{p\times q}$ be an arbitrary matrix pencil. A subspace $\mathcal{V} \subset \mathbf{R}^q$ of dimension ρ will be called a proper deflating subspace to the right if*

$$NV = MVS \tag{14}$$

and

$$MV \text{ is monic,}$$

where $V \in \mathbf{R}^{q\times\rho}$ is any basis matrix for $\mathcal{V}$ and S is an appropriate $\rho \times \rho$ matrix. A subspace $\mathcal{W} \subset \mathbf{R}^p$ of dimension σ is a proper deflating subspace to the left if

$$WN = TWM \tag{15}$$

and

$$WM \text{ is epic,}$$

where $W^T \in \mathbf{R}^{p\times\sigma}$ is any basis matrix for $\mathcal{W}$, and T is an appropriate $\sigma \times \sigma$ matrix. A proper deflating subspace to the right is said to be stable (antistable) if $\Lambda(S) \subset \mathbf{D}_1(0)(\Lambda(S) \subset \mathbf{C} - \overline{\mathbf{D}}_1(0))$. A proper deflating subspace to the left is said to be stable (antistable) if $\Lambda(T) \subset \mathbf{D}_1(0)(\Lambda(T) \subset \mathbf{C} - \overline{\mathbf{D}}_1(0))$. A proper deflating subspace to the right (left) is said strictly stable if $\Lambda(S)(\Lambda(T))$ is in $\overline{\mathbf{D}}_1(0) - \{0\}$.

For deflating and reducing subspaces of a matrix pencil see also [4], [7], [8], [10], [24].

Recall from [9] that any matrix pencil is strictly equivalent to the Kronecker

canonical form

$$\lambda M - N \sim \begin{bmatrix} L_{\epsilon_1} & & & & & & & \\ & \ddots & & & & & & \\ & & L_{\epsilon_{\nu_r}} & & & & & \\ & & & {L_{\eta_1}}^T & & & & \\ & & & & \ddots & & & \\ & & & & & {L_{\eta_{\nu_l}}}^T & & \\ & & & & & & \lambda I - J & \\ & & & & & & & \lambda E - I \end{bmatrix},$$

where $L_{\epsilon_i} \in \mathbf{R}^{\epsilon_i \times (\epsilon_i+1)}, i = 1, \ldots, \nu_r$ (${L_{\eta_j}}^T \in \mathbf{R}^{(\eta_j+1)\times\eta_j}, j = 1, \ldots, \nu_l$) are the right (left) Kronecker blocks with $\epsilon_i \geq 0 (\eta_j \geq 0)$ the right (left) Kronecker indices and $\nu_r(\nu_l)$ the number of right (left) Kronecker blocks, respectively; J is in the real Jordan canonical form; and E is block diagonal, each block consisting of units placed on the first upper diagonal and zeros in the rest.

Let $n_r := \Sigma_{i=1}^{\nu_r} \epsilon_i$, $n_l := \Sigma_{j=1}^{\nu_l} \eta_j$ and let n_f and n_∞ be the sizes of J and E, respectively. Then n_f equals the number of finite generalized eigenvalues and n_∞ equals the number of infinite generalized eigenvalues. Clearly

$$\operatorname{rank}_{\mathbf{R}(\lambda)}(\lambda M - N) = n_r + n_l + n_f + n_\infty \leq \min(p, q).$$

Now write $n_f = n_f^- + n_f^+ + n_f^0$, where $n_f^-(n_f^+)$ is the number of finite generalized eigenvalues inside the unit disk (outside the closed unit disk) and n_f^0 is the number of finite generalized eigenvalues placed on the unit circle. Also write $n_f^- = \tilde{n}_f^- + \pi_0$, where π_0 is the multiplicity of the generalized eigenvalues located in the origin and $\tilde{n}_f^-$ is the number of nonzero generalized eigenvalues in $\mathbf{D}_1(0)$.

THEOREM 2.2. 1. *A matrix pencil has a stable (antistable) proper deflating subspace to the right if and only if* $n_r + n_f^- > 0$ $(n_r + n_f^+ > 0)$. *Moreover the maximal dimension of a stable (antistable) proper deflating subspace is* $n_r + n_f^-$ $(n_r + n_f^+)$, *and such a dimension can be effectively attained.*

2. *A matrix pencil has a stable (antistable) proper deflating subspace to the left if and only if* $n_l + n_f^- > 0$ $(n_l + n_f^+ > 0)$. *Moreover the maximal dimension of a stable (antistable) proper deflating subspace is* $n_l + n_f^-$ $(n_l + n_f^+)$, *and such a dimension can be effectively attained.*

3. *A matrix pencil has a strictly stable proper deflating subspace to the right (left) if and only if* $n_r + \tilde{n}_f^- > 0$ $(n_l + \tilde{n}_f^- > 0)$. *Moreover the maximal dimension of a strictly stable proper deflating subspace to the right (left) is* $n_r + \tilde{n}_f^-$ $(n_l + \tilde{n}_f^-)$, *and such a dimension can be effectively attained.*

The proof runs similarly to that given for Theorem 1 in [17] by changing the role of $\mathbf{C}^-$ and $\mathbf{C}^+$ with $\mathbf{D}_1(0)$ and $\mathbf{C} - \overline{\mathbf{D}}_1(0)$. Here $\mathbf{C}^-(\mathbf{C}^+)$ stands for the open left (right) half of the complex plane.

3. The EHP. Using Theorem 2.2 some remarkable properties of the EHP will be emphasized.

PROPOSITION 3.1. *If the EHP has a strictly stable (antistable) proper deflating subspace to the right, then it has an antistable (strictly stable) proper deflating subspace to the left of the same dimension.*

Proof. Let

$$V = [\underbrace{V_1^T}_{n} \quad \underbrace{V_2^T}_{n} \quad \underbrace{V_3^T}_{m}]^T$$

be monic and such that (14) holds, with $\Lambda(S) \subset \mathbf{D}_1(0) - \{0\}$ and MV monic. Then making (14) explicit, we have

$$AV_1 + BV_3 = V_1 S,$$

$$QV_1 - V_2 + LV_3 = -A^T V_2 S, \tag{16}$$

$$L^T V_1 + RV_3 = -B^T V_2 S$$

with

$$MV = \begin{bmatrix} V_1 \\ -A^T V_2 \\ -B^T V_2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & -A^T \\ 0 & -B^T \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \tag{17}$$

monic. By transposing (16) and then interchanging the first two equations, one obtains

$$S^T V_2^T A + V_1^T Q + V_3^T L^T = S^{-T}(S^T V_2^T),$$

$$-V_1^T = S^{-T}(-V_1^T A^T - V_3^T B^T), \tag{18}$$

$$S^T V_2^T B + V_1^T L + V_3^T R = 0,$$

where S is clearly nonsingular. Thus if we define

$$W := [S^T V_2^T \quad V_1^T \quad V_3^T], \tag{19}$$

then (15) is fulfilled with $T = S^{-T}$. Moreover from (17) it follows that $[V_1^T V_2^T]^T$ must be also monic; otherwise the left-hand side of (17) is not monic, which yields a contradiction. Hence

$$\begin{aligned} WM &= [S^T V_2^T \quad -V_1^T A^T - V_3^T B^T \quad 0] = [S^T V_2^T \quad -S^T V_1^T \quad 0] \\ &= S^T [V_2^T \quad -V_1^T \quad 0], \end{aligned} \tag{20}$$

where (19) and the second equation in (18) have been used. But the rightmost term in (20) is clearly epic (as $[V_1^T V_2^T]$ is epic and S nonsingular), and consequently WM and W are epic. Thus (19) defines, via $\mathcal{W} = \mathrm{Im} W^T$, an antistable proper deflating subspace to the left of the same dimension as $\mathcal{V}$. For the parenthesized text the same scheme of proof works. □

Remark 3. It has been shown in [17] that in the continuous-time case Proposition 3.1 can be stated in the "if and only if" form; that is, the proposition is also true if in its statement the order of the words *right* and *left* is changed. In the discrete-time case such a scheme does not work. To be more specific, assume that for $W = [W_1 \; W_2 \; W_3]$

epic, (15) is fulfilled, with $\Lambda(T) \subset \mathbf{C} - \mathbf{D}_1(0)$, and WM is epic. Then it can be easily checked that

$$V = \begin{bmatrix} W_2^T \\ W_1^T T^T \\ W_3^T \end{bmatrix}$$

is monic and (14) is fulfilled for such V with $S = T^{-T}$. Hence $\Lambda(S) \subset \mathbf{D}_1(0)$ and S is nonsingular. Moreover, since WM is epic, it follows that $[W_1 \quad -T^{-1}W_2 \quad 0]$ is epic. Thus

$$MV = \begin{bmatrix} I & 0 \\ 0 & -A^T \\ 0 & -B^T \end{bmatrix} \begin{bmatrix} W_2^T T^{-T} \\ W_1^T \end{bmatrix} T^T, \tag{21}$$

where the last two matrices in the right-hand side of (21) are clearly both monic. Hence the whole right-hand side of (21) is monic, provided that

$$\begin{bmatrix} I & 0 \\ 0 & -A^T \\ 0 & -B^T \end{bmatrix}$$

is also monic. But this clearly happens if the pair (A, B) has no uncontrollable modes in the origin. Thus we can conclude that the converse of Proposition 3.1 is also true if the minor hypothesis on the pair (A, B) is in addition assumed to be true. Notice also that if the EHP is regular, then the converse of Proposition 3.1 is true (see Remark 4 in [17]). □

Now introduce the discrete-time Popov function

$$\Pi(\lambda) = [B^T(\frac{1}{\lambda}I - A^T)^{-1} \; I] \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} \begin{bmatrix} (\lambda I - A)^{-1}B \\ I \end{bmatrix} \tag{22}$$

associated with Σ (see [16]). Let $\rho = \operatorname{rank}_{\mathbf{R}(\lambda)}\Pi(\lambda)$. Then we have the main result of this section.

THEOREM 3.2. *For the EHP the following hold:*

1. $n_r \leq n_l$.
2. $\tilde{n}_f^- = n_f^+$.
3. $\operatorname{rank}_{\mathbf{R}(\lambda)}(\lambda M - N) = 2n + \rho$.
4. $n_\infty \geq \rho + \pi_0$.
5. $n_r + n_f^- \leq n$, *and the equality holds if and only if* $n_f^0 = 0$, $n_r = n_l$, *and* $n_\infty = \rho + \pi_0$.

Proof. 1. Using 3. of Theorem 2.2 followed by Proposition 3.1 and then 2. of Theorem 2.2, we get $n_r + \tilde{n}_f^- \leq n_l + n_f^+$. Using 1. of Theorem 2.2 followed by Proposition 3.1 and then 3. of Theorem 2.2, we have that $n_r + n_f^+ \leq n_l + \tilde{n}_f^-$. By adding the above two inequalities, 1. follows.

2. The following identity can be easily checked:

$$\frac{1}{\lambda}M - N = \begin{bmatrix} 0 & \frac{1}{\lambda}I_n & 0 \\ I_n & 0 & 0 \\ 0 & 0 & I_m \end{bmatrix} (\lambda M - N)^T \begin{bmatrix} 0 & \frac{1}{\lambda}I_n & 0 \\ I_n & 0 & 0 \\ 0 & 0 & I_m \end{bmatrix}. \tag{23}$$

This shows that λ is a nonzero generalized eigenvalue for the EHP if and only if $\frac{1}{\lambda}$ is a generalized eigenvalue for the EHP.

3. By simple computation we may write the identity

$$\lambda M - N = \begin{bmatrix} I_n & 0 & 0 \\ -Q(\lambda I - A)^{-1} & I_n & 0 \\ -L^T(\lambda I - A)^{-1} & -\lambda B^T(I - \lambda A^T)^{-1} & I_m \end{bmatrix}$$

$$\times \begin{bmatrix} \lambda I - A & 0 & 0 \\ 0 & I - \lambda A^T & 0 \\ 0 & 0 & \Pi(\lambda) \end{bmatrix} \begin{bmatrix} I_n & 0 & -(\lambda I - A)^{-1}B \\ 0 & I_n & -(I - \lambda A^T)^{-1}(L + Q(\lambda I - A)^{-1}B) \\ 0 & 0 & I_m \end{bmatrix},$$

which proves assertion 3.

4. The following identity can be directly checked:

$$(M - \lambda N)\Xi(\lambda) = G(\lambda)\lambda(\lambda M - N)^T H(\lambda), \tag{24}$$

where

$$\Xi(\lambda) := \begin{bmatrix} \lambda I_n & 0 & 0 \\ 0 & \lambda I_n & 0 \\ 0 & 0 & I_m \end{bmatrix}, \qquad G(\lambda) := \begin{bmatrix} 0 & I_n & 0 \\ I_n & (1-\lambda)Q(I - \lambda A)^{-1} & 0 \\ 0 & (1-\lambda)L^T(I - \lambda A)^{-1} & I_m \end{bmatrix},$$

$$H(\lambda) := \begin{bmatrix} 0 & I_n & 0 \\ I_n & 0 & (\lambda - 1)(I - \lambda A)^{-1}B \\ 0 & 0 & I_m \end{bmatrix}.$$

Clearly $G(\lambda)$ and $H(\lambda)$ are regular at 0 (i.e., without null poles and zeroes). Denote by $n_{K(\lambda)}$ the number of null zeroes of the rational matrix $K(\lambda)$. Using the Smith–McMillan form of a rational matrix it can be easily seen (see also [26]) that

$$n_{\lambda(\lambda M - N)^T} = \pi_0 + \operatorname{rank}_{\mathbf{R}(\lambda)}(\lambda M - N) = \pi_0 + 2n + \rho, \tag{25}$$

where assertion 3. has been used. As both $G(\lambda)$ and $H(\lambda)$ are regular at 0, with (24) we get that

$$n_{\lambda(\lambda M - N)^T} \leq n_{M - \lambda N} + n_{\Xi(\lambda)},$$

and since $n_{\Xi(\lambda)} = 2n$ and $n_\infty = n_{M-\lambda N}$, we obtain with (25)

$$\pi_0 + 2n + \rho \leq n_\infty + 2n. \tag{26}$$

From (26) the conclusion follows immediately.

5. Since $\operatorname{rank}_{R(\lambda)}(\lambda M - N) = n_r + n_l + n_f^- + n_f^+ + n_f^0 + n_\infty = n_r + n_l + \tilde{n}_f^- + \pi_0 + n_f^+ + n_f^0 + n_\infty$ always, it follows from 1., 2., and 3. just proved above that $2n + \rho \geq 2n_r + 2n_f^- + \pi_0 + \rho + n_f^0$, from where $n \geq n_r + n_f^-$ and the equality holds if and only if $n_f^0 = 0$, $n_r = n_l$ and $n_\infty = \pi_0 + \rho$. □

COROLLARY 3.3. *The maximal dimension of a stable proper deflating subspace to the right of the EHP does not exceed n.*

Proof. This follows directly from assertion 4. of Theorem 3.2 combined with 1. of Theorem 2.2. □

Remark 4. If the EHP is regular, then assertion 4. of Theorem 3.2 reads $n_\infty = \pi_0 + m$, and thus the result stated in Proposition 3 in [13] is recovered. Indeed, (24) can be rewritten as

$$M - \lambda N = G(\lambda)(\lambda M - N)^T H(\lambda)\tilde{\Xi}(\lambda),$$

where

$$\tilde{\Xi}(\lambda) := \begin{bmatrix} I_n & 0 & 0 \\ 0 & I_n & 0 \\ 0 & 0 & \lambda I_m \end{bmatrix}.$$

Using arguments similar to those for assertion 4. in Theorem 3.2 we get

$$n_\infty \leq \pi_0 + n_{\tilde{\Xi}(\lambda)} = \pi_0 + m.$$

Since the EHP is regular, $\operatorname{rank}(\lambda M - N) = 2n + m = 2n + \rho$ and consequently $m = \rho$. Hence the conclusion follows by using assertion 4. in Theorem 3.2 in conjunction with the above inequality. □

Remark 5. Unlike the continuous case, where the corresponding assertions 1. and 4. stated in Theorem 3.2 are always expressed via equalities (see Theorem 2 in [17]), the discrete case is quite different. This is emphasized by the fact that sometimes 1. and 4. become strict inequalities, as is shown below.

Example. Let the Popov triplet be defined as $n = m = 1$, $A = 0, B = 0, Q = 0, L = 1, R = 0$. Then (see (13))

$$M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad N = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & 0 \end{bmatrix},$$

and by elementary row and column operations we get

$$\lambda M - N \sim \begin{bmatrix} \lambda & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

So $n_r = 0, n_l = 1, n_\infty = 1, \pi_0 = 0$ (in fact $n_f = 0$), and $\rho = 0$ because of $\Pi(\lambda) = 0$ (see (22)). Hence $n_l > n_r$ and $n_\infty > \rho + \pi_0$. □

Now a remarkable property of the stable proper deflating subspaces will be pointed out.

PROPOSITION 3.4. *Let $\mathcal{V}$ be any stable proper deflating subspace to the right for the EHP and*

$$V = [\underbrace{V_1^T}_{n} \; \underbrace{V_2^T}_{n} \; \underbrace{V_3^T}_{m}]^T$$

be any basis matrix for it. Then

$$V_1^T V_2 = V_2^T V_1. \tag{27}$$

Proof. The proof runs similarly to that in [13]. □

Suppose temporarily that the pair (A, B) has no null uncontrollable modes. Then we are in the position to formulate the following unicity result.

PROPOSITION 3.5. *Let $\mathcal{V}$ be a stable proper deflating subspace to the right of maximal dimension, i.e., $\dim \mathcal{V} = n_r + n_f^- > 0$, and let*

$$V = [\underbrace{V_1^T}_{n} \quad \underbrace{V_2^T}_{n} \quad \underbrace{V_3^T}_{m}]^T$$

be any basis matrix for $\mathcal{V}$. Then

$$\bar{\mathcal{V}} := \operatorname{Im} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \subset \mathbf{R}^{2n} \tag{28}$$

does not depend on the choice of the maximal stable proper deflating subspace $\mathcal{V}$, i.e., it is uniquely determined.

Proof. Let

$$\begin{gathered} \bar{V} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}, \quad \bar{A} = \begin{bmatrix} A & 0 \\ -Q & -I \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} B \\ L \end{bmatrix}, \quad \bar{C} = \begin{bmatrix} L^T & 0 \end{bmatrix}, \quad \bar{D} = R, \\ \bar{E} = \begin{bmatrix} I & 0 \\ 0 & -A^T \end{bmatrix}, \quad \bar{G} = \begin{bmatrix} 0 & -B^T \end{bmatrix}, \end{gathered} \tag{29}$$

where $\bar{V}$ is clearly monic, as has been shown in the proof of Proposition 3.1 (see (17)). Then the EHP (13) can be written as

$$M = \begin{bmatrix} \bar{E} & 0 \\ \bar{G} & 0 \end{bmatrix}, \qquad N = \begin{bmatrix} \bar{A} & \bar{B} \\ \bar{C} & \bar{D} \end{bmatrix}, \tag{30}$$

and $[\frac{\bar{E}}{\bar{G}}]$ is monic. From (14) we get

$$\begin{bmatrix} \bar{A} \\ \bar{C} \end{bmatrix} \bar{V} + \begin{bmatrix} \bar{B} \\ \bar{D} \end{bmatrix} V_3 = \begin{bmatrix} \bar{E} \\ \bar{G} \end{bmatrix} \bar{V} S, \tag{31}$$

where obviously (see Definition 3)

$$\begin{bmatrix} \bar{E} \\ \bar{G} \end{bmatrix} \bar{V} \tag{32}$$

is monic and

$$\Lambda(S) \subset \mathbf{D}_1(0). \tag{33}$$

If $\bar{F} := V_3 \bar{V}^{\#}$, where $\bar{V}^{\#}$ is any left inverse of $\bar{V}$ ($\bar{V}$ is monic), then (31) finally yields

$$\begin{bmatrix} \bar{A} + \bar{B}\bar{F} \\ \bar{C} + \bar{D}\bar{F} \end{bmatrix} \bar{V} = \begin{bmatrix} E \\ G \end{bmatrix} \bar{V} S. \tag{34}$$

Note now that if $\bar{E} = I$ and $\bar{G} = 0$, then (34) with (33) expresses the usual condition that $\bar{\mathcal{V}}$ (defined by (28)) is a stable null output $(\bar{A}, \bar{B})$ invariant subspace of the system $(\bar{A}, \bar{B}, \bar{C}, \bar{D})$. Clearly $\bar{\mathcal{V}} \subset \mathbf{R}^{2n}$ is of maximal dimension since otherwise (34) would hold for a monic $\bar{V}$ with the rank greater than $n_f + n_f^-$, which contradicts the maximality of $\mathcal{V} \subset \mathbf{R}^{2n+m}$. As is well known, such a subspace is unique (see [27], [23]). Generalizing the above notion, we shall say that $\bar{\mathcal{V}}$, satisfying (34) with

(33), is a stable null output invariant subspace of the generalized descriptor system representation (M, N), with M and N written as in (30). Based on similar arguments as above, $\bar{\mathcal{V}}$ is of maximal dimension. As in the usual case, mentioned above, such a subspace is the supremal one of the family of subspaces with the properties (32), (33), (34) and consequently is unique (see Appendix A). □

DEFINITION 4. *The EHP is said to be dichotomic if* $n_f^0 = 0, n_r = n_l$, *and* $n_\infty = \rho + \pi_0$.

We have immediately the following proposition.

PROPOSITION 3.6. *The EHP is dichotomic if and only if it has a stable proper deflating subspace to the right of dimension* n.

Proof. Only if: According to 5. of Theorem 3.2 we have $n_r + n_f^- = n$. Following 1. of Theorem 2.2 a stable proper deflating subspace of dimension n for the EHP exists.

If: Following 1. of Theorem 2.2 we have $n \leq n_r + n_f^-$. Hence by 5. of Theorem 3.2 the equality holds, and $n_f^0 = 0$, $n_r = n_l$, and $n_\infty = \rho + \pi_0$. □

DEFINITION 5. *Let* $\mathcal{V}$ *be any stable proper deflating subspace to the right of dimension* σ *and*

$$V = [\underbrace{V_1^T}_{n} \quad \underbrace{V_2^T}_{n} \quad \underbrace{V_3^T}_{m}]^T$$

be any basis matrix for it. We call $\mathcal{V}$ *disconjugate if* V_1 *is monic.*

Note that

a) The notion of disconjugacy is well defined since Corollary 3.3 asserts that the number of columns of V_1 does not exceed n.

b) The notion of disconjugacy is independent of the choice of the basis matrix V for the (stable) proper deflating subspace. Indeed, if $\hat{V} = [\hat{V}_1^T \ \hat{V}_2^T \ \hat{V}_3^T]^T$ is another basis matrix, then clearly $\hat{V} = VG$ for an appropriate nonsingular $\sigma \times \sigma$ matrix G. Hence $\hat{V}_1 = V_1 G$ and the conclusion follows. Based on Proposition 3.6 we can introduce the following definition.

DEFINITION 6. *A dichotomic EHP is said to be disconjugate if it has a disconjugate (stable) proper deflating subspace to the right of dimension* n.

Remark 6. According to Proposition 3.6 the dichotomy of the EHP implies the existence of a stable proper deflating subspace to the right $\mathcal{V}$ of dimension n. Hence in accordance with Corollary 3.3 such a subspace is of maximal dimension. If

$$V = [\underbrace{V_1^T}_{n} \quad \underbrace{V_2^T}_{n} \quad \underbrace{V_3^T}_{m}]^T$$

is any basis matrix for $\mathcal{V}$, then, following Proposition 3.5, $\bar{\mathcal{V}} = \mathrm{Im}[V_1^T V_2^T]^T$ is unique. Consequently the disconjugacy is checked by checking the invertibility of V_1. If V_1 is singular, there is no other basis matrix V with V_1 nonsingular because of the uniqueness of $\bar{\mathcal{V}}$. Thus we can effectively check the disconjugacy of a dichotomic EHP. □

4. Main result. Now we are ready to state and prove our main result.

THEOREM 4.1. *The GDTRE* (1) *has a stabilizing solution* (r, X, V, F) *if and only if the EHP has a disconjugate proper deflating subspace.*

Proof. If: Let $\mathcal{V}$ be any disconjugate proper deflating subspace of dimension r with basis matrix

$$V = [\underbrace{V_1^T}_{n} \quad \underbrace{V_2^T}_{n} \quad \underbrace{V_3^T}_{m}]^T.$$

Then equations (16) are fulfilled for $S \in \mathbf{R}^{r\times r}$ stable since disconjugacy *pre*supposes the proper deflating subspace to be stable (see Definition 5). Let $X := V_2V_1^{\#}$ and $F := V_3V_1^{\#}$, where V_1 is any left inverse of V_1, which is monic because of disconjugacy. Then (16) provides

$$\begin{aligned} (A+BF)V_1 &= V_1S, \\ (Q-X+LF)V_1 &= -A^TXV_1S, \\ (L^T+RF)V_1 &= -B^TXV_1S. \end{aligned} \tag{35}$$

By eliminating V_1S in the first two equations (35), equation (1) with $V = V_1$ is easily obtained. This, together with the first equation in (35) and Proposition 3.4, which implies (see (27)) $V_1^TXV_1 = V_1^TV_2V_1^{\#}V_1 = V_1^TV_2 = V_2^TV_1 = V_1^T(V_1^T)^{\#}V_2^TV_1 = V_1^TX^TV_1$, prove the if part.

Only if: Let (r, X, V_1, F) be a stabilizing solution to (1). Then by taking

$$\begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} = \begin{bmatrix} V_1 \\ XV_1 \\ FV_1 \end{bmatrix}$$

it can be easily checked that (16) holds, i.e., (14) holds and MV is monic since V_1 is monic as follows from

$$MV = \begin{bmatrix} V_1 \\ -A^TV_2 \\ -B^TV_2 \end{bmatrix}. \qquad \square$$

From Theorem 4.1 we immediately derive the following corollary.

COROLLARY 4.2. *The CDTARE* (3) (*or* (4)) *has a stabilizing solution if and only if the EHP is disconjugate.*

5. Constrained dynamics with quadratic cost. Let Σ be a Popov triplet and associate with it the following:

1) the linear (discrete) system

$$\sigma x = Ax + Bu \tag{36}$$

and the quadratic cost

$$J = \sum_{k=0}^{\infty} \begin{bmatrix} x_k \\ u_k \end{bmatrix}^T \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \tag{37}$$

with x_k and u_k linked by (36);

2) the GDTARE (1) written in explicit form, for $V = V_1$:

$$(A^TXA - X + Q)V_1 + (A^TXB + L)FV_1 = 0, \tag{38}$$

$$(L^T + B^TXA)V_1 + (R + B^TXB)FV_1 = 0. \tag{39}$$

Assume that the GDTARE has a stabilizing solution (r, X, V_1, F); that is, (38) and (39) both hold

$$X_1 := V_1^T X V_1 = V_1^T X^T V_1 = X_1^T \tag{40}$$

and

$$(A + BF)V_1 = V_1 S \tag{41}$$

with $S \in \mathbf{R}^{r\times r}$ stable. Following Remark 1, $\mathcal{V}_1 = \mathrm{Im}V_1$ is a stable (A, B) invariant subspace of dimension r. Hence, for any initial condition placed in $\mathcal{V}_1$ and an appropriate control input $u \in l^2(\mathbf{N}; \mathbf{R}^m)$ we can force the system (36) to have its evolution entirely contained in $\mathcal{V}_1$ and exponentially approaching zero as k approaches infinity. Thus we can write

$$V_1\sigma\xi = AV_1\xi + Bu. \tag{42}$$

In this section, we shall evaluate, in terms of the GDTARE (1), the quadratic cost (37) along the trajectories described by (42) with $u \in l^2(\mathbf{N}, \mathbf{R}^m)$, i.e., along those trajectories located in $\mathcal{V}_1$ and which exponentially approach the origin. To this end, let us introduce the following notations:

$$Q_1 = V_1^T Q V_1, \quad L_1 = V_1^T L, \quad F_1 = FV_1. \tag{43}$$

With (43) and (41), (38) and (39) become

$$A^T X V_1 S - X V_1 + Q V_1 + L F_1 = 0, \tag{44}$$

$$L_1^T + B^T X V_1 S + R F_1 = 0. \tag{45}$$

Premultiplying (44) by V_1^T and taking into account the transpose of (41) we get

$$S^T X_1 S - F_1^T B^T X V_1 S - X_1 + Q_1 + L_1 F_1 = 0. \tag{46}$$

Premultiplying (45) by F_1^T one obtains

$$F_1^T L_1^T + F_1^T B^T X V_1 S + F_1^T R F_1 = 0. \tag{47}$$

With $F_1^T B^T X V_1 S$ substituted from (47) in (46) we have further

$$S^T X_1 S - X_1 + Q_1 + L_1 F_1 + F_1^T L_1^T + F_1^T R F_1 = 0. \tag{48}$$

Remark 7. Let V_1 be nonsingular. Then without loss of generality we may take $V_1 = I_n$ in (38), (39), and (41), and (48) becomes

$$(A + BF)^T X (A + BF) - X + Q + LF + F^T L^T + F^T R F = 0 \tag{49}$$

with $A + BF$ stable. But (49) is exactly the "closed loop" form of the (constrained) Riccati equation. □

Now we are ready to state the main result of this section.

PROPOSITION 5.1. *Suppose that the GDTARE has a stabilizing solution. Then the quadratic cost* (37) *evaluated for those pairs* $(\xi, u) \in l^2(\mathbf{N}; \mathbf{R}^r) \times l^2(\mathbf{N}; \mathbf{R}^m)$ *linked by* (42) *has the expression*

$$J = \xi_0^T X_1 \xi_0 + \langle u - F_1\xi, (R + B^T X B)(u - F_1\xi)\rangle, \tag{50}$$

where $\xi_0 := \xi(0)$. *If the "positivity condition"* $R + B^T X B \geq 0$ *holds, then the minimum of* J *over* u *is attained (nonuniquely) for* $u = F_1\xi$. *Such a feedback provides also stabilization with respect to the stable* (A, B) *invariant subspace* $\mathcal{V}_1$.

Proof. See Appendix B for the proof. □

6. Conclusions. Necessary and sufficient conditions for the existence of the stabilizing solution to the GDTARE have been given. These conditions are expressed in terms of the disconjugacy of proper (stable) deflating subspaces associated with the EHP. The basic results concerning the CDTARE, intensively used in nonstandard factorizations, are easily recovered and enlarged. An evaluation of the quadratic index along constrained stable trajectories was also given in terms of the stabilizing solution of the associated GDTARE. Based on the above results, a numerically reliable method for computing the stabilizing solution of the GDTARE and CDTARE is proposed in [22]. The method uses the algorithms for computing the generalized Schur form of a singular pencil developed in [7], [8] and the refinements in [4]. Applications of the developed theory can be found in [18] and in a forthcoming paper dealing with singular H^2 optimal control.

Appendix A. Supremal stable null output invariant subspace of the generalized descriptor system representation. Let (M, N) be a matrix pair defined by

$$M = \begin{bmatrix} E & 0 \\ G & 0 \end{bmatrix}, \qquad N = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{A.1}$$

and termed a *generalized descriptor system representation.* Assume that $\left[\begin{smallmatrix} E \\ G \end{smallmatrix}\right]$ is monic.

DEFINITION A1. *A subspace $\mathcal{V} \subset \mathbf{R}^n$ is called a null output invariant subspace of the system pair (M, N), represented by* (A.1), *if*

$$\begin{bmatrix} A \\ C \end{bmatrix} \mathcal{V} \subset \begin{bmatrix} E \\ G \end{bmatrix} \mathcal{V} + \operatorname{Im} \begin{bmatrix} B \\ D \end{bmatrix}. \tag{A.2}$$

The family of all such subspaces will be denoted by $\mathbf{I}(M, N)$.

Clearly

$$\begin{bmatrix} E \\ G \end{bmatrix} V \text{ is monic} \tag{A.3}$$

for any basis matrix V of $\mathcal{V}$ and $\mathbf{I}(M, N) \neq \emptyset$ because it contains $\mathbf{R}^n$.

Note that

a) If $E = I$ and $G = 0$, then the usual definition of the null output (A, B) invariant subspace encountered in [23] and [27] is recovered. It corresponds to the usual system representation $\sigma x = Ax + Bu, y = Cx + Du$.

b) If $G = 0$, then the above notion is adapted to the usual descriptor system representation $E\sigma x = Ax + Bu, y = Cx + Du$.

As has been noted in the previous sections, the notion introduced by Definition A1 is associated with the (discrete) Hamiltonian systems (12) (with $v = 0$), which simultaneously incorporates both forward and backward time evolutions.

Let $\mathcal{V} \subset \mathbf{I}(M, N)$, with $\dim \mathcal{V} = \rho$, and $V \in \mathbf{R}^{n \times \rho}$ be any basis matrix of $\mathcal{V}$. Then (A.2) is equivalent to

$$\begin{bmatrix} A \\ C \end{bmatrix} V = \begin{bmatrix} E \\ G \end{bmatrix} VS - \begin{bmatrix} B \\ D \end{bmatrix} H \tag{A.4}$$

for adequate $S \in \mathbf{R}^{\rho \times \rho}$ and $H \in \mathbf{R}^{m \times \rho}$. Since V is monic, let

$$F = HV^{\#}, \tag{A.5}$$

where $V^{\#}$ is any left inverse of V. Then (A.4) reduces to

$$\begin{bmatrix} A+BF \\ C+DF \end{bmatrix} V = \begin{bmatrix} EV \\ GV \end{bmatrix} S, \tag{A.6}$$

which is also equivalent to

$$\begin{bmatrix} A+BF \\ C+DF \end{bmatrix} \mathcal{V} \subset \begin{bmatrix} E \\ G \end{bmatrix} \mathcal{V} \tag{A.7}$$

and where (A.3) automatically holds.

Let $\mathbf{F}(M,N,\mathcal{V}) = \{ F : \text{(A.7) is true} \}$. Clearly $\mathcal{V} \subset \mathbf{I}(M,N)$ if and only if $\mathbf{F}(M,N,\mathcal{V}) \neq \emptyset$.

In order to emphasize the structural aspects related to Definition A1, we have to perform several strict equivalence operations on the (singular) matrix pencil $\lambda M - N$. Let $F_1 \in F(M,N,\mathcal{V})$, define the nonsingular matrix

$$Z_1 = \begin{bmatrix} I_n & 0 \\ F_1 & I_m \end{bmatrix}, \tag{A.8}$$

and consider the strict equivalent pencil $\lambda M_1 - N_1 \sim \lambda M - N$ defined via

$$M_1 = MZ_1 = \begin{bmatrix} E & 0 \\ G & 0 \end{bmatrix}, \qquad N_1 = NZ_1 = \begin{bmatrix} A+BF_1 & B \\ C+DF_1 & D \end{bmatrix}. \tag{A.9}$$

Introduce now the nonsingular matrix

$$Z_2 = \begin{bmatrix} V & W & 0 \\ 0 & 0 & I_m \end{bmatrix} = \begin{bmatrix} \tilde{V} & 0 \\ 0 & I_m \end{bmatrix}, \tag{A.10}$$

where W is any completion of V (which is monic) up to a nonsingular matrix $\tilde{V} = [V\ W]$. Consider the decomposition

$$\operatorname{Im} \begin{bmatrix} B \\ D \end{bmatrix} = \operatorname{Im} \begin{bmatrix} B \\ D \end{bmatrix} \cap \operatorname{Im} \begin{bmatrix} EV \\ GV \end{bmatrix} \oplus \operatorname{Im} \begin{bmatrix} \tilde{B} \\ \tilde{D} \end{bmatrix}, \tag{A.11}$$

and introduce the nonsingular matrix

$$Q_2 = \begin{bmatrix} EV & E_1 \\ GV & G_1 \end{bmatrix}, \tag{A.12}$$

where the completion $\left[\begin{smallmatrix} E_1 \\ G_1 \end{smallmatrix}\right]$ of the monic matrix $\left[\begin{smallmatrix} EV \\ GV \end{smallmatrix}\right]$ (see (A.3)) is chosen such that

$$\operatorname{Im} \begin{bmatrix} \tilde{B} \\ \tilde{D} \end{bmatrix} \subset \operatorname{Im} \begin{bmatrix} E_1 \\ G_1 \end{bmatrix}. \tag{A.13}$$

Let now $\lambda M_2 - N_2 \sim \lambda M_1 - N_1$ be defined as $M_2 = Q_2 M_1 Z_2$ and $N_2 = Q_2 N_1 Z_2$. Then according to (A.6), where $F = F_1$, (A.11) and (A.13), we get

$$M_2 = \begin{bmatrix} & \overbrace{}^{n-\rho} & \overbrace{}^{m} \\ I_\rho & x & 0 \\ 0 & x & 0 \end{bmatrix}, \qquad N_2 = \begin{bmatrix} \overbrace{}^{\rho} & \overbrace{}^{n-\rho} & \overbrace{}^{m_1} & \overbrace{}^{m_2} \\ S & x & B_1 & 0 \\ 0 & T & 0 & D_1 \end{bmatrix}, \tag{A.14}$$

where D_1 can be chosen to be monic as easily can be remarked. It can be also easily checked that for

$$\hat{Z}_3 = \begin{bmatrix} I_n & 0 \\ F_3 & I_m \end{bmatrix} \tag{A.15}$$

the strict equivalence $\lambda M_2 - N_2 \sim \lambda M_2 \hat{Z}_3 - N_2 \hat{Z}_3$ preserves the structure of (A.14) if and only if

$$F_3 = \begin{bmatrix} F_{3,11} & F_{3,12} \\ 0 & F_{3,22} \end{bmatrix}, \tag{A.16}$$

as follows from the fact that $D_1 F_{3,21} = 0 \Leftrightarrow F_{3,21} = 0$ since D_1 is monic. Note now that with (A.8), (A.10), and (A.15) we have

$$Z_1 Z_2 \hat{Z}_3 = \begin{bmatrix} I & 0 \\ F_1 & I_m \end{bmatrix} \begin{bmatrix} \tilde{V} & 0 \\ 0 & I_m \end{bmatrix} \begin{bmatrix} I & 0 \\ F_3 & I \end{bmatrix} = \begin{bmatrix} \tilde{V} & 0 \\ F_1 \tilde{V} + F_3 & I_m \end{bmatrix}$$

$$= \begin{bmatrix} I_n & 0 \\ F & I_m \end{bmatrix} \begin{bmatrix} \tilde{V} & 0 \\ 0 & I_m \end{bmatrix}, \tag{A.17}$$

where $F\tilde{V} = F_1 \tilde{V} + F_3$, i.e., $F = F_1 + F_3 \tilde{V}^{-1}$. Thus $\mathbf{F} = \{F_1 + F_3 \tilde{V}^{-1} : F_3$ of the form (A.16) $\}$ and for any F_1 for which (A.6) holds, i.e., determined via (A.5).

Remark A1. Equality (A.17) shows that any successive product of matrices of type (A.8) and (A.10) reduces to the product of *two* matrices of type (A.8) and (A.10). In the usual cases this expresses the well-known feedback and coordinate changing operations. □

Since $\lambda M_2 - N_2 \sim \lambda M - N$, (A.14) shows that $\Lambda(S)$ is a subset of the set of finite generalized eigenvalues of the pencil $\lambda M - N$, and we shall write $\Lambda(S) = \Lambda((M, N)|\mathcal{V})$.

DEFINITION A2. *$\mathcal{V} \in \mathbf{I}(M, N)$ is said to be stable if $\Lambda((M, N)|\mathcal{V})$ is located in $\mathbf{D}_1(0)$.*

Let $\mathbf{I}_s = \{\mathcal{V} : \mathcal{V} \in \mathbf{I}$ and stable$\}$. Since if $\mathcal{V}_1, \mathcal{V}_2 \in \mathbf{I}(M, N)$, then $\mathcal{V}_1 + \mathcal{V}_2 \in \mathbf{I}(M, N)$, as directly follows from (A.2), we conclude that $\mathbf{I}(M, N)$ has a (unique) supremal element denoted $\mathcal{V}^* = \sup \mathbf{I}(M, N)$. Following a scheme of proof similar to that in [27] we have that $\lim \mathcal{V}_k = \mathcal{V}^*$ for the sequence $\mathcal{V}_k$ defined by

$$\mathcal{V}_{k+1} = \begin{bmatrix} A \\ C \end{bmatrix}^{-1} \left(\begin{bmatrix} E \\ G \end{bmatrix} \mathcal{V}_k + \operatorname{Im} \begin{bmatrix} B \\ D \end{bmatrix} \right), \quad \mathcal{V}_0 = \mathbf{R}^n, \quad k \geq 0. \tag{A.18}$$

We are now interested in finding $\mathcal{V}_s^* := \sup \mathbf{I}_s(M, N)$. To this end consider the structure (A.14), where $\rho = \dim \mathcal{V}^*$. For the pair (S, B_1) apply the controllable decomposition, that is,

$$TST^{-1} = \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{bmatrix}, \qquad TB_1 = \begin{bmatrix} B_{11} \\ 0 \end{bmatrix}, \tag{A.19}$$

where (S_{11}, B_{11}) is controllable. Let

$$Q_3 = \begin{bmatrix} T & 0 \\ 0 & I_{n+p-\rho} \end{bmatrix}, \qquad Z_3 = \begin{bmatrix} T & 0 \\ 0 & I_{n+m-\rho} \end{bmatrix}, \tag{A.20}$$

and consider $\lambda M_3 - N_3 \sim \lambda M_2 - N_2$ with $M_3 = Q_3 M_2 Z_3$ and $N_3 = Q_3 N_2 Z_3$. Then

$$\text{(A.21)}\quad M_3 = \begin{bmatrix} I_{\rho_1} & 0 & x & 0 \\ 0 & I_{\rho_2} & x & 0 \\ 0 & 0 & x & 0 \end{bmatrix}, \qquad N_3 = \begin{bmatrix} S_{11} & S_{12} & x & B_{11} & 0 \\ 0 & S_{22} & x & 0 & 0 \\ 0 & 0 & x & 0 & D_1 \end{bmatrix}.$$

Now, following the same argument as in [27], we have

$$\text{(A.22)}\qquad \mathcal{V}^* \supset V_s^* = \langle S_{11} | \mathrm{Im} B_{11} \rangle \oplus \chi^-(S_{22}),$$

where $\langle S_{11} | \mathrm{Im} B_{11} \rangle$ is the controllable subspace of the pair (S_{11}, B_{11}) and $\chi^-(S_{22})$ is the stable subspace of S_{22}, i.e., $\chi^-(S_{22}) = \ker \mu^-(\lambda)$, where $\mu(\lambda) = \mu^-(\lambda)\mu^+(\lambda)$ is the factorization of the minimal polynomial $\mu(\lambda)$ of S_{22} with respect to the unit circle. Here the roots of $\mu^-(\lambda)$ are in $\mathbf{D}_1(0)$.

Appendix B. Proof of Proposition 5.1. Suppose the GDTARE (1) (or equivalently the system (38)–(39)) has a stabilizing solution (r, X, V_1, F). Then using (41), (42) becomes

$$\text{(B.1)}\qquad V_1 \sigma\xi = V_1 S\xi - BF_1\xi + Bu,$$

and premultiplying both sides of the above equation by $V_1^T X^T$ we get (see (40))

$$\text{(B.2)}\qquad X_1 \sigma\xi = X_1 S\xi - V_1^T X^T BF_1\xi + V_1^T X^T Bu.$$

Now we are ready to evaluate the quadratic cost (37) for those pairs $(\xi, u) \in l^2(\mathbf{N}; \mathbf{R}^r) \times l^2(\mathbf{N}; \mathbf{R}^m)$ linked by (42) (or equivalently (B.2)). First we can write

$$J = \sum_{k=0}^{\infty} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^T \begin{bmatrix} Q_1 & L_1 \\ L_1^T & R \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} = \left\langle \begin{bmatrix} \xi \\ u \end{bmatrix}, \begin{bmatrix} Q_1 & L_1 \\ L_1^T & R \end{bmatrix} \begin{bmatrix} \xi \\ u \end{bmatrix} \right\rangle,$$

where $x = V_1\xi$ and (43) have been used. Then with (48) and (45), we have the evaluation

$$\begin{bmatrix} \xi \\ u \end{bmatrix}^T \begin{bmatrix} Q_1 & L_1 \\ L_1^T & R \end{bmatrix} \begin{bmatrix} \xi \\ u \end{bmatrix} = \xi^T Q_1 \xi + 2u^T L_1^T \xi + u^T Ru$$

$$= \xi^T(-S^T X_1 S + X_1 - L_1 F_1 - F_1^T L_1^T - F_1^T R F_1)\xi + 2u^T(-B^T X V_1 S - RF_1)\xi + u^T Ru$$

$$= -\xi^T S^T X_1 S\xi + \xi^T X_1 \xi + \xi^T (F_1^T R + S^T V_1^T X^T B) F_1 \xi + \xi^T F_1^T (B^T X V_1 S + RF_1)\xi$$

$$-2u^T(B^T X V_1 S + RF_1)\xi + u^T Ru - \xi^T F_1^T R F_1 \xi$$

$$= -\xi^T S^T X_1 S\xi + \xi^T X_1 \xi + 2\xi^T S^T V_1^T X^T B F_1 \xi - 2u^T B^T X V_1 S\xi + (u - F_1\xi)^T R(u - F_1\xi)$$

$$\overset{(B2)}{=} (-(\sigma\xi)^T X_1 - \xi^T F_1^T B^T X V_1 + u^T B^T X V_1) S\xi + \xi^T X_1 \xi + 2\xi^T S^T V_1^T X^T B F_1 \xi$$

$$-2u^T B^T X V_1 S\xi + (u - F_1\xi)^T R(u - F_1\xi)$$

$$= -(\sigma\xi)^T X_1 S\xi + \xi^T F_1^T B^T X V_1 S\xi - u^T B^T X V_1 S\xi + \xi^T X_1 \xi + (u - F_1\xi)^T R(u - F_1\xi)$$

$$\overset{(B2)}{=} -(\sigma\xi)^T (X_1\sigma\xi + V_1^T X^T B F_1\xi - V_1^T X^T B u) + \xi^T F_1^T B^T X V_1 S\xi - u^T B^T X V_1 S\xi$$

$$+\xi^T X_1 \xi + (u - F_1\xi)^T R(u - F_1\xi)$$

$$= -(\sigma\xi)^T X_1\sigma\xi + \xi^T X_1\xi + \xi^T F_1^T B^T X(-V_1\sigma\xi + V_1 S\xi) + u^T B^T X(V_1\sigma\xi - V_1 S\xi)$$

$$+(u - F_1\xi)^T R(u - F_1\xi)$$

$$\overset{(B1)}{=} -(\sigma\xi)^T X_1\sigma\xi + \xi^T X_1\xi + \xi^T F_1^T B^T X(BF_1\xi - Bu) + u^T B^T X(Bu - BF_1\xi)$$

$$+(u - F_1\xi)^T R(u - F_1\xi)$$

$$= -(\sigma\xi)^T X_1\sigma\xi + \xi^T X_1\xi + \xi^T F_1^T B^T X B F_1\xi - \xi^T F_1^T B^T X B u - u^T B^T X B F_1\xi$$

$$+u^T B^T X B u + (u - F_1\xi)^T R(u - F_1\xi)$$

$$= -(\sigma\xi)^T X_1\sigma\xi + \xi^T X_1\xi + (u - F_1\xi)^T R(u - F_1\xi) + (u - F_1\xi)^T (B^T X B)(u - F_1\xi)$$

$$= -(\sigma\xi)^T X_1\sigma\xi + \xi^T X_1\xi + (u - F_1\xi)^T (R + B^T X B)(u - F_1\xi). \tag{B.3}$$

As $\xi \in l^2(\mathbf{N}; \mathbf{R}^r)$, then $\langle \xi, X_1\xi\rangle - \langle \sigma\xi, X_1\sigma\xi\rangle = \xi_0^T X_1 \xi_0$. Hence it follows from (B.3) that

$$J = \xi_0^T X_1 \xi_0 + \langle u - F_1\xi, (R + B^T X B)(u - F_1\xi)\rangle. \tag{B.4}$$

From (B.4) the conclusion follows immediately.

Acknowledgments. After the review process was completed, Dr. Martin Weiss drew our attention to the fact that the absence of null uncontrollable modes for the pair (A, B) is in fact not necessary for proving that the disconjugacy of a maximal proper deflating subspace to the EHP does not depend on the particular choice of the maximal proper deflating subspace. Moreover, the proof of Proposition 3.5 can be considerably simplified.

REFERENCES

[1] M. A. AIZERMAN AND F. R. GANTMACHER, *Absolute Stability of Control Systems*, Soviet Academy Publishing House, Moscow, 1963. (In Russian.)

[2] L. R. ANDERSON, D. W. BREWER, AND A. R. BAYKAM, *Numerical solution of the symmetric Riccati equation through iteration*, in Proc. American Control Conference, Arlington, June 1982, pp. 1010–1015.

[3] T. BASAR AND P. BERNHARD, *H^∞-Optimal Control and Related Minmax Design Problems: A Dynamic Game Approach*, Birkhäuser-Verlag, Basel, 1991.

[4] TH. G. BEELEN, *New algorithms for computing the Kronecker structure of a pencil with applications to systems and control theory*, Ph.D. thesis, Technical University Eindhoven, Eindhoven, 1987.
[5] S. BITTANTI, A. LAUB, AND J. C. WILLEMS, EDS., *The Riccati Equation*, Springer-Verlag, Berlin, 1991.
[6] T. CHEN AND B. A. FRANCIS, *Spectral and inner-outer factorization of rational matrices*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 1–17.
[7] P. VAN DOOREN, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–141.
[8] ———, *Reducing Subspaces: Definitions, Properties and Algorithms*, Lecture Notes in Mathematics 973, Springer-Verlag, Berlin, 1983, pp. 58–73.
[9] F. R. GANTMACHER, *Theory of Matrices*, Vols. I and II, Chelsea, New York, 1959.
[10] I. GOHBERG AND M. A. KAASHOEK, *Block Toeplitz operators with rational symbols*, in Contributions to Operator Theory and Its Applications, I. Gohberg, et al., eds., Operator Theory 35, Birkhäuser-Verlag, Basel, 1988, pp. 385–440.
[11] A. HALANAY AND V. IONESCU, *Generalized discrete-time Popov-Yakubovich theory*, Systems Control Lett., 20 (1993), pp. 1–6.
[12] ———, *Linear Time Varying Discrete Systems*, Operator Theory 67, Birkhauser-Verlag, Basel, 1994.
[13] V. IONESCU AND M. WEISS, *On computing the stabilizing solution of the discrete-time Riccati equation*, Linear Algebra Appl., 174 (1992), pp. 229–238.
[14] ———, *The constrained continuous-time algebraic Riccati equation*, in Proc. 12th World Congress IFAC, Sydney, 1993.
[15] ———, *The constrained discrete-time algebraic Riccati equation*, in Proc. 2nd European Control Conference, Groningen, 1993, pp. 1800–1804.
[16] ———, *Continuous and discrete-time Riccati theory: A Popov function approach*, Linear Algebra Appl., 193 (1993), pp. 173–209.
[17] V. IONESCU AND C. OARĂ, *Generalized continuous-time Riccati theory*, Linear Algebra Appl., (1996), to appear.
[18] ———, *Discrete singular Riccati theory and nonstandard factorizations*, in Proc. 1st Asian Control Conference, July 27–30, 1994, Tokyo.
[19] T. KAILATH, *Linear Systems*, Prentice-Hall, New York, 1980.
[20] A. I. LURIE, *Several Nonlinear Problems in Automatic Control*, Gostehizdat, Moscow, 1951. (In Russian.)
[21] A. LAUB, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Control, AC 24 (1979), pp. 913–921.
[22] C. OARĂ, *Proper deflating subspaces: Properties, algorithms and applications*, Numer. Algorithms, 7 (1994), pp. 355–373.
[23] D. Y. OHM, J. W. HOWZE, AND S. P. BHATTACHARYYA, *Structural synthesis of multivariable controllers*, Automatica, 21 (1975), pp. 35–56.
[24] F. STUMMEL, *Diskrete Konvergenz linearer Operatoren*, Math. Z., 120 (1971), pp. 231–264.
[25] G. TADMOR, *Worst case design in the time-domain: The maximum principle and the standard H^∞ problem*, Math. Control Signals Systems, 3 (1990), pp. 301–324.
[26] M. WEISS, *Spectral and inner-outer factorization in the general case through the constrained Riccati equation*, IEEE Trans. Automat. Control, AC 39 (1994), pp. 677–681.
[27] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Lecture Notes in Economics and Mathematical Systems 101, Springer-Verlag, Berlin, 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 620–634, March 1996

012

APPROXIMATION OF THE ZAKAI EQUATION FOR NONLINEAR FILTERING*

KAZUFUMI ITO†

Abstract. In this paper we consider numerical approximations of solutions to the Zakai equation. Time discretization based on the implicit Milshtein and Euler methods and Galerkin approximation in the spatial coordinates are investigated. Convergence and rate of convergence of approximation methods are established.

Key words. Zakai equation, numerical approximations

AMS subject classifications. 60H15, 65M10, 93E11

1. Introduction.

In this paper we consider approximations of solutions to the Zakai equation of the form

$$dp(t) + A(y(t))p(t)\,dt = B(y(t))p(t)\,dy(t), \quad p(0) = p_0 \in L^2(R^d), \tag{1.1}$$

where $p(t) = p(t,x)$ is the nonnormalized conditional probability density function appearing in nonlinear filtering problem as follows. A signal process $x(t) \in R^d$ satisfies the Ito stochastic differential equation

$$dx(t) = g(x(t), y(t))\,dt + \sigma(x(t))\,dw_1(t), \quad x(0) = x, \tag{1.2}$$

and the observation process $y(t) \in R^p$ is given by

$$dy(t) = h(x(t))\,dt + b(y(t))\,dw_1(t) + dw_2(t), \quad y(0) = 0. \tag{1.3}$$

The dependency of the drift term g on y accounts for output feedback of the observation $y(t)$. Let $(\Omega, \mathcal{F}, P)$ be the probability space with an increasing family of sub-σ-algebras $\{\mathcal{F}_t\}$ of $\mathcal{F}$ is right continuous and complete with respect to the probability measure P. Assume that $w_1(t)$, $w_2(t)$ are $\mathcal{F}_t$-adapted independent Wiener processes with covariance I and R, respectively. The initial condition x is a R^d-valued, $\mathcal{F}_0$-measurable random variable with probability density $p_0(x)$. Suppose $b(y) = 0$; then the signal process and the observation process are uncorrelated. Assume that the functions g, σ, h, and b are bounded and that g, σ, and b are Lipschtz. A core of the Zakai theory (e.g., see [Be], [Pa], [Ro]) is that the conditional expectation of $\phi(x(t))$ based on the observations $\{y(s),\ 0 \le s \le T\}$ is given by

$$E\left[\phi(x(t))|y(s),\ 0 \le s \le t\right] = \frac{\int \phi(x)p(t,x)\,dx}{\int p(t,x)\,dx}. \tag{1.4}$$

The Zakai theory is based on the change of probability measure [Be], [Ro]. Let $\eta(t)$ be a stochastic process defined by

$$\eta(t) = \exp\left(-\int_0^t h^*(x)D(y)^{-1}(b(y)\,dw_1(s) + dw_2(s)) \right.$$
$$\left. -\frac{1}{2}\int_0^t h^*(x)D(y)^{-1}h(x)\,ds\right),$$

* Received by the editors August 30, 1993; accepted for publication (in revised form) December 12, 1994. This research was supported in part by Air Force Office of Scientific Research grant AFOSR-90-0091 and National Science Foundation grant DMS-8818530.

† Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695-8205.

where $D(y) = R + b(y)b(y)^*$. Define a probability measure $\tilde{P}$ on $(\Omega, \mathcal{F})$ by

$$\left.\frac{d\tilde{P}}{dP}\right|_{\mathcal{F}_t} = \eta(t).$$

Then the observation process $y(t)$ becomes an $\mathcal{F}_t$-adapted Wiener process with covariance D on $(\Omega, \mathcal{F}, \tilde{P})$. Henceforth, $\tilde{E}$ denotes the expectation with respect to $\tilde{P}$. The linear operators $A = A(y)$ and $B = B(y)$ appearing in (1.1) are defined by

$$-A\phi = \frac{\partial}{\partial x_i}\left(a_{i,j}\frac{\partial}{\partial x_j}\phi\right) - \frac{\partial}{\partial x_i}(a_i\phi) \tag{1.5}$$

and

$$B\phi = hD^{-1}\phi - \frac{\partial}{\partial x_i}(c_i\phi),$$

where

$$c = \sigma b^* D^{-1}, \quad a = \frac{1}{2}\sigma\sigma^* + \frac{1}{2}cDc^*, \quad \text{and} \quad a_i = g_i - \frac{\partial}{\partial x_j}a_{i,j}.$$

As in [Be], [Pa], [Ro] we employ the variational formulation of (1.1). Let $H = L^2(R^d)$ and $V = H^1(R^d)$, and V^* denotes the strong dual space of V. H^* is identified with H so that $V \subset H = H^* \subset V^*$. We define the inner product $(\cdot, \cdot)$ of H by

$$(\phi, \psi) = \int_{R^d} \phi(x)\psi(x)\,dx.$$

The dual product of $V^* \times V$ is denoted by $\langle \cdot, \cdot \rangle$. Let $H_k(x)$, $k \geq 0$, be the Hermite polynomials on R. Then a family of functions

$$e_k(x) = \prod_{i=0}^{d} \exp(-x_i^2/2)H_{k_i}(x_i), \quad k = (k_1, \ldots, k_d) \in N^d \tag{1.6}$$

forms the orthogonal basis of $L^2(R^d)$. If for $\phi \in V$ we define

$$\phi^n(x) = \sum_{k \in J} (\phi, e_k)\, e_k(x),$$

where $J = [0, n]^d \in N^d$, then $|\phi^n - \phi|_V \to 0$ as $n \to \infty$. Then we have that for $\phi, \psi \in V$

$$\langle A\phi, \psi \rangle = \lim (A\phi^n, \psi) = \lim (A_0\nabla\phi^n - a\phi^n, \nabla\psi) = \langle A_0\nabla\phi - a\phi, \nabla\psi \rangle, \tag{1.7}$$

where A_0 denotes the symmetric matrix $\{a_{i,j}\}$ on R^d. We assume that there exists a constant $\alpha > 0$ such that

$$x^* A_0 x \geq \alpha |x|^2 \quad \text{for all } x \in R^d,\ y \in R^p. \tag{1.8}$$

Hence $A = A(y) \in \mathcal{L}(V, V^*)$, and if $\sigma\sigma^*$ is coercive uniformly in x and y, then there exist positive constants ρ and β such that

(1.9)

$$\langle A(y)\phi, \phi \rangle - \frac{1}{2}|B(y)\phi|_D^2 + \rho|\phi|_H^2 \geq \frac{1}{2}\beta|\phi|_V^2 \quad \text{for all } \phi \in V \text{ and uniformly in } y,$$

where $|\phi|_V^2 = (\nabla\phi, \nabla\phi) + |\phi|_H^2$ and $|B\phi|_D^2 = \int (B\phi)D(y)(B\phi)^* \, dx$.

An outline of our paper is as follows. In §2 we consider an approximation scheme (2.1) in which the backward Euler scheme is used to discretize (1.1) in time and the operators A, B are approximated by a sequence of stable and consistent approximations (see (A1)–(A3)), and its convergence properties are established. For the uncorrelated case a higher-order approximation based on the Milshtein scheme [Mi] is discussed in §3. The convergence rate of both schemes is established in §4. In §3 we also consider the robust form of the Zakai equation for the uncorrelated case. Finally, we apply our results to the spectral method based on the Hermite polynomials.

We refer to [BGR], [FL], [Pi] for time-discretization schemes for the Zakai equation. In [BGR], [FL] the operator splitting method is used to split up the deterministic and stochastic evolutions (see also §3). Numerical experiments based on the methods described in this paper will be reported in a forthcoming paper.

2. Implicit Euler approximation. In this section we consider the following approximation method of (1.1); for $n, m \in N$ the sequence $\{p_{m,n}^k\}$ is generated by

$$p^k - p^{k-1} + \lambda (A^n(y_{k-1}) + \rho I)p^k = \lambda\rho\, p^{k-1} + B^n(y_{k-1})p^{k-1}(y_k - y_{k-1}), \tag{2.1}$$

where $\lambda = T/m$, $y_k = y(k\lambda)$, and the pair of linear operators $(A^n(y), B^n(y))$ satisfies the approximation conditions:

$$\text{(A1)} \quad |A^n(y)\phi|_{V^*} \le \gamma\,|\phi|_V \quad \text{and} \quad |B^n(y)\phi|_H \le \gamma\,|\phi|_V \quad \text{uniformly in } n \text{ and } y \in R^p,$$

$$\text{(A2)} \quad \langle A^n(y)\phi, \phi\rangle - \frac{1}{2}|B^n(y)\phi|_D^2 + \rho\,|\phi|_H^2 \ge \frac{1}{2}\beta\,|\phi|_V^2 \quad \text{for all } \phi \in V \text{ and } y \in R^p,$$

$$\text{(A3)} \quad |A^n(y)\phi - A(y)\phi|_{V^*} + |B^n(y)\phi - B(y)\phi|_H \to 0 \text{ as } n \to \infty \text{ for all } \phi \in V \text{ and } y \in R^p.$$

For ease of presentation $A(y_{k-1})$ and $B(y_{k-1})$ will be denoted simply by A and B, respectively, throughout the paper. Similarly, the dependency of (A^n, B^n) on $y(\cdot)$ will be suppressed and understood from the context of our discussions. Let $S = (\Omega \times [0,T], \mathcal{F} \times \mathcal{B}, d\tilde{P} \times dt)$. Then we have the convergence theory.

THEOREM 2.1. *Assume that a family of linear operators (A^n, B^n) and initial conditions p_0^n satisfy the conditions* (A1)–(A3) *and $|p_0^n - p_0|_H \to 0$ as $n \to \infty$, respectively. Then the sequence $\{p_{m,n}^k\}$ generated by* (2.1) *with initial condition p_0^n converges to the unique solution $p(t)$ of* (1.1) *as $m,\ n \to \infty$ in the sense that the function $p_{m,n}(t)$ defined by*

$$p_{m,n}(t) = p_{m,n}^k \quad \text{on } [k\lambda, (k+1)\lambda)$$

converges to $p(t)$ strongly in $L^2(S;V)$ and weakly star in $L^\infty(0,T;L^2(\Omega,H,d\tilde{P}))$.

Proof. We prove the theorem in two steps. First we show convergence of the sequence $\{p^k\}$ generated by

$$p^k - p^{k-1} + \lambda (A + \rho I)p^k = \lambda\rho\, p^{k-1} + Bp^{k-1}(y_k - y_{k-1}) \tag{2.2}$$

to the unique solution $p(t)$ of (1.1) provided that (1.9) holds. Define a linear operator A_ρ by $A + \rho I$. Then from (1.9) $J_\lambda = (I + \lambda A_\rho)^{-1} \in \mathcal{L}(V^*, V)$. Thus, given V-valued random variable x, the equation

$$\hat{x} - x + \lambda A_\rho \hat{x} = \lambda\rho\, x + Bx\,\Delta y \tag{2.3}$$

has a unique solution $\hat{x} = J_\lambda((1+\lambda\rho)x + Bx\,\Delta y))$, where $\Delta y = y_k - y_{k-1}$ is a R^p-valued random variable with mean 0 and covariance λD that is independent of x. Multiplying (2.3) by $\hat{x}$,

$$(\hat{x} - (1+\lambda\rho)x,\, \hat{x}) + \lambda\,\langle A_\rho \hat{x},\, \hat{x}\rangle = (Bx\,\Delta y,\, \hat{x}).$$

Completing the square,

$$\frac{1}{2}(\,|\hat{x}|_H^2 - (1+\lambda\rho)^2\,|x|_H^2 + |\hat{x} - (1+\lambda\rho)x|_H^2\,) + \lambda\,\langle A_\rho\hat{x},\, \hat{x}\rangle = (Bx\,\Delta y,\, \hat{x}).$$

Let $z = (1+\lambda\rho)x + Bx\,\Delta y$. Then $\hat{x} = J_\lambda z$ and

$$\begin{aligned} |J_\lambda z - z|_H^2 &= |\hat{x} - (1+\lambda\rho)x - Bx\,\Delta y|_H^2 \\ &= |\hat{x} - (1+\lambda\rho)x|_H^2 - 2\,(Bx\,\Delta y,\, \hat{x} - (1+\lambda\rho)x) + |Bx\,\Delta y|_H^2. \end{aligned}$$

Note that $\tilde{E}\,(Bx\,\Delta y,\, x) = 0$ since x and Δy are independent. Hence, combining the above equalities, we obtain

$$\frac{1}{2}\,(\tilde{E}|\hat{x}|_H^2 - (1+\lambda\rho)^2\,\tilde{E}|x|_H^2) + \lambda\,\tilde{E}\left(\langle A_\rho\hat{x},\, \hat{x}\rangle - \frac{1}{2}|Bx|_D^2\right) \le 0. \tag{2.4}$$

Setting $\hat{x} = p^k$ and $x = p^{k-1}$ in (2.4) and multiplying it by $(1+\lambda\rho)^{-2k}$, it follows from (1.9) that for $k \ge 1$

$$c^{-k}\,\tilde{E}\,(|p^k|_H^2 + \lambda\,|Bp^k|_D^2) + \lambda\beta c^{-k}\,\tilde{E}\,|p^k|_V^2 \le c^{-(k-1)}\,(\tilde{E}\,(|p^{k-1}|_H^2 + \lambda\,|Bp^{k-1}|_D^2)),$$

where $c = (1+\lambda\rho)^2$. Summing up this in k, we obtain

$$c^{-m}\,\tilde{E}\,(|p^m|_H^2 + \lambda\,|Bp^m|_D^2) + \beta\,\tilde{E}\sum_{k=1}^{m}\lambda c^{-k}\,|p^k|_V^2 \le \tilde{E}(|p^0|_H^2 + \lambda\,|Bp^0|_D^2).$$

Thus, for $\lambda\rho$ sufficiently small,

$$\tilde{E}|p^m|_H^2 + \beta\,\tilde{E}\sum_{k=1}^{m}\lambda c^{m-k}\,|p^k|_V^2 \le e^{2\rho T}\,\tilde{E}\,(|p^0|_H^2 + \lambda\,|Bp^0|_D^2). \tag{2.5}$$

Let Δ be the Laplacian and $p^0 = (I - \lambda\,\Delta)^{-1}p_0$. Here, for $x \in H$, we have

$$|(I - \lambda\,\Delta)^{-1}x - x|_H \to 0 \quad \text{and} \quad \lambda\,|B(I - \lambda\,\Delta)^{-1}x|_D^2 \to 0$$

as $\lambda \to 0$. Define a function $p_\lambda(t)$ by

$$p_\lambda(t) = p^k \quad \text{if} \quad t \in [k\lambda, (k+1)\lambda),\ k \ge 0.$$

Then from (2.5) we obtain

$$\sup_{t\in[0,T]}\tilde{E}\,|p_\lambda(t)|_H^2 + \int_0^T \tilde{E}\,|p_\lambda(t)|_V^2\,dt \le e^{2\rho T}\,\tilde{E}\,(|p^0|_H^2 + \lambda\,|Bp^0|_D^2), \tag{2.6}$$

and thus $p_\lambda(t) \in L^2(S;V) \cap L^\infty(0,T;L^2(\Omega;H))$ uniformly in $\lambda > 0$. Consequently, there exists a subsequence of $\{p_\lambda\}$, which will be denoted by the same symbol, such

that $p_\lambda \to p$ weakly in $L^2(S;V)$ and weakly star in $L^\infty(0,T;L^2(\Omega;H))$, where $p(t) \in L^2(S;V) \cap L^\infty(0,T;L^2(\Omega,H))$ can be chosen to be progressively measurable. Note that

$$\tilde{E}\,|\,(A_\rho(y(k\lambda+s)) - A_\rho(y(k\lambda)))\,p_\lambda(k\lambda)|_{V^*}^2$$

$$\le M\tilde{E}\,|y(k\lambda+s)-y(k\lambda)|^2\,\tilde{E}\,|p_\lambda(k\lambda)|_V^2 \le M\lambda\tilde{E}\,|p_\lambda(k\lambda)|_V^2$$

for $s \in (0,\lambda)$ and some constant $M > 0$. Similarly, we have

$$\tilde{E}\,|\,(B(y(k\lambda+s)) - B(y(k\lambda)))\,p_\lambda(k\lambda)|_H^2 \le M\lambda\tilde{E}\,|p_\lambda(k\lambda)|_V^2.$$

Thus we have

(2.7)

$A_\rho p_\lambda(t) \to A_\rho p(t)$ weakly in $L^2(S;V^*)$ and $Bp_\lambda(t) \to Bp(t)$ weakly in $L^2(S;\mathcal{L}(R^p,H))$.

From (2.2) we have

$$p_\lambda(t) = p^0 - \int_0^{[t/\lambda]\lambda} (A_\rho p_\lambda(s+\lambda) - \rho\, p_\lambda(s))\,ds + \int_0^{[t/\lambda]\lambda} Bp_\lambda(s)\,dy(s).$$

Thus, for almost all (t,ω), a continuous modification of $p(t)$ in H (see [KR, Thm. 3.2]) satisfies

$$p(t) = p_0 - \int_0^t (A_\rho p(s) - \rho\, p(s))\,ds + \int_0^t Bp(s)\,dy(s). \tag{2.8}$$

Define

$$S_\lambda = \int_0^T e_\lambda(t)\tilde{E}\left(2\langle A_\rho p_\lambda(t+\lambda) - A_\rho p(t),\, p_\lambda(t+\lambda) - p(t)\rangle - |Bp_\lambda(t+\lambda) - Bp(t)|_D^2\right)\,dt,$$

where $e_\lambda(t) = c^{-k}$ on $[(k-1)\lambda, k\lambda)$, $k \ge 1$. It then follows from (1.9) that

$$S_\lambda \ge \beta \int_0^T e_\lambda(t)\,\tilde{E}\,|p_\lambda(t+\lambda) - p(t)|_V^2\,dt. \tag{2.9}$$

Setting $\hat{x} = p^k$ and $x = p^{k-1}$ in (2.4) and multiplying it by $c^{-k} = (1+\lambda\rho)^{-2k}$, we obtain for $k \ge 1$

$$c^{-k}\tilde{E}(|p^k|_H^2 + \lambda\,|Bp^k|^2) - c^{k-1}\tilde{E}(|p^{k-1}|_H^2 + \lambda\,|Bp^{k-1}|^2) + \lambda c^{-k}\,\tilde{E}(2\langle A_\rho p^k, p^k\rangle - |Bp^k|_D^2) \le 0.$$

Summing up this in k, we obtain

$$c^{-m}\tilde{E}\,|p_\lambda(T)|_H^2 - \tilde{E}\,(|p^0|_H^2 + \lambda\,|Bp^0|_D^2) + T_\lambda \;\le\; 0, \tag{2.10}$$

where T_λ is defined by

$$T_\lambda = \int_0^T e_\lambda(t)\,\tilde{E}\,(2\langle A_\rho p_\lambda(t+\lambda),\, p_\lambda(t+\lambda)\rangle - |Bp_\lambda(t+\lambda)|_D^2)\,dt.$$

On the other hand, it follows from (2.8) and the Ito lemma (Thm. I.3.2 in [KR]) that

$$e^{-2\rho T}\,\tilde{E}\,|p(T)|_H^2 - \tilde{E}\,|p_0|_H^2 + \int_0^T e^{-2\rho t}\tilde{E}\,(\,2\langle A_\rho p(t),\, p(t)\rangle - |Bp(t)|_D^2\,)\,dt = 0. \tag{2.11}$$

Note that $|e_\lambda(t) - e^{-2\rho t}| \to 0$ as $\lambda \to 0$, uniformly on $[0,T]$. Hence, since from (2.7)

$$\liminf\ S_\lambda = \int_0^T e^{-2\rho t}\tilde{E}\left(-2\langle A_\rho p(t),\, p(t)\rangle + |Bp(t)|_D^2\right) dt + \liminf\ T_\lambda \geq 0,$$

it follows from (2.10) that

(2.12)

$$\int_0^T e^{2\rho(t-s)}\tilde{E}(-2\langle A_\rho p(t),\, p(t)\rangle + |Bp(t)|_D^2)\, dt + \tilde{E}|p_0|_H^2 - e^{-2\rho T}\ \limsup\ \tilde{E}|p_\lambda(T)|_H^2 \geq 0.$$

Combining (2.11) – (2.12), we obtain

$$e^{-2\rho T}\left(\tilde{E}\,|p(T)|_H^2 - \limsup\ \tilde{E}\,|p_\lambda(T)|_H^2\right) \geq 0.$$

Since $p_\lambda(T)$ converges weakly to $p(T)$ (without loss of generality),

$$\tilde{E}\,|p(T)|_H^2 - \liminf\ \tilde{E}\,|p_\lambda(T)|_H^2 \leq 0,$$

and thus we have

$$\tilde{E}\,|p_\lambda(T) - p(T)|_H^2 \to 0 \quad \text{as } \lambda \to 0.$$

Moreover, from (2.9)

$$\tilde{E}\,|p_\lambda(t) - p(t)|_{L^2(0,T;V)} \to 0 \quad \text{as } \lambda \to 0.$$

Suppose there exist two solutions $p(t)$, $\hat{p}(t) \in L^2(S;V)$ to (2.8) satisfying $p(0) = \hat{p}(0) = p_0$. It then follows from the Ito lemma that

$$\tilde{E}\,|p(t) - \hat{p}(t)|_H^2 + \int_0^t e^{-2\rho s}\tilde{E}\left(2\langle A_\rho(p(s) - \hat{p}(s)),\, p(s) - \hat{p}(s)\rangle - |B(p(s) - \hat{p}(s))|_D^2\right) ds = 0$$

for $t \in [0,T]$. Hence from (1.9) $\tilde{E}\,|p(t) - \hat{p}(t)|_H^2 = 0$ on $[0,T]$, and thus (2.8) has a unique solution. Since the above arguments do not depend on $T > 0$, $p_\lambda(t)$ converges strongly to $p(t)$, a unique solution of (2.8) in $L^2(\Omega;H)$ everywhere on $[0,T]$. Hence $p_\lambda(t)$ converges to $p(t)$ in $L^2(S;V)$ and strongly in $L^2(\Omega,H)$ everywhere on $[0,T]$.

Next we show that if for each n, $p_n(t)$ is the solution to

$$p_n(t) = p_0^n - \int_0^t A^n p_n(s)\, ds + \int_0^t B^n p_n(s)\, dy(s) \quad \text{in } V^*,$$

then $p_n(t)$ converges to $p(t)$ strongly in $L^2(S;V)$ and $L^\infty(0,T;L^2(\Omega;H,d\tilde{P}))$ as $n \to \infty$. It follows from the Ito lemma that

$$\tilde{E}\,|p_n(t) - p(t)|_H^2 + \int_0^t e^{2\rho(t-s)}\tilde{E}\left(2\langle A_\rho^n(p_n(s) - p(s)) + A_\rho^n p(s) - A_\rho p(s),\, p_n(s) - p(s)\rangle\right.$$

$$\left. - |B^n(p_n(s) - p(s)) + B^n p(s) - Bp(s)|_D^2\right) ds = e^{2\rho t}\,\tilde{E}\,|p_0^n - p_0|_H^2.$$

From (A1) we have

$$\langle A_\rho^n p_n(s) - A_\rho p(s),\, p_n(s) - p(s)\rangle \leq \frac{\beta}{4}\,|p_n(s) - p(s)|_V^2 + \frac{1}{\beta}\,|A_\rho^n p_n(s) - A_\rho p(s)|_{V^*}^2$$

and

$$(B^n(p_n(s)-p(s)),\, B^n p(s)-Bp(s))_D \le \frac{\beta}{4}\,|p_n(s)-p(s)|_V^2 + \frac{\gamma^2}{\beta}|B^n p_n(s)-Bp(s)|_D^2.$$

It thus follows from (A2) that

(2.13)

$$\tilde{E}\,|p_n(t)-p(t)|_H^2 + \frac{\beta}{4}\int_0^t e^{2\rho(t-s)}\tilde{E}\,|p_n(s)-p(s)|_V^2\,ds \le e^{2\rho t}\,\tilde{E}\,|p_0^n-p_0|_H^2$$

$$+\int_0^t e^{2\rho(t-s)}\tilde{E}\left(\frac{2}{\beta}\,|A_\rho^n p(s)-A_\rho p(s)|_{V^*}^2\,ds + \left(\frac{\gamma^2}{\beta}+1\right)|B^n p(s)-Bp(s)|_D^2\right)ds.$$

It then follows from (A1) and (A3) that $|A_\rho^n p(s)-A_\rho p(s)|_{V^*}^2$ and $|B^n p(s)-Bp(s)|_D^2$ converge to zero as $n\to\infty$, almost all (t,ω) and are uniformly integrable. By Lebesgue-dominated convergence theory,

$$\int_0^T e^{2\rho(T-t)}\tilde{E}\,(|A_\rho^n p(t)-A_\rho p(t)|_{V^*}^2 + |B^n p(t)-Bp(t)|_D^2)\,dt$$

converges to zero as $n\to\infty$. Thus it follows from (2.13) that $p_n(t)$ converges to $p(t)$ strongly in $L^2(S;V)$ and $L^\infty(0,T;L^2(\Omega;H,d\tilde{P}))$ as $n\to\infty$. Hence, combining the two steps we obtain the desired convergence. □

3. Uncorrelated case: Milshtein approximation and robust form. In this section we consider the case when $b(y)=0$. Then $B\in\mathcal{L}(H)$ is given by $(B\phi)(x)=h(x)\phi(x)$. Here, without loss of generality, one can assume that $R=I$ and thus $D=I$ on R^p. First, we consider the time discretization of (1.1) based on the Milshtein approximation of the Ito stochastic integral:

$$p^k-p^{k-1}+\lambda\,(A+\rho\,I)p^k = \lambda\rho\,p^{k-1}+Bp^{k-1}\,\Delta y_k+\frac{1}{2}\,BBp^{k-1}(\Delta y_k^2-\lambda), \tag{3.1}$$

where $A=A(y_{k-1})$, $B=B(y_{k-1})$, and $\Delta y_k = y(k\lambda)-y((k-1)\lambda)$. The last term in (3.1) is understood as

$$\sum_{i=1}^p B_iB_i\phi(\,(\Delta y_k^i)^2-\lambda) \quad \text{for } \phi\in H,$$

where $(B_i\phi)(x)=h_i(x)\phi(x)$. Note that B_i is self-adjoint and commutes with B_j for all i, j. As shown in the next section, the Milshtein scheme (3.1) provides a higher-order approximation than the Euler scheme (2.2). Assume that x is a V-valued random variable and x and $\Delta y\in R^p$ are independent. Then the equation for $\hat{x}\in V$

$$\hat{x}-x+\lambda\,(A+\rho\,I)\hat{x} = \lambda\rho\,x+Bx\,\Delta y+\frac{1}{2}BBx\,(\Delta y^2-\lambda) \tag{3.2}$$

has a unique solution $\hat{x}=J_\lambda z$, where $z=(1+\lambda\rho)\,x+Bx\,\Delta y+\frac{1}{2}BBx\,(\Delta y^2-\lambda)$. Multiplying (3.2) by $\hat{x}$ and completing the square, we obtain

$$\frac{1}{2}\,(\,|\hat{x}|_H^2-(1+\lambda\rho)^2\,|x|_H^2+|\hat{x}-(1+\lambda\rho)\,x|_H^2\,)+\lambda\,\langle A_\rho\hat{x},\,\hat{x}\rangle = (\xi,\hat{x}),$$

where $\xi = Bx\,\Delta y + \frac{1}{2}BBx\,(\Delta y^2 - \lambda)$. Then

$$|J_\lambda z - z|_H^2 = |\hat{x} - (1+\lambda\rho)\,x - \xi|_H^2$$
$$= |\hat{x} - (1+\lambda\rho)\,x|_H^2 - 2\,(\xi,\ \hat{x} - (1+\lambda\rho)\,x) + |\xi|_H^2.$$

Note that $\tilde{E}\,(\xi,\ x) = 0$ since x and Δy are independent and that

$$\tilde{E}\,|\xi|_H^2 = \tilde{E}\,|Bx\,\Delta y|_H^2 + \tilde{E}\,\left|\frac{1}{2}BBx\,(\Delta y^2 - \lambda)\right|_H^2$$
$$= \lambda\left(\tilde{E}\,|Bx|_H^2 + \frac{1}{2}\lambda\,\tilde{E}\,|BBx|_H^2\right).$$

Hence, combining the above equalities, we obtain

$$\frac{1}{2}\,(\tilde{E}\,|\hat{x}|_H^2 - (1+\lambda\rho)^2\,\tilde{E}\,|x|_H^2) + \lambda\,\tilde{E}\left(\langle A_\rho\hat{x},\ \hat{x}\rangle - \frac{1}{2}\,|Bx|_H^2\right) \le \frac{1}{4}\lambda^2\,\tilde{E}\,|BBx|_H^2. \tag{3.3}$$

Setting $\hat{x} = p^k$ and $x = p^{k-1}$ in (3.3) and multiplying it by $(1+\lambda\rho)^{-2k}$, it follows from (3.1) and (1.9) that, for $k \ge 1$,

$$c^{-k}\,\tilde{E}\,(|p^k|_H^2 + \lambda\,|Bp^k|_H^2) - c^{k-1}\,\tilde{E}\,(|p^{k-1}|_H^2 + \lambda\,|Bp^{k-1}|_H^2)$$
$$+\lambda\beta c^{-k}\,\tilde{E}\,|p^k|_V^2 \le \frac{1}{2}\lambda^2 c^{-k}\,\tilde{E}\,|BBp^{k-1}|_H^2, \tag{3.4}$$

where $c = (1+\lambda\rho)^2$. Summing up this in k, we have

$$c^{-m}\,\tilde{E}\,(|p^m|_H^2 + \lambda\,|Bp^m|_H^2) + \beta\,\tilde{E}\sum_{k=1}^{m}\lambda c^{-k}\,|p^k|_V^2$$
$$\le \tilde{E}\,(|p_0|_H^2 + \lambda\,|Bp_0|_H^2) + \frac{1}{2}\lambda|B|^2\,\tilde{E}\sum_{k=1}^{m}\lambda c^{-k}\,|p^{k-1}|_H^2.$$

Thus, for $\lambda\rho$ sufficiently small,

$$\tilde{E}\,|p^m|_H^2 + \frac{\beta}{2}\,\tilde{E}\sum_{k=1}^{m}\lambda c^{m-k}\,|p^k|_V^2 \le e^{2\rho T}\,\tilde{E}\,(|p_0|_H^2 + \lambda\,|Bp_0|_H^2),$$

where we assume that $\beta \le 2\lambda|B|^2$. Define a function $p_\lambda(t)$ by

$$p_\lambda(t) = p^k \quad \text{if } t \in [k\lambda, (k+1)\lambda),\ k \ge 0.$$

Then from (3.4) instead of (2.10) we have

$$c^{-m}\,\tilde{E}\,|p_\lambda(T)|_H^2 - \tilde{E}\,(|p_0|_H^2 + \lambda\,|Bp_0|_H^2) + T_\lambda \le \frac{1}{2}\lambda|B|^2\,\tilde{E}\sum_{k=1}^{m}\lambda c^{-k}\,|p^{k-1}|_H^2,$$

where the right-hand side converges to zero as $\lambda \to 0$. Thus, using the same arguments as in the proof of Theorem 2.1, one can show that $p_\lambda(t)$ converges to $p(t)$ strongly

in $L^2(S;V)$ and weakly star in $L^\infty(0,T;L^2(\Omega,H))$. Hence we obtain the following theorem.

THEOREM 3.1. *Assume that* $b(y) = 0$ *and that a family of linear operators* (A^n, B^n) *and initial conditions* p_0^n *satisfy the conditions* (A1)–(A3) *and* $|p_0^n - p_0|_H \to 0$ *as* $n \to \infty$, *respectively. Then the sequence* $\{p_{m,n}^k\}$ *generated by*

$$(3.5)\quad p^k - p^{k-1} + \lambda(A^n + \rho I)p^k = \lambda\rho p^{k-1} + B^n p^{k-1}\,\Delta y_k + \frac{1}{2}B^n B^n p^{k-1}(\Delta y_k^2 - \lambda)$$

converges to the unique solution $p(t)$ *of* (1.1) *as* $m,\ n \to \infty$ *in the sense that the function* $p_{m,n}(t)$ *defined by*

$$p_{m,n}(t) = p_{m,n}^k \quad \text{on } [k\lambda, (k+1)\lambda),\ k \geq 0,$$

converges $p(t)$ *strongly in* $L^2(S;V)$ *and weakly star in* $L^\infty(0,T;L^2(\Omega;H,d\tilde{P}))$.

Remark 3.2. Consider the approximation scheme based on the Trotter product formula (e.g., [BGR], [FL], [Pi]):

$$(3.6)\quad p^k - p^{k-1} + \lambda(A + \rho I)p^k = \lambda\rho p^{k-1} + \left(\exp\left(B\,\Delta y_k - \frac{\lambda}{2}BB\right) - I\right)p^{k-1},$$

where $\hat{p}(t) = (\exp(B\,\Delta y_k - \frac{\lambda}{2}BB) - I)p^{k-1}$ satisfies

$$\hat{p}(t) - p^{k-1} = \int_{k-1\lambda}^{t} B\hat{p}(s)dy(s), \quad t \geq (k-1)\lambda.$$

For the scheme (3.6), in steps (3.1)–(3.3) ξ is defined by

$$\xi = \left(\exp\left(B\,\Delta y_k - \frac{\lambda}{2}BB\right) - I\right)p^{k-1}.$$

Note that

$$\hat{p}(t) - p^{k-1} = Bp^{k-1}\Delta y_k + \int_{k-1\lambda}^{t} B(\hat{\rho}(s) - p^{k-1})dy(s),$$

where $x = p^{k-1}$ in (3.2). Hence

$$\tilde{E}\,|\xi|_H^2 \leq \lambda(1 + M\lambda)\,\tilde{E}\,|Bx|^2$$

for some $M > 0$ independent of $\lambda > 0$ and $x \in L^2(\Omega;H)$. Thus exactly the same arguments as above are applied to show that Theorem 3.1 also holds for the sequence generated by (3.6)

Next we consider the robust form of (1.1) in which the Ito stochastic differential equation is transformed into an ordinary partial differential equation with random coefficients [Cl], [Da], [Be], [Ro]. Define a function $\eta(t)$ by

$$\eta(t,x) = \exp(-h(x)^* y(t)).$$

Then $\eta(t)$ satisfies

$$d\eta(t,x) = \frac{1}{2}|h(x)|^2\eta(t,x)\,dt - h(x)\eta(t,x)\,dy(t).$$

If we define a function $q(t)$ by $q(t,x) = \eta(t,x)p(t,x)$, then it follows from the Ito stochastic differential rule (e.g., see [KR]) that $q(t)$ satisfies

$$\frac{d}{dt}q(t) + \eta A(\eta^{-1}q(t)) + \frac{1}{2}|h|^2 q(t) = 0, \quad q(0) = p_0 \in H. \tag{3.7}$$

The weak or variational form of (3.7) is given by

$$\left\langle \frac{d}{dt}q(t), \psi \right\rangle + \langle A_0(\nabla q(t) + q(t)\nabla\chi) - aq(t), \nabla\psi - \psi\nabla\chi\rangle + \frac{1}{2}(|h|^2 q(t), \psi) = 0$$

for all $\psi \in V$, where $\chi(t,x) = h(x)^* y(t)$ is a pathwise continuous random function. Define a sesquilinear form μ on $V \times V$ by

$$\mu(t,\phi,\psi) = \langle A_0(\nabla\phi + \phi\nabla\chi) - a\phi, \nabla\psi - \psi\nabla\chi\rangle + \frac{1}{2}(|h|^2\phi, \psi)_H.$$

Then there exist functions $M(t,\omega)$ and $\rho(t,\omega)$ such that

$$|\mu(t,\phi,\psi)| \le M\,|\phi|_V|\psi|_V, \quad \phi, \psi \in V,$$

and

$$\operatorname{Re}\mu(t,\phi) + \rho\,|\phi|_H^2 \ge \frac{1}{2}\alpha\,|\phi|_V^2, \quad \phi \in V,$$

for almost all (t,ω). Hence the standard theory of the parabolic equation in [Li], [KR] shows that there exists a pathwise unique solution $q(t) \in L^2(0,T;V) \cap H^1(0,T;V^*) \cap C(0,T;H)$. Let us consider a full approximation scheme of (3.7); i.e., a sequence of functions $\{q^k\}$ that takes values in a finite dimensional subspace V^n of V is defined by

$$\left(\frac{q^k - q^{k-1}}{\lambda}, \psi\right) + (A_0(\nabla q^k + q^k\nabla\chi^k) - a\,q^k, \nabla\psi - \psi\nabla\chi^k) + \frac{1}{2}(|h|^2 q^k, \psi) = 0$$

for all $\psi \in V^n$, where $\lambda = T/m$ and $\chi^k = \chi(k\lambda)$. Assume the approximation condition:

for each ϕ in V there exists a sequence of function ϕ^n in V^n such that $|\phi^n - \phi|_V \to 0$ as $n \to \infty$.

Then it is easy to prove (e.g, see [Li] and Theorem 2.1) that the function $q_\lambda^n(t)$ defined by

$$q_\lambda^n(t) = q^k + \frac{q^{k+1} - q^k}{\lambda}(t - k\lambda) \quad \text{on } [k\lambda, (k+1)\lambda]$$

converges to $q(t)$ almost surely in $L^2(0,T;V) \cap H^1(0,T;V^*) \cap C(0,T;H)$ as $m,\ n \to \infty$.

4. Convergence rate. In this section we establish convergence rate of the time discretization schemes (2.1) and (3.1). Assume that $g = g(x)$ does not depend on y. First we consider the scheme (2.1). For $k \ge 0$ define the approximation error $\epsilon_k = \epsilon_k^{(1)} - \epsilon_k^{(2)}$ by

$$\begin{aligned} \epsilon_k^{(1)} &= \int_{t_{k-1}}^{t_k} A(p(t) - p(t_k))\,dt + \lambda\rho\,(p(t_k)) - p(t_{k-1})), \\ \epsilon_k^{(2)} &= \int_{t_{k-1}}^{t_k} B(p(t) - p(t_{k-1}))\,dy(t), \end{aligned} \tag{4.1}$$

where $p(t)$ is the unique solution to (1.1) and $t_k = k\lambda$. Then since (1.1) is equivalently written as

$$p(t_k) - p(t_{k-1}) + \int_{t_{k-1}}^{t_k} Ap(t)\,dt = \int_{t_{k-1}}^{t_k} Bp(t)\,dy(t), \tag{4.2}$$

the error function $\delta p_k = p^k - p(t_k)$ satisfies

$$\delta p_k - \delta p_{k-1} + \lambda(A + \rho I)\delta p_k = \lambda\rho\,\delta p_{k-1} + B\delta p_{k-1}\,\Delta y_k + \epsilon_k, \tag{4.3}$$

where $\lambda = T/m$, $\Delta y_k = y(t_k) - y(t_{k-1})$, and $\{p^k\}$ is the approximate sequence generated by (2.2). Note that $\tilde{E}(\delta p_{k-1}, \epsilon_k^{(2)}) = 0$. Thus, multiplying (4.2) by δp_k and using exactly the same arguments as in the proof of Theorem 2.1, we obtain

$$\tilde{E}\,|\delta p_k|_H^2 - c\,\tilde{E}\,|\delta p_{k-1}|_H^2 + 2\,\tilde{E}\langle\lambda A_\rho \delta p_k - \epsilon_k^{(1)},\, \delta p_k\rangle \le \tilde{E}\,|B\delta p_{k-1}\,\Delta y_k - \epsilon_k^{(2)}|_H^2,$$

where $c = (1+\lambda\rho)^2$. Since $B_i \in \mathcal{L}(V,H)$ it follows from (1.9) that there exists an $\varepsilon > 0$ such that

$$2\,\langle A_\rho\phi,\,\phi\rangle - (1+\varepsilon)\,|B\phi|_D^2 \ge \frac{3\beta}{4}\,|\phi|_V^2 \quad \text{for all } \phi \in V.$$

Note that

$$\tilde{E}\,|B\delta p_{k-1}\,\Delta y_k - \epsilon_k^{(2)}|_H^2 \le (1+\varepsilon)\,\tilde{E}\,|B\delta p_{k-1}|_D^2 + \left(1 + \frac{1}{\varepsilon}\right)\tilde{E}\,|\epsilon_k^{(2)}|_H^2.$$

Since $2\,(\delta p_k, \epsilon_k^{(1)}) \le \frac{\lambda\beta}{4}\,|\delta p_k|_V^2 + \frac{4}{\lambda\beta}|\epsilon_k^{(1)}|_{V^*}^2$ we have

$$\begin{aligned}
&\tilde{E}\,(|\delta p_k|_H^2 + \lambda\,|B\delta p_k|_D^2) + \frac{\lambda\beta}{2}\,\tilde{E}|\delta p_k|_V^2 \\
&\le c\,\tilde{E}\,(|\delta p_{k-1}|_H^2 + \lambda\,|B\delta p_{k-1}|_D^2) + \frac{4\lambda}{\beta}\,\tilde{E}\left|\frac{\epsilon_k^{(1)}}{\lambda}\right|_{V^*}^2 + \left(1 + \frac{1}{\varepsilon}\right)\tilde{E}\,|\epsilon_k^{(2)}|_H^2.
\end{aligned}$$

Multiplying this by c^{-k} and summing it up in k,

$$c^{-k}\,\tilde{E}\,|\delta p_k|_H^2 + \frac{\beta}{2}\,\tilde{E}\sum_{j=1}^{k}\lambda c^{-j}\,|\delta p_j|_V^2 \le \tilde{E}\sum_{j=1}^{k}\lambda c^{-j}\left(\frac{4}{\beta}\left|\frac{\epsilon_j^{(1)}}{\lambda}\right|_{V^*}^2 + \frac{1+\frac{1}{\varepsilon}}{\lambda}\,\tilde{E}\,|\epsilon_j^{(2)}|_H^2\right) \tag{4.4}$$

for $1 \le k \le m$.

THEOREM 4.1. *Assume that the solution $p(t)$ of (1.1) satisfies the following regularity:*

$$\tilde{E}\,|Bp(t)|_V^2 \le M_1 \quad \textit{and} \quad \int_0^T \tilde{E}\,|Ap(t)|_V^2\,dt \le M_1 \tag{4.5}$$

for some $M_1 > 0$, independent of $t \in [0,T]$. Then

$$\tilde{E}\,|\delta p_k|_H^2 + \frac{\beta}{2}\,\tilde{E}\sum_{j=1}^{k}\lambda c^{k-j}\,|\delta p_j|_V^2 \le M_2\,\frac{1 - e^{-2\rho T}}{2\rho}\lambda$$

for $1 \le k \le m$ *and some* $M_2 > 0$.

Proof. Note that $p(\cdot)$ satisfies

$$p(t) - p(s) + \int_s^t Ap(\tau)\, d\tau = \int_s^t Bp(\tau)\, dy(\tau) \tag{4.6}$$

for $s \le t$. Thus,

$$\tilde{E}\,|p(t_k) - p(t)|_V^2 \le 2|t - t_k| \int_t^{t_k} \tilde{E}\,|Ap(\tau)|_V^2\, d\tau + 2|D| \int_t^{t_k} \tilde{E}\,|Bp(\tau)|_V^2\, d\tau$$

for $t \le t_k$. From (4.5) we have $\tilde{E}\,|p(t_k) - p(t)|_V^2 \le 2M_1(1+|D|)\,|t - t_k|$. It then follows from (4.1) that there exists a constant $M_2 > 0$ such that

$$\frac{4}{\beta}\,\tilde{E}\left|\frac{\epsilon_k^{(1)}}{\lambda}\right|_{V^*}^2 + \frac{1+\frac{1}{\varepsilon}}{\lambda}\,\tilde{E}\,|\epsilon_k^{(2)}|_H^2 \le M_2\lambda$$

for $1 \le k \le m$. Hence the theorem follows from (4.4). □

Next we discuss the convergence rate of the Milshtein scheme (3.1) for the uncorrelated case. In this case the approximation error $\epsilon_k^{(2)}$ is defined by

$$\epsilon_k^{(2)} = \int_{t_{k-1}}^{t_k} B(\, p(t) - p(t_{k-1}) - Bp(t_{k-1})(y(t) - y(t_{k-1}))\,)\, dy(t). \tag{4.7}$$

Note that

$$\int_{t_{k-1}}^{t_k} (y^i(t) - y^i(t_{k-1}))\, dy^j(t) = \frac{1}{2}\delta_{i,j}\,((\Delta y^i)^2 - \lambda)$$

since we assumed that $R = I$. Thus, the error function $\delta p_k = p^k - p(t_k)$ satisfies

$$\delta p_k - \delta p_{k-1} + \lambda\,(A + \rho\, I)\delta p_k = \lambda\rho\,\delta p_{k-1} + B\delta p_{k-1}\,\Delta y_k + \frac{1}{2}\,BB\delta p_{k-1}\,((\Delta y_k)^2 - \lambda) + \epsilon_k, \tag{4.8}$$

where $\{p^k\}$ is generated by (3.1). Note that $\tilde{E}\,(\delta p_{k-1},\, \epsilon_k^{(2)}) = 0$. Multiplying (4.8) by δp_k and using exactly the same arguments as in §3, we obtain

$$\begin{aligned} &\tilde{E}\,|\delta p_k|_H^2 - c\,\tilde{E}\,|\delta p_{k-1}|_H^2 + 2\lambda\,\tilde{E}\,\langle A_\rho \delta p_k,\, \delta p_k\rangle \\ &\quad -2(1+\lambda\rho)\,\tilde{E}\,(\delta p_{k-1},\, \epsilon_k^{(1)}) \le \tilde{E}\,|\xi_k - \epsilon_k^{(2))}|_H^2, \end{aligned} \tag{4.9}$$

where $\xi_k = B\delta p_{k-1}\,\Delta y_k + \frac{1}{2}BB\delta p_{k-1}\,((\Delta y_k)^2 - \lambda)$. Here

$$\begin{aligned} \tilde{E}\,(\delta p_{k-1},\, \epsilon_k^{(1)}) = \tilde{E}\,\Bigg(&\delta p_{k-1},\, \int_{t_{k-1}}^{t_k} A(p(t_k) - p(t)) - Bp(t)\,(y(t_k) - y(t))\, dt \\ &+\lambda\rho\,(p(t_k) - p(t_{k-1})) - Bp(t_{k-1})\,\Delta y_k\Bigg). \end{aligned} \tag{4.10}$$

Assume that the solution $p(t)$ of (1.1) satisfies the following regularity:

$$\tilde{E}\,|Bp(t)|_V^2 \le M_3 \quad \text{and} \quad \tilde{E}\,|Ap(t)|_V^2 \le M_3 \tag{4.11}$$

for some $M_3 > 0$ and all $t \in [0, T]$. It then follows from the proof of Theorem 4.1 that $\tilde{E}\,|p(t) - p(s)|^2 \leq M_3\,|t - s|$ for $t \geq s$. From (4.6) we have that for $t \geq s$

(4.12)
$$\tilde{E}\,|p(t) - p(s) - Bp(s)\,(y(t) - y(s))|_V^2 \leq 2M_3\,|t - s|^2 + \int_s^t |B(p(\tau) - p(s))|_V^2\,d\tau$$
$$\leq M\,|t - s|^2$$

for some $M > 0$. Hence it follows from (4.10)–(4.12) that

$$2(1 + \lambda\rho)\,\tilde{E}\,(\delta p_{k-1},\, \epsilon_k^{(1)}) \leq \frac{\beta\lambda}{4}\,|\delta p_{k-1}|_V^2 + \frac{4c\lambda^3}{\beta}\,(|A|_{\mathcal{L}(V,V^*)} + \rho)^2 M^2. \tag{4.13}$$

It follows from (1.9) that there exists an $\varepsilon > 0$ such that for λ sufficiently small

$$2\,\langle A_\rho\phi,\, \phi\rangle - (1 + \varepsilon)|B\phi|_H^2 \geq \frac{3\beta}{4}\,|\phi|_V^2 \quad \text{for all } \phi \in V.$$

Note that

$$\tilde{E}\,|\xi_k - \epsilon_k^{(2)}|_H^2 \leq (1 + \varepsilon)\,\tilde{E}\,|\xi_k|_H^2 + \left(1 + \frac{1}{\varepsilon}\right)\,\tilde{E}\,|\epsilon_k^{(2)}|_H^2.$$

From (4.7) and (4.12) we have

$$\tilde{E}\,|\epsilon_k^{(2)}|_H^2 \leq M^2|B|^2\lambda^3. \tag{4.14}$$

Since

$$|\xi_k|_H^2 = \lambda\left(|B\delta p_{k-1}|_H^2 + \frac{\lambda}{2}\,|BB\delta p_{k-1}|_H^2\right),$$

it follows from (4.9), (4.13), and (4.14) that

$$\tilde{E}\,(|\delta p_k|_H^2 + \lambda(1 + \varepsilon)\,|B\delta p_k|_H^2) + \frac{\lambda\beta}{2}\,\tilde{E}\,|\delta p_k|_V^2$$
$$\leq c\,\tilde{E}\,(|\delta p_{k-1}|_H^2 + \lambda(1 + \varepsilon)\,|B\delta p_{k-1}|_H^2) + \frac{1 + \varepsilon}{2}\lambda^2|B|^2\,\tilde{E}\,|\delta p_{k-1}|_H^2 + M_4\lambda^3$$

for some $M_4 > 0$. Multiplying this by c^{-k} and summing it up in k, we obtain

$$\tilde{E}\,|\delta p_k|_H^2 + \frac{\beta}{4}\,\tilde{E}\sum_{j=1}^{k}\lambda c^{k-j}|\delta p_k|_V^2 \leq M_4\frac{1 - e^{-2\rho T}}{2\rho}\lambda^2 \tag{4.15}$$

for $1 \leq k \leq m$, where we assumed that $2(1 + \varepsilon)\lambda|B|^2 \leq \beta$. Hence one obtains the error estimate.

THEOREM 4.2. *Assume that the solution $p(t)$ of* (1.1) *satisfies the regularity* (4.11). *Then rate of convergence of the Milshtein scheme* (3.1) *applied to the uncorrelated case is the first order in the sense of* (4.15).

Remark 4.3. The regularity assumptions (4.5) and (4.11) of the solution to (1.1) can be verified under certain smoothness assumptions on the function g, σ, h, and b and the initial condition p_0 (e.g., see [Ro]).

5. Hermite polynomial–based spectral method. In this section we discuss the so-called spectral method [GO] based on the Hermite polynomial to approximate the solution to (1.1). Consider the orthogonal functions $e_k(x)$, $k \geq 0$, of $L^2(R)$:

$$e_k(x) = \exp(-x^2/2)H_k(x),$$

where $H_k(x)$ is the normalized Hermite polynomial of degree k, i.e., the original Hermite polynomial divided by $\sqrt{k!}$. A family of functions $\{e_k\}$ has the orthogonality

$$(e_k, e_j)_H = \sqrt{\pi}\,\delta_{k,j}$$

and satisfies the recursive formula

$$\sqrt{k+1}\,e_{k+1}(x) = \sqrt{2}x\,e_k(x) - \sqrt{k}e_{k-1}(x).$$

The differential rule is given by

$$e_k'(x) = \frac{1}{\sqrt{2}}\left(\sqrt{k}\,e_{k-1}(x) - \sqrt{k+1}\,e_{k+1}(x)\right)$$

since $H_k'(x) = \sqrt{2k}\,H_{k-1}(x)$. The Gauss quadrature rule is given by

$$\int_R f(x)e^{-x^2}\,dx = \sum_{i=1}^{m} w_i f(x_i),$$

where the equality holds for all polynomials of degree up to $2m+1$ and the quadrature points x_i and the weights w_i are determined (e.g., see [G]) as follows. Let J be the symmetric tridiagonal matrix with zero diagonals and $J_{i,i+1} = \sqrt{i/2}$, $1 \leq i \leq m-1$. Then $\{x_i\}$ are eigenvalues of J and w_i equal to $(v_i)_1^2$, where $(v_i)_1$ is the first element of the ith normalized eigenvector of J.

For $\sigma > 0$ let $\phi_k(x) = \prod_{i=1}^n e_{k_i}(x_i/\sigma)$ for $k = (k_1, \ldots, k_d) \in N^d$ and $K = [0, n]^d \subset N^d$. The Galerkin approximation of (1.1) involves representing the approximate solution $p_n(t, x)$ by

$$p_n(t, x) = \sum_{k \in K} \alpha_k(t)\phi_k(x), \quad \alpha_k \in R,$$

and projecting the equation (1.1) onto $V^n = \text{span}\{\phi_k,\ k \in K\}$ in the sense that

$$(p_n(t) - p_0,\, \psi) + \int_0^t \langle Ap_n(s),\, \psi\rangle\, ds = \int_0^t (Bp_n(s),\, \psi)\, dy(s),$$

for all $\psi \in V^n$. Define a pair of linear operators (A^n, B^n) by

$$\langle A^n\phi,\, \psi\rangle = \langle A_0\phi - a\,\phi,\, \psi\rangle,$$

$$(B^n\phi,\, \psi) = (c\phi,\, \nabla\psi) + (h\phi,\, \psi)$$

for all $\phi,\ \psi \in V^n$, where $A_0 = a_{i,j}$, a_i and c_i are defined in (1.6). Then $p_n(t, \cdot)$ satisfies

$$dp_n(t) + A^n p_n(t)\,dt = B^n p_n(t)\,dy(t), \quad p_n(0) = p_0^n,$$

where p_0^n is the orthogonal projection of p_0 onto V^n of H. Condition (A2) follows from (1.9) since the left-hand side of (A2) is the restriction of (1.9) onto V^n. Condition (A3) follows from the fact that $|P_V^n\phi - \phi|_V \to 0$ as $n \to \infty$, where P_V^n is the orthogonal projection of V onto V^n. Hence Theorems 2.1 and 3.1 are applied to the Galerkin approximation based on the Hermite polynomials.

Acknowledgment. The author thanks Professor B. Rozovskii for many helpful discussions and the referees for careful reviews and useful suggestions.

REFERENCES

[Be] A. BENSOUSSAN, *Nonlinear filtering theory*, in Progress in Automation and Information Systems, Recent Advances in Stochastic Calculus, J. S.Baras and V. Mirelli, eds., Springer-Verlag, New York, 1990.

[BGR] A. BENSOUSSAN, R. GLOWINSKI, AND A. RASCANU, *Approximation of the Zakai equation by the splitting up method*, SIAM J. Control Optim., 28 (1990), pp. 1420–1431.

[Cl] J. M. C. CLARK, *The design of robust approximation to the stochastic differential equation of nonlinear filtering*, in Communication Systems and Random Process Theory, J. Skwirzynski, ed., Sijthoff and Noordhoof, Alphen aun den Rijn, 1978.

[Da] M. H. A. DAVIS, *Pathwise nonlinear filtering*, in Stochastic Systems, the Mathematics of Filtering and Identification and Application, Hazenwinkel and Williams, eds., Reidel, Dordrecht, 1981.

[FL] P. FLORCHINGER AND F. LE GLAND, *Time-discretization of the Zakai equation for diffusion processes observed in correlated noise*, in 9th Conference on Analysis and Optimization of Systems, Lecture Notes in Control and Inform. Sci., 144, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, New York, 1990.

[G] G. H. GOLUB, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.

[GO] D. GOTTLIEB AND S. A. ORSZAG, *Numerical Analysis of Spectral Methods: Theory and Applications*, Society for Industrial and Applied Mathematics, Philadelphia, 1981.

[KR] N. V. KRYLOV AND B. L. ROZOVSKII, *Stochastic evolution equations*, J. Soviet Math., 16 (1981), pp. 1233–1276.

[Li] J. L. LIONS, *Quelques Methodes de Resolution des Problemes aux Limites Non Lineaires*, Dunod, Paris, 1969.

[Mi] G. N. MILSHTEIN, *Approximate integration of stochastic differential equations*, Theory Probab. Appl., 19 (1974), pp. 557–562.

[Pi] M. PICCIONI, *Convergence of implicit discretization schemes for linear differential equtions with application to filtering*, in Stochastic Partial Differential Equations and Applications, Trento, 1985, Lecture Notes Math. 1236, G. Da Prato and L. Tabaro, eds., Marcel Dekker, New York, 1994.

[Pa] E. PARDOUX, *Equations du filtrage non linéaire, de la prédiction et du lissage*, Stochastics, 3 (1979), pp. 127–168.

[Ro] B. L. ROZOVSKII, *Stochastic Evolution Systems, Linear Theory and Application to Nonlinear Filtering*, Math. Appl., Kluwer Academic Publishers, Norwell, MA, 1991.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 635–659, March 1996

013

OPTIMALITY CONDITIONS FOR A CONSTRAINED CONTROL PROBLEM*

GIANNA STEFANI† AND PIERLUIGI ZEZZA‡

Abstract. This paper is devoted to the study of necessary or sufficient second-order conditions for a weak local minimum in an optimal control problem. The problem is stated in the Mayer form and includes equality constraints both on the endpoints and on the state-control trajectory. The second-order conditions are stated through an associated linear–quadratic problem.

Key words. optimal control, second-order conditions, necessary and sufficient conditions, state-control constraints

AMS subject classifications. 49K15, 49K27, 93C10, 93C50

1. Introduction. This paper is devoted to obtaining necessary or sufficient optimality conditions for an optimal control problem in the Mayer form which is characterized by the presence of two types of equality constraints. The first one concerns the endpoints of the trajectory, and it is a finite-dimensional constraint, while the second one is a time-dependent state-control constraint and it can be regarded as being infinite dimensional. We look for necessary or sufficient second-order conditions for a weak local minimum in the framework of extremal problems, that is, our approach is to transform the original optimal control problem into an abstract constrained optimization problem in the Banach space $E = \mathbf{R}^n \times L^\infty$, which is the space of the couples (*initial condition*, *control*). The transformed problem is the following:

$$\text{Minimize} \quad \phi_0(e)$$

subject to

$$\chi(e) = 0,$$

where $\chi \equiv (\phi_1, \dots, \phi_p) + \psi$, the ϕ_i's are scalar functions, and ψ has range in an infinite-dimensional space F. The optimality conditions are obtained by studying the range of the map $\tilde{\chi} \equiv (\phi_0, \dots, \phi_p) + \psi : E \to \mathbf{R}^{p+1} \oplus F$ locally around a reference admissible point $\hat{e}$. In fact $\hat{e}$ is a local constrained minimizer if and only if locally the range of $\tilde{\chi}$ does not intersect the vertical line below $\phi_0(\hat{e})$. These results (Lemma 4.3 and Lemma 4.5) are expressed by the first and second derivatives of $\tilde{\chi}$ and χ. In particular in the *normal* case, when the multiplier λ_0 corresponding to the cost ϕ_0 is positive, the second-order conditions depend on the *hessian* of $\tilde{\chi}$ while in the abnormal one they depend only on the hessian of χ (see Remark 4.6). Recall that the hessian is the restriction of the second derivative to the kernel of the first one modulo the range of the first derivative.

A crucial role will be played by the assumptions on the constraints. We assume that the reference point is a *regular* point for the infinite-dimensional constraint ψ, i.e., $D\psi(\hat{e})$ is onto (Assumption 2.4 and Lemma 5.2). This assumption allows us to reduce

* Received by the editors January 3, 1994; accepted for publication (in revised form) December 12, 1994. This research was supported in part by Ministerio Università e Ricerca Scientifica e Tecnologica research grants "Teoria dei sistemi e del controllo" and "Equazioni differenziali ordinarie e applicazioni."

† Dipartimento di Matematica e Applicazioni, Via Mezzocannone 8, 80134 Napoli, Italy.

‡ Dipartimento di Matematica per le Decisioni Economiche, Finanziarie, Attuariali, e Sociali, Università degli Studi di Firenze, Via C. Lombroso 6/17, 50134 Firenze, Italy.

the constraint to a finite-dimensional one (see Lemma 4.2) and it ensures that the multiplier associated with the state-control constraint belongs to L^∞. This is the only rank assumption we make and unlike many authors we do not make any controllability assumption on the linearized problem. We prove that the controllability assumption is equivalent to the regularity of the constraint χ at the reference point (Lemma 5.4).

In our approach the regularity assumptions on the data (Assumption 2.2) are in some sense minimal to obtain the C^2 dependence of the flow of the differential equation from the control in the appropriate function spaces. This is needed to prove the smoothness of the function $\tilde{\chi}$. In this setting we determine that the multiplier belongs to L^∞ for a problem whose data have a t-dependence that is only locally bounded and measurable while other authors [9] assume a continuous t-dependence. We should mention that in [9] a maximum principle is derived, while we are interested in second-order necessary or sufficient conditions and we obtain only the weak version of the maximum principle.

In stating the second-order necessary conditions (Theorem 2.5), we assume that the dimension of the space of the multipliers which satisfies the first-order conditions is 1. The case when there are independent multipliers satisfying the first-order conditions has been addressed by some authors [4], [18] while studying the case of inequality finite-dimensional constraints. In the case studied here we show (Corollary 4.4) that the second-order necessary conditions are trivial when there exist independent multipliers, in the sense that they are satisfied for every cost. Also in the case when the space of multipliers has dimension 1 we have to distinguish between the so-called normal case when the reference point is regular for the constraint χ and the abnormal one. In this last case, although the cost does not appear in the necessary conditions we still have some information on the constraints (see Remark 2.7).

To derive the sufficient conditions it is crucial to consider different norms on L^∞, because a coerciveness condition on the second derivative is needed but it cannot be imposed on L^∞ endowed with its norm because it is not isomorphic to a Hilbert space (see, e.g., [3]). The possibility of considering the L^p norms is suggested by Volterra expansions which give us the ith derivative of the flow of the system in an integral form which is defined on L^i (Theorem 3.3). The regularity Assumption 2.2 needs to be strengthened to achieve the needed regularity with respect to the required norms (Theorem 3.5), it cannot be weakened as is shown by Example 3.6. In the statement of the second-order sufficient conditions (Theorem 2.6), we need not assume that the space of multipliers has dimension 1. Also for the sufficient conditions the abnormal case has a special meaning: it means that the reference point is an isolated admissible one (Remark 4.8). Theorem 2.6 gives full information in the normal case and also when there are independent multipliers. Many authors, see, e.g., [3], assume controllability of the linearized system, which implies normality but which is not equivalent to it.

Most of the literature addressing state-control constraints considers the case of equality and inequality constraints, so that it is difficult to compare those results with ours because of the constraint qualification assumptions. For example, some authors impose a constraint qualification assumption that cannot be satisfied by pure equality constraints [10].

The Russian school uses a different approach which obtains powerful abstract results (for a survey see [8], and also the recent book [1] in Russian). An application of this theory to control problems with mixed equality and inequality constraints is stated but not proved in [11]. The same results are quoted in the survey paper [8] where, in the supplement to Chapter VI, it is said that "the derivation of the conditions

is very special and difficult." This survey paper does not contain the proofs and it again quotes [11] and Osmolovskiĭ's thesis [12]. In any case the methods used should be completely different from ours because the results for the equality case are derived from those in the case of mixed equality and inequality by adding an extra inequality constraint. It is our opinion that it is interesting to have a clear and readable proof of the results which also gives an explicit expression of the infinite-dimensional multiplier.

Other authors have addressed the same problem, i.e., to obtain stability results for the numerical solution of optimal control problems. In [3], sufficient conditions for weak local optimality are stated in a problem with equality and inequality constraints on the control but mixed state-control constraints are not considered. Moreover they assume that the reference trajectory satisfies the maximum principle.

Riccati-type techniques are used in [19] to prove sufficient conditions for weak and strong local minima for a problem in the calculus of variations with separate constraints on the endpoints but without other restrictions on the control. While preparing this revised version, a further paper by Zeidan has been published [20], where Riccati-type techniques are used for control problems with fixed initial point equality constraint on the final one and mixed state-control inequality constraints. The specific constraint qualification assumption used does not allow the inclusion of equality constraints.

We have studied separately the reduction to the accessory problem and the nonnegativity (coercivity) of the corresponding quadratic form. Preliminary results concerning second-order conditions have been presented in [14], [15], while an extension of the conjugate point theory which applies to the accessory problem is in progress [17] and it has been presented in [16].

This paper is organized as follows. In §2 we state our main results. In §3 we derive the regularity properties of the flow of the control system. In §4 we prove the abstract lemmata from which, in §5, we derive the proofs of the main results.

2. An optimal control problem. Let us first introduce some notation and definitions needed to properly describe the problem, the assumptions, and the main results.

X, Y, Z are Banach spaces with norm $\|\cdot\|$. Let $\phi : X \to Y$ be a C^2 map; we write $D^i\phi(x)$ for the ith Fréchet derivative of the map ϕ evaluated at the point $x \in X$. By definition $D^0\phi \equiv \phi$, $D\phi(x) \in \mathcal{L}(X,Y)$, and $D^2\phi(x) \in \mathcal{L}^2(X,Y)$. For $\Gamma \in \mathcal{L}^2(X,Y)$ and $\Phi \in \mathcal{L}(Z,X)$ we write

$$\Gamma(x)^2 \equiv \Gamma(x,x), \ (\Gamma \otimes \Phi)(z)^2 \equiv \Gamma(\Phi z)^2.$$

If $X \equiv X_1 \times \cdots \times X_s$, with $X_1, \ldots, X_s$ normed spaces, we denote by $D_i\phi(x)$ the Fréchet derivative of ϕ with respect to the ith variable and by $D^2_{ij}\phi(x) \equiv D_i \circ D_j\phi(x) \in \mathcal{L}^2(X_i \times X_j, Y)$.

The next definition describes the main regularity assumption which is a strengthening of the usual Carathéodory-type assumption. This hypothesis will be used to ensure that the solution of the system (1) depends regularly on the control. Its role is explained in §3.

DEFINITION 2.1 (see [5]). *Assume that X, Y are finite-dimensional vector spaces. We will say that the map $G : \mathbf{R} \times X \to Y$ is quasi-C^k if it satisfies the following:*

(i) *for each $t \in \mathbf{R}$ the map $x \mapsto G(t,x)$ is C^k,*

(ii) *the maps $D^i_2 G$ are locally essentially bounded and measurable in their variables, for $i = 0, \ldots, k$.*

Moreover we will say that the map G is uniformly quasi-C^k if

(iii) *the map $D_2^k G$ is continuous in x uniformly with respect to t in any compact interval J, i.e., for all $x_0 \in X$, $\epsilon > 0$, there exists $\delta > 0$ such that*

$$\|x - x_0\| \leq \delta \Longrightarrow \|D_2^k G(t,x) - D_2^k G(t,x_0)\| \leq \epsilon, \text{ a.e. } t \in J.$$

Property (ii) implies that for any compact interval J and any point $x_0 \in X$ there is a neighborhood $\mathcal{U}$ of x_0 and a constant $h > 0$ such that

$$\|D_2^k G(t,x)\| \leq h, \ x \in \mathcal{U}, \text{ a.e. } t \in J.$$

Therefore, from the intermediate value theorem it is easy to prove that if a function is quasi-C^k then it is uniformly quasi-C^{k-1}.

Since we are interested in local properties, we could assume that the domains of all the maps are not the whole space but just open sets; we prefer the above notation to emphasize the space where we are working.

On a given interval $J = [t_0, t_1]$, let us consider the following optimal control problem:

$$\text{Minimize } a_0(\xi(t_0), \xi(t_1))$$

over all ξ satisfying the following control problem with constraints:

$$\begin{aligned} \dot{\xi}(t) &= F(t, \xi(t), u(t)), \quad \text{a.e. } t \in J, \\ a_i(\xi(t_0), \xi(t_1)) &= 0, \ i = 1, \ldots, p, \\ \alpha(t, \xi(t), u(t)) &= 0, \quad \text{a.e. } t \in J, \end{aligned} \tag{1}$$

where the data satisfy the following regularity assumptions.

Assumption 2.2. The map $F : \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$ is quasi-C^2, the map $\alpha : \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^r$ is uniformly quasi-C^2, and the maps $a_i : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$, $i = 0, \ldots, p$, are C^2.

By Assumption 2.2, equation (1) has uniqueness of solutions so that we can identify a couple (ξ, u), satisfying equation (1), with the couple (x, u) given by the initial condition and the control which will always be our control variable. Hence, we will minimize on the space

$$\mathbf{R}^n \times L^\infty(J, \mathbf{R}^m),$$

which is a Banach space with the topology τ_∞ induced by the norm

$$\|(x,u)\|_\infty \equiv \|x\| + \|u\|_\infty.$$

On this space we will also consider other topologies, namely τ_p, $p \geq 1$, will be the topologies induced by the norm $\|(x,u)\|_p \equiv \|x\| + \|u\|_p$, where $\|\cdot\|_p$ denotes the L^p norm. Pointwise equalities between functions in L^p spaces are always assumed to hold almost everywhere.

We are now interested in necessary or sufficient conditions for a weak local minimum.

DEFINITION 2.3. *A reference couple $(x_0, \hat{u})$, satisfying the above constrained control system, is a weak local minimizer for the optimal control problem if it is a local constrained minimizer in $\mathbf{R}^n \times L^\infty(J, \mathbf{R}^m)$ with respect to the τ_∞ topology.*

In the following we will consider a given $(x_0, \hat{u})$ which satisfies the constraints and we denote by $\hat{\xi}$ the corresponding solution of (1) and by x_1 the endpoint value $\hat{\xi}(t_1)$. We denote by $\hat{}$ the evaluation along the reference objects and by $^\top$ the transpose.

An assumption which will play a crucial role concerns the infinite-dimensional constraint α, and it corresponds to the regularity of this constraint at $(x_0, \hat{u})$; see Lemma 5.2.

Assumption 2.4. The constraint α satisfies the following rank condition at $(x_0, \hat{u})$

$$\det\left(D_3\hat{\alpha}(t)D_3\hat{\alpha}(t)^{\top}\right) \geq k > 0,$$

for some positive $k \in \mathbf{R}$.

This assumption allows us to reduce the set of constraints to a finite-dimensional one so that we can give an explicit (through the data) expression of the modified Hamiltonian associated to the constrained problem. Let

$$A(t) \equiv D_2\hat{F}(t),\ B(t) \equiv D_3\hat{F}(t),\ C(t) \equiv D_2\hat{\alpha}(t),\ D(t) \equiv D_3\hat{\alpha}(t). \tag{2}$$

For the sake of simplicity we will denote by ∇ the derivative with respect to the coupled variables $(x, w) \in \mathbf{R}^n \times \mathbf{R}^m$, so that, for example, $\nabla\hat{\alpha}(t) = (C(t), D(t))$.

The second-order conditions will hold on the space of *critical directions*, i.e., the space of couples (x, u) which satisfies the following system obtained by linearizing (1) and the constraints along the reference trajectory:

$$\dot{\xi}_L(t) = A(t)\xi_L(t) + B(t)u(t),\ \xi_L(t_0) = x, \tag{3}$$

$$C(t)\xi_L(t) + D(t)u(t) = 0, \tag{4}$$

$$Da_i(x_0, x_1)(x, \xi_L(t_1)) = 0,\ i = 1, \ldots, p. \tag{5}$$

We denote the solutions of equation (3) by $\xi_L(\cdot, x, u)$. Assumption 2.4 ensures the existence of a right inverse of $D(t)$ which can be taken as

$$D^{\sharp}(t) \equiv D^{\top}(t)\left(D(t)D^{\top}(t)\right)^{-1}.$$

The next theorems are our main results and give first- and second-order necessary or sufficient weak optimality conditions for this kind of constraint. The results are expressed through the Hamiltonian $\mathcal{H} : J \times (\mathbf{R}^n)^* \times \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ modified to take into account the infinite-dimensional constraint and defined by

$$\mathcal{H}(t, \omega, x, w) = \omega(F(t, x, w) - B(t)D^{\sharp}(t)\alpha(t, x, w)). \tag{6}$$

The first result concerns necessary optimality conditions which are derived under two main assumptions. The infinite-dimensional constraint is regular (the rank condition) and the multiplier associated with the finite-dimensional part (cost and endpoint constraints) is unique up to a positive constant.

THEOREM 2.5. *Assume that*

(i) *F is quasi-C^2, α is uniformly quasi-C^2, and the a_i, $i = 0, \ldots, p$, are C^2;*

(ii) *the rank condition* $\det\left(D(t)D^{\top}(t)\right) \geq k > 0$ *is satisfied.*

Assume that $(x_0, \hat{u})$ is a weak local minimizer for the optimal control problem; then there exist $\lambda = (\lambda_0, \ldots, \lambda_p) \neq 0$ with $\lambda_0 \geq 0$ and a solution $\hat{p}$ of the adjoint equation (7) *satisfying the transversality conditions* (8)*:*

$$-\dot{p}(t) = D_3\mathcal{H}(t, p(t), \hat{\xi}(t), \hat{u}(t)), \tag{7}$$

$$(-p(t_0), p(t_1)) = \sum_{i=0}^{p} \lambda_i Da_i(x_0, x_1) \tag{8}$$

such that

$$D_4\hat{\mathcal{H}}(t) = D_4\mathcal{H}(t, \hat{p}(t), \hat{\xi}(t), \hat{u}(t)) = 0. \tag{9}$$

Assume moreover that

(iii) *the above multiplier* λ *is unique up to a positive constant. Then for each* (x, u) *satisfying the linearized system* (3)–(5) *we have*

$$\sum_{i=0}^{p} \lambda_i D^2 a_i(x_0, x_1)((x, \xi_L(t_1, x, u)))^2 + \int_{t_0}^{t_1} \nabla^2 \hat{\mathcal{H}}(s)((\xi_L(s, x, u), u(s)))^2 ds \geq 0. \tag{10}$$

In the literature the case when the assumption (iii) is dropped has also been considered [4], [18]. Their results specialized to our case give trivial conditions in the following sense. They would say that for each critical direction (x, u) there is a multiplier λ such that (9) and (10) hold. It is a consequence of Corollary 4.4 that this is true independently from the cost and it depends only on the existence of independent multipliers.

The second result concerns second-order sufficient conditions and it does not require the uniqueness of the multiplier. It is stated under stronger regularity assumptions on F.

THEOREM 2.6. *Assume that*

(i) F *and* α *are uniformly quasi-*C^2 *and the* a_i, $i = 0, \dots, p$, *are* C^2,

(ii) *the rank condition* $\det(D(t)D^{\top}(t)) \geq k > 0$ *is satisfied,*

(iii) *there exist* $\lambda = (\lambda_0, \dots, \lambda_p) \neq 0$ *with* $\lambda_0 \geq 0$ *and a solution* $\hat{p}$ *of the adjoint equation* (7) *satisfying the transversality conditions* (8) *for which the first-order conditions* (9) *hold true,*

(iv) *there is* $K > 0$ *such that for each* (x, u) *satisfying the linearized system* (3)–(5) *one has*

$$\sum_{i=0}^{p} \lambda_i D^2 a_i(x_0, x_1)((x, \xi_L(t_1, x, u)))^2 + \int_{t_0}^{t_1} \nabla^2 \hat{\mathcal{H}}(s)((\xi_L(s, x, u), u(s)))^2 ds \geq K\|(x, u)\|_2^2;$$

hence, as a result, $(x_0, \hat{u})$ *is a weak local minimizer for the optimal control problem.*

Remark 2.7. Although Theorems 2.5 and 2.6 are stated without any normality assumption, the abnormal case, i.e., $\lambda_0 = 0$, has a particular meaning. Let us first remark that the multiplier is normal and unique if and only if the point $(x_0, \hat{u})$ is regular for the constraints and if and only if the input–output system

$$\dot{\eta}(t) = (A(t) - B(t)D^{\sharp}(t)C(t))\eta(t) + B(t)(Id - D^{\sharp}(t)D(t))u(t), \ \eta(t_0) = x,$$
$$y_i(t) = Da_i(x_0, x_1)(x, \eta(t, x, u)), \ i = 1, \dots, p,$$

is controllable at time t_1 (see Lemma 5.4), that is, the input–output map $(x, u) \mapsto (y_1(t_1), \dots, y_p(t_1))$ is surjective. The second-order necessary conditions are stated under the assumption that there are not independent multipliers λ. Hence, if the point $(x_0, \hat{u})$ is not regular for the constraints, the multiplier is abnormal and the

statements do not involve the cost a_0 so that they concern mainly the constraints. Nevertheless if (10) is not satisfied then, for the reference point to be an extremum for a cost a_0, it is necessary to have independent multipliers λ satisfying the first-order conditions.

The codimension assumption does not play any role in the sufficient conditions but if Theorem 2.6 holds for an abnormal multiplier, then the reference point $(x_0, \hat{u})$ is an isolated admissible point (Remark 4.8).

Finally let us remark that from the proof of Lemma 3.1 in [15] it follows that if Assumption 2.4 is not satisfied then the codimension of the closure of the range of $\tilde{\chi}'$ is infinite. Consequently, by using the Hahn–Banach theorem, we can prove that the space of multipliers is infinite dimensional and they belong to the dual of L^∞.

3. Properties of the flow. This section is devoted to the study of some local properties of the flow of (1). We denote by $\xi(t, s, x, u)$ the solution of system (1) at time t with initial condition $\xi(s) = x$ and control u. For a fixed reference initial point x_0 and reference control function $\hat{u}$, let the corresponding solution of (1), $\hat{\xi}(\cdot) \equiv \xi(\cdot, t_0, x_0, \hat{u})$, be defined on the compact interval J.

For more details on the properties of the flow described below we refer to [5] where complete proofs are given. If we assume that F is quasi-C^2 then there is an open set $\mathcal{D}(\xi) \subset \mathbf{R} \times \mathbf{R} \times \mathbf{R}^n \times L^\infty(J, \mathbf{R}^m)$ where the flow of system (1) is defined, namely

$$\xi : \mathcal{D}(\xi) \to \mathbf{R}^n, \ (t, s, x, u) \mapsto \xi(t, s, x, u).$$

$\xi(\cdot, s, x, u)$ is a locally Lipschitz map and we look at its restriction to any compact interval J as an element of the Banach space $L^\infty(J, \mathbf{R}^n)$. There is a neighborhood $\mathcal{V}$ of $(x_0, \hat{u}) \in \mathbf{R}^n \times L^\infty(J, \mathbf{R}^m)$ such that $J \times \{t_0\} \times \mathcal{V} \subseteq \mathcal{D}(\xi)$ and the flow of the system, which can be seen as the map

$$\Xi : \mathcal{V} \subset \mathbf{R}^n \times L^\infty(J, \mathbf{R}^m) \to L^\infty(J, \mathbf{R}^n) \ , \ (x, u) \mapsto \xi(\cdot, t_0, x, u), \tag{11}$$

is C^2. When it is clear from the context, we will drop the variable denoting the initial time and we will write $\xi(t, x, u)$ instead of $\xi(t, t_0, x, u)$.

In [15] we describe the second-order Taylor approximation of the map Ξ with respect to u. The results therein are based on [6], [7], [13]. Here we are going to give the approximation with respect to both variables and a sharper estimate of the remainder.

To simplify the computations, we consider the system pulled back by the reference flow which is suitable for studying the system in a neighborhood of the reference trajectory. Let $\gamma : \mathcal{D}(\gamma) \to \Omega$ be the flow of the reference time–dependent vector field $(t, x) \mapsto F(t, x, \hat{u}(t))$. If we define

$$\sigma(t, t_0, x, u) = \gamma(t_0, t, \xi(t, t_0, x, \hat{u} + u)),$$
$$\Phi(t, x, w) = D_3\gamma(t_0, t, \gamma(t, t_0, x))[F(t, \gamma(t, t_0, x), \hat{u}(t) + w) - F(t, \gamma(t, t_0, x), \hat{u}(t))],$$

then $\sigma(\cdot, t_0, x, u)$ is the solution of

$$\dot{\sigma}(t) = \Phi(t, \sigma(t), u(t)), \ \sigma(t_0) = x. \tag{12}$$

Although this pull-back system is easier to handle, we have lost some regularity. If F is (uniformly) quasi-C^2 then Φ is not necessarily (uniformly) quasi-C^2 in both variables. Nevertheless it is (uniformly) quasi-C^1, and the second derivatives with respect to w are (uniformly) quasi-C.

Notice that $\Phi(t,x,w)$ is defined for $t \in J$ and (x,w) in a suitable neighborhood of $(x_0,0) \in \mathbf{R}^n \times \mathbf{R}^m$. For this new system the reference control function is zero, and the reference control vector field $(t,x) \mapsto \Phi(t,x,0)$ is identically zero, so that the reference trajectory $t \mapsto \hat{\sigma}(t)$ is constant, equal to x_0. Moreover by taking a suitable neighborhood $\mathcal{W}$ of $(x_0,0)$ in $\mathbf{R}^n \times L^\infty(J,\mathbf{R}^m)$ as the set of admissible couples, all the trajectories remain close to the reference constant trajectory x_0. The corresponding flow will be denoted by

$$\Sigma : \mathcal{W} \subset \mathbf{R}^n \times L^\infty(J,\mathbf{R}^m) \to L^\infty(J,\mathbf{R}^n).$$

We are now going to define a linear–quadratic system which gives the second-order approximation of the flow of (12):

$$\begin{aligned} \dot{\sigma}_L(t) &= D_3\hat{\Phi}(t)u(t),\ \sigma_L(t_0) = x, \\ \dot{\sigma}_Q(t) &= \nabla^2\hat{\Phi}(t)((\sigma_L(t),u(t)))^2,\ \sigma_Q(t_0) = 0. \end{aligned} \tag{13}$$

The above system is a cascade of integrators whose solution has components $\sigma_L(t,x,u)$ and $\sigma_Q(t,x,u)$. We write x in the second argument of σ_Q because we want to recall that the initial condition of σ_L is x, while the initial condition of σ_Q will always be zero.

We are now going to express the second derivative of the flow of (12) by means of the solutions of the above linear–quadratic system (13).

THEOREM 3.1. *Let F be quasi-C^2 then the second-order expansion of the flow of* (12) *at $(x_0,0)$ can be written as*

$$\sigma(t,t_0,x_0+x,u) = x_0 + \sigma_L(t,x,u) + \frac{1}{2}\sigma_Q(t,x,u) + \mathcal{R}_2(x,u)(t),$$

where

$$\left\|\mathcal{R}_2(x,u)(t)\right\| \le o(\|(x,u)\|_\infty^2).$$

Proof. Let us first take the second-order approximation of Φ with respect to $w \in \mathbf{R}^m$,

$$\Phi_2(t,y,w) = D_3\Phi(t,y,0)w + \frac{1}{2}D_{33}^2\Phi(t,y,0)(w)^2,$$

and consider the corresponding differential equation

$$\dot{\sigma}_2(t) = \Phi_2(t,\sigma_2(t),u(t)). \tag{14}$$

By taking the approximate flow we may lose some properties of the original flow; for example, the new flow may not be uniquely defined, as is shown by Example 3.2. We want to prove that

$$\|\sigma(\cdot,x_0+x,u) - \sigma_2(\cdot,x_0+x,u)\|_\infty = o(\|(x,u)\|_\infty^2), \tag{15}$$

for any solution σ_2 of (14) with initial condition x_0+x. Let us prove that (15) holds for any sequence $(x_i,u_i) \in \mathbf{R}^n \times L^\infty(J,\mathbf{R}^m)$ converging to zero. Using the notation $\sigma^i(t) = \sigma(t,x_0+x_i,u_i)$, $\sigma_2^i(t) = \sigma_2(t,x_0+x_i,u_i)$, we have

$$\|\sigma^i(t) - \sigma_2^i(t)\| \le \int_{t_0}^t \|\Phi(s,\sigma^i(s),u_i(s)) - \Phi_2(s,\sigma_2^i(s),u_i(s))\|ds$$

$$\leq \int_{t_0}^{t} \|\Phi(s,\sigma^i(s),u_i(s)) - \Phi(s,\sigma_2^i(s),u_i(s))\|ds$$
$$+ \int_{t_0}^{t} \|\Phi(s,\sigma_2^i(s),u_i(s)) - \Phi_2(s,\sigma_2^i(s),u_i(s))\|ds.$$

Set $h_i(s) = \|\Phi(s,\sigma_2^i(s),u_i(s)) - \Phi_2(s,\sigma_2^i(s),u_i(s))\|/\|(x_i,u_i)\|_\infty^2$, then there are $\theta_i(s) \in [0,1]$ such that

$$h_i(s) \leq \frac{1}{2}\|D_{33}^2\Phi(s,\sigma_2^i(s),\theta_i(s)u_i(s)) - D_{33}^2\Phi(s,\sigma_2^i(s),0)\|. \tag{16}$$

Thanks to their definition, the h_i's are integrable. Moreover from (16), since $D_{33}^2\Phi$ is quasi-C, then $\{h_i\}$ is a bounded sequence converging pointwise to zero. Therefore the L^∞ norm of the integral of h_i tends to zero by the Lebesgue theorem on dominated convergence. From the above estimate, since Φ is quasi-C^1, we deduce

$$\frac{\|\sigma^i(t) - \sigma_2^i(t)\|}{\|(x_i,u_i)\|_\infty^2} \leq H_1 \int_{t_0}^{t} \frac{\|\sigma^i(s) - \sigma_2^i(s)\|}{\|(x_i,u_i)\|_\infty^2}ds + \int_{t_0}^{t} h_i(s)ds.$$

By the Gronwall inequality we obtain (15).

Volterra expansions will give us the approximation of σ_2. For any scalar C^2 map ρ and any vector field h we define the new function $h \cdot \rho : y \mapsto D\rho(y)h(y)$. Let us denote the derivatives of Φ as time-dependent vector fields by

$$D_3\Phi(t,y,0) = (g_1(t)(y),\ldots,g_m(t)(y)),\ D_{33}^2\Phi(t,y,0) = (g_{ij}(t)(y))_{i,j=1,\ldots,m}.$$

The g_i's are quasi-C^1 and the g_{ij}'s are quasi-C. We will first take the expansion of the flow of σ_2 with respect to u and then take the expansion with respect to x. Set $y = x_0 + x$ and let ρ be a coordinate function. Then

$$\rho(\sigma_2(t)) = \rho(y) + \int_{t_0}^{t} \sum_{i=1}^{m} u_i(s)(g_i(s)\cdot\rho)(\sigma_2(s))ds$$
$$+ \frac{1}{2}\int_{t_0}^{t} \sum_{i,j=1}^{m} u_i(s)u_j(s)(g_{ij}(s)\cdot\rho)(\sigma_2(s))ds$$
$$= \rho(y) + \int_{t_0}^{t} \sum_{i=1}^{m} u_i(s)(g_i(s)\cdot\rho)(y)ds$$
$$+ \frac{1}{2}\int_{t_0}^{t} \sum_{i,j=1}^{m} u_i(s)u_j(s)(g_{ij}(s)\cdot\rho)(\sigma_2(s))ds$$
$$+ \int_{t_0}^{t}\int_{t_0}^{s} \sum_{i,j=1}^{m} u_i(s)u_j(\tau)(g_j(\tau)\cdot g_i(s)\cdot\rho)(\sigma_2(s))d\tau ds.$$

If we proceed as before, we can estimate the remainder in the expansion to obtain

$$\rho(\sigma_2(t)) = \rho(y) + \int_{t_0}^{t} \sum_{i=1}^{m} u_i(s)(g_i(s)\cdot\rho)(y)ds + \frac{1}{2}\int_{t_0}^{t} \sum_{i,j=1}^{m} u_i(s)u_j(s)(g_{ij}(s)\cdot\rho)(y)ds$$
$$+ \int_{t_0}^{t}\int_{t_0}^{s} \sum_{i,j=1}^{m} u_i(s)u_j(\tau)(g_j(\tau)\cdot g_i(s)\cdot\rho)(y)d\tau ds + R(t,x,u),$$

where $|R(t,x,u)| \le o(\|(x,u)\|_\infty^2)$. Let us now take the expansion with respect to x:

$$\rho(x_0+x) = \rho(x_0) + D\rho(x_0)x + S(x) \text{ with } |S(x)| = o(\|x\|^2),$$
$$g_i(s)\cdot\rho(x_0+x) = g_i(s)\cdot\rho(x_0) + D(g_i(s)\cdot\rho)(x_0)x + T(s,x)$$

with $|T(s,x)| \le o(\|x\|)$. Then

$$\int_{t_0}^t \sum_{i=1}^m u_i(s)(g_i(s)\cdot\rho)(x_0+x)ds$$
$$= \int_{t_0}^t \sum_{i=1}^m u_i(s)\left(g_i(s)\cdot\rho(x_0) + (x\cdot g_i(s)\cdot\rho)(x_0)\right)ds + W(t,x,u),$$

with $|W(t,x,u)| \le o(\|(x,u)\|_\infty^2)$. Combining all the previous estimates we obtain for the second derivative

$$\sigma(t,x_0+x,u) = x_0 + x + \int_{t_0}^t \left(\sum_{i=1}^m u_i(s)g_i(s,x_0)\right)ds$$
$$+ \int_{t_0}^t \left(\sum_{i=1}^m u_i(s)D_2g_i(s,x_0)x\right)ds + \frac{1}{2}\int_{t_0}^t \left(\sum_{i,j=1}^m u_i(s)u_j(s)g_{ij}(s,x_0)\right)ds$$
$$+ \int_{t_0}^t\int_{t_0}^s \left(\sum_{i,j=1}^m u_i(s)u_j(\tau)D_2g_i(s,x_0)g_j(\tau,x_0)\right)d\tau ds + \mathcal{R}_2(x,u)(t)$$
$$= x_0 + \sigma_L(t,x,u) + \int_{t_0}^t D^2_{23}\hat\Phi(s)(x,u(s))ds$$
$$+ \frac{1}{2}\int_{t_0}^t D^2_{33}\hat\Phi(s)(u(s))^2ds + \int_{t_0}^t D^2_{23}\hat\Phi(s)(\sigma_L(s,0,u),u(s))ds + \mathcal{R}_2(x,u)(t)$$
$$= x_0 + \sigma_L(t,x,u) + \frac{1}{2}\int_{t_0}^t D^2_{33}\hat\Phi(s)(u(s))^2ds$$
$$+ \int_{t_0}^t D^2_{23}\hat\Phi(s)(\sigma_L(s,x,u),u(s))ds + \mathcal{R}_2(x,u)(t)$$
$$= x_0 + \sigma_L(t,x,u) + \frac{1}{2}\sigma_Q(t,x,u) + \mathcal{R}_2(x,u)(t). \qquad \square$$

The next example shows that the approximating flow may not have uniqueness of solutions.

Example 3.2. Let us consider

$$\Phi(y,w) = (y+w)^{\frac{7}{3}} - y^{\frac{7}{3}};$$

Φ is quasi-C^2. For the reference control $u \equiv 0$, the reference vector field is also zero and the second-order approximation is

$$\Phi_2(y,w) = D_2\Phi(y,0)w + \frac{1}{2}D^2_{22}\Phi(y,0)w^2 = \frac{7}{3}y^{\frac{4}{3}}w + \frac{14}{9}y^{\frac{1}{3}}w^2,$$

which is not Lipschitz continuous with respect to y at $y=0$, and the corresponding flow, defined through

$$\dot\sigma_2(t) = \frac{7}{3}\sigma_2^{\frac{4}{3}}(t)u(t) + \frac{14}{9}\sigma_2^{\frac{1}{3}}(t)u^2(t)\ ,\ \sigma_2(t_0) = x,$$

does not have a unique solution at $x = 0$.

Let us now express the results of Theorem 3.1 for the original system (1). Let A, B be the matrices describing the linearized system introduced in §2 and consider the following system:

$$\begin{aligned} \dot{\xi}_L(t) &= A(t)\xi_L(t) + B(t)u(t),\ \xi_L(t_0) = x, \\ \dot{\xi}_Q(t) &= A(t)\xi_Q(t) + \nabla^2\hat{F}(t)((\xi_L(t), u(t)))^2,\ \xi_Q(t_0) = 0. \end{aligned} \tag{17}$$

We denote by $\xi_L(t, x, u)$ and by $\xi_Q(t, x, u)$ the solutions of the above system, following the notation used for the solutions of (13).

Theorem 3.1 is equivalent to the following theorem.

THEOREM 3.3. *Let F be quasi-C^2, then the second-order expansion of the flow of* (1) *at $(x_0, \hat{u})$ can be written as*

$$\xi(t, x_0 + x, \hat{u} + u) = \hat{\xi}(t) + \xi_L(t, x, u) + \frac{1}{2}\xi_Q(t, x, u) + \mathcal{R}_2(x, u)(t),$$

where

$$\|\mathcal{R}_2(x, u)(t)\| \leq o(\|(x, u)\|_\infty^2).$$

Proof. Since $\xi(t, x, \hat{u} + u) = \gamma(t, t_0, \sigma(t, t_0, x, u))$ then

$$\begin{aligned} D\Xi(x_0, \hat{u})(x, u)(t) &= D_3\gamma(t, t_0, x_0)D\Sigma(x_0, 0)(x, u)(t), \\ D^2\Xi(x_0, \hat{u})((x, u))^2(t) &= D_{33}^2\gamma(t, t_0, x_0)(D\Sigma(x_0, 0)(x, u)(t))^2 \\ &\quad + D_3\gamma(t, t_0, x_0)D^2\Sigma(x_0, 0)((x, u))^2(t). \end{aligned}$$

The maps $\Gamma_L(t) \equiv D_3\gamma(t, t_0, x_0)$, $\Gamma_Q(t) \equiv D_{33}^2\gamma(t, t_0, x_0)$ satisfy the following differential system:

$$\begin{aligned} \dot{\Gamma}_L(t) &= A(t)\Gamma_L(t),\ \Gamma_L(t_0) = Id, \\ \dot{\Gamma}_Q(t) &= A(t)\Gamma_Q(t) + D_{22}^2\hat{F}(t) \otimes \Gamma_L(t),\ \Gamma_Q(t_0) = 0. \end{aligned}$$

Moreover from $\gamma(t_0, t, \gamma(t, t_0, x_0)) = x_0$ one can derive

$$\begin{aligned} D_3\gamma(t_0, t, \gamma(t, t_0, x_0)) &= \Gamma_L^{-1}(t), \\ D_{33}^2\gamma(t_0, t, \gamma(t, t_0, x_0)) \otimes \Gamma_L(t) &= -\Gamma_L^{-1}(t)\Gamma_Q(t). \end{aligned}$$

Hence $D\Xi(x_0, \hat{u})(x, u)$ is the solution of the system

$$\dot{\xi}_L(t) = \dot{\Gamma}_L(t)\sigma_L(t) + \Gamma_L(t)\dot{\sigma}_L(t) = A(t)\xi_L(t) + \Gamma_L D_3\hat{\Phi}(t)u(t).$$

Taking into account that $D_3\hat{\Phi}(t) = \Gamma_L^{-1}(t)D_3F(t, \gamma(t, t_0, x_0), \hat{u}(t))$, we determine that the first derivative of Ξ is ξ_L as expressed in (17). $D^2\Xi(x_0, \hat{u})(x, u)$ is the solution of the system

$$\begin{aligned} \dot{\xi}_Q(t) &= \dot{\Gamma}_Q(t)(\sigma_L(t))^2 + 2\Gamma_Q(t)(\dot{\sigma}_L(t), \sigma_L(t)) + \dot{\Gamma}_L(t)\sigma_Q(t) + \Gamma_L(t)\dot{\sigma}_Q(t) \\ &= A(t)\Gamma_Q(t)(\sigma_L(t))^2 + D_{22}^2\hat{F}(t)(\Gamma_L(t)\sigma_L(t))^2 + 2\Gamma_Q(t)(\dot{\sigma}_L(t), \sigma_L(t)) \\ &\quad + A(t)\Gamma_L(t)\sigma_Q(t) + \Gamma_L(t)\dot{\sigma}_Q(t) \\ &= A(t)\xi_Q(t) + D_{22}^2\hat{F}(t)(\xi_L(t))^2 + 2\Gamma_Q(t)(\dot{\sigma}_L(t), \sigma_L(t)) + \Gamma_L(t)\dot{\sigma}_Q(t). \end{aligned}$$

Since

$$D^2_{33}\hat{\Phi}(t) = \Gamma_L^{-1}(t) D^2_{33} F(t, \gamma(t, t_0, x_0), \hat{u}(t))$$

and

$$\begin{aligned} &D^2_{23}\hat{\Phi}(t)(\sigma_L(t), u(t)) \\ &\quad = -\Gamma_L^{-1}(t)\Gamma_Q(t)(\sigma_L(t), \Gamma_L^{-1}(t) D_3\hat{F}(t)u(t)) + \Gamma_L^{-1}(t) D^2_{33}\hat{F}(t)(\Gamma_L(t)\sigma_L(t), u(t)), \end{aligned}$$

then

$$\begin{aligned} &2\Gamma_Q(t)(\dot{\sigma}_L(t), \sigma_L(t)) + \Gamma_L(t)\dot{\sigma}_Q(t) \\ &\quad = 2\Gamma_Q(t)(D_3\hat{\Phi}(t)u(t), \sigma_L(t)) - 2\Gamma_Q(t)(\sigma_L(t), \Gamma_L^{-1}(t) D_3\hat{F}(t)(u(t))) \\ &\qquad + D^2_{23}\hat{F}(t)(\Gamma_L(t)\sigma_L(t), u(t)) + D^2_{33}\hat{F}(t)(u(t))^2 \\ &\quad = 2D^2_{23}\hat{F}(t)(\xi_L(t), u(t)) + D^2_{33}\hat{F}(t)(u(t))^2. \end{aligned}$$

Thus we obtain the second derivative of Ξ:

$$\dot{\xi}_Q(t) = A(t)\xi_Q(t) + D^2_{22}\hat{F}(t)(\xi_L(t))^2 + 2D^2_{23}\hat{F}(t)\,(\xi_L(t), u(t)) + D^2_{33}\hat{F}(t)(u(t))^2,$$

which is equivalent to the expression in (17). □

Remark 3.4. It is important to notice that the first and second derivatives of the flow of (12), at a given point, are defined and continuous also on the larger spaces $\mathbf{R}^n \times L^1(J, \mathbf{R}^m)$ and $\mathbf{R}^n \times L^2(J, \mathbf{R}^m)$, respectively.

The next theorem states the smooth dependence of the derivatives of the flow of (12) with respect to the point (x, u).

THEOREM 3.5. *Assume that the map F is uniformly quasi-C^2; then the maps*

$$\begin{aligned} D\Xi &: \mathbf{R}^n \times L^\infty(J, \mathbf{R}^m) \to \mathcal{L}\left(\mathbf{R}^n \times L^1(J, \mathbf{R}^m), L^\infty(J, \mathbf{R}^n)\right), \\ D^2\Xi &: \mathbf{R}^n \times L^\infty(J, \mathbf{R}^m) \to \mathcal{L}^2\left(\mathbf{R}^n \times L^2(J, \mathbf{R}^m), L^\infty(J, \mathbf{R}^m)\right) \end{aligned}$$

are continuous on a neighborhood of $(x_0, \hat{u})$.

Proof. We notice that for the solution of a linear system

$$\dot{\varphi}(t) = A(t)\varphi(t) + f(t), \ \varphi(t_0) = y$$

if $\|A\|_\infty \le L$ then there is $M > 0$ such that

$$\|\varphi\|_\infty \le M\|(y, f)\|_1. \tag{18}$$

Let (x, u) and $(\bar{x}, \bar{u}) = (x + \Delta x, u + \Delta u)$ be two points in a neighborhood of $(x_0, \hat{u})$ and let $\xi_i \equiv D^i\Xi(x, u)\big((y, v)\big)^i$ and $\eta_i \equiv D^i\Xi(\bar{x}, \bar{u})\big((y, v)\big)^i$ for $i = 1, 2$.

By the previous results we have

$$\begin{aligned} \dot{\xi}_1(t) &= A(t)\xi_1(t) + B(t)v(t), \ \xi_1(t_0) = y, \\ \dot{\xi}_2(t) &= A(t)\xi_2(t) + H(t)(\xi_1(t), v(t))^2, \ \xi_2(t_0) = 0, \\ \dot{\eta}_1(t) &= \bar{A}(t)\eta_1(t) + \bar{B}(t)v(t), \ \eta_1(t_0) = y, \\ \dot{\eta}_2(t) &= \bar{A}(t)\eta_2(t) + \bar{H}(t)(\eta_1(t), v(t))^2, \ \eta_2(t_0) = 0, \end{aligned}$$

where $H, \bar{H}$ are the second derivatives of F evaluated along the corresponding solutions of (1). We investigate the behavior of $\epsilon_i \equiv \xi_i - \eta_i$, $i = 1, 2$, as follows:

$$\dot{\epsilon}_1(t) = A(t)\epsilon_1(t) + (A(t) - \bar{A}(t))\eta_1(t) + (B(t) - \bar{B}(t))v(t), \ \epsilon_1(t_0) = 0.$$

Applying (18) first to η_1 and then to ϵ_1 we obtain

$$\|\epsilon_1\|_\infty \le \left(h_1\|A-\bar{A}\|_\infty + h_2\|B-\bar{B}\|_\infty\right)\|(y,v)\|_1 = \phi_1(\Delta x,\Delta u)\|(y,v)\|_1, \tag{19}$$

where $\phi_1(\Delta x,\Delta u)\to 0$ as $\|(\Delta x,\Delta u)\|_\infty \to 0$ thanks to the regularity assumptions on the data.

With

$$\begin{aligned}\dot{\epsilon}_2(t) = {} & A(t)\epsilon_2(t) + (A(t)-\bar{A}(t))\eta_2(t) + H(t)(\xi_1(t),v(t))^2 \\ & - \bar{H}(t)(\xi_1(t)-\epsilon_1(t),v(t))^2, \quad \epsilon_2(t_0)=0,\end{aligned}$$

let us estimate ϵ_2. By (18) we have

$$\|(\xi_1,v)\|_2^2 \le M(\|\xi_1\|_\infty^2 + \|v\|_2^2) \le M_1\|(y,v)\|_2^2. \tag{20}$$

Since an analogous estimate holds for η_1, we can estimate the forcing term involving η_2. For the other terms

$$\begin{aligned}&\left\|H(t)(\xi_1(t),v(t))^2 - \bar{H}(t)(\xi_1(t)-\epsilon_1(t),v(t))^2\right\| \\ &\quad \le \left\|H(t)-\bar{H}(t)\right\|\|(\xi_1(t),v(t))\|^2 + \left\|\bar{H}(t)\right\|\|(\epsilon_1(t),0)\|^2 \\ &\qquad + 2\left\|\bar{H}(t)\right\|\|(\xi_1(t),v(t))\|\,\|(\epsilon_1(t),0)\|,\end{aligned}$$

from estimates (19) and (20) and taking into account the uniform continuity of $\|\nabla^2 F\|$ with respect to t. Applying (18) we obtain

$$\|\epsilon_2\|_\infty \le \phi_2(\Delta x,\Delta u)\|(y,v)\|_2^2,$$

where $\phi_2(\Delta x,\Delta u)\to 0$ as $\|(\Delta x,\Delta u)\|_\infty \to 0$. This ends the proof. □

The assumption that the vector field is uniformly quasi-C^2 is essential to obtaining the smoothness of Theorem 3.5. For the sake of simplicity we give a first-order example where a flow whose first derivative does not have the required continuity properties corresponds to a quasi-C^1 vector field.

Example 3.6. For $t\in[0,T]$, let us consider the following control system:

$$\dot{\xi}(t) = -t\cos\left(\frac{u(t)}{t}\right), \quad \xi(0)=x.$$

Take $(x_0,\hat{u})\equiv(0,0)$ to obtain $\hat{\xi}=-\frac{1}{2}t^2$ as a reference trajectory. By taking the sequence $\{w_n\equiv 1/n\}$ it is easy to check that the nonlinear function $(t,w)\mapsto -t\cos(w/t)$ is quasi-C^1 but not uniformly quasi-C^1. The first derivative of the flow of this system at $(x_0,\hat{u})$ is clearly $\xi_L(t,x,u)\equiv x$, and if we take the derivative at a δ-neighboring point $(0,u_\delta)$ we obtain

$$\dot{\eta}(t) = \sin\left(\frac{u_\delta(t)}{t}\right)v(t), \quad \eta(0)=0.$$

Let us compute the error

$$\epsilon_1(t) = \int_0^t \sin\left(\frac{u_\delta(s)}{s}\right)v(s)ds.$$

Choosing $u_\delta(t) = t\rho_{[0,\delta]}(t)$, where ρ_I is the characteristic function of the set I, we obtain

$$\epsilon_1(t) = \begin{cases} \sin(1)\displaystyle\int_0^t v(s)ds, & \text{if } t \le \delta, \\ \sin(1)\displaystyle\int_0^\delta v(s)ds, & \text{if } t \ge \delta. \end{cases}$$

We want to show that

$$\lim_{\delta\to 0}\sup_v \frac{\|\epsilon_1\|_\infty}{\|v\|_1} \neq 0.$$

To do this let us take

$$v_\delta = \frac{1}{\delta}\rho_{[0,\delta]}$$

to obtain

$$\epsilon_1(t) = \begin{cases} \frac{t}{\delta}\sin(1), & \text{if } t \le \delta, \\ \sin(1), & \text{if } t \ge \delta, \end{cases}$$

so that

$$\lim_{\delta\to 0} \frac{\|\epsilon_1\|_\infty}{\|v_\delta\|_1} = \sin(1) \neq 0.$$

4. Abstract optimization results. In this section we state some abstract results which will be applied to obtain necessary or sufficient conditions for our original optimization problem. Let us first transform the control problem described in §2 into an abstract problem on the Banach space $E = \mathbf{R}^n \times L^\infty(J, \mathbf{R}^m)$. Define $\phi \equiv (\phi_0, \dots, \phi_p) : E \to \mathbf{R}^{p+1}$ as

$$\phi_i(x,u) = a_i(x, \xi(t_1, x, u)),\ i = 0, \dots, p,$$

and $\psi : E \to F \equiv L^\infty(J, \mathbf{R}^r)$ as

$$\psi(x,u)(t) = \alpha(t, \xi(t,x,u), u(t)).$$

The study of the problem will be pursued through the analysis of the range of the map

$$\tilde{\chi} \equiv (\phi + \psi) : E \to Z \equiv \mathbf{R}^{p+1} \oplus F.$$

The first component of $\tilde{\chi}$ is the cost and it has a special role. For this reason we decompose $\tilde{\chi}$ in the direct sum of two components: the cost ϕ_0 and the constraint $\chi \equiv (\phi_1, \dots, \phi_p) + \psi$. We indicate with $\mathbf{z_0}$ the unit vector of the cost axis, i.e., the vector in Z which has the first component equal to one and all the others equal to zero. A point $e = (x,u) \in E$ satisfies the constraints if and only if it has an image on the straight line through the origin, parallel to $\mathbf{z_0}$. These points will be called admissible. The point e is said to be *regular* for the constraints χ if and only if $D\chi(e)$ is onto. An element $\Lambda \in Z^*$, called *multiplier*, is said to be *normal* if and only if $\Lambda \mathbf{z_0} \neq 0$, that is, its cost component λ_0 is not zero.

An admissible reference couple $\hat{e} \equiv (x_0, \hat{u})$ is a weak local minimizer for the original problem if and only if there exists a neighborhood Θ of $\hat{e}$ in E such that

$$\tilde{\chi}(\Theta) \bigcap \{\tilde{\chi}(\hat{e}) - a\mathbf{z_0} : a > 0\} = \emptyset. \tag{21}$$

The properties of the range of $\tilde{\chi}$ will be described by the first and second derivatives at the reference point $\hat{e} \in E$, which will be denoted by

$$\tilde{\chi}' \equiv D\tilde{\chi}(\hat{e}),\ \tilde{\chi}'' \equiv D^2\tilde{\chi}(\hat{e}).$$

The same notation will also be used for other maps. The main assumption we will make on the function $\tilde{\chi}$ is as follows.

Assumption 4.1. The map $\tilde{\chi}$ satisfies

(i) $\tilde{\chi}$ is a C^2 map,

(ii) ψ' is onto and it has a continuous right inverse $\psi^\sharp : F \to E$.

Let us consider a complement Z_2 of $\operatorname{Im} \phi'_{|\operatorname{Ker}\psi'} = Y$ in $\mathbf{R}^{p+1}$. When $\mathbf{z_0} \notin Y$, Z_2 will be chosen so that $\mathbf{z_0} \in Z_2$. We denote by p_1, p_2 the projections from $\mathbf{R}^{p+1}$ onto Y and Z_2, respectively. There exists a continuous right inverse of ϕ' defined on Y with values in $\operatorname{Ker}\psi'$. Let $\phi^\sharp : \mathbf{R}^{p+1} \to \operatorname{Ker}\psi'$ be such an inverse extended zero to Z_2 so that $\phi'\phi^\sharp = p_1$ and $\phi^\sharp_{|Z_2} = 0$. The choice of Z_2 implies that

$$\text{either } \mathbf{z_0} \in Y \text{ or } \phi^\sharp \mathbf{z_0} = 0. \tag{22}$$

Notice that if $Z = X \oplus Y$ then we consider X, Y as subspaces of Z. From this point of view we consider $\mathbf{z_0}$ as an element of $\mathbf{R}, \mathbf{R}^{p+1}$ or Z.

In the following lemma we describe some properties of $\tilde{\chi}'$.

LEMMA 4.2. *If Assumption* 4.1 *is satisfied, then*

(i) $\operatorname{Im}\tilde{\chi}'$ *is closed in* Z *and it has* Z_2 *as a finite-dimensional complement,*

(ii) *the map* $\tilde{\chi}^\sharp : Z \to E$ *defined as*

$$\tilde{\chi}^\sharp(y + \omega) = \phi^\sharp(y - \phi'\psi^\sharp\omega) + \psi^\sharp\omega$$

is such that $\tilde{\chi}^\sharp_{|\operatorname{Im}\tilde{\chi}'}$ *is a continuous right inverse of* $\tilde{\chi}'$ *and* $\tilde{\chi}^\sharp_{|Z_2} = 0$,

(iii) *for* $\Lambda \in Z^*$ *let* $\lambda = \Lambda_{|\mathbf{R}^{p+1}}$. *Then* $\Lambda\tilde{\chi}' \equiv 0$ *if and only if*

$$\lambda\phi'_{|\operatorname{Ker}\psi'} = 0,\ \Lambda(y + \omega) = \lambda(y - \phi'\psi^\sharp\omega).$$

Proof. If we define $P_2 : Z \to Z_2$ as

$$P_2 : (y + \omega) \mapsto p_2\left(y - \phi'\psi^\sharp\omega\right),$$

then P_2 is a continuous projection.

Let us show that $\operatorname{Im}\tilde{\chi}' = \operatorname{Ker} P_2$. By definition $P_2\tilde{\chi}' = p_2\left(\phi' - \phi'\psi^\sharp\psi'\right)$. Part (ii) of Assumption 4.1 implies that $E = \operatorname{Ker}\psi' \oplus \operatorname{Im}\psi^\sharp$, hence $Id - \psi^\sharp\psi'$ is the projection onto $\operatorname{Ker}\psi'$ and we obtain

$$P_2\tilde{\chi}' = p_2\phi'_{|\operatorname{Ker}\psi'} = 0.$$

On the other hand if $P_2(y+\omega) = p_2\left(y - \phi'\psi^\sharp\omega\right) = 0$, then there exists $e \in \operatorname{Ker}\tilde{\chi}'$ such that $\phi' e = y - \phi'\psi^\sharp\omega$. Hence $\tilde{\chi}'(e + \psi^\sharp\omega) = \phi' e + \phi'\psi^\sharp\omega + \psi'\psi^\sharp\omega = y + \omega$, and statement (i) is proved.

Since $\phi^\sharp p_2 = 0$, from the definitions of P_2 and of $\tilde{\chi}^\sharp$ it follows that $\tilde{\chi}^\sharp P_2 = 0$. Moreover

$$\tilde{\chi}'\tilde{\chi}^\sharp\tilde{\chi}' = \tilde{\chi}'(\phi^\sharp\phi'(Id - \psi^\sharp\psi') + \psi^\sharp\psi').$$

Recalling that $\psi'\phi^\sharp = 0$, we obtain that $\tilde{\chi}'\tilde{\chi}^\sharp\tilde{\chi}' = \tilde{\chi}'$, which is equivalent to (ii).

Set $\Lambda = \lambda + \Theta$,

$$\Lambda\tilde{\chi}' = 0 \iff \Lambda = \Lambda P_2 \iff \lambda y + \Theta\omega = \lambda p_2(y - \phi'\psi^\sharp\omega).$$

The last equivalence implies that $\lambda = \lambda p_2$ and statement (iii) is easily verified. □

Let $P_2 : Z \to Z_2$ be defined as in Lemma 4.2 and define $Z_1 = \operatorname{Im}\tilde{\chi}'$, $P_1 = Id - P_2$. With this notation

$$\tilde{\chi}'\tilde{\chi}^\sharp = P_1,\ \tilde{\chi}^\sharp P_1 \equiv \tilde{\chi}^\sharp.$$

If we set $\operatorname{Ker}\tilde{\chi}' \equiv E_0$, then $E_1 \equiv \operatorname{Im}\tilde{\chi}^\sharp$ is a topological complement of E_0, and $E = E_0 \oplus E_1$. For $i = 0, 1$, denote by Π_i the canonical projections onto E_i so that $\Pi_1 \equiv \tilde{\chi}^\sharp\tilde{\chi}'$.

The next lemma gives necessary conditions for (21) to hold under the assumption that *codim* $\operatorname{Im}\tilde{\chi}' = 1$, hence this lemma gives necessary optimality conditions for our problem.

LEMMA 4.3. *Assume that the map $\tilde{\chi}$ satisfies Assumption* 4.1 *and that*

1. *there exists a neighborhood Θ of $\hat{e}$ such that* (21) *holds,*
2. *codim* $\operatorname{Im}\tilde{\chi}' = 1$;

then there exists a nonzero multiplier $\Lambda \in Z^*$ such that

(i) $\Lambda\tilde{\chi}' = 0$,

(ii) $\Lambda\mathbf{z_0} \geq 0$,

(iii) $\Lambda\tilde{\chi}''(e)^2 \geq 0$ *for all* $e \in \operatorname{Ker}\chi' = \operatorname{Ker}\tilde{\chi}' \oplus \mathbf{R}\chi^\sharp\mathbf{z_0}$.

Proof. (21) can be written as $\phi(\psi^{-1}(0)\bigcap\Theta)\bigcap\{\phi(\hat{e}) - a\mathbf{z_0} : a > 0\} = \emptyset$. The implicit function theorem applied to ψ yields that there is a neighborhood $\mathcal{V}$ of 0 in E_0 and a C^2 map $\nu : \mathcal{V} \to E$ such that, for a possibly smaller Θ, $\psi^{-1}(0)\bigcap\Theta = \{\hat{e} + v + \nu(v) : v \in \mathcal{V}\}$. Moreover $D\nu(0) = 0$, $D^2\nu(0) = -\psi^\sharp\psi''$. If we let $\Phi(v) = \phi(\hat{e} + v + \nu(v))$ then (21) is equivalent to

$$\Phi(\mathcal{V})\bigcap\{\phi(\hat{e}) - a\mathbf{z_0} : a > 0\} = \emptyset. \tag{23}$$

Notice that $\Phi' \equiv D\Phi(0) = \phi'_{|E_0}$ and without loss of generality we may assume that

$$\Phi'' \equiv D^2\Phi(0) = p_2\Phi''. \tag{24}$$

In fact on E_0 we can make the local change of coordinates given by $\beta \equiv Id - \frac{1}{2}\phi^\sharp\Phi''$ to obtain

$$(\Phi \circ \beta)' = \Phi',\ (\Phi \circ \beta)'' = p_2(\Phi \circ \beta)'' = p_2\Phi''.$$

By Assumption 2 and Lemma 4.2 there is a nonzero $\lambda \in (\mathbf{R}^{p+1})^*$, unique up to a nonzero factor, such that

$$\lambda\phi'_{|\operatorname{Ker}\psi'} = 0.$$

Without loss of generality we may assume that $\lambda\mathbf{z_0}$ is either one or zero. If we define $\Lambda = \lambda P_2$, we have that $\Lambda\tilde{\chi}' = 0$ and $\Lambda\tilde{\chi}'' = \lambda(\phi'' - \phi'\psi^\sharp\psi'') = \lambda\Phi''$, moreover $\Lambda\mathbf{z_0}$ is either one or zero. Let us consider these two cases separately.

Case I: $\lambda \mathbf{z_0} = \Lambda \mathbf{z_0} = 1$. In this case $\mathbf{z_0} \notin \operatorname{Im} \tilde{\chi}'$. Hence by (22), $p_2 = \mathbf{z_0}\lambda$ and $\operatorname{Ker} \chi' = \operatorname{Ker} \tilde{\chi}'$. By contradiction assume that there is $e_0 \in \operatorname{Ker} \tilde{\chi}' = \operatorname{Ker} \psi' \bigcap \operatorname{Ker} \phi'$ such that

$$\Lambda \tilde{\chi}''(e_0)^2 = \lambda \Phi''(e_0)^2 = -2.$$

We want to show that this contradicts (23). Let $H = \phi^\sharp(\mathbf{R}^{p+1}) \subset E_0$ and, locally around zero, define $\rho : \mathbf{R} \times H \times \mathbf{R} \to \mathbf{R}^{p+1}$ by

$$\rho(\epsilon, e, c) \equiv \rho_\epsilon(e, c) = \begin{cases} \dfrac{1}{\epsilon^2}\left[\Phi(\epsilon c e_0 + \epsilon^2 e) - \phi(\hat{e})\right], & \text{if } \epsilon \neq 0, \\ \phi' e + \dfrac{1}{2} c^2 \Phi''(e_0)^2, & \text{if } \epsilon = 0. \end{cases}$$

For $\epsilon \neq 0$ we obtain $\rho_\epsilon(e, c) = \rho_0(e, c) + \frac{1}{\epsilon^2} o(\epsilon^2)$, so ρ is continuous with respect to ϵ. By (24) $\rho_0(0, 1) = -\mathbf{z_0}$. It is easy to check that $D\rho_0(0, 1)$ is an isomorphism so that by standard arguments of degree theory, we determine that, for ϵ sufficiently small, $\operatorname{Im} \rho_\epsilon$ contains a ball with center at $-\mathbf{z_0}$ and radius r. As a consequence $\Phi(\mathcal{V})$ contains $\phi(\hat{e}) - \epsilon^2 \mathbf{z_0}$, which is a contradiction.

Case II: $\lambda \mathbf{z_0} = \Lambda \mathbf{z_0} = 0$. In this case $\phi^\sharp \mathbf{z_0} \neq 0$. Let $\bar{z}$ be such that $p_2 = \bar{z}\lambda$. If $\lambda \Phi''(\tilde{\chi}^\sharp \mathbf{z_0})^2 \neq 0$ then we may assume that it is negative. Assume by contradiction that there is $e_0 \in E_0$ such that

$$\lambda \Phi''(e_0)^2 > 0.$$

Decompose H as $H = K \oplus \mathbf{R}\phi^\sharp \mathbf{z_0}$ and define, for $\epsilon > 0$, $\rho_\epsilon : K \times \mathbf{R}^2 \to \mathbf{R}^{p+1}$ by

$$\rho_\epsilon(e, a, b) = \Phi' e + a \mathbf{z_0} + \frac{1}{2} \epsilon \lambda \Phi''(e + a\phi^\sharp \mathbf{z_0} + b e_0)^2 \bar{z}.$$

We have

$$\frac{1}{\epsilon}[\Phi(\epsilon(e + a\phi^\sharp \mathbf{z_0} + b e_0)) - \phi(\hat{e})] = \rho_\epsilon(e, a, b) + o(\epsilon).$$

For ϵ sufficiently small let us now define the homotopy $h_\epsilon : [0, 1] \times K \times \mathbf{R}^2 \to \mathbf{R}^{p+1} = \Phi' K \oplus \mathbf{R}\mathbf{z_0} \oplus \mathbf{R}\bar{z}$ by

$$h_\epsilon(t, e, a, b) = \rho_\epsilon(e, a, b) + t\, o(\epsilon).$$

By the assumptions there is $\bar{b}$ such that $\lambda \Phi''(-\phi^\sharp \mathbf{z_0} + \bar{b} e_0)^2 = 0$, so $\rho_\epsilon(0, -1, \bar{b}) = -\mathbf{z_0}$. Moreover we obtain

$$D\rho_\epsilon(0, -1, \bar{b}) = \begin{pmatrix} \Phi' & 0 & 0 \\ 0 & 1 & 0 \\ * & * & \epsilon \lambda \Phi''(-\phi^\sharp \mathbf{z_0} + \bar{b} e_0, e_0) \end{pmatrix}.$$

Since the quadratic form $\lambda \Phi''$ is not degenerate on $Span\{e_0, \phi^\sharp \mathbf{z_0}\}$ then $\lambda \Phi''(-\phi^\sharp \mathbf{z_0} + \bar{b} e_0, e_0) = \kappa \neq 0$. Hence there is $\bar{\epsilon}$, such that for $\epsilon < \bar{\epsilon}$

$$\left\| D\rho_\epsilon(0, -1, \bar{b}) x \right\| \geq \epsilon \kappa \|x\|$$

and there is $r > 0$ such that if $\|f\| \leq r$, then

$$\|\rho_\epsilon((0, -1, \bar{b}) + f) + \mathbf{z_0}\| \geq \frac{\epsilon \kappa}{2} \|f\|.$$

For $\|f\| = \|(e,a,b)\| = r$ we obtain the following estimate:

$$\|h_\epsilon(t,e,-1+a,\bar{b}+b)) + \mathbf{z_0}\| = \|\rho_\epsilon(e,-1+a,\bar{b}+b)) + t\,o(\epsilon) + \mathbf{z_0}\|$$
$$\geq \epsilon\Big(\frac{\kappa r}{2} - \frac{\|o(\epsilon)\|}{\epsilon}\Big) > 0$$

for ϵ sufficiently small. Using the homotopy invariance property of the Brower degree (see [2]), the above inequality implies that the degree of the maps $h_\epsilon(1,\cdot)$ and $h_\epsilon(0,\cdot)$ on the ball $B(r,(0,-1,\bar{b}))$ with respect to $\mathbf{z_0}$ is the same. Since the second map is a local homeomorphism these degrees are not zero; therefore $\phi(\hat{e}) - \epsilon\mathbf{z_0} \in \Phi(\mathcal{V})$, which contradicts (23).

To end the analysis of this case let $\lambda\Phi''(\phi^\sharp\mathbf{z_0})^2 = 0$ and assume by contradiction that there are $e_0,\ f_0 \in E_0$ such that

$$\lambda\Phi''(e_0)^2 > 0,\ \lambda\Phi''(f_0)^2 < 0.$$

Let us prove that this contradicts (23). With the same notation as in Case I define, locally around zero, $\rho : \mathbf{R}\times H\times\mathbf{R}\to\mathbf{R}^{p+1}$ by

$$\rho(\epsilon,e,c) \equiv \rho_\epsilon(e,c) = \begin{cases} \dfrac{1}{\epsilon^2}\left[\Phi(\epsilon c e_0 + \epsilon f_0 + \epsilon^2 e) - \phi(\hat{e})\right], & \text{if } \epsilon \neq 0, \\ \phi' e + \dfrac{1}{2}\Phi''(ce_0+f_0)^2, & \text{if } \epsilon = 0. \end{cases}$$

By (24) ρ is continuous with respect to ϵ and $\rho_0(e,c) = \phi' e + \frac{1}{2}\lambda\Phi''(ce_0+f_0)^2\bar{z}$. By the assumptions, there exists $\bar{c}$ such that $\lambda\Phi''(\bar{c}e_0+f_0)^2 = 0$. Moreover for the same reasons as before $\lambda\Phi''(\bar{c}e_0+f_0,e_0) \neq 0$. We have that $\rho_0(0,\bar{c}) = 0$ and

$$D\rho_0(0,\bar{c}) = \begin{pmatrix} \phi' & 0 \\ 0 & \lambda\Phi''(\bar{c}e_0+f_0,e_0) \end{pmatrix}.$$

Reasoning as in Case I we obtain a contradiction to (23). □

If we drop the assumption *codim* $\operatorname{Im}\tilde{\chi}' = 1$, the statement of Lemma 4.3 could be false but one can obtain the following trivial corollary.

COROLLARY 4.4. *Assume that the map $\tilde{\chi}$ satisfies Assumption* 4.1 *and that there exists a neighborhood Θ of $\hat{e}$ such that* (21) *holds. Then for all $e \in \operatorname{Ker}\chi' = \operatorname{Ker}\tilde{\chi}' \oplus \mathbf{R}\chi^\sharp\mathbf{z_0}$ there exists a nonzero multiplier $\Lambda \in Z^*$ such that*

(i) $\Lambda\tilde{\chi}' = 0$,
(ii) $\Lambda\mathbf{z_0} \geq 0$,
(iii) $\Lambda\tilde{\chi}''(e)^2 \geq 0$

Proof. If *codim* $\operatorname{Im}\tilde{\chi}' = 1$ we can apply the previous Lemma 4.3. If it is greater than one then we can consider the space $\bar{Z}$ spanned by $\operatorname{Im}\tilde{\chi}'$ and $\tilde{\chi}''(e)^2$. Take a nonzero multiplier $\bar{\Lambda}$ which is zero on $\bar{Z}$ to obtain the statement. □

To state conditions sufficient for (21) to hold, we now consider the Banach space E endowed also with another, possibly different, norm $\|\cdot\|_2$. We denote by τ the topology under which E is a Banach space and by τ_2, the other.

The next lemma provides an abstract framework to prove sufficient conditions for weak local optimality for our optimization problem. Let us emphasize that it is necessary that the completion of E_0 under τ_2 be a Hilbert space, otherwise no continuous positive quadratic form on (E_0,τ_2) could be coercive.

LEMMA 4.5. *Assume that $\tilde{\chi} : E \to Z$ satisfies Assumption* 4.1 *and*

1. $\tilde{\chi}^\sharp D\tilde{\chi} : (E,\tau) \to \mathcal{L}((E,\tau_2),(E,\tau_2))$ *is continuous.*

Assume moreover that there exists a multiplier $\Lambda \in Z^*$ *and* $K > 0$ *such that*

2. $\Lambda D^2\tilde{\chi} : (E,\tau) \to \mathcal{L}^2((E,\tau_2),\mathbf{R})$ *is continuous,*
3. $\Lambda\tilde{\chi}' = 0$,
4. $\Lambda\mathbf{z_0} \geq 0$,
5. $\Lambda\tilde{\chi}''(e)^2 \geq K\|e\|_2^2$ *for every* $e \in \operatorname{Ker}\chi' = \operatorname{Ker}\tilde{\chi}' \oplus \mathbf{R}\tilde{\chi}^\sharp\mathbf{z_0}$.

Then there exists a neighborhood Θ *of* $\hat{e}$ *in* (E,τ) *such that*

$$\tilde{\chi}(\Theta) \cap \{\tilde{\chi}(\hat{e}) - a\mathbf{z_0} : a > 0\} = \emptyset.$$

Proof. Let us prove the statement by contradiction. Assume that there exist sequences $\{a_i\} \subset \mathbf{R}_+$ and $\{e_i\} \subset E$ such that

$$\|e_i\| \to 0 \text{ and } \tilde{\chi}(\hat{e} + e_i) = \tilde{\chi}(\hat{e}) - a_i\mathbf{z_0}.$$

There are $\theta_i \in [0,1]$ such that

$$P_1(\tilde{\chi}(\hat{e} + e_i) - \tilde{\chi}(\hat{e})) = \tilde{\chi}'e_i + P_1\big(D\tilde{\chi}(\hat{e} + \theta_i e_i) - \tilde{\chi}'\big)e_i = -a_i P_1\mathbf{z_0}.$$

Hence applying $\tilde{\chi}^\sharp$ we obtain

$$\tilde{\chi}^\sharp\tilde{\chi}'e_i = \Pi_1 e_i = -a_i\tilde{\chi}^\sharp\mathbf{z_0} - \tilde{\chi}^\sharp(D\tilde{\chi}(\hat{e} + \theta_i e_i) - D\tilde{\chi}(\hat{e}))e_i.$$

Let us consider two different cases. If $\Lambda\mathbf{z_0} \neq 0$, then by assumption 4 we can assume that $\Lambda\mathbf{z_0} = 1$ and by (22) $\tilde{\chi}^\sharp\mathbf{z_0} = 0$. Then assumption 1 yields

$$\frac{\|\Pi_1 e_i\|_2}{\|e_i\|_2} \to 0. \tag{25}$$

Moreover, using assumptions 3 and 5, for suitable $\theta_i \in [0,1]$, we obtain

$$\begin{aligned} -\frac{1}{\|e_i\|_2^2}a_i &= \frac{1}{\|e_i\|_2^2}\Lambda\big(\tilde{\chi}(\hat{e} + e_i) - \tilde{\chi}(\hat{e})\big) \\ &= \frac{1}{2}\Lambda\tilde{\chi}''\left(\frac{e_i}{\|e_i\|_2}\right)^2 + \frac{1}{2}\Lambda\big(D^2\tilde{\chi}(\hat{e} + \theta_i e_i) - D^2\tilde{\chi}(\hat{e})\big)\left(\frac{e_i}{\|e_i\|_2}\right)^2 \\ &\geq K\frac{\|\Pi_0 e_i\|_2^2}{\|e_i\|_2^2} - \frac{1}{2}\frac{2\|\Pi_0 e_i\|_2\|\Pi_1 e_i\|_2 + \|\Pi_1 e_i\|_2^2}{\|e_i\|_2^2}\|\Lambda\tilde{\chi}''\|_2 \\ &\quad - \frac{1}{2}|\Lambda(D^2\tilde{\chi}(\hat{e} + \theta_i e_i) - D^2\tilde{\chi}(\hat{e}))|. \end{aligned} \tag{26}$$

Let us examine the three addends in (26). From assumption 1 evaluated at $\hat{e}$ we also have that Π_0 is continuous with respect to τ_2. By (25) and assumption 2 the second and the third addend tend to zero. Since K and a_i have the same sign it follows that

$$\frac{\|\Pi_0 e_i\|_2}{\|e_i\|_2} \to 0,$$

which together with (25) yields a contradiction.

If $\Lambda\mathbf{z_0} = 0$, then assumption 1 yields

$$\Pi_1 e_i = -a_i\tilde{\chi}^\sharp\mathbf{z_0} + o_i$$

with

$$\frac{\|o_i\|_2}{\|e_i\|_2} \to 0. \tag{27}$$

From this equation one can deduce that $f_i = -a_i\tilde{\chi}^\sharp \mathbf{z_0} + \Pi_2 e_i = e_i - o_i \in \operatorname{Ker}\chi'$ is such that

$$\frac{\|f_i\|_2}{\|e_i\|_2} \text{ is bounded.} \tag{28}$$

Moreover, using assumptions 3 and 5, for suitable $\theta_i \in [0,1]$, we obtain

$$\begin{aligned} 0 &= \Lambda\tilde{\chi}''\left(\frac{e_i}{\|e_i\|_2}\right)^2 + \Lambda\big(D^2\tilde{\chi}(\hat{e}+\theta_i e_i) - D^2\tilde{\chi}(\hat{e})\big)\left(\frac{e_i}{\|e_i\|_2}\right)^2 \\ &\geq K\frac{\|f_i\|_2^2}{\|e_i\|_2^2} - \frac{2\|f_i\|_2\|o_i\|_2 + \|o_i\|_2^2}{\|e_i\|_2^2}\|\Lambda\tilde{\chi}''\|_2 - |\Lambda(D^2\tilde{\chi}(\hat{e}+\theta_i e_i) - D^2\tilde{\chi}(\hat{e}))|. \end{aligned} \tag{29}$$

By (27), (28), and assumption 2 the second and the third addend in (29) tend to zero, so

$$\frac{\|f_i\|_2}{\|e_i\|_2} \to 0,$$

which together with (27) yields a contradiction. □

Remark 4.6. It is important to emphasize the difference between the normal case ($\Lambda\mathbf{z_0} \neq 0$) and the abnormal one ($\Lambda\mathbf{z_0} = 0$). In the normal case $\operatorname{Ker}\tilde{\chi}' = \operatorname{Ker}\chi'$ and the second order optimality conditions concern the quadratic form $\Lambda\tilde{\chi}''_{|\operatorname{Ker}\tilde{\chi}'}$ which depends only on the *hessian* of $\tilde{\chi}$. In the abnormal case $\Lambda \in (\mathbf{R}^p \oplus F)^*$ and the quadratic form is $\Lambda\chi''_{|\operatorname{Ker}\chi'}$ which depends only on the hessian of the constraints χ. In this last case the optimality conditions essentially give information on the constraints.

The next remarks explain the difference between these two cases.

Remark 4.7. Lemma 4.3 is stated under the assumption that $codim\,\operatorname{Im}\tilde{\chi}' = 1$ so that in this case a normal multiplier corresponds to a regular point $\hat{e}$ and the optimality conditions directly involve both the cost ϕ_0 and the constraints χ. When the multiplier Λ is abnormal the cost does not appear in the optimality conditions. The information we have is that if $\Lambda\chi''_{|\operatorname{Ker}\chi'}$ is indefinite, then $\hat{e}$ cannot be an extremum for any cost satisfying $codim\,\operatorname{Im}\tilde{\chi}' = codim\,\operatorname{Im}\chi' = 1$.

Remark 4.8. If the multiplier in Lemma 4.5 is abnormal then the sufficient conditions hold true for any cost so that the point $\hat{e}$ is isolated among the admissible points.

In the normal case the coerciveness of $\Lambda\tilde{\chi}''$ is assumed on $\operatorname{Ker}\tilde{\chi}' = \operatorname{Ker}\chi'$. We have seen in Remark 4.8 that, in the abnormal case, if one imposes the coerciveness on $\operatorname{Ker}\chi'$ the sufficient conditions hold only at isolated points. One could think that, in this case, the coerciveness has to be imposed on $\operatorname{Ker}\tilde{\chi}'$, which is smaller, but this is not sufficient for optimality as is shown by the following example.

Example 4.9. Let us consider the following function $\tilde{\chi} : \mathbf{R}^2 \to \mathbf{R}^2$, defined as

$$\tilde{\chi}(x,y) = \begin{pmatrix} x \\ \frac{1}{2}(y^2 - x^2) \end{pmatrix}.$$

All the smoothness assumptions are verified and we have $\tilde{\chi}' = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ so that

$$E_0 = Z_2 = span\left\{\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right\}, \quad E_1 = Z_1 = span\left\{\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right\}.$$

By taking $\Lambda = (0,1) \in (\mathbf{R}^2)^*$ we obtain $\operatorname{Im}\tilde{\chi}' = \operatorname{Ker}\Lambda$. Since $\Lambda\tilde{\chi}'' \otimes \Pi_0$ is the restriction to E_0 of the quadratic form $\left(\begin{smallmatrix} -2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$ then $\Lambda\tilde{\chi}''_{|\operatorname{Ker}\tilde{\chi}'}$ is coercive, while studying the image of $\tilde{\chi}$ by the image of straight lines through the origin we have

$$\operatorname{Im}\tilde{\chi} = \left\{\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbf{R}^2 : x \geq -y^2\right\},$$

and the image contains the whole y-axis.

5. Proof of the main results. In this section we suppose that our main Assumptions 2.2 and 2.4 are satisfied. Necessary or sufficient conditions which characterize a weak local minimum of the optimal control problem will be deduced from the abstract results described in §4.

For the derivatives at our reference point $\hat{e} \in E$, we will use the same notation as in §4. Moreover let $a \equiv (a_0, \ldots, a_p)$, $a' = Da(x_0, x_1)$, and $a'' = D^2a(x_0, x_1)$.

Since in the definition of ψ a Nemytskii-type operator is involved, we introduce a specific notation for the *superposition operator* and we state some of its properties. Let X, Y be finite-dimensional vector spaces. If $\mu : J \times X \to Y$, we denote by $\tilde{\mu} : L^\infty(J,X) \to L^\infty(J,Y)$ the superposition operator

$$\tilde{\mu}(\xi)(t) = \mu(t, \xi(t))$$

so, for example, $\psi(x,u) \equiv \tilde{\alpha} \circ (\Xi(x,u), u)$.

LEMMA 5.1. *Assume that the map μ is uniformly quasi-C^2, then the operator $\tilde{\mu}$ is C^2. Moreover the operator*

$$D\tilde{\mu} : L^\infty(J,X) \to \mathcal{L}((L^\infty(J,X), \tau_p), (L^\infty(J,Y), \tau_p))$$

is continuous for $1 \leq p \leq \infty$ and

$$D^2\tilde{\mu} : L^\infty(J,X) \to \mathcal{L}^2((L^\infty(J,X), \tau_2), (L^\infty(J,Y), \tau_1))$$

is continuous.

The proof of the above lemma is a straightforward consequence of the regularity assumption on the map μ.

Let us describe some properties of the infinite-dimensional constraint ψ that are needed to apply the results obtained in §4. Under our assumptions the map ψ' admits a continuous right inverse and we can apply Lemma 4.2.

LEMMA 5.2. *If the smoothness Assumption 2.2 is satisfied then ψ is a C^2 map and $\psi' : E \to L^\infty(J, \mathbf{R}^r)$ is given by*

$$\psi'(x,u)(t) = C(t)\xi_L(t,x,u) + D(t)u(t), \tag{30}$$

where ξ_L is the solution of the linearized system (3) *and C, D are defined in* (2).

Moreover ψ' is onto and it admits a continuous right inverse $\psi^\sharp$ if and only if Assumption 2.4 *holds. For $\omega \in L^\infty(J, \mathbf{R}^r)$, $\psi^\sharp\omega$ is given by the output of the system*

$$\begin{aligned} \dot{\zeta}(t) &= (A(t) - K(t))\zeta(t) + B(t)D^\sharp(t)\omega(t), \quad \zeta(t_0) = 0 \\ \psi^\sharp\omega(t) &= (0, D^\sharp(t)(\omega(t) - C(t)\zeta(t))), \end{aligned} \tag{31}$$

where $D^\sharp = D^\top(DD^\top)^{-1}$ *and* $K = BD^\sharp C$.

Proof. The smoothness of ψ follows from the properties of the flow and of the superposition operator described in Lemma 5.1. The proof of the other statements of this lemma is not significantly different from that of Lemma 3.1 in [15]. □

LEMMA 5.3. *If the smoothness Assumption* 2.2 *is satisfied, then the map* $\tilde{\chi}$ *satisfies Assumption* 4.1 *and*

$$\tilde{\chi}'(x,u) = a'(x,\xi_L(t_1,x,u)) + C(\cdot)\xi_L(\cdot,x,u) + D(\cdot)u(\cdot). \tag{32}$$

Proof. The smoothness of ϕ is a consequence of the properties of the flow and the chain rule so the statement is a consequence of Lemma 5.2. □

LEMMA 5.4. *If the smoothness Assumption* 2.2 *and the rank Assumption* 2.4 *are satisfied then the following statements are equivalent:*

(i) $(x_0,\hat{u})$ *is a regular point for the constraint* χ,

(ii) *the input–output system*

$$\begin{aligned} \dot{\eta}(t) &= (A(t) - B(t)D^\sharp(t)C(t))\eta(t) + B(t)(Id - D^\sharp(t)D(t))u(t),\ \eta(t_0) = x \\ y_i(t) &= Da_i(x_0,x_1)(x,\eta(t,x,u)),\ i = 1,\dots,p, \end{aligned} \tag{33}$$

is controllable at time t_1, *i.e.,* $(x,u) \mapsto (y_1(t_1),\dots,y_p(t_1))$ *is surjective.*

Proof. Applying (iii) of Lemma 4.2 to χ we determine that $(x_0,\hat{u})$ is regular for the constraints if and only if the map $(\phi_1',\dots,\phi_p')_{|Ker\,\psi'}$ is onto. The properties of the right inverse give that $Id - \psi^\sharp\psi'$ is the projection onto $\operatorname{Ker}\psi'$ so

$$\operatorname{Im}\phi'_{i|\operatorname{Ker}\psi'} = \operatorname{Im}\ \phi_i'(Id - \psi^\sharp\psi').$$

Let us first compute $\phi_i'\psi^\sharp$. By Lemma 5.2 and by (32) we have that, for $\omega \in L^\infty(J,\mathbf{R}^r)$, $\phi_i'\psi^\sharp\omega$ is given by

$$\begin{aligned} \dot{\zeta}(t) &= (A(t) - K(t))\zeta(t) + B(t)D^\sharp(t)\omega(t),\ \zeta(t_0) = 0, \\ \dot{\xi}_L(t) &= A(t)\xi_L(t) + B(t)D^\sharp(t)(\omega(t) - C(t)\zeta(t)),\ \xi_L(t_0) = 0, \\ \phi_i'\psi^\sharp\omega &= a_i'(x,\xi_L(t_1)). \end{aligned}$$

Taking the difference between the two above differential equations we can immediately verify that $\zeta - \xi_L$ is a solution of a linear differential equation with zero initial condition and hence it is zero itself. Thus

$$\phi_i'\psi^\sharp\omega = a_i'(x,\zeta(t_1,0,\omega)), \tag{34}$$

where ζ is defined in (31). Let us set $\rho_i(x,u) \equiv (\phi_i'(x,u) - \phi_i'\psi^\sharp\psi'(x,u))$, for $i = 1,\dots,p$. Then from (30) it follows that we can define ρ_i through the following cascade of systems:

$$\begin{aligned} \dot{\xi}_L(t) &= A(t)\xi_L(t) + B(t)u(t),\ \xi_L(t_0) = x, \\ \dot{\zeta}(t) &= (A(t) - K(t))\zeta(t) + B(t)D^\sharp(t)(C(t)\xi_L(t) + D(t)u(t)),\ \zeta(t_0) = 0, \\ \rho_i(x,u) &= a_i'(x,\xi_L(t_1) - \zeta(t_1)). \end{aligned}$$

Setting $\eta = \xi_L - \zeta$ one can readily verify that $\rho_i(x,u) = a_i'(x,\eta(t_1,x,u))$, where η is defined in (33). □

In the following lemma we describe some relations linking the adjoint covector, the Hamiltonian, and the ranges of $\tilde{\chi}'$ and $\tilde{\chi}''$.

LEMMA 5.5. *For* $\lambda \equiv (\lambda_0, \dots, \lambda_p) \in (\mathbf{R}^{p+1})^*$ *we have that* $\lambda \phi'_{|\operatorname{Ker} \psi'} = 0$ *if and only if there exists a solution* $\hat{p} : J \to (\mathbf{R}^n)^*$ *of the adjoint equation* (7) *satisfying the transversality conditions* (8) *such that*

$$D_4 \mathcal{H}(t, \hat{p}(t), \hat{\xi}(t), \hat{u}(t)) = 0,$$

where the Hamiltonian $\mathcal{H}$ *is defined in* (6).

If we define $\Lambda \in Z^*$ *as* $\Lambda(y + \omega) = \lambda(y - \phi' \psi^\sharp \omega)$, *then*

$$\begin{aligned} \Lambda \tilde{\chi}''((x,u))^2 &= \lambda a''((x, \xi_L(t_1, x, u)))^2 \\ &\quad + \int_{t_0}^{t_1} \nabla^2 \hat{\mathcal{H}}(s)((\xi_L(s, x, u), u(s)))^2 ds. \end{aligned} \tag{35}$$

Proof. Using the same arguments as in Lemma 5.4, one can show that $\lambda \phi'_{|\operatorname{Ker} \psi'} = 0$ if and only if $\lambda \phi'(Id - \psi^\sharp \psi') = 0$. From the proof of Lemma 5.4 it follows that

$$\lambda \phi'(Id - \psi^\sharp \psi')(x,u) = \lambda a'(x, \eta(t_1, x, u)),$$

where η is defined in (33). Therefore we have

$$\lambda \phi'_{|\operatorname{Ker} \psi'} = 0 \iff \lambda D_1 a(x_0, x_1) x + \lambda D_2 a(x_0, x_1) \eta(t_1, x, u) = 0, \ \forall (x,u) \in E. \tag{36}$$

To express the above relations by means of the appropriate Hamiltonian let us denote by Ω the solution of the matrix equation

$$\dot{\Omega}(t) = (A(t) - K(t))\Omega(t), \ \Omega(t_0) = Id.$$

Then $\eta(t_1, x, u) = \Omega(t_1)(x + \int_{t_0}^{t_1} \Omega^{-1}(t) B(t)(Id - D^\sharp(t) D(t)) u(t) dt)$, so that equation (36) is equivalent to the system

$$\lambda D_1 a(x_0, x_1) = -\lambda D_2 a(x_0, x_1) \Omega(t_1), \tag{37}$$

$$\lambda D_2 a(x_0, x_1) \Omega(t_1) \Omega^{-1}(t) B(t)(Id - D^\sharp(t) C(t)) = 0. \tag{38}$$

If $\hat{p}(t)$ is the solution of $\dot{p}(t) = -D_3 \mathcal{H}(t, p(t), \hat{\xi}(t), \hat{u}(t)) = -p(t)(A(t) - K(t))$, which satisfies $p(t_1) = \lambda D_2 a(x_0, x_1)$, then (37) is equivalent to the transversality condition (8). Taking into account the definitions of B, C, D, and the Hamiltonian $\mathcal{H}$, it follows that condition (38) can be expressed as

$$\hat{p}(t) D_3 \hat{F}(t) - \hat{p}(t) B(t) D^\sharp(t) D_3 \hat{\alpha}(t) = D_4 \hat{\mathcal{H}}(t) = 0.$$

This ends the first part of the proof.

From the definition of Λ it follows that $\Lambda \tilde{\chi}'' = \lambda(\phi'' - \phi' \psi^\sharp \psi'')$. Since

$$\begin{aligned} \phi''(x,u) &= a''((x, \xi_L(t_1, x, u)))^2 + D_2 a(x_0, x_1) \ \xi_Q(t_1, x, u), \\ \psi''(x,u)(t) &= \nabla^2 \hat{\alpha}(t)(\xi_L(t, x, u), u(t))^2 + D_2 \hat{\alpha}(t) \xi_Q(t, x, u), \end{aligned}$$

then by (34), $\Lambda \tilde{\chi}''$ an be written as the output of the following cascade of systems:

$$\begin{aligned} \dot{\xi}_L(t) &= A(t) \xi_L(t) + B(t) u(t), \ \xi_L(t_0) = x, \\ \dot{\xi}_Q(t) &= A(t) \xi_Q(t) + \nabla^2 \hat{F}(t)((\xi_L(t), u(t)))^2, \ \xi_Q(t_0) = 0, \\ \dot{\zeta}(t) &= (A(t) - K(t)) \zeta(t) \\ &\quad + B(t) D^\sharp(t)[\nabla^2 \hat{\alpha}(t)((\xi_L(t), u(t)))^2 + C(t) \xi_Q(t)], \ \zeta(t_0) = 0, \\ \lambda(\phi'' - \phi' \psi^\sharp \psi'')(x,u) &= \lambda \, a''((x, \xi_L(t_1)))^2 + \lambda \, D_2 a(x_0, x_1) \, [\xi_Q(t_1) - \zeta(t_1)]. \end{aligned}$$

If we define $z_Q(t) = \xi_Q(t) - \zeta(t)$, then one can readily verify that it is the solution of

$$\dot{z}_Q(t) = (A(t) - K(t))z_Q(t) + H(t)((\xi_L(t), u(t)))^2, \ z_Q(t_0) = 0,$$

where $H(t) = \nabla^2 \hat{F}(t) - B(t)D^\sharp(t)\nabla^2\hat{\alpha}(t)$. Therefore we obtain

$$\Lambda\tilde{\chi}''((x,u))^2 = \lambda a''((x, \xi_L(t_1, x, u)))^2 + \lambda\, D_2 a(x_0, x_1) z_Q(t_1, x, u).$$

Writing z_Q explicitly by means of the fundamental matrix Ω, substituting $\hat{p}$, and taking into account the expressions of H and $\mathcal{H}$, we easily obtain the second statement. □

Proof of Theorem 2.5. If $\hat{e} \equiv (x_0, \hat{u})$ is a weak local minimizer, then $\tilde{\chi}$ is not locally onto at that point and the same holds for $\tilde{\chi}'$. Lemma 4.2 (i) yields that *codim* $\operatorname{Im}\tilde{\chi}' = \dim Z_2$. Since Z_2 is a complement of $\operatorname{Im}\phi'_{|\operatorname{Ker}\psi'}$, then there exists $\lambda \equiv (\lambda_0, \ldots, \lambda_p) \neq 0$ such that $\lambda\,\phi'_{|\operatorname{Ker}\psi'} = 0$ and we can always take $\lambda_0 \geq 0$. Lemma 5.5 proves the first-order conditions (9).

The uniqueness of λ up to a positive constant and Lemma 4.2 imply that the codimension of $\operatorname{Im}\tilde{\chi}'$ is one. Therefore Lemma 4.3 applies and $\Lambda\tilde{\chi}''_{|\operatorname{Ker}\chi'} \geq 0$. Since $(x,u) \in \operatorname{Ker}\chi'$ if and only if (x,u) satisfies the linearized problem (3), (4), (5), then from (35) we obtain the second-order conditions. □

Proof of Theorem 2.6. To prove the theorem we only have to verify that the map $\tilde{\chi}$ satisfies the assumptions of Lemma 4.5. Let λ be the one given in (ii) of Theorem 2.6 and define $\Lambda \in Z^*$ as in Lemma 5.5. Then $\Lambda\tilde{\chi}' = \lambda\phi'(Id - \psi^\sharp\psi') = 0$, so assumption (3) in Lemma 4.5 holds. From Lemma 4.2 we obtain

$$\tilde{\chi}^\sharp(y + \omega) = \phi^\sharp(y - \phi'\psi^\sharp\omega) + \psi^\sharp\omega.$$

By (31) the map $\psi^\sharp$ belongs to $\mathcal{L}(L^q(J, \mathbf{R}^r), (E, \tau_q))$, for all $q \geq 1$. Since $\phi' \in \mathcal{L}((E, \tau_q), \mathbf{R}^{p+1})$ and $\phi^\sharp$ is a linear map between finite-dimensional vector spaces, $\tilde{\chi}^\sharp \in \mathcal{L}((E, \tau_q), (E, \tau_q))$, for all $q \geq 1$. Denoting by $\pi_2 : E \to L^\infty(J, \mathbf{R}^m)$ the projection on the second factor, we can write $D\tilde{\chi} = D\phi + D\tilde{\alpha} \circ (D\Xi, \pi_2)$; therefore by (i) in Theorem 2.6, Theorem 3.5, and Lemma 5.1, we can deduce that $\tilde{\chi}^\sharp D\tilde{\chi} : (E, \tau_\infty) \to \mathcal{L}((E, \tau_2), (E, \tau_2))$ is continuous. Assumption (1) of Lemma 4.5 is verified.

By definition $\Lambda D^2\tilde{\chi} = \lambda(D^2\phi - \phi'\psi^\sharp D^2\psi)$. $\lambda D^2\phi : (E, \tau_\infty) \to \mathcal{L}^2((E, \tau_2), \mathbf{R})$ is continuous by Theorem 3.5. Again using the results in Theorem 3.5 and in Lemma 5.1 it is not difficult to prove that $D^2\psi = D^2\tilde{\alpha} \otimes (D\Xi, \pi_2) + D\tilde{\alpha}(D^2\Xi, 0)$ is a continuous map from (E, τ_∞) to $\mathcal{L}^2((E, \tau_2), L^1(J, \mathbf{R}^r)$. From the regularity properties of $\psi^\sharp$ and ϕ' discussed above, we obtain assumption (2) of Lemma 4.5. Finally (iii) of Theorem 2.6 and Lemma 5.5 imply assumption (5) of Lemma 4.5, which now yields the desired result. □

REFERENCES

[1] A. Afanasev, V. Dikusar, A. Milyutin, and S. Chukanov, *A necessary condition in optimal control*, Nauka, Moscow, 1990. (In Russian.)

[2] R. F. Brown, *A Topological Introduction to Nonlinear Analysis*, Birkhäuser, Boston, 1993.

[3] A. L. Dontchev, W. W. Hager, A. B. Poore, and B. Yang, *Optimal stability and convergence in nonlinear control.* Preprint, 1992.

[4] E. G. Gilbert and D. S. Bernstein, *Second order necessary conditions in optimal control: Accessory-problem results without normality conditions*, J. Optim. Theory Appl., 41 (1983), pp. 75–106.

[5] K. A. GRASSE, *Controllability and Accessibility in Nonlinear Control Systems*, Ph.D. thesis, University of Illinois at Urbana-Champaign, Champaign, IL, 1979.

[6] F. LAMNABHI LAGARRIGUE, *Series de Volterra et commande optimale singuliere*, Ph.D. thesis, Université Paris XI, Paris, France, 1985.

[7] C. LESIAK AND A. J. KRENER, *The existence and uniqueness of Volterra series for nonlinear systems*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 1090–1095.

[8] E. LEVITIN, A. MILYUTIN, AND N. OSMOLOVSKIĬ, *Conditions of high order for a local minimum in problems with constraints*, Uspekhi Mat. Nauk, 33 (1978), pp. 85–148. (English transl. in Russian Math. Surveys, 33 (1978), pp. 97–168.)

[9] K. MAKOWSKI AND L. NEUSTADT, *Optimal control problems with mixed control-phase variable equality and inequality constraints*, SIAM J. Control Optim., 12 (1974), pp. 184–228.

[10] H. MAURER, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Study, 14 (1981), pp. 163–177.

[11] N. OSMOLOVSKIĬ, *Second order conditions for a weak local minimum in an optimal control problem (necessity, sufficiency)*, Dokl. Akad. Nauk SSSR, 225 (1975), pp. 259–262. (English transl. in Soviet Math. Dokl., 16 (1975), pp. 1480–1484.)

[12] ———, *Second order conditions for a weak extremum in an optimal control problem*, Ph.D. thesis, Moscow State University, 1976.

[13] G. STEFANI, *On Volterra approximations*, in New Trends in Nonlinear Control Theory, Lectures Notes in Control and Information Sciences 122, Springer Verlag, New York, 1989, pp. 212–221.

[14] G. STEFANI AND P. ZEZZA, *A new type of sufficient optimality conditions for a nonlinear constrained optimal control problem*, in Proc. Nonlinear Control System Design Symposium, M. Fliess, ed., Bordeaux, 1992, pp. 713–719.

[15] ———, *Optimal control problems with mixed state-control constraints: necessary conditions*, J. Math. Systems Estim. Control, 2 (1992), pp. 155–189.

[16] ———, *The Jacobi condition for LQ-control problems with constraints*, in Proc. Second European Control Conference, J. W. Nieuwenhuis, et al., eds., 1993, pp. 1003–1007.

[17] ———, *Regular constrained LQ-control problems.* University of Florence, preprint, 1994.

[18] J. WARGA, *Second order necessary conditions in optimization*, SIAM J. Control Optim., 22 (1984), pp. 524–528.

[19] V. ZEIDAN, *Sufficiency conditions for variational problems with variable endpoints: Coupled points*, Appl. Math. Optim., 27 (1993), pp. 191–209.

[20] ———, *The Riccati equation for optimal control problems with mixed state-control constraints: necessity and sufficiency*, SIAM J. Control Optim., 32 (1994), pp. 1297–1321.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 660–676, March 1996

014

THE EFFICIENCY OF SUBGRADIENT PROJECTION METHODS FOR CONVEX OPTIMIZATION, PART I: GENERAL LEVEL METHODS*

KRZYSZTOF C. KIWIEL†

Abstract. We study subgradient methods for convex optimization that use projections onto successive approximations of level sets of the objective corresponding to estimates of the optimal value. We present several variants and show that they enjoy almost optimal efficiency estimates. In another paper we discuss possible implementations of such methods. In particular, their projection subproblems may be solved inexactly via relaxation methods, thus opening the way for parallel implementations. They can also exploit accelerations of relaxation methods based on simultaneous projections, surrogate constraints, and conjugate and projected (conditional) subgradient techniques.

Key words. nondifferentiable (nonsmooth) optimization, convex programming, relaxation methods, subgradient optimization, successive projections, linear inequalities, parallel computing

AMS subject classifications. 65K05, 90C25

1. Introduction.

This is the first of two papers in which we study various modifications of Polyak's [Pol69] subgradient projection algorithm (SPA) and the recently proposed level method of [LNN95, LNN91] for solving the convex program

$$f^* = \min\{ f(x) : x \in S \} \tag{1.1}$$

under the following assumptions. S is a nonempty compact convex subset of $\mathbb{R}^N$; f is a convex function Lipschitz continuous on S with Lipschitz constant L_f; for each $x \in S$ we can compute $f(x)$ and a subgradient $g_f(x) \in \partial f(x)$ of f at x such that $|g_f(x)| \le L_f$; and for each $x \in \mathbb{R}^N$ we can find $P_S(x) = \arg\min\{|x - y| : y \in S\}$, its orthogonal projection on S, where $|\cdot|$ denotes the Euclidean norm.

If f^* is known, the simplest version of the SPA generates successive iterates

$$x^{k+1} = P_S(x^k - t_k(f(x^k) - f^*)g_f(x^k)/|g_f(x^k)|^2) \quad \text{for } k = 1, 2, \ldots, \tag{1.2}$$

until $g_f(x^k) = 0$, where $x^1 \in S$ and t_k are scalars in the set of *admissible stepsizes*

$$T = [\, t_{\min}, t_{\max} \,] \quad \text{for some fixed} \quad 0 < t_{\min} \le t_{\max} < 2. \tag{1.3}$$

It has the following *efficiency estimate* for any (absolute) accuracy $\epsilon > 0$:

$$\begin{aligned} &k > c_{\mathrm{SPA}}(t_{\min}, t_{\max})(\operatorname{diam}(S)L_f/\epsilon)^2 \quad \Rightarrow \quad \min\{ f(x^j) : j = 1{:}k \} - f^* < \epsilon, \\ &c_{\mathrm{SPA}}(t_{\min}, t_{\max}) = 1/[t_{\min}(2 - t_{\max})] \quad \text{and} \quad \min c_{\mathrm{SPA}}(\cdot,\cdot) = c_{\mathrm{SPA}}(1,1) = 1, \end{aligned} \tag{1.4}$$

where $\operatorname{diam}(S) = \sup_{x,y \in S} |x - y|$ denotes the diameter of S. This estimate (see §5) seems to be a folklore result, but it is less well known that it is optimal in a certain sense [LNN95, NeY79]: if S is a ball and $N \ge (\operatorname{diam}(S)L_f/\epsilon)^2/4$, then for any method that uses at most $(\operatorname{diam}(S)L_f/\epsilon)^2/4$ objective and subgradient evaluations, there exists a function for which this method does not obtain an accuracy better than ϵ.

* Received by the editors January 12, 1994; accepted for publication (in revised form) December 12, 1994. This research was supported by Polish State Committee for Scientific Research grant 8S50502206.

† Systems Research Institute, Newelska 6, 01–447 Warsaw, Poland (kiwiel@ibspan.waw.pl).

We present three schemes for estimating f^* in (1.2) that extend the ideas in [KAC91, KuF90, LNN95]. Two of them employ an overestimate $\bar{D} \geq \operatorname{diam}(S)$, which replaces $\operatorname{diam}(S)$ in (1.4); the third does not involve $\bar{D}$ but is much more difficult to implement.

To enable faster convergence, we give algorithms that use projections onto successive approximations of level sets of f derived from *several accumulated subgradient linearizations* of f or their *aggregates* (convex combinations) as in descent bundle methods for nondifferentiable optimization [HUL93, Kiw85, Lem89]. Such algorithms provide freedom to trade off storage requirements and work per iteration for speed of convergence.

In the accompanying paper [Kiw96] we discuss implementations of such algorithms, based on accelerations of the relaxation method for linear inequalities [Agm54, MoS54], and provide a unified perspective on various modifications proposed in the literature.

In effect, we show that several versions of subgradient projection methods share efficiency estimates similar to (1.4). Since this estimate cannot, in general, be improved uniformly with respect to the dimension N by more than an absolute constant factor, all these methods are optimal in the sense of [NeY79]. We note, however, that this estimate can be attained only for really large N. We may also expect that for "most" functions encountered in applications, the methods should be much more efficient than the worst-case estimates suggest. Indeed, preliminary numerical experience with the level method of [LNN95] has been very encouraging. Yet this method is not readily implementable because it requires unbounded storage (at least of order $k(N+1)$ at iteration k). Thus the main aim of our work has been to derive methods which have comparable efficiency but are more easily implementable.

We may add that the alternative extension [Kiw95] of the level methods of [LNN95, LNN91] is less suitable for parallel computing.

The paper is organized as follows. In §2 we introduce a general relaxation level algorithm. Its efficiency is analyzed in §3. In §4 we extend the nested ball principle of [Dre83]. Some useful modifications are given in §§5 and 6. Two alternative techniques for generating lower bounds f^k_{low} via fixed level gaps and full model minimizations are described in §7 and §8, respectively. Dual level methods are the subject of §9.

We use the following notation. We denote by $\langle \cdot, \cdot \rangle$ and $|\cdot|$, respectively, the usual inner product and norm in $\mathbb{R}^N$. $B(x,r) = \{y : |y-x| \leq r\}$ denotes the ball with center x and radius $r \geq 0$. For $\epsilon \geq 0$, the ϵ-subdifferential of f at x is defined by $\partial_\epsilon f(x) = \{p \in \mathbb{R}^N : f(y) \geq f(x) + \langle p, y-x \rangle - \epsilon \quad \forall y \in \mathbb{R}^N\}$. We denote by ∂f the ordinary subdifferential $\partial_0 f$. The natural logarithm with base e is denoted by $\ln(\cdot)$. We let $1{:}k$ denote $1, 2, \ldots, k$. For brevity, we let $a/bc = a/(bc)$. The convex hull is denoted by co.

2. The relaxation level algorithm. In this section we describe our first modification of the SPA. As in [BaS81, KAC91, LNN95], when the optimal value f^* is unknown, it may be replaced in (1.2) by a *variable target (level) value*

$$f^k_{\text{lev}} = f^k_{\text{up}} - \kappa(f^k_{\text{up}} - f^k_{\text{low}}) = \kappa f^k_{\text{low}} + (1-\kappa) f^k_{\text{up}}, \tag{2.1}$$

where $0 < \kappa < 1$ is fixed, $f^k_{\text{up}} = \min_{j=1:k} f(x^j)$ is an *upper bound* on f^*, and the *lower bound* $f^k_{\text{low}} \leq f^*$ is chosen to ensure $f^k_{\text{lev}} \to f^*$ as $k \to \infty$. Thus we obtain the subgradient projection level algorithm (SPLA):

$$x^{k+1} = P_S(x^k - t_k(f(x^k) - f^k_{\text{lev}}) g_f(x^k)/|g_f(x^k)|^2), \; t_k \in T, \; k = 1, 2, \ldots; \tag{2.2}$$

if $g_f(x^k) = 0$, then, of course, the method stops with an optimal x^k in $S^* = \operatorname{Arg\,min}_S f$. To get some feeling about possible updates of f^k_{low}, it is instructive to consider first the following ideal bisection method (cf. [MTA81]). Let $\mathcal{L}(f, f^k_{\text{lev}}) = \{x : f(x) \le f^k_{\text{lev}}\}$.

ALGORITHM 2.1 (*ideal level method for* (1.1)).

Step 0. Choose $0 < \kappa < 1$, $z^1 \in S$, and $f^1_{\text{low}} \le f^*$. Set $k = 1$.

Step 1. Set $f^k_{\text{up}} = f(z^k)$, f^k_{lev} by (2.1), and the optimality gap $\Delta^k = f^k_{\text{up}} - f^k_{\text{low}}$.

Step 2. If $\mathcal{L}(f, f^k_{\text{lev}}) \cap S = \emptyset$, go to step 4.

Step 3. Find $z^{k+1} \in \mathcal{L}(f, f^k_{\text{lev}}) \cap S$, set $f^{k+1}_{\text{low}} = f^k_{\text{low}}$, increase k by 1, and go to step 1.

Step 4. Set $z^{k+1} = z^k$, $f^{k+1}_{\text{low}} = f^k_{\text{lev}}$, increase k by 1, and go to step 1.

Clearly, the method produces $f^k_{\text{low}} \le f^*$, $f^k_{\text{up}} - f^* \le \Delta^k$, and $\Delta^{k+1} \le \max\{\kappa, (1-\kappa)\}\Delta^k$ for all k. The crucial property is that $\mathcal{L}(f, f^k_{\text{lev}}) \cap S = \emptyset$ implies $f^k_{\text{lev}} < f^*$.

To make Algorithm 2.1 implementable, we need a submethod for finding a point in $\mathcal{L}(f, f^k_{\text{lev}}) \cap S$ or detecting that $\mathcal{L}(f, f^k_{\text{lev}}) \cap S = \emptyset$. For this kth *set intersection problem*, an iteration of the successive projections method [GPR67] of the form $\hat{x}^{k+1} = P_S(P_{\mathcal{L}(f, f^k_{\text{lev}})}(x^k))$ can be implemented approximately as follows: letting $\bar{f}(\cdot; y) = f(y) + \langle g_f(y), \cdot - y\rangle$ denote the *linearization* of f at any $y \in S$, with $\bar{f}(\cdot; y) \le f(\cdot)$ and $\bar{f}(y; y) = f(y)$ by convexity, we have

$$S^* = \{ x : \bar{f}(x; y) \le f^* \ \forall y \in S \} \cap S \tag{2.3}$$

and

$$\mathcal{L}(f, f^k_{\text{lev}}) = \{ x : \bar{f}(x; y) \le f^k_{\text{lev}} \ \forall y \in S \}. \tag{2.4}$$

We may use some accumulated linearizations $f^j(\cdot) = \bar{f}(\cdot; x^j)$, $j \le k$, in the kth model of f,

$$\hat{f}^k(x) = \max\{ f^j(x) : j \in J^k \} \quad \text{with} \quad k \in J^k \subset \{1{:}\,k\}, \tag{2.5}$$

and let

$$x^{k+1} = P_S(x^k + t_k[P_{\mathcal{L}(\hat{f}^k, f^k_{\text{lev}})}(x^k) - x^k]), \tag{2.6}$$

where we have *underprojection* if $t_k < 1$ or *overprojection* if $t_k > 1$. For instance, (2.6) gives (1.2) when $f^k_{\text{lev}} = f^*$, $J^k = \{k\}$, and $\mathcal{L}(\hat{f}^k, f^k_{\text{lev}})$ is the halfspace $H^k = \{x : f^k(x) \le f^k_{\text{lev}}\}$ given by the inequality of (2.3) most violated at x^k; of course, $P_{H^k}(x^k) = x^k - (f(x^k) - f^k_{\text{lev}}) g_f(x^k)/|g_f(x^k)|^2$. This is just an iteration of a relaxation method for solving the inequalities of (2.4), followed by a projection on S. As for (2.2), *f^k_{low} may be increased to f^k_{lev} when it is discovered that these inequalities do not have a solution in S.* Hence we shall exploit the fact that certain versions of successive projections methods can detect in finite time that a given set intersection problem is unsolvable (although they need not find a solution in finite time when it exists). As will be seen below, the main idea of such methods is to reduce the distance of the iterates from S^*. They may be painfully slow, even in the most favorable case of $f^k_{\text{lev}} = f^*$, when only one inequality of (2.3) is considered at a time. To accelerate convergence, we may use a larger J^k, i.e., a tighter approximation $\hat{f}^k$ of f.

To illustrate these facts we need a result of Agmon [Agm54]. Given a closed convex set $C \subset \mathbb{R}^N$ and an admissible stepsize $t \in T$, we define the *relaxation operator*

$$\mathcal{R}_{C,t}(x) = x + t(P_C(x) - x) \tag{2.7}$$

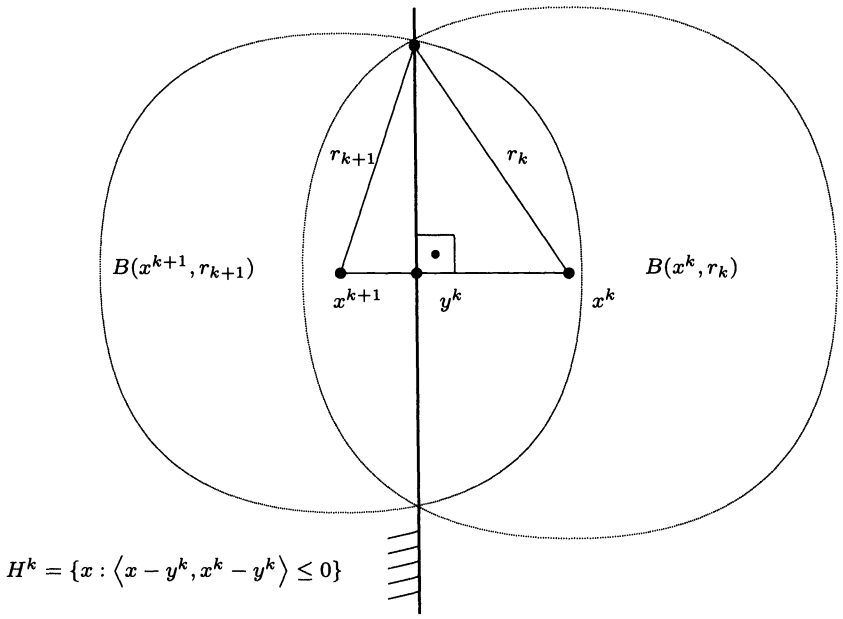


FIG. 2.1. *Illustration of the Fejér property of* $x^{k+1} = x^k + t_k(y^k - x^k)$ *with* $y^k = P_{H^k}(x^k)$ *when* $x^{k+1} \in S$. *By Pythagoras' theorem,* $r_{k+1}^2 - |x^{k+1} - y^k|^2 = r_k^2 - |y^k - x^k|^2$, *where* $x^{k+1} - y^k = (t_k - 1)(y^k - x^k)$, *so* $r_{k+1}^2 = r_k^2 - t_k(2 - t_k)|y^k - x^k|^2$. *Clearly,* $B(x^k, r_k) \cap H^k \subset B(x^{k+1}, r_{k+1}) \cap H^k$.

(where $P_C(x) = x$ if $C = \emptyset$) that has the Fejér contraction property

$$
\begin{aligned}
|y - \mathcal{R}_{C,t}(x)|^2 &\leq |y - x|^2 - t(2 - t)|x - P_C(x)|^2 \\
&\leq |y - x|^2 - t_{\min}(2 - t_{\max}) d_C^2(x) \quad \forall y \in C, x \in \mathbb{R}^N,
\end{aligned} \tag{2.8}
$$

where $d_C(x) = \inf_{y \in C} |x - y|$. Indeed, if $y \in C$, $P = P_C$, and $z = x + t(P(x) - x)$, then

$$
\begin{aligned}
|y - z|^2 &= |y - x|^2 + (t|P(x) - x|)^2 - 2t \langle y - x, P(x) - x \rangle \\
&= |y - x|^2 + (t|P(x) - x|)^2 - 2t \langle P(x) - x, P(x) - x \rangle \\
&\qquad - 2t \langle y - P(x), P(x) - x \rangle \\
&\leq |y - x|^2 - t(2 - t)|P(x) - x|^2 \leq |y - x|^2 - t_{\min}(2 - t_{\max})|P(x) - x|^2
\end{aligned}
$$

from the projection property $\langle y - P(x), P(x) - x \rangle \geq 0$ and (1.3). Note that $t_{\min}(2 - t_{\max})$ in (2.8) can be replaced by $\min_{t \in T} t(2 - t)$.

Figure 2.1 illustrates the Fejér property of (2.2) with $H^k = \mathcal{L}(f^k, f_{\text{lev}}^k)$. For motivation, we now state some facts that will be proved later. Suppose we have generated some $r_k \geq d_{S^*}(x^k)$ (starting, e.g., from $r_1 = \bar{D} \geq \text{diam}(S)$) so that $B(x^k, r_k) \cap S^* \neq \emptyset$. If $f_{\text{lev}}^k \geq f^*$, then $S^* \subset H^k$, so setting $y^k = P_{H^k}(x^k)$, finding r_{k+1} from

$$
r_{k+1}^2 = r_k^2 - t_k(2 - t_k)|y^k - x^k|^2, \tag{2.9}
$$

and applying (2.8) twice, as in the proof of Lemma 3.2 below, we deduce that $S^* \cap B(x^k, r_k) \subset S^* \cap B(x^{k+1}, r_{k+1})$. Thus we improve our localization of the solution (since $r_{k+1} < r_k$ due to $x^k \notin H^k$ from $f(x^k) > f_{\text{lev}}^k$). On the other hand, if $t_k(2 -$

$t_k)d^2_{H^k}(x^k) > r_k^2$, then $f^k_{\text{lev}} < f^*$ (by contradiction), so we may increase f^k_{low} to f^k_{lev} and reset r_{k+1} to $\bar{D}$. To sum up, if $f^k_{\text{lev}} \geq f^*$, then progress towards the solution is measured by the magnitude of $d_{H^k}(x^k)$, otherwise $d_{H^k}(x^k)$ may be used to shrink r_k until $f^k_{\text{lev}} < f^*$ is discovered; thus $d_{H^k}(x^k)$ should be as large as possible in both cases. Hence the algorithm may be accelerated by choosing a smaller $\mathcal{L}(\hat{f}^k, f^k_{\text{lev}})$ to produce $d_{\mathcal{L}(\hat{f}^k, f^k_{\text{lev}})}(x^k) > d_{H^k}(x^k)$. However, a large J^k in (2.5) would create difficulties with storage and work per iteration. This raises the following basic questions. Is it possible to select J^k so that $\hat{f}^k$ approximates f tightly in the region of interest without J^k becoming inordinately large? Can we reduce J^k by replacing some f^j with their convex combinations, i.e., by aggregating some constraints in $\mathcal{L}(\hat{f}^k, f^k_{\text{lev}})$? Should not $\mathcal{L}(\hat{f}^k, f^k_{\text{lev}})$ be augmented with some inequalities related to S? Instead of finding $P_{\mathcal{L}(\hat{f}^k, f^k_{\text{lev}})}(x^k)$, can we perform several "simpler" projections (possibly inexactly and in parallel) and combine their solutions? Our partial answers to these questions will involve a combination of some quite technical properties of relaxation methods. For instance, note that, in view of the outer projection in (2.2), r^2_{k+1} in (2.9) could be further reduced by $d^2_S(z^k)$, where $z^k = x^k + t_k(y^k - x^k)$. In fact more than two successive projections could be employed to reduce r_{k+1}. We shall need rather abstract notation to make such concepts precise.

Let δ_S denote the indicator of S ($\delta_S(x) = 0$ if $x \in S$, ∞ otherwise) and $f_S = f + \delta_S$ the extended objective. Let $\check{f}^k = \max_{j=1:k} f^j$ denote the kth "best" model of f (which we would *not* like to store). Note that $\check{f}^k_S = \check{f}^k + \delta_S$ is the largest convex minorant of f_S compatible with the accumulated information about f. Clearly, f^k, $\hat{f}^k$, $\check{f}^k$, and $\check{f}^k_S$ belong to the following *set of admissible models of* f_S:

$$\Phi = \{\, \phi\colon \mathbb{R}^N \to (-\infty, \infty] : \phi \text{ is closed convex and } \phi(x) \leq f^* \ \forall x \in S^* \,\}. \tag{2.10}$$

At iteration k, we may choose a model $\phi^k \in \Phi$ such that $\phi^k \geq f^k$ (to exploit the latest subgradient information) and a stepsize $t_k \in T$. Then the iteration

$$x^{k+1} = P_S(\mathcal{R}_{\mathcal{L}(\phi^k, f^k_{\text{lev}}), t_k}(x^k)) \tag{2.11}$$

is a generalization of (2.2) and (2.6), which have $\phi^k = f^k$ and $\phi^k = \hat{f}^k$, respectively. This notation is convenient for the implementations discussed later, in which each ϕ^k may be the maximum of several accumulated linearizations f^j, $j \leq k$, or their convex combinations, possibly augmented with δ_S or its convex minorants. It will also prepare ground for extensions which use several models from Φ at each iteration for successive or parallel relaxations. (For the first reading, one may assume $\phi^k = f^k$.) We should, of course, ensure that $\mathcal{L}(\phi^k, f^k_{\text{lev}}) \neq \emptyset$ in (2.11). (Detecting this may require calculating $\inf \phi^k$ approximately.) Since, by (2.10),

$$\emptyset \neq S^* \subset \mathcal{L}(\phi, f^k_{\text{lev}}) \quad \text{if} \quad \phi \in \Phi \text{ and } f^k_{\text{lev}} \geq f^*, \tag{2.12}$$

$\mathcal{L}(\phi^k, f^k_{\text{lev}}) = \emptyset$ means that we may repeat (2.11) with f^k_{lev} increased to a new value by increasing f^k_{low} to the old value of f^k_{lev}. Note that $\mathcal{L}(\phi^k, f^k_{\text{lev}}) = \emptyset$ cannot occur in the simplest method (2.2), for which the test based on r_k must be employed.

We may now state the first general subgradient projection algorithm with relaxation and target level updating. Its notation is slightly redundant, being geared toward subsequent convergence proofs and modifications.

ALGORITHM 2.2.

Step 0 (Initialization). Select an initial point $x^1 \in S$, a final optimality tolerance

$\epsilon_{\text{opt}} \geq 0$, a level parameter $0 < \kappa < 1$, and stepsize parameters $0 < t_{\min} \leq t_{\max} < 2$. Choose $\bar{D} \geq \operatorname{diam}(S)$ and $f^1_{\text{low}} \leq f^*$. Set $\rho_1 = 0$ and $f^0_{\text{up}} = \infty$. Set the counters $k = 1$, $l = 0$, and $k(0) = 0$. ($k(l)$ will denote the iteration number of the lth increase of f^k_{low}.)

Step 1 (Objective evaluation). Calculate $f(x^k)$ and $g_f(x^k)$.

Step 2 (Level update). Set $f^k_{\text{up}} = \min\{f(x^k), f^{k-1}_{\text{up}}\}$, f^k_{lev} by (2.1), and the *gap* $\Delta^k = f^k_{\text{up}} - f^k_{\text{low}}$.

Step 3 (Stopping criterion). If $\min\{\Delta^k, |g_f(x^k)|/\bar{D}\} \leq \epsilon_{\text{opt}}$, terminate.

Step 4 (Projections). Perform (2.11), checking if it is well defined, as follows:

(i) Choose an admissible model $\phi^k \in \Phi$ such that $\phi^k \geq f^k$ and a stepsize $t_k \in T$.

(ii) If $\mathcal{L}(\phi^k, f^k_{\text{lev}}) = \emptyset$, go to step 5. Otherwise, set $y^k = P_{\mathcal{L}(\phi^k, f^k_{\text{lev}})}(x^k)$, $z^k = x^k + t_k(y^k - x^k)$, $x^{k+1} = P_S(z^k)$, $\rho^k_\phi = t_k(2 - t_k)|y^k - x^k|^2$, and $\rho^k_S = |x^{k+1} - z^k|^2$.

(iii) If $\rho_k + \rho^k_\phi + \rho^k_S > \bar{D}^2$, go to step 5; otherwise, go to step 6.

Step 5 (Update lower bound).

(i) Choose a lower bound $\hat{f}^k_{\text{low}} \in [\max\{f^k_{\text{low}}, f^k_{\text{lev}}\}, f^*]$ (e.g., $\hat{f}^k_{\text{low}} = \max\{f^k_{\text{low}}, f^k_{\text{lev}}\}$). Set $f^{k+1}_{\text{low}} = \hat{f}^k_{\text{low}}$, $\rho_{k+1} = 0$, and $\hat{\Delta}^k = f^k_{\text{up}} - \hat{f}^k_{\text{low}}$.

(ii) If $\hat{\Delta}^k \leq \epsilon_{\text{opt}}$, terminate; otherwise, continue.

(iii) Set $x^{k+1} = x^k$ (*null step*), $k(l+1) = k$, and increase k and l by 1. Go to step 2.

Step 6 (Serious step). Set $f^{k+1}_{\text{low}} = f^k_{\text{low}}$, $\hat{f}^k_{\text{low}} = f^k_{\text{low}}$, $\hat{\Delta}^k = \Delta^k$, and $\rho_{k+1} = \rho_k + \rho^k_\phi + \rho^k_S$. Increase k by 1 and go to step 1.

A few comments on the method are in order.

At step 0, f^1_{low} may be obtained, e.g., from a relaxation of (1.1) or from the relations

$$f^* \geq \min_S f^1 \geq f(x^1) - |g_f(x^1)| \operatorname{diam}(S) \geq f(x^1) - |g_f(x^1)|\bar{D}, \tag{2.13}$$

since $f(\cdot) \geq f^1(\cdot) = f(x^1) + \langle g_f(x^1), \cdot - x^1 \rangle \geq f(x^1) - |g_f(x^1)||\cdot - x^1|$ by the Cauchy–Schwarz inequality. In many applications one may find a "simple" set (e.g., a box or a ball) that contains S; the diameter of this set may serve as $\bar{D}$. (Choosing f^1_{low} and $\bar{D}$ when f is strongly convex on S is discussed in [KAC91]; see also [KuF90].) In general, the algorithm should perform better the closer f^1_{low} and $\bar{D}$ are to f^* and $\operatorname{diam}(S)$, respectively.

Note that the f-evaluation at step 1 is skipped if $x^k = x^{k-1}$ after a null step at step 5, i.e., if $k = k(l) + 1$. The current number of f-evaluations is $k - l$.

Step 3 is justified by the optimality estimates (2.14) and and the fact that $f^* \geq f(x^k) - |g_f(x^k)| \operatorname{diam}(S)$.

Step 4 performs the two successive relaxations of (2.11), unless an exit to step 5 occurs with $f^k_{\text{lev}} < f^*$. (The empty intersection test of step 2 in Algorithm 2.1 is done in two separate tests in step 4 of Algorithm 2.2.) The exit from step 4(ii) is justified by (2.12) and from step 4(iii) by Lemma 3.3, which formalizes our argument concerning (2.9). Specifically, with $r^2_k = \bar{D}^2 - \rho_k$, step 6 replaces (2.9) by $r^2_{k+1} = r^2_k - \rho^k_\phi - \rho^k_S$, whereas steps 0 and 5 ensure $r_{k(l)+1} = \bar{D}$.

Let us split the iterations into groups $K_0 = \{1 : k(1) - 1\}$ and $K_l = \{k(l) : k(l+1) - 1\}$ if $l \geq 1$. Each group K_l ends by discovering that the target level is unattainable. Then an increase of the lower bound reduces the gap between the bounds by at least a fraction of $\kappa < 1$. The remaining level and gap decreases within each group occur only when the objective improves, with the lower bound staying fixed. These simple

properties of the method may by derived inductively from the following observations. By construction, $f^k_{\rm up} \ge f^{k+1}_{\rm up} \ge f^* \ge f^{k+1}_{\rm low} = \hat f^k_{\rm low} \ge f^k_{\rm low}$, $\hat\Delta^k = f^k_{\rm up} - \hat f^k_{\rm low}$, and $\Delta^k = f^k_{\rm up} - f^k_{\rm low}$, so the gaps $\hat\Delta^k \le \Delta^k$ overestimate the optimality gap

$$f^k_{\rm up} - f^* = \min\{ f(x^j) : j = 1{:}k \} - f^* \le \hat\Delta^k \le \Delta^k \tag{2.14}$$

and $\Delta^{k+1} \le \hat\Delta^k \le \Delta^k$ for all k. In fact, if $k(l) < k < k(l+1)$, then $f^k_{\rm low} = \hat f^{k(l)}_{\rm low}$ $(= f^1_{\rm low}$ if $l = 0)$; therefore, the level $f^k_{\rm lev} = f^k_{\rm low} + (1-\kappa)\Delta^k$ cannot increase:

$$f^{k(l)+1}_{\rm lev} \ge f^j_{\rm lev} \ge f^k_{\rm lev} \quad \text{if} \quad k(l) < j \le k \le k(l+1) \tag{2.15}$$

and $\Delta^k = \hat\Delta^k$ if $k(l) < k < k(l+1)$. Hence $\hat f^k_{\rm low}$ and $\hat\Delta^k$ only reflect the improvement in $f^k_{\rm low}$ and Δ^k at iterations $k = k(l+1)$, $l \ge 0$. Then at step 5, $\hat f^k_{\rm low} \ge f^k_{\rm lev} = f^k_{\rm up} - \kappa\Delta^k$ implies $\hat\Delta^k = f^k_{\rm up} - \hat f^k_{\rm low} \le \kappa\Delta^k$. Thus we have the useful relations

$$\Delta^k \ge \hat\Delta^k \ge \hat\Delta^{k(l+1)}/\kappa \quad \text{if} \quad k \in K_l \text{ and } l \ge 0, \tag{2.16}$$

$$\hat\Delta^{k(l)} \le \kappa^l \Delta^1 \quad \text{if} \quad l \ge 1. \tag{2.17}$$

3. Efficiency. Our aim is to show that the SPLA of (2.2) has the following efficiency estimate for any $\epsilon > 0$:

$$k > c_{\rm SPLA}(t_{\min}, t_{\max}, \kappa)(\bar D L_f/\epsilon)^2 \quad \Rightarrow \quad \min\{ f(x^j) : j = 1{:}k \} - f^* < \epsilon, \tag{3.1a}$$

$$c_{\rm SPLA}(t_{\min}, t_{\max}, \kappa) = 1/t_{\min}(2 - t_{\max})\kappa^2(1-\kappa^2), \tag{3.1b}$$

$$\min c_{\rm SPLA}(\cdot,\cdot,\cdot) = c_{\rm SPLA}(1, 1, 1/\sqrt{2}) = 4, \tag{3.1c}$$

and to establish a modified form of this estimate for Algorithm 2.2. We assume, with no loss of generality, that the tolerance $\epsilon_{\rm opt} = 0$ and that the algorithm does not terminate, i.e., $\Delta^k \ge \hat\Delta^k > 0$ for all k.

We start by showing that each first relaxation at step 4 provides a significant growth of ρ_k related to Fejér contractions. Note that with $H^k = \{x : f^k(x) \le f^k_{\rm lev}\}$ we have

$$d_C(x^k) \ge d_{H^k}(x^k) = (f(x^k) - f^k_{\rm lev})/|g_f(x^k)| \quad \text{if} \quad C \subset H^k. \tag{3.2}$$

LEMMA 3.1. *If $\mathcal{L}(\phi^k, f^k_{\rm lev}) \ne \emptyset$ at step 4, then $\rho^k_\phi \ge t_{\min}(2 - t_{\max})(\kappa\Delta^k/L_f)^2$.*

Proof. Use (3.2) with $\mathcal{L}(\phi^k, f^k_{\rm lev}) \subset H^k$ (from $\phi^k \ge f^k$), $|g_f(x^k)| \le L_f$, and $f(x^k) - f^k_{\rm lev} \ge f^k_{\rm up} - (f^k_{\rm up} - \kappa\Delta^k) = \kappa\Delta^k$ (from $f(x^k) \ge f^k_{\rm up}$); recall step 4 and (1.3). □

LEMMA 3.2. *Suppose $y \in \mathcal{L}(\phi^k, f^k_{\rm lev}) \cap S$ for iterations $k = k_1{:}k_2$ that do not execute step 5 (i.e., y is a common point of all the sets involved in the successive relaxations (2.11) at step 4 for such k). Then*

$$\begin{aligned} \rho_{k_2+1} - \rho_{k_1} &= \sum_{k=k_1}^{k_2} [t_k(2-t_k)|y^k - x^k|^2 + |x^{k+1} - z^k|^2] \\ &\le |y - x^{k_1}|^2 - |y - x^{k_2+1}|^2. \end{aligned} \tag{3.3}$$

Proof. Fix $k \in [k_1, k_2]$. Use (2.8) with $C = \mathcal{L}(\phi^k, f^k_{\text{lev}})$, $t = t_k$, and $x = x^k$ and next with $C = S$, $t = 1$, and $x = z^k$ to get

$$\rho^k_\phi = t_k(2 - t_k)|y^k - x^k|^2 \le |y - x^k|^2 - |y - z^k|^2, \tag{3.4a}$$

$$\rho^k_S = |x^{k+1} - z^k|^2 \le |y - z^k|^2 - |y - x^{k+1}|^2. \tag{3.4b}$$

Add the inequalities above to get $\rho_{k+1} - \rho_k = \rho^k_\phi + \rho^k_S \le |y - x^k|^2 - |y - x^{k+1}|^2$. Adding these inequalities for $k = k_1 : k_2$ yields (3.3). □

The next result validates the test at step 4(iii) for increasing f^k_{low} at step 5. Recall that $k(l)$ is the iteration number of the lth increase of f^k_{low}; these quantities change only at step 5, and we *always* have $k(l) < k$ (cf. step 0). Note that Lemma 3.2 assumes that step 5 is not executed, but this is to be proved for the lemma below.

LEMMA 3.3. *If $f^k_{\text{lev}} \ge f^*$ at step 4, then step 5 is not entered and*

$$\begin{aligned} \rho_{k+1} = \rho_k + \rho^k_\phi + \rho^k_S &\le |y - x^{k(l)+1}|^2 - |y - x^{k+1}|^2 \\ &\le \operatorname{diam}(S)^2 \le \bar{D}^2 \quad \forall y \in S^*. \end{aligned} \tag{3.5}$$

Proof. First, suppose $k > k(l) + 1$. Since $f^k_{\text{lev}} \ge f^*$, (2.12) and (2.15) imply that the assumptions of Lemma 3.2 hold for any fixed $y \in S^*$, $k_1 = k(l)+1$, and $k_2 = k-1$. Then, due to the rules of steps 5 and 6, (3.3) becomes

$$\rho_k \le |y - x^{k(l)+1}|^2 - |y - x^k|^2. \tag{3.6}$$

Adding (3.6) to (3.4) we get (3.5), noting that $\rho_k + \rho^k_\phi + \rho^k_S \le |y - x^{k(l)+1}|^2 \le \bar{D}^2$; i.e., no null step occurs. Next, if $k = k(l) + 1$, then $\rho_k = 0$ (cf. steps 0 and 5), so (3.6) holds again and the conclusion follows as before. □

We may now estimate the rate of decrease of the gap Δ^k within each group K_l.

LEMMA 3.4. *If $k(l) < k < k(l+1)$ and $\Delta^k > 0$, then*

$$k - k(l) \le (\bar{D}L_f/\kappa\Delta^k)^2/t_{\min}(2 - t_{\max}). \tag{3.7}$$

Proof. Note that $\Delta^j \ge \Delta^k$ for $j = 1 : k$ because Δ^j never increases. By the rules of steps 4 and 5, we have $\rho_{k+1} \le \bar{D}^2$ (otherwise $k(l+1) = k$ would occur, a contradiction) and

$$\begin{aligned} \bar{D}^2 \ge \rho_{k+1} &\ge \sum_{j=k(l)+1}^{k} \rho^j_\phi \ge t_{\min}(2 - t_{\max}) \sum_{j=k(l)+1}^{k} (\kappa\Delta^j/L_f)^2 \\ &\ge t_{\min}(2 - t_{\max})(\kappa\Delta^k/L_f)^2(k - k(l)) \end{aligned}$$

from Lemma 3.1. Rearranging, we get (3.7). □

At step 2, let $n^k_f = k - l$ and $l^k = l$ denote the total numbers of f-evaluations and lower bound increases, respectively. In (3.8) below we in fact relate n^k_f to the gap $\hat{\Delta}^k$.

LEMMA 3.5. *If $\hat{\Delta}^{k_\epsilon} \ge \epsilon > 0$ for some $k_\epsilon \in K_m$ and $m \ge 0$, then $n^{k_\epsilon}_f = k_\epsilon - m$ and*

$$k_\epsilon \le m + (\bar{D}L_f/\epsilon)^2/\kappa^2(1 - \kappa^2)t_{\min}(2 - t_{\max}). \tag{3.8}$$

If additionally $\Delta^1 \le \bar{D}L_f$, then $m \le -\ln(\bar{D}L_f/\epsilon)/\ln(\kappa)$ and

$$k_\epsilon \le (\bar{D}L_f/\epsilon)^2[1/\kappa^2(1 - \kappa^2)t_{\min}(2 - t_{\max}) - 1/2e\ln(\kappa)]. \tag{3.9}$$

Proof. (i) Let $K(\epsilon) = \{1{:}\, k_\epsilon\}$. Since $\hat{\Delta}^{k_\epsilon} \geq \epsilon > 0$ and $\Delta^{k+1} \leq \hat{\Delta}^k \leq \Delta^k$ for all k, use (2.16) and induction to obtain $\Delta^k \geq \epsilon/\kappa^{m-l}$ for all $k \in K_l \cap K(\epsilon)$ and $l = 0{:}\, m$.

(ii) Let $c = (\bar{D}L_f/\kappa)^2/t_{\min}(2 - t_{\max})$. By (i) and Lemma 3.4, $|K_l \cap K(\epsilon)| \leq 1 + c\kappa^{2(m-l)}/\epsilon^2$ for $l = 1{:}\, m$ and $|K_0 \cap K(\epsilon)| \leq c\kappa^{2m}/\epsilon^2$. Since $0 < \kappa < 1$, we get (3.8) from

$$k_\epsilon = \sum_{l=0}^{m} |K_l \cap K(\epsilon)| \leq m + \sum_{l=0}^{m} (c/\epsilon^2)\kappa^{2(m-l)} \leq m + c/\epsilon^2(1 - \kappa^2).$$

(iii) If $\Delta^1 \leq \bar{D}L_f$ and $m > 0$, then (2.17) yields $\epsilon \leq \hat{\Delta}^{k(m)} \leq \kappa^m \bar{D}L_f$, so $m \leq -\ln(\bar{D}L_f/\epsilon)/\ln(\kappa)$ in (3.8). Thus, to get (3.9), it suffices to prove that $-\ln(t)/\ln(\kappa) \leq -t^2/2e\ln(\kappa)$ for all $t > 0$. Indeed, $t^2 - 2e\ln(t) \geq 0$ for all $t > 0$ (minimize it!). □

We may now state our principal result. Notice that, in view of (2.13), we may always ensure that $\Delta^1 \leq \bar{D}L_f$ by taking $f^1_{\text{low}} \geq f(x^1) - |g_f(x^1)|\bar{D}$, and recall (2.14).

THEOREM 3.6. *If $\Delta^1 \leq \bar{D}L_f$, then the following efficiency estimate holds for each $\epsilon > 0$:*

$$k > c_{\text{RLA}}(t_{\min}, t_{\max}, \kappa)(\bar{D}L_f/\epsilon)^2 \quad \Rightarrow \quad f^k_{\text{up}} - f^* \leq \hat{\Delta}^k < \epsilon, \tag{3.10a}$$

$$c_{\text{RLA}}(t_{\min}, t_{\max}, \kappa) = 1/t_{\min}(2 - t_{\max})\kappa^2(1 - \kappa^2) - 1/2e\ln(\kappa), \tag{3.10b}$$

$$\min c_{\text{RLA}}(\cdot,\cdot,\cdot) = c_{\text{RLA}}(1, 1, 0.677653\ldots) \approx 4.49950. \tag{3.10c}$$

Proof. This is an immediate consequence of Lemma 3.5. □

Let $x^k_{\text{rec}} \in \{x^j\}_{j=1}^k$ be such that $f(x^k_{\text{rec}}) = f^k_{\text{up}}$ $(= \min_{j=1:k} f(x^j))$, for all k.

COROLLARY 3.7. *If $\epsilon_{\text{opt}} = \epsilon > 0$ and $\Delta^1 \leq \bar{D}L_f$, then the algorithm will terminate with $f(x^k_{\text{rec}}) \leq f^* + \epsilon$ in $k = 1 + k_\epsilon$ iterations after $n^k_f = 1 + n^{k_\epsilon}_f$ f-evaluations, where k_ϵ and $n^{k_\epsilon}_f = k_\epsilon - m$ satisfy the bounds of Lemma 3.5.* □

For completeness, we include an asymptotic result.

THEOREM 3.8. *If the algorithm does not terminate, then f^k_{up}, f^k_{low}, and f^k_{lev} converge to f^*, and Δ^k and $\hat{\Delta}^k$ converge to zero as $k \to \infty$. Moreover, $\{x^k_{\text{rec}}\}$ converges to S^*.*

Proof. Since $\hat{\Delta}^k > 0$ never increases, $\hat{\Delta}^k \downarrow 0$ either by (2.17) if $l \to \infty$ or by Lemma 3.5 otherwise. (Then m would be bounded in (3.8).) Hence the facts that $\Delta^{k+1} \leq \hat{\Delta}^k \leq \Delta^k$ and $\max\{|f^k_{\text{low}} - f^*|, |f^k_{\text{up}} - f^*|, |f^k_{\text{lev}} - f^*|\} \leq \Delta^k$ for all k imply the first assertion. The second one follows from $f(x^k_{\text{rec}}) \to f^*$, the continuity of f, and the compactness of S. □

Remark 3.9. In view of the preceding results, we again emphasize the crucial role of $\rho^k_\phi = t_k(2 - t_k)d^2_{\mathcal{L}(\phi^k, f^k_{\text{lev}})}(x^k)$ in our efficiency analysis. The algorithm may be accelerated (locally) by choosing ϕ^k and t_k to generate a large ρ^k_ϕ. (In fact we should try to increase the less easily manageable quantity $\rho^k_\phi + \rho^k_S$ instead of just ρ^k_ϕ.) Our efficiency estimates are best when $t_{\min} = t_{\max} = 1$; also $t_k = 1$ maximizes each ρ^k_ϕ. However, as in other relaxation methods, other choices of t_k may be preferable in practice.

4. The nested ball principle. We shall need the following reformulation of Lemma 3.3 in terms of $r^2_k = \bar{D}^2 - \rho_k$. It generalizes similar results of [Gof81, Tel82] obtained for classical relaxation methods.

LEMMA 4.1 (the ball induction principle). *If $f^k_{\text{lev}} \geq f^*$ at step 4, then $\emptyset \neq S^* \cap B(x^k, r_k) \subset S^* \cap B(z^k, (r^2_k - \rho^k_\phi)^{1/2}) \subset S^* \cap B(x^{k+1}, r_{k+1})$.*

Proof. Letting $y \in S^*$, use (3.4) and (3.6). □

The following result extends one of Drezner [Dre83] (and simplifies its proof).

LEMMA 4.2 (the nested ball principle). *If* $(\bar{D} - |z^k - x^{k(l)+1}|)^2 > r_k^2 - \rho_\phi^k$ *or* $(\bar{D} - |x^{k+1} - x^{k(l)+1}|)^2 > r_k^2 - \rho_\phi^k - \rho_S^k$ *at step* 4, *then* $f_{\text{lev}}^k < f^*$.

Proof. For contradiction suppose $f_{\text{lev}}^k \geq f^*$. Let $x = x^{k(l)+1}$, $z = z^k$, $\hat{r} = (r_k^2 - \rho_\phi^k)^{1/2}$, and $y \in S^* \cap B(x, \bar{D}) \cap B(z, \hat{r})$ (cf. Lemma 4.1). Suppose $\bar{D} > \hat{r} + |z - x|$. By construction, (3.4a) and (3.6), $|y-z|^2 \leq |y-x|^2 + \hat{r}^2 - \bar{D}^2 \leq (|y-z| + |z-x|)^2 + \hat{r}^2 - \bar{D}^2$ with $|z-x| \neq 0$ due to $\hat{r} < \bar{D}$, so $|y-z| \geq (\bar{D}^2 - \hat{r}^2 - |z-x|^2)/2|z-x| > \hat{r}$ contradicts $y \in B(z, \hat{r})$. Hence $\bar{D} \leq \hat{r} + |z - x|$ and, since $|z - x| \leq |z - y| + |y - x| \leq \hat{r} + \bar{D}$, we have $\left|\bar{D} - |z - x|\right| \leq \hat{r}$. Next, obtain the same inequality with $z = x^{k+1}$ and $\hat{r} = (r_k^2 - \rho_\phi^k - \rho_S^k)^{1/2}$ to get a contradiction. □

Lemma 4.2 says that for each group there is a growing ball $B(x^{k(l)+1}, \bar{D} - r_k)$ such that if x^k enters this ball then $f_{\text{lev}}^k < f^*$. Hence Lemma 4.2 may be used at step 4(iii) to detect $f_{\text{lev}}^k < f^*$. Following [Dre83], one may argue that the conditions of Lemma 4.2 will be activated earlier than the usual condition $r_k^2 < \rho_\phi^k + \rho_S^k$. Indeed, r_k decreases from $\bar{D}$ to zero, whereas usually $|x^{k+1} - x^{k(l)+1}| \ll \bar{D}$, e.g., if $\bar{D}$ is a generous overestimate of diam(S).

5. Simple modifications. We shall now describe some simple modifications of Algorithm 2.2.

At step 5(iii) one may set $x^{k+1} = x_{\text{rec}}^k$; i.e., each group K_l may start from the best point found so far (if $g_f(x_{\text{rec}}^k)$ is stored). Alternatively, as in [KAC91], x^{k+1} could be chosen arbitrarily in S, but then step 1 would have to evaluate f and g_f at this point, leading to a slight deterioration in efficiency estimates such as Lemma 3.5 and Corollary 3.7 (where we would have $n_f^{k_\epsilon} = k_\epsilon$).

By suppressing null steps in our notation we may express the efficiency estimates in terms of the number of f-evaluations alone (as is customary in, e.g., [NeY79]).

THEOREM 5.1. *Suppose step* 5(iii) *sets* $f_{\text{low}}^k = \hat{f}_{\text{low}}^k$ *and* $\rho_k = 0$ *without increasing* k. *Then the total number of* f*-evaluations always equals* k, *and the efficiency estimate* (3.1) *holds.*

Proof. For contradiction, consider the *unmodified* algorithm. At step 0 set $n = 1$ and $\hat{z}^1 = x^1$. At step 6 set $\hat{z}^{n+1} = x^{k+1}$ and $\Delta_z^n = \hat{\Delta}^k$, and increase n by 1. Then at steps 2 and 6 we always have $n = k - l$ for the current values of k, l and n, and at step 2 $n_f^k = n$. Suppose $\Delta_z^{n_\epsilon} \geq \epsilon > 0$ for some $n_\epsilon = k_\epsilon - l^{k_\epsilon}$ at step 6. By Lemma 3.5 and (3.1b), we have $n_\epsilon \leq c_{\text{SPLA}}(t_{\min}, t_{\max}, \kappa)(\bar{D}L_f/\epsilon)^2$. Hence, for any $\epsilon > 0$, (3.1a) holds with k and $\{x^j\}_{j=1}^k$ replaced by n and $\{\hat{z}^j\}_{j=1}^n$, respectively. It remains to identify $\{\hat{z}^n\}$ with the sequence $\{x^k\}$ generated when step 5 does not increase k. □

We conclude, in particular, that by letting $\phi^k \equiv f^k$ at step 4, we obtain the simple SPLA of (2.2) that enjoys the efficiency estimate (3.1). One must, however, be cautious in interpreting such results, because Algorithm 2.2 could loop infinitely between steps 2 and 5.

COROLLARY 5.2. *Suppose* $\epsilon_{\text{opt}} > 0$ *and step* 5(iii) *sets* $f_{\text{low}}^k = \hat{f}_{\text{low}}^k$ *and* $\rho_k = 0$ *without increasing* k. *Then the algorithm will terminate with* $f(x_{\text{rec}}^k) \leq f^* + \epsilon_{\text{opt}}$ *after* $k = 1 + k_{\epsilon_{\text{opt}}}$ *iterations and* f*-evaluations, where*

$$k_{\epsilon_{\text{opt}}} \leq c_{\text{SPLA}}(t_{\min}, t_{\max}, \kappa)(\bar{D}L_f/\epsilon_{\text{opt}})^2$$

with c_{SPLA} *given by* (3.1b). *Moreover,* (3.1) *holds for any* $\epsilon > \epsilon_{\text{opt}}$.

Proof. Arguing by contradiction, use Theorem 3.8 to deduce that any loop between steps 2 and 5 must be finite when $\epsilon_{\text{opt}} > 0$, and then apply Theorem 5.1. □

As in [KAC91], let us consider setting $f^k_{\text{lev}} = f^k_{\text{up}} - \kappa_k \Delta^k$ at step 2, where $\kappa_k \in [\kappa_{\min}, \kappa_{\max}]$ for some fixed $0 < \kappa_{\min} \le \kappa_{\max} < 1$. We only require κ_k to produce $f^k_{\text{lev}} \le f^{k-1}_{\text{lev}}$ if $k > k(l) + 1$ (e.g., let $\kappa_k \in [\kappa_{k-1}, \kappa_{\max}]$ for such k). Then, as before, the level can increase only after the lower bound increases (i.e., (2.15) holds). Clearly, we must replace κ by $\kappa_{\max}$ in (2.16) and (2.17), and by $\kappa_{\min}$ in Lemmas 3.1 and 3.4. Similar replacements should be made in the remaining efficiency results. For instance, (3.10b) becomes

$$\hat{c}_{\text{RLA}}(t_{\min}, t_{\max}, \kappa_{\min}, \kappa_{\max}) = 1/t_{\min}(2 - t_{\max})\kappa^2_{\min}(1 - \kappa^2_{\max}) - 1/2e \ln(\kappa_{\max}),$$

where again the "best" $\kappa_{\min} = \kappa_{\max} \approx 0.677653$. We conclude that this modification cannot improve the preceding efficiency estimates. It may, however, be useful in practice to choose small κ_k at initial iterations in order to reduce the dependence on f^k_{low} until it is improved.

6. Using the known optimal value. Let us now consider the case when f^* is known.

THEOREM 6.1. *If $f^1_{\text{low}} = f^*$, then $l \equiv 0$ and the following efficiency estimate holds*:

$$\text{(6.1a)} \qquad k > c^*_{\text{RLA}}(t_{\min}, t_{\max}, \kappa)(\text{diam}(S) L_f/\epsilon)^2 \quad \Rightarrow \quad f^k_{\text{up}} - f^* \le \hat{\Delta}^k < \epsilon,$$

$$\text{(6.1b)} \qquad c^*_{\text{RLA}}(t_{\min}, t_{\max}, \kappa) = 1/t_{\min}(2 - t_{\max})\kappa^2,$$

$$\text{(6.1c)} \qquad \min c^*_{\text{RLA}}(\cdot, \cdot, \kappa) = c^*_{\text{RLA}}(1, 1, \kappa) = 1/\kappa^2.$$

Moreover, one may use $\kappa = 1$ and $f^k_{\text{lev}} \equiv f^*$, in which case (6.1) *reduces to* (1.4).

Proof. Use (2.1), (2.12), and Lemma 3.3 to deduce that step 5 cannot be entered (i.e., $l \equiv 0$) and $f^k_{\text{lev}} \ge f^k_{\text{low}} = f^*$ for all k. Next, invoke (3.5) in the proof of Lemma 3.4 in order to replace $\bar{D}$ by $\text{diam}(S)$ in Lemmas 3.4 and 3.5. Finally, observe that m and $(1 - \kappa^2)$ may be dropped from (3.8) to give (6.1), since $m = 0$ and $k_\epsilon \le c/\epsilon^2$ in part (ii) of the proof of Lemma 3.5, which remains valid even if $\kappa = 1$ because no summation is required. □

We conclude that if f^* is known then step 5 and the tests of step 4 may be omitted, so that $\bar{D}$ is not required. Moreover, setting $f^k_{\text{lev}} \equiv f^*$ ($\kappa = 1$ in (6.1)) gives the "best" efficiency estimate (1.4). In particular, (1.4) holds for for the simplest method of (1.2) (using $\phi^k \equiv f^k$ at step 4), as well as for Polyak's accelerated method from [Pol69] (with $\phi^k \equiv \hat{f}^k$; cf. (2.5)).

Remark 6.2. Note that, by Lemma 3.3, $f^k_{\text{low}} \equiv f^*$ ensures Fejér monotonicity $|x^* - x^{k+1}| \le |x^* - x^k|$ for all k and $x^* \in S^*$. Hence one easily checks that $\text{diam}(S)$ and L_f in (6.1) may be replaced by $D^* = |x^* - x^1|$ and $L^*_f = \sup\{|g_f(x)| : |x^* - x| \le D^*\}$ for any $x^* \in S^*$. Thus one may get an efficiency estimate even for *unbounded* S if S^* is nonempty! Also Fejér monotonicity and Theorem 3.8 imply that $\{x^k\}$ *converges* to an optimal point. (Let x^* be a cluster point of $\{x^k_{\text{rec}}\}$.) The question whether $\{x^k\}$ converges for other level controls is left open for future research.

The same argument also shows that if we chose $f^1_{\text{low}} > f^*$, then either termination would occur with $f^k_{\text{up}} \le f^1_{\text{low}} + \epsilon_{\text{opt}}$ or (6.1) would hold with f^* replaced by f^1_{low} (as if f were replaced by $\max\{f, f^1_{\text{low}}\}$).

7. Level control via frozen level gaps. In Algorithm 2.2 we have $f^k_{\rm up} - f^k_{\rm lev} = \kappa\Delta^k$; i.e., the desired objective reduction is a fraction of the current gap. An alternative technique consists in freezing the level gap $\Delta^k_{\rm lev} = f^k_{\rm up} - f^k_{\rm lev}$ at $\kappa\Delta^{k(l)}$ between iterations $k(l)$ and $k(l+1)$ that increase the lower bound.

Thus we modify Algorithm 2.2 as follows. Step 2 sets $f^k_{\rm lev} = f^k_{\rm up} - \Delta^k_{\rm lev}$, with $\Delta^1_{\rm lev} = \kappa\Delta^1$; step 5 sets $\Delta^{k+1}_{\rm lev} = \kappa\hat\Delta^k$, whereas step 6 sets $\Delta^{k+1}_{\rm lev} = \Delta^k_{\rm lev}$.

It is easy to check that the relations that ensure (2.14) continue to hold, whereas (2.15) follows from the fact that $f^{k+1}_{\rm up} \le f^k_{\rm up}$, while $\Delta^k_{\rm lev} = \kappa\hat\Delta^{k(l)}$ if $k(l) < k \le k(l+1)$ and $l \ge 0$, where $\hat\Delta^0 = \Delta^1$. Next, for $k = k(l+1)$ at step 5 we have $\hat f^k_{\rm low} \ge f^k_{\rm lev} = f^k_{\rm up} - \Delta^k_{\rm lev}$, so

$$\hat\Delta^{k(l+1)} \le \Delta^k_{\rm lev} = \kappa\hat\Delta^{k(l)} \quad \text{if} \quad k(l) < k \le k(l+1) \text{ and } l \ge 0 \tag{7.1}$$

and (2.17) follow by induction. Notice that the algorithm may also go to step 5 from step 2 if $f^k_{\rm lev} \le f^k_{\rm low}$. In other words: each group K_l ends by discovering that the target level is unattainable (and possibly that the lower bound may be increased). Then the level is raised by setting the level gap to a fraction of the "true" gap (between the bounds). The remaining level reductions within each group occur only when the objective improves, with the level gap and the lower bound staying fixed.

The efficiency analysis for the modified algorithm is similar to that for Algorithm 2.2, so we shall only indicate changes. Lemmas 3.2 and 3.3 remain valid. In Lemma 3.1 we may replace $\kappa\Delta^k$ by $\Delta^k_{\rm lev}$ (using $f(x^k) - f^k_{\rm lev} \ge f^k_{\rm up} - f^k_{\rm up} + \Delta^k_{\rm lev} = \Delta^k_{\rm lev}$), so (3.7) is replaced by

$$k - k(l) \le (\bar D L_f/\Delta^k_{\rm lev})^2/t_{\min}(2 - t_{\max}) \text{ if } k(l) < k < k(l+1) \text{ and } \Delta^k_{\rm lev} > 0. \tag{7.2}$$

In part (i) of the proof of Lemma 3.5 refer to (7.1) (instead of (2.16)) to get $\Delta^k_{\rm lev} \ge \epsilon/\kappa^{m-l-1}$ for all $k \in K_l \cap K(\epsilon)$ and $l = 0{:}m$, and use this relation and (7.2) in part (ii) to get the previous bounds. The remaining convergence results of §3 are easy to verify.

Another interesting modification is described in the following theorem.

THEOREM 7.1. *If we set $\epsilon_{\rm opt} = \epsilon$ and $\Delta^1_{\rm lev} = \epsilon$ for a given $\epsilon > 0$ (and possibly $f^1_{\rm low} = -\infty$), then the modified algorithm will terminate with $f^k_{\rm up} - f^* \le \epsilon$ and $l = 0$ at iteration $k = 1 + k_\epsilon$, where*

$$k_\epsilon \le (\bar D L_f/\epsilon)^2/t_{\min}(2 - t_{\max}). \tag{7.3}$$

Proof. If step 5 is not entered for $k = 1{:}k_\epsilon$ then (7.2) with $\Delta^k_{\rm lev} = \epsilon$ and $l = 0$ implies (7.3). Iteration $k = k_\epsilon + 1$ terminates at step 2, or at step 5 with $\hat\Delta^k \le \Delta^1_{\rm lev} = \epsilon$ (cf. (7.1)). □

A result essentially equivalent to the above theorem is given in [KuF90] for the simplest case of $\phi^k \equiv f^k$ at step 4. A comparison with all the preceding efficiency estimates (especially Corollary 5.2) suggests that, for a given accuracy $\epsilon_{\rm opt} > 0$, the strategy of Theorem 7.1 yields the best estimate. We believe, however, that in practice a "small" $\Delta^1_{\rm lev} = \epsilon_{\rm opt}$ might result in a slow "short-step" method, whose behavior would be close to the worst-case estimate even for "well-behaved" objectives. On the other hand, one may set $\Delta^1_{\rm lev}$ to an estimate of $f(x^1) - f^*$ (if any), so as to exploit any extra information at initial iterations; once a "reasonable" $f^k_{\rm low}$ is obtained then a switch to the original level strategy of Algorithm 2.2 may occur. (A similar idea is used in [LNN95].)

8. Level control via full model minimization. As in [LNN95], the *best underestimate* of f^* at iteration k is given by $\check{f}^k_{\min} = \min_S \check{f}^k$ with $\check{f}^k = \max_{j=1:k} f^j$. Let us, therefore, consider a version of Algorithm 2.2 in which step 2 sets $f^k_{\rm low} = \check{f}^k_{\min}$, step 4(i) chooses $\phi^k \le \check{f}^k$, and steps 4(iii) and 5 are deleted because $\bar{D}$ is no longer required for updating $f^k_{\rm low}$. (Note that $\mathcal{L}(\phi^k, f^k_{\rm lev}) \neq \emptyset$ at step 4(ii) due to $f^k_{\rm lev} > \check{f}^k_{\min} \ge \inf \phi^k$ by (2.1).)

Since $\check{f}^k \le \max\{\check{f}^k, f^{k+1}\} = \check{f}^{k+1} \le f$, we still have $f^{k+1}_{\rm low} \ge f^k_{\rm low}$, $\Delta^{k+1} \le \Delta^k$ and (2.14) for all k. Next,

$$\Delta^k < \kappa\Delta^j \quad \text{if} \quad \check{f}^k_{\min} > f^j_{\rm lev} \text{ and } j < k, \tag{8.1}$$

since then $\check{f}^k_{\min} > f^j_{\rm up} - \kappa\Delta^j \ge f^k_{\rm up} - \kappa\Delta^j$ by (2.1) with $\Delta^k = f^k_{\rm up} - \check{f}^k_{\min}$.

THEOREM 8.1. *The following efficiency estimate holds for each $\epsilon > 0$:*

$$k > c_{\rm LNN}(t_{\min}, t_{\max}, \kappa)(\operatorname{diam}(S)L_f/\epsilon)^2 \quad \Rightarrow \quad f^k_{\rm up} - f^* \le \Delta^k < \epsilon, \tag{8.2a}$$

$$c_{\rm LNN}(t_{\min}, t_{\max}, \kappa) = 1/t_{\min}(2 - t_{\max})\kappa^2(1 - \kappa^2), \tag{8.2b}$$

$$\min c_{\rm LNN}(\cdot,\cdot,\cdot) = c_{\rm LNN}(1, 1, 1/\sqrt{2}) = 4. \tag{8.2c}$$

Proof. (i) Suppose $\Delta^{k_\epsilon} \ge \epsilon > 0$ for some k_ϵ. Let us split $K(\epsilon) = \{1{:}\,k_\epsilon\}$ into groups $\check{K}_l$, $l = 1{:}\,m$ as follows. Let $\check{k}(1) = k_\epsilon$. For $l = 1, 2, \ldots$, set $\check{K}_l = \{k \le \check{k}(l) : \Delta^k \le \Delta^{\check{k}(l)}/\kappa\}$ and $\check{k}(l+1) = \min\{k : k \in \check{K}_l\} - 1$ until $\check{k}(l+1) = 0$, and then set $m = l$. By construction, $\Delta^k \ge \epsilon/\kappa^{l-1}$ for all $k \in \check{K}_l = \{\check{k}(l+1)+1{:}\,\check{k}(l)\}$ and $l = 1{:}\,m$.

(ii) Fix $1 \le l \le m$ and let $\check{y}^l \in \operatorname{Arg\,min}_S \check{f}^{\check{k}(l)}$. By (i) and (8.1), $\check{f}^{\check{k}(l)}_{\min} \le f^k_{\rm lev}$ for all $k \in \check{K}_l$. Hence, since $\check{f}^k$ are nondecreasing and $\phi^k \le \check{f}^k$, we have $\check{y}^l \in \mathcal{L}(\phi^k, f^k_{\rm lev})$ at step 4 for all $k \in \check{K}_l$. Therefore, $|\check{K}_l| \le (\operatorname{diam}(S)L_f/\kappa\Delta^{\check{k}(l)})^2/t_{\min}(2 - t_{\max})$ by Lemmas 3.1 and 3.2, as in the proof of Lemma 3.4, with $k_1 = \check{k}(l+1) + 1$, $k_2 = \check{k}(l)$ and $y = \check{y}^l \in S$.

(iii) Let $c = (\operatorname{diam}(S)L_f/\kappa)^2/t_{\min}(2 - t_{\max})$. By (i) and (ii),

$$k_\epsilon = \sum_{l=1}^{m} |\check{K}_l| \le \sum_{l=1}^{m} (c/\epsilon^2)\kappa^{2(l-1)} \le c/\epsilon^2(1 - \kappa^2). \qquad \square$$

Theorem 8.1 subsumes a result in [LNN95] obtained for $\phi^k \equiv \check{f}^k + \delta_S$ and $t_k \equiv 1$. Thus it shows that the good efficiency estimate for the level method of [LNN95] comes from level control, rather than from full projection subproblems.

It is easy to verify Theorem 3.8 and Corollary 5.2 (with $\bar{D} = \operatorname{diam}(S)$) for the modified method. Moreover, we may consider setting $f^k_{\rm lev} = f^k_{\rm up} - \kappa_k\Delta^k$, where $\kappa_k \in [\kappa_{\min}, \kappa_{\max}] \subset (0, 1)$. Then (8.2) involves $\hat{c}_{\rm LNN}(t_{\min}, t_{\max}, \kappa_{\min}, \kappa_{\max}) = 1/t_{\min}(2 - t_{\max})\kappa^2_{\min}(1 - \kappa^2_{\max})$, where again the "best" $\kappa_{\min} = \kappa_{\max} = 1/\sqrt{2}$. To check this, replace κ with $\kappa_{\max}$ in (8.1) and part (i) of the proof of Theorem 8.1 and with $\kappa_{\min}$ in part (ii).

Although it eliminates the need for $\bar{D}$, finding $\check{f}^k_{\min} = \min_S \max_{j=1:k} f^j$ may require too much storage and work per iteration when k is large. Let us therefore consider the following *partial model minimization strategy.* At step 2 find $\hat{f}^k_{\min} \le \min_S \hat{f}^k$ (cf. (2.5)) and set $f^k_{\min} = \max\{\hat{f}^k_{\min}, f^{k-1}_{\min}\}$ (with $f^0_{\min} = f^1_{\rm low}$). If $f^k_{\rm lev} \le f^k_{\min}$, go to step 5, choosing $\hat{f}^k_{\rm low} \ge f^k_{\min}$. Clearly, the efficiency results of the preceding sections remain true. Although this technique does not eliminate the dependence on $\bar{D}$ in theory, we believe that when $\hat{f}^k$ are chosen "rich enough," it will ensure better performance in practice. Also note that, to save work, $\min_S \hat{f}^k$ need not be found exactly.

9. Dual level methods. We shall now show how to use dual (ϵ-subgradient) techniques for constructing models of f that generalize those in [KuF90, LNN95]. In the simplest case such models are aggregate linearizations of f that are convex combinations of the ordinary linearizations f^j. We shall later relate them to surrogate inequalities used in relaxation methods. We start with an abstract framework that will cover several examples motivated by the the following representation:

$$\partial_\epsilon \hat{f}^k(x^k) = \left\{ \sum_{j\in J^k} \lambda_j g_f(x^j) : \lambda_j \geq 0, j \in J^k, \sum_{j\in J^k} \lambda_j = 1, \right. \tag{9.1}$$
$$\left. \sum_{j\in J^k} \lambda_j [\hat{f}^k(x^k) - f^j(x^k)] \leq \epsilon \right\}.$$

DEFINITION 9.1. *Let $\mu > 0$ be an additional fixed parameter associated with Algorithm 2.2. At step 4 let Φ^k_μ denote the set of all closed proper convex functions $\phi^k\colon \mathbb{R}^N \to (-\infty, \infty]$ that satisfy*

$$S^* \subset \mathcal{L}(\phi^k, f^k_{\text{lev}}) \quad \text{and} \quad d_{\mathcal{L}(\phi^k, f^k_{\text{lev}})}(x^k) \geq \mu\kappa\Delta^k/L_f \quad \text{if} \quad f^k_{\text{lev}} \geq f^*. \tag{9.2}$$

LEMMA 9.2. *Let $0 < \mu_\epsilon \leq 1$ and $\mu_g > 0$ be additional fixed parameters associated with Algorithm 2.2. At step 4 let $\Phi^k_{\mu_\epsilon,\mu_g}$ denote the set of all functions of the form $\phi^k(\cdot) = \hat{\phi}^k(x^k) - \epsilon_k + \langle p^k, \cdot - x^k\rangle$, where $\hat{\phi}^k\colon \mathbb{R}^N \to (-\infty, \infty]$ is closed, proper and convex, $\hat{\phi}^k \in \Phi$ if $f^k_{\text{lev}} \geq f^*$, $p^k \in \partial_{\epsilon_k}\hat{\phi}^k(x^k)$ satisfies $|p^k| \leq L_f/\mu_g$, and $\epsilon_k \in [0, \epsilon^k_{\max}]$ with*

$$\epsilon^k_{\max} = \hat{\phi}^k(x^k) - f^k_{\text{up}} + (1-\mu_\epsilon)(f^k_{\text{up}} - f^k_{\text{lev}}) = \hat{\phi}^k(x^k) - f^k_{\text{lev}} - \mu_\epsilon\kappa\Delta^k. \tag{9.3}$$

Suppose step 4(ii) uses such a ϕ^k (although it need not majorize f^k). If $f^k_{\text{lev}} \geq f^*$ then $S^* \subset \mathcal{L}(\phi^k, f^k_{\text{lev}})$, $y^k = P_{\mathcal{L}(\phi^k, f^k_{\text{lev}})}(x^k) = x^k - (\hat{\phi}^k(x^k) - \epsilon_k - f^k_{\text{lev}})p^k/|p^k|^2$ and

$$\begin{aligned} \rho^k_\phi/t_{\min}(2 - t_{\max}) &\geq d^2_{\mathcal{L}(\phi^k, f^k_{\text{lev}})}(x^k) = (\hat{\phi}^k(x^k) - \epsilon_k - f^k_{\text{lev}})^2/|p^k|^2 \\ &\geq (\mu_\epsilon\mu_g\kappa\Delta^k/L_f)^2, \end{aligned} \tag{9.4}$$

whereas if $\mathcal{L}(\phi^k, f^k_{\text{lev}}) = \emptyset$, then $f^k_{\text{lev}} < f^*$. Moreover, $\Phi^k_{\mu_\epsilon,\mu_g} \subset \Phi^k_\mu$ if $\mu = \mu_\epsilon\mu_g$.

Proof. Suppose $f^k_{\text{lev}} \geq f^*$. Let $x \in S^*$. By construction and (2.10), $\phi^k(x) \leq \hat{\phi}^k(x) \leq f^*$ yield $S^* \subset \mathcal{L}(\phi^k, f^k_{\text{lev}}) \neq \emptyset$. By (9.3), $p^k = 0$ would imply $f^* \geq \phi^k(\cdot) = \hat{\phi}^k(x^k) - \epsilon_k \geq f^k_{\text{lev}} + \mu_\epsilon\kappa\Delta^k > f^k_{\text{lev}}$, a contradiction. Thus $y^k - x^k = -(\phi^k(x^k) - f^k_{\text{lev}})p^k/|p^k|^2$. Since $\phi^k(x^k) - f^k_{\text{lev}} = \hat{\phi}^k(x^k) - \epsilon_k - f^k_{\text{lev}} \geq \mu_\epsilon\kappa\Delta^k$ from (9.3) and $|p^k| \leq L_f/\mu_g$ by the choice of p^k, we have (9.4), which yields (9.2) if $\mu = \mu_\epsilon\mu_g$. □

Let us consider efficiency before examples. The preceding proofs hinge on $\phi^k \in \Phi^k_1$ *only* (cf. (2.15) and the proof of Lemma 3.3), so $\phi^k \notin \Phi$ is admissible if $f^k_{\text{lev}} < f^*$. (This is used in [Kiw96, §6].) Hence step 4 may use any $\phi^k \in \Phi^k_{\mu_\epsilon,\mu_g} \cup \Phi^k_\mu$ with $\mu = \mu_\epsilon\mu_g > 0$. Then, comparing (9.2) and (9.4) with Lemma 3.1, we see that *only the first terms* of the constants in all the preceding efficiency estimates and the right side of (7.3) need be *divided by* μ^2; of course, Δ^k_{lev} replaces $\kappa\Delta^k$ in (9.2), (9.3) and (9.4) for the frozen level gaps of §7.

As in §8, suppose $f^k_{\text{low}} \equiv \check{f}^k_{\min}$ at step 2, step 5 is deleted, and step 4 chooses $\hat{\phi}^k \le \check{f}^k$ and ϕ^k as in Lemma 9.2. Then Theorem 8.1 holds with c_{LNN} divided by $\mu^2 = \mu_\epsilon^2$, since $\mu_g = 1$. (In part (ii) of its proof use (9.4) to replace κ by $\mu\kappa$.)

It should be clear that, as in §§5 and 8, we may use variable $\kappa_k \in [\kappa_{\min}, \kappa_{\max}]$ and $\mu_{\epsilon,k} \in [\mu_{\min}, 1] \subset (0,1]$ in (9.3). Then the efficiency constants are divided by $(\mu_{\min}\mu_g)^2$.

For choosing $\hat{\phi}^k$ in Lemma 9.2, note that any $\hat{\phi}^k \in \Phi$ such that $f^k \le \hat{\phi}^k \le \check{f}^k$ is admissible. Indeed, $f^k(x^k) = \check{f}^k(x^k) = f(x^k)$ yield $\epsilon^k_{\max} \ge 0$ in (9.3), and we may always ensure $|p^k| \le L_f$ because $g_f(x^k) \in \partial\hat{\phi}^k(x^k)$ has $|g_f(x^k)| \le L_f$. In view of Remark 3.9, to enlarge ρ^k_ϕ in (9.4), we may let

$$p^k = \arg\min\{\, |p|^2/2 : p \in \partial_{\epsilon_k}\hat{\phi}^k(x^k)\,\}. \tag{9.5}$$

For example, if we use $\hat{\phi}^k = \hat{f}^k$ and (9.1), then p^k may by found by quadratic programming (QP). Since there is no need to solve (9.5) exactly, iterative QP methods (e.g., parallel relaxation-type methods) and various heuristics may by employed to save work. Note that (9.5) minimizes the denominator in (9.4) with a fixed numerator. An alternative construction is described in the following lemma (in which one may assume $\hat{\phi}^k = \hat{f}^k$ for the first reading). It implies that we may use convex combinations of linearizations with quite arbitrary weights without destroying the preceding efficiency estimates. Possible advantages of such combinations are discussed later.

LEMMA 9.3. *Let $\hat{\phi}^k = \max_{i\in I^k}\phi^k_i$, where $|I^k| < \infty$, each ϕ^k_i is an affine function of the form $\phi^k_i(x) = \phi^k_i(x^k) + \langle p^k_{\phi_i}, x - x^k\rangle$ with $|p^k_{\phi_i}| \le L_f/\mu_g$, and $\phi^k_i \in \Phi$ if $f^k_{\text{lev}} \ge f^*$. Suppose $\hat{\phi}^k(x^k) > f^k_{\text{lev}}$, $\hat{I}^k \subset \{i \in I^k : \phi^k_i(x^k) \ge f^k_{\text{lev}}\}$, $\bar{I}^k = \{i \in \hat{I}^k : \phi^k_i(x^k) = \hat{\phi}^k(x^k)\} \ne \emptyset$, $\lambda_i > 0$ for $i \in \hat{I}^k$, $\lambda_i = 0$ for $i \in I^k \setminus \hat{I}^k$, $\sum_{i\in\hat{I}^k}\lambda_i = 1$, $(p^k, \epsilon_k) = \sum_{i\in\hat{I}^k}\lambda_i(p^k_{\phi_i}, \hat{\phi}^k(x^k) - \phi^k_i(x^k))$, and $\phi^k(\cdot) = \hat{\phi}^k(x^k) - \epsilon_k + \langle p^k, \cdot - x^k\rangle$. If $f^k_{\text{lev}} \ge f^*$, then $d_{\mathcal{L}(\phi^k, f^k_{\text{lev}})}(x^k) \ge \sum_{i\in\bar{I}^k}\lambda_i(\hat{\phi}^k(x^k) - f^k_{\text{lev}})\mu_g/L_f$, whereas if $p^k = 0$, then $f^k_{\text{lev}} < f^*$. In particular, $\phi^k \in \Phi^k_{\mu_\epsilon,\mu_g}$ if $\lambda_j \ge \mu_\epsilon\kappa\Delta^k/(\hat{\phi}^k(x^k) - f^k_{\text{lev}})$ for some $j \in \bar{I}^k$, e.g., if $\lambda_j \ge \mu_\epsilon$ and $\hat{\phi}^k(x^k) \ge f^k_{\text{up}}$.*

Proof. Suppose $f^k_{\text{lev}} \ge f^*$, $x \in S^*$, and $j \in \bar{I}^k$. By construction, $\hat{\phi}^k(x^k) - \epsilon_k - f^k_{\text{lev}} = \sum_{i\in\hat{I}^k}\lambda_i(\phi^k_i(x^k) - f^k_{\text{lev}}) \ge \sum_{i\in\bar{I}^k}\lambda_i(\hat{\phi}^k(x^k) - f^k_{\text{lev}}) > 0$. Hence with $\phi^k(x^k) = \hat{\phi}^k(x^k) - \epsilon_k$, $p^k = 0$ would give $f^* \ge \sum_{i\in\hat{I}^k}\lambda_i\phi^k_i(x) = \phi^k(x) = \phi^k(x^k) > f^k_{\text{lev}}$, a contradiction. Thus $d_{\mathcal{L}(\phi^k, f^k_{\text{lev}})}(x^k) = (\phi^k(x^k) - f^k_{\text{lev}})/|p^k| \ge \sum_{i\in\bar{I}^k}\lambda_i(\hat{\phi}^k(x^k) - f^k_{\text{lev}})/|p^k|$, where $|p^k| \le \sum_{i\in\hat{I}^k}\lambda_i|p^k_{\phi_i}| \le L_f/\mu_g$. Recall Lemma 9.2 and (2.1) to complete the proof. □

Supposing $f^k_{\text{lev}} \ge f^*$, let us now compare the dual approach based on (9.4) (and possibly (9.5)) using $y^k = P_{\mathcal{L}(\phi^k, f^k_{\text{lev}})}(x^k)$ with a primal one that employs $\hat{\phi}^k$ directly to find

$$\hat{y}^k = P_{\mathcal{L}(\hat{\phi}^k, f^k_{\text{lev}})}(x^k) = \arg\min\{\, |x - x^k|^2/2 : \hat{\phi}^k(x) \le f^k_{\text{lev}}\,\}. \tag{9.6}$$

LEMMA 9.4. *We have $|y - \hat{y}^k|^2 \le |y - x^k|^2 - |\hat{y}^k - x^k|^2$ and $|y - y^k|^2 \le |y - x^k|^2 - |y^k - x^k|^2$ for all $y \in \mathcal{L}(\hat{\phi}^k, f^k_{\text{lev}})$, where $|y^k - x^k|^2 \le |\hat{y}^k - x^k|^2 - |\hat{y}^k - y^k|^2$. Moreover,*

$$\begin{aligned}\sup\{\, (\hat{\phi}^k(x^k) - \epsilon - f^k_{\text{lev}})/|p| : \epsilon \in [0, \hat{\phi}^k(x^k) - f^k_{\text{lev}}),\ p \in \partial_\epsilon\hat{\phi}^k(x^k)\,\} \\ = |\hat{y}^k - x^k|,\end{aligned} \tag{9.7}$$

with the supremum attained at some $\hat{\epsilon}_k$ *and* $\hat{p}^k$ *if* $\hat{\phi}^k$ *is polyhedral or* $\inf \hat{\phi}^k < f^k_{\text{lev}}$.

Proof. The first assertion follows from (2.8) and $\phi^k \le \hat{\phi}^k$. Hence, recalling (9.4), $|\hat{y}^k - x^k|$ majorizes the left side of (9.7). To establish equality, suppose initially that $\hat{\phi}^k$ is polyhedral or $\inf \hat{\phi}^k < f^k_{\text{lev}}$. Then, by the Karush–Kuhn–Tucker conditions for (9.6), there exist $\hat{p}^k \in \partial\hat{\phi}^k(\hat{y}^k)$ and a multiplier $\hat{\lambda} \ge 0$ such that $\hat{y}^k - x^k = -\hat{\lambda}\hat{p}^k$. Clearly, $\hat{\phi}^k(\hat{y}^k) = f^k_{\text{lev}}$ and $\hat{\lambda} > 0$ because $x^k \notin \mathcal{L}(\phi^k, f^k_{\text{lev}}) \supset \mathcal{L}(\hat{\phi}^k, f^k_{\text{lev}})$ $(\phi^k \le \hat{\phi}^k)$. Letting $\tilde{\phi}^k(\cdot) = \hat{\phi}^k(\hat{y}^k) + \langle \hat{p}^k, \cdot - \hat{y}^k \rangle$ and $\hat{\epsilon}_k = \hat{\phi}^k(x^k) - \tilde{\phi}^k(x^k) = \hat{\phi}^k(x^k) - f^k_{\text{lev}} - \hat{\lambda}|\hat{p}^k|^2$, we get (9.7). In the general case, replace f^k_{lev} in (9.6) by $t > f^k_{\text{lev}}$ so that the Slater condition holds, define $\hat{y}(t)$, $\hat{p}(t)$, and $\hat{\epsilon}(t)$, as above, and let $t \downarrow f^k_{\text{lev}}$ with $\hat{y}(t) \to \hat{y}^k$. □

In view of Remark 3.9, the first bound of Lemma 9.4 is, in general, better than the second one. On the other hand, (9.7) says that the dual approach can, in principle, be as good as the primal one if (9.5) is used with a carefully chosen ϵ_k. Thus such bounds seem to favor the primal approach. However, they are local and the dual one may employ inexact QP solvers, so it may be easier to implement.

Choosing $(\epsilon_k, p^k) = (\hat{\epsilon}_k, \hat{p}^k)$ to solve (9.7) gives an *"optimal" dual method* that does not need $\mu_\epsilon > 0$ in (9.3) if $\hat{\phi}^k \ge f^k$ and $\hat{\phi}^k \in \Phi$. It is, however, more difficult to implement than the equivalent primal method that may solve (9.6) via QP when $\hat{\phi}^k$ is polyhedral.

We may add that [LNN95] employs $\hat{\phi}^k = \check{f}^k$, $t_k = 1$, $f^k_{\text{lev}} = \check{f}^k_{\min}$ and constant κ, $\mu_\epsilon \in (0,1)$ and $\mu_g = 1$, whereas [KuF90] proceeds as in Theorem 6.1 with $\mu_\epsilon = 1$ and $\hat{\phi}^k = f$ without specifying any models of f (but $\mu_\epsilon = 1$ may severely restrict the choice of p^k; cf. (9.3)).

Acknowledgments. This work was started during my six-month stay at INRIA, Rocquencourt, in 1992, owing to the kind invitation by C. Lemaréchal and the French Ministry for Research and Technology, whose financial support is gratefully acknowledged. I would also like to thank the two anonymous referees for their valuable comments.

REFERENCES

[Agm54] S. Agmon, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 382–392.

[BaS81] M. S. Bazaraa and H. D. Sherali, *On the choice of step size in subgradient optimization*, European J. Oper. Res., 7 (1981), pp. 380–388.

[Dre83] Z. Drezner, *The nested ball principle for the relaxation method*, Oper. Res., 31 (1983), pp. 587–590.

[Gof81] J.-L. Goffin, *Convergence results in a class of variable metric subgradient methods*, in Nonlinear Programming 4, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1981, pp. 283–326.

[GPR67] L. G. Gurin, B. T. Polyak, and E. V. Raik, *The method of projections for finding a common point of convex sets*, Zh. Vychisl. Mat. i Mat. Fiz., 7 (1967), pp. 1211–1228. (In Russian.) English transl. in U.S.S.R. Comput. Math. and Math. Phys. 7 (1967), pp. 1–24.

[HUL93] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

[KAC91] S. Kim, H. Ahn, and S.-C. Cho, *Variable target value subgradient method*, Math. Programming, 49 (1991), pp. 359–369.

[Kiw85] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Mathematics 1133, Springer-Verlag, Berlin, 1985.

[Kiw95] K. C. KIWIEL, *Proximal level bundle methods for convex nondifferentiable optimization, saddle-point problems and variational inequalities*, Math. Programming, 69 (1995), pp. 89–109.

[Kiw96] ———, *The efficiency of subgradient projection methods for convex optimization, part* II: *Implementations and extensions*, SIAM J. Control Optim., 34 (1996), pp. 677–697.

[KuF90] A. N. KULIKOV AND V. R. FAZILOV, *Convex optimization with prescribed accuracy*, Zh. Vychisl. Mat. i Mat. Fiz., 30 (1990), pp. 663–671. (In Russian.)

[Lem89] C. LEMARÉCHAL, *Nondifferentiable optimization*, in Optimization, Handbooks in Operations Research and Management Science, vol. 1, G. L. Nemhauser, A. H. G. Rinnooy-Kan, and M. J. Todd, eds., North-Holland, Amsterdam, 1989, pp. 529–572.

[LNN91] C. LEMARÉCHAL, A. S. NEMIROVSKII, AND YU. E. NESTEROV, *New variants of bundle methods*, Research report 1508, INRIA, Rocquencourt, 1991.

[LNN95] ———, *New variants of bundle methods*, Math. Programming, 69 (1995), pp. 111-147.

[MoS54] T. MOTZKIN AND I. J. SCHOENBERG, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 393–404.

[MTA81] J. F. MAURRAS, K. TRUMPER, AND M. AKGÜL, *Polynomial algorithms for a class of linear programs*, Math. Programming, 21 (1981), pp. 121–136.

[NeY79] A. S. NEMIROVSKII AND D. B. YUDIN, *Problem Complexity and Method Efficiency in Optimization*, Nauka, Moscow, 1979. (In Russian.) English translation Wiley, New York, 1983.

[Pol69] B. T. POLYAK, *Minimization of unsmooth functionals*, Zh. Vychisl. Mat. i Mat. Fiz., 9 (1969), pp. 509–521. (In Russian.) English transl. in U.S.S.R. Comput. Math. and Math. Phys. 9 (1969), pp. 14–29.

[Tel82] J. TELGEN, *On relaxation methods for systems of linear inequalities*, European J. Oper. Res., 9 (1982), pp. 184–189.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 677–697, March 1996

015

THE EFFICIENCY OF SUBGRADIENT PROJECTION METHODS FOR CONVEX OPTIMIZATION, PART II: IMPLEMENTATIONS AND EXTENSIONS*

KRZYSZTOF C. KIWIEL†

Abstract. In the first part of this paper we studied subgradient methods for convex optimization that use projections onto successive approximations of level sets of the objective corresponding to estimates of the optimal value. We presented several variants and showed that they enjoy almost optimal efficiency estimates. In this part of the paper we discuss possible implementations of such methods. In particular, their projection subproblems may be solved inexactly via relaxation methods, thus opening the way for parallel implementations. We discuss accelerations of relaxation methods based on simultaneous projections, surrogate constraints, and conjugate and projected (conditional) subgradient techniques.

Key words. nondifferentiable (nonsmooth) optimization, convex programming, relaxation methods, subgradient optimization, successive projections, linear inequalities, parallel computing

AMS subject classifications. 65K05, 90C25

1. Introduction.

This is the second of two papers in which we study variants of Polyak's [Pol69] subgradient projection algorithm (SPA) and the level method of [LNN95] for solving the convex program

$$f^* = \min\{ f(x) : x \in S \} \tag{1.1}$$

under the following assumptions. S is a nonempty compact convex subset of $\mathbb{R}^N$; f is a convex function Lipschitz continuous on S with Lipschitz constant L_f; for each $x \in S$ we can compute $f(x)$ and a subgradient $g_f(x) \in \partial f(x)$ of f at x such that $|g_f(x)| \leq L_f$; and for each $x \in \mathbb{R}^N$ we can find $P_S(x) = \arg\min\{|x - y| : y \in S\}$, its orthogonal projection on S, where $|\cdot|$ denotes the Euclidean norm. The SPA generates successive iterates

$$x^{k+1} = P_S(x^k - t_k(f(x^k) - f^*)g_f(x^k)/|g_f(x^k)|^2) \quad \text{for } k = 1, 2, \ldots, \tag{1.2}$$

where $x^1 \in S$ and t_k are stepsizes in $T = [t_{\min}, t_{\max}]$ with $0 < t_{\min} \leq t_{\max} < 2$.

In part I [Kiw96] we gave efficiency estimates for three schemes for estimating f^* in (1.2), stemming from [KAC91, KuF90, LNN95]. Moreover, to enable faster convergence, we studied algorithms that use projections onto successive approximations of level sets of f derived from *several accumulated subgradient linearizations* of f or their *aggregates* as in descent bundle methods for nondifferentiable optimization (NDO) [HUL93, Kiw85, Lem89].

In this paper we discuss possible implementations of such algorithms that provide freedom to trade off storage requirements and work per iteration for speed of convergence. Their projection subproblems can be solved efficiently even in the large-scale case by a variety of methods, especially those that can benefit from parallel computation; see, e.g., [AhC89, BeT89, IDP91, Kiw95, LoH88, Oko92, Spi87, Tse90a, Tse90b, YaM92] and the references therein. The ability to use *inexact* projections makes such

* Received by the editors January 12, 1994; accepted for publication (in revised form) December 12, 1994. This research was supported by Polish State Committee for Scientific Research grant 8S50502206.

† Systems Research Institute, Newelska 6, 01–447 Warsaw, Poland (kiwiel@ibspan.waw.pl).

algorithms very attractive in large-scale applications. In contrast, the existing bundle methods (cf. [Kiw89, ScZ92]) employ nonstrictly convex quadratic programming (QP) subproblems, and it is not clear how to solve such QP subproblems exactly via parallel computation.

It is fruitful to view subgradient methods as extensions of relaxation methods for linear inequalities [Agm54, Gof78, MoS54]. Motivated by possible applications to the subgradient methods of this paper, we have introduced in [Kiw95] several parallelizable relaxation methods. Here we provide a unified perspective on acceleration techniques for such methods, including simultaneous projections [Tod79], surrogate cuts [BGT81, GoT82, Oko92], surrogate constraints [YaM92], conjugate subgradients (CSs) [CFM75, KKA87, Shc92, ShU89, Sho79], and projected (conditional) subgradients [BaG79, KiU89]. In contrast to their usual interpretations, we show that such methods hinge on *implicitly* generated affine (or polyhedral) models of f. *Explicit* use of such models allows various modifications and extensions that seem more efficient. It turns out that some of these methods are simplified versions of others that trade speed of convergence for ease of implementation. Further, our framework shows how to modify their models to account for the constraint $x \in S$. For instance, it suggests the following simple modification of (1.2) (proposed independently in [KiU93, Kiw93]):

$$x^{k+1} = \arg\min\{\, |x - x^k|^2/2 : f(x^k) + \langle g_f(x^k), x - x^k \rangle \le f^*, x \in S \,\}, \tag{1.3}$$

which seems to be more efficient in general [KiU93].

The paper is organized as follows. In §2 we extend the level algorithm of [Kiw96] and give conditions that allow efficient implementations via general relaxation and QP methods discussed in §§3 and 4, as well as "cheap" surrogate projection methods developed in §5. Extensions of conjugate subgradient implementations are given in §6. In §7 we argue that subgradient relaxation should also include inequalities related to S. Finally, we have a concluding section.

We use the following notation. We denote by $\langle \cdot, \cdot \rangle$ the usual inner product in $\mathbb{R}^N$. For $\epsilon \ge 0$, the ϵ-subdifferential of f at x is defined by $\partial_\epsilon f(x) = \{p \in \mathbb{R}^N : f(y) \ge f(x) + \langle p, y - x \rangle - \epsilon \quad \forall y \in \mathbb{R}^N\}$. $\mathcal{L}(f, \alpha) = \{x : f(x) \le \alpha\}$ is the α-level set of f, and $\operatorname{diam}(S) = \sup_{x,y \in S} |x - y|$ is the diameter of S. Given a closed convex set $C \subset \mathbb{R}^N$ and a stepsize $t \in T$, the *relaxation operator* $\mathcal{R}_{C,t}(x) = x + t(P_C(x) - x)$ has the Fejér contraction property [Agm54, Kiw96]

$$|y - \mathcal{R}_{C,t}(x)|^2 \le |y - x|^2 - t(2 - t)d_C^2(x) \quad \forall y \in C, x \in \mathbb{R}^N, \tag{1.4}$$

where $d_C(x) = |x - P_C(x)|$. We let $1{:}k$ denote $1, 2, \ldots, k$. For brevity, we let $a/bc = a/(bc)$. The convex hull is denoted by co.

2. The generalized relaxation level algorithm. We start by presenting an extension of Algorithm 2.2 from [Kiw96].

ALGORITHM 2.1.

Step 0 (Initialization). Select an initial point $x^1 \in S$, a final optimality tolerance $\epsilon_{\text{opt}} \ge 0$, a level parameter $0 < \kappa < 1$, and stepsize parameters $0 < t_{\min} \le t_{\max} < 2$. Choose $\bar{D} \ge \operatorname{diam}(S)$ and $f_{\text{low}}^1 \le f^*$. Set $\rho_1 = 0$ and $f_{\text{up}}^0 = \infty$. Set the counters $k = 1$, $l = 0$, and $k(0) = 0$. ($k(l)$ will denote the iteration number of the lth increase of f_{low}^k.)

Step 1 (Objective evaluation). Calculate $f(x^k)$ and $g_f(x^k)$.

Step 2 (Level update). Using the current *lower bound* f_{low}^k on f^*, update the *upper bound* $f_{\text{up}}^k = \min\{f(x^k), f_{\text{up}}^{k-1}\}$, the *gap* $\Delta^k = f_{\text{up}}^k - f_{\text{low}}^k$, and the *target* (*level*) $f_{\text{lev}}^k = f_{\text{up}}^k - \kappa\Delta^k$.

Step 3 (Stopping criterion). If $\min\{\Delta^k, |g_f(x^k)|/\bar{D}\} \le \epsilon_{\text{opt}}$, terminate.

Step 4 (Relaxations). Let $f^k(\cdot) = f(x^k) + \langle g_f(x^k), \cdot - x^k\rangle$.

(i) Find $z^k \in \mathbb{R}^N$ and $\rho_\phi^k \ge 0$ such that $\rho_\phi^k \ge t_{\min}(2 - t_{\max})d^2_{\mathcal{L}(f^k, f_{\text{lev}}^k)}(x^k)$ and

$$|y - z^k|^2 \le |y - x^k|^2 - \rho_\phi^k \quad \forall y \in S^* \quad \text{if} \quad f_{\text{lev}}^k \ge f^*; \tag{2.1}$$

if $f_{\text{lev}}^k < f^*$, then z^k and ρ_ϕ^k are arbitrary (even $\rho_\phi^k = \infty$ is admissible). If $\rho_k + \rho_\phi^k > \bar{D}^2$ or it is discovered by another test that $f_{\text{lev}}^k < f^*$, go to step 5.

(ii) Find $x^{k+1} \in S$ and $\rho_S^k \ge 0$ such that $|y - x^{k+1}|^2 \le |y - z^k|^2 - \rho_S^k$ for all $y \in S$. If $\rho_k + \rho_\phi^k + \rho_S^k > \bar{D}^2$, go to step 5; otherwise, go to step 6.

Step 5 (Update lower bound).

(i) Choose a lower bound $\hat{f}_{\text{low}}^k \in [\max\{f_{\text{low}}^k, f_{\text{lev}}^k\}, f^*]$ (e.g., $\hat{f}_{\text{low}}^k = \max\{f_{\text{low}}^k, f_{\text{lev}}^k\}$). Set $f_{\text{low}}^{k+1} = \hat{f}_{\text{low}}^k$, $\rho_{k+1} = 0$ and $\hat{\Delta}^k = f_{\text{up}}^k - \hat{f}_{\text{low}}^k$.

(ii) If $\hat{\Delta}^k \le \epsilon_{\text{opt}}$, terminate; otherwise, continue.

(iii) Set $x^{k+1} = x^k$ (*null step*), $k(l+1) = k$, and increase k and l by 1. Go to step 2.

Step 6 (Serious step). Set $f_{\text{low}}^{k+1} = f_{\text{low}}^k$, $\hat{f}_{\text{low}}^k = f_{\text{low}}^k$, $\hat{\Delta}^k = \Delta^k$ and $\rho_{k+1} = \rho_k + \rho_\phi^k + \rho_S^k$. Increase k by 1 and go to step 1.

We need only discuss possible implementations of step 4. First, as in [Kiw96], we may choose a stepsize $t_k \in T$ and an *admissible model* ϕ^k of f in the set

$$\Phi = \{\, \phi\colon \mathbb{R}^N \to (-\infty, \infty] : \phi \text{ is closed convex and } \phi(x) \le f^* \ \forall x \in S^* \,\}$$

such that $\phi^k \ge f^k$, discover that $f_{\text{lev}}^k < f^*$ if $\mathcal{L}(\phi^k, f_{\text{lev}}^k) = \emptyset$, and otherwise set $y^k = P_{\mathcal{L}(\phi^k, f_{\text{lev}}^k)}(x^k)$, $z^k = x^k + t_k(y^k - x^k)$, $x^{k+1} = P_S(z^k)$, $\rho_\phi^k = t_k(2 - t_k)|y^k - x^k|^2$, and $\rho_S^k = |x^{k+1} - z^k|^2$. (Then $x^{k+1} = P_S(\mathcal{R}_{\mathcal{L}(\phi^k, f_{\text{lev}}^k), t_k}(x^k))$, and ρ_ϕ^k and ρ_S^k stem from Fejér estimates; cf. (1.4).) For instance, we may let ϕ^k be the kth polyhedral model of f

$$\hat{f}^k(x) = \max\{\, f^j(x) : j \in J^k \,\} \quad \text{with} \quad k \in J^k \subset \{1\colon k\};$$

e.g., $J^k \equiv \{k\}$ yields the subgradient projection level algorithm (SPLA):

$$x^{k+1} = P_S(x^k - t_k(f(x^k) - f_{\text{lev}}^k)g_f(x^k)/|g_f(x^k)|^2) \quad \text{for } k = 1, 2, \ldots. \tag{2.2}$$

Alternatively, for the *dual level methods* of [Kiw96, §9], we may choose $\phi^k \in \Phi_\mu^k \cup \Phi_{\mu_\epsilon, \mu_g}^k$ and replace f^k with ϕ^k at step 4(i), where $\mu > 0$, $0 < \mu_\epsilon \le 1$, and $\mu_g > 0$ are additional fixed parameters, and Φ_μ^k and $\Phi_{\mu_\epsilon, \mu_g}^k$ are defined as follows. Φ_μ^k denotes the set of all closed proper convex functions $\phi^k\colon \mathbb{R}^N \to (-\infty, \infty]$ that satisfy

$$S^* \subset \mathcal{L}(\phi^k, f_{\text{lev}}^k) \quad \text{and} \quad d_{\mathcal{L}(\phi^k, f_{\text{lev}}^k)}(x^k) \ge \mu\kappa\Delta^k/L_f \quad \text{if} \quad f_{\text{lev}}^k \ge f^*. \tag{2.3}$$

$\Phi_{\mu_\epsilon, \mu_g}^k$ denotes the set of all functions of the form $\phi^k(\cdot) = \hat{\phi}^k(x^k) - \epsilon_k + \langle p^k, \cdot - x^k\rangle$, where $\hat{\phi}^k\colon \mathbb{R}^N \to (-\infty, \infty]$ is closed, proper and convex, $\hat{\phi}^k \in \Phi$ if $f_{\text{lev}}^k \ge f^*$, $p^k \in$

$\partial_{\epsilon_k}\hat{\phi}^k(x^k)$ satisfies $|p^k| \le L_f/\mu_g$, and $\epsilon_k \in [0, \epsilon^k_{\max}]$ with $\epsilon^k_{\max} = \hat{\phi}^k(x^k) - f^k_{\text{lev}} - \mu_\epsilon \kappa \Delta^k$. In particular, we may choose $\hat{\phi}^k = \hat{f}^k$, $\epsilon_k \in [0, \epsilon^k_{\max}]$ and

$$p^k = \arg\min\{\, |p|^2/2 : p \in \partial_{\epsilon_k}\hat{\phi}^k(x^k)\,\}, \tag{2.4}$$

using

$$\partial_\epsilon \hat{f}^k(x^k) = \left\{ \sum_{j\in J^k} \lambda_j g_f(x^j) : \lambda_j \ge 0, j \in J^k, \sum_{j\in J^k} \lambda_j = 1, \right. \tag{2.5}$$
$$\left. \sum_{j\in J^k} \lambda_j [\hat{f}^k(x^k) - f^j(x^k)] \le \epsilon \right\}.$$

Remark 2.2. For the methods from [Kiw96, §8], which use $f^k_{\text{lev}} \equiv \min_S \check{f}^k$ with $\check{f}^k = \max_{j=1:k} f^j$, S^* in (2.1) should be replaced by $\{y \in S : \check{f}^k(y) \le f^k_{\text{lev}}\}$. It is easy to verify all the efficiency results of [Kiw96] for such modifications. (Hint: let $y \in S^*$ in [Kiw96, Lem. 3.2], with S^* replaced by $\mathcal{L}(\check{f}^k, f^k_{\text{lev}})$ in [Kiw96, §8].)

3. Using general relaxation methods. We now show how to implement step 4 via general relaxation methods for linear inequalities; see, e.g., [AhC89, Kiw95] and the references therein.

Suppose ϕ^k is polyhedral, so that $\mathcal{L}(\phi^k, f^k_{\text{lev}})$ has the form $\{x : \langle a^i, x\rangle \le b_i, i \in I\}$. Let $C_i = \{x : \langle a^i, x\rangle \le b_i\}$, $i \in I$. Given a starting point $\tilde{x}^1 \notin \cap_{i\in I} C_i$, many relaxation methods attempt to find a point in $\cap_i C_i$ via the iteration $\tilde{x}^{n+1} = \sum_{i\in I} \tilde{\lambda}^n_i \mathcal{R}_{C_i,\tilde{t}_n}(\tilde{x}^n)$, $n = 1, 2, \ldots$, where the weights $\tilde{\lambda}^n_i \ge 0$, $i \in I$, satisfy $\sum_i \tilde{\lambda}^n_i = 1$, and $0 < \tilde{t}_n \le 2$. By (1.4),

$$|y - \mathcal{R}_{C_i,\tilde{t}_n}(\tilde{x}^n)|^2 \le |y - \tilde{x}^n|^2 - \tilde{t}_n(2 - \tilde{t}_n) d^2_{C_i}(\tilde{x}^n) \quad \forall y \in C_i;$$

multiply this by $\tilde{\lambda}^n_i$, sum over i, and use $\sum_i \tilde{\lambda}^n_i = 1$ and the convexity of $|\cdot|^2$ to get

$$\left| y - \sum_i \tilde{\lambda}^n_i \mathcal{R}_{C_i,\tilde{t}_n}(\tilde{x}^n) \right|^2 \le |y - \tilde{x}^n|^2 - \tilde{t}_n(2 - \tilde{t}_n) \sum_i \tilde{\lambda}^n_i d^2_{C_i}(\tilde{x}^n) \quad \forall y \in \cap_i C_i.$$

In other words, letting $\tilde{\rho}_n = \tilde{t}_n(2 - \tilde{t}_n) \sum_i \tilde{\lambda}^n_i d^2_{C_i}(\tilde{x}^n)$, we have the Fejér estimates $|y - \tilde{x}^1|^2 - |y - \tilde{x}^n|^2 \ge \sum_{j=1}^{n-1} \tilde{\rho}_j$ $\forall y \in \cap_i C_i$. Therefore, if we start from $\tilde{x}^1 = P_{\mathcal{L}(f^k, f^k_{\text{lev}})}(x^k)$ and terminate *for any* $n \ge 1$, then $z^k = \tilde{x}^n$ and $\rho^k_\phi = d^2_{\mathcal{L}(f^k, f^k_{\text{lev}})}(x^k) + \sum_{j=1}^{n-1} \tilde{\rho}_j$ will satisfy the requirements of step 4. (In particular we may stop if $\rho_k + \rho^k_\phi > \bar{D}^2$ for such ρ^k_ϕ.) Moreover, ρ^k_ϕ may be increased by using more refined Fejér estimates to replace $\tilde{\rho}_j$ with some larger quantities [Kiw95]. In fact [Kiw95] shows that other relaxation methods have much better Fejér estimates; hence they could provide more efficient implementations of step 4(i).

Similar ideas may be used for implementing step 4(ii) via finite iterative methods that do not necessarily compute x^{k+1} as the projection of z^k onto S; see [KuF90] for details.

It is worth observing that many relaxation methods are highly amenable to parallel computation; see [AhC89, Kiw95]. Since we do not require exact projections, various heuristics may limit the work spent on relaxations.

4. QP-based implementations. We shall now discuss possible implementations of our methods that employ subgradient selection and aggregation. These two techniques have proved to be highly useful in implementations of other NDO bundle methods; see, e.g., [Kiw85, Kiw89, Kiw90] for details.

First, we describe *subgradient selection*. If $\phi^k = \hat{f}^k$ and $\mathcal{L}(\phi^k, f^k_{\text{lev}}) \neq \emptyset$ then

$$y^k = \arg\min\{\, |x - x^k|^2/2 : f^j(x) \leq f^k_{\text{lev}}, j \in J^k \,\}. \tag{4.1}$$

Denote the Lagrange multipliers of (4.1) by λ^k_j, $j \in J^k$. Let $\hat{J}^k = \{j \in J^k : \lambda^k_j > 0\}$. By the Karush–Kuhn–Tucker (KKT) conditions, if we select $J^k_s \subset J^k$ such that $\hat{J}^k \subset J^k_s$, then J^k_s may replace J^k in (4.1) without changing its solution. This suggests that only the linearizations f^j, $j \in J^k_s$, that have contributed to y^k should be retained for the next iteration. Moreover, many QP methods will automatically produce $|\hat{J}^k| \leq N$. Hence we may choose $J^{k+1} = J^k_s \cup \{k+1\}$ such that $|J^{k+1}| \leq N+1$. Storing the subgradients $g^j = g_f(x^j)$ for the representation $f^j = f^j(x^k) + \langle g^j, \cdot - x^k \rangle$, we do not need x^j to update $f^j(x^{k+1}) = f^j(x^k) + \langle g^j, x^{k+1} - x^k \rangle$ for $j \in J^k_s$. Thus the required storage is of order $(N+1)^2$ (plus the QP workspace).

Since subgradient selection may require excessive storage for large N, we now turn to *subgradient aggregation*, in which aggregate linearizations are produced recursively by taking convex combinations of the "ordinary" linearizations. Suppose $\phi^k = \max\{\hat{f}^k, \psi^{k-1}\}$ for some affine $\psi^{k-1} \in \text{co}\{f^j\}_{j=1}^{k-1}$ of the form $\psi^{k-1}(\cdot) = \psi^{k-1}(x^k) + \langle g^{k-1}_\psi, \cdot - x^k \rangle$ ($\psi^0 = f^1$). Let us add to (4.1) the constraint $\psi^{k-1}(x) \leq f^k_{\text{lev}}$ with Lagrange multiplier λ^k_ψ. Equivalently, in terms of $d^k = y^k - x^k$, $\alpha^k_j = f^k_{\text{lev}} - f^j(x^k)$, $j \in J^k$, and $\alpha^k_\psi = f^k_{\text{lev}} - \psi^{k-1}(x^k)$, we must find

$$d^k = \arg\min\{\, |d|^2/2 : \langle g^j, d \rangle \leq \alpha^k_j, j \in J^k, \langle g^{k-1}_\psi, d \rangle \leq \alpha^k_\psi \,\}. \tag{4.2}$$

Letting $\lambda^k_s = \sum_{j \in J^k} \lambda^k_j + \lambda^k_\psi$, we define "normalized" multipliers $\hat{\lambda}^k_j = \lambda^k_j / \lambda^k_s$, $j \in J^k$, $\hat{\lambda}^k_\psi = \lambda^k_\psi / \lambda^k_s$ that form a convex combination. Then, by the KKT conditions, $d^k = -\lambda^k_s g^k_\psi$, where $g^k_\psi = \sum_{j \in J^k} \hat{\lambda}^k_j g^j + \hat{\lambda}^k_\psi g^{k-1}_\psi \in \partial \phi^k(y^k)$. (Incidentally, $\lambda^k_s > 0$ because $y^k \neq x^k$ due to $f(x^k) > f^k_{\text{lev}}$.) Defining the next *aggregate linearization* $\psi^k(\cdot) = \phi^k(y^k) + \langle g^k_\psi, \cdot - y^k \rangle$, we observe that $\psi^k = \sum_{j \in J^k} \hat{\lambda}^k_j f^j + \hat{\lambda}^k_\psi \psi^{k-1} \in \text{co}\{f^j\}_{j=1}^k$ and $y^k = P_{\mathcal{L}(\psi^k, f^k_{\text{lev}})}(x^k)$. In effect, ψ^k embodies all the past subgradient information that determined y^k. (Equivalently, this amounts to replacing the constraints of (4.2) by their convex combination with normalized multipliers.) With such motivation, the next iteration may use $\phi^{k+1} = \max\{\hat{f}^{k+1}, \psi^k\}$ with J^{k+1} containing $k+1$ and, e.g., all but one of the elements of J^k to ensure bounded storage.

An alternative *selective aggregation* consists in aggregating just two linearizations. Specifically, if we pick $i, j \in J^k$ with $\lambda^k_i, \lambda^k_j > 0$, replace f^j with $(\lambda^k_i f^i + \lambda^k_j f^j)/(\lambda^k_i + \lambda^k_j)$, and drop i from J^k, then the solution of (4.1) is unchanged and the new $f^j \in \Phi$ ($f^j \leq \hat{f}^k$). In other words, we may replace f^j with the aggregate of f^i and f^j and destroy f^i to make room for the next f^{k+1}. (Here aggregation limits only the loss of information necessary to ensure bounded storage. In other bundle methods [Kiw85], it is crucial for convergence.)

Remark 4.1. The simplest case of aggregating just two linearizations, i.e., $J^k = \{k\}$ in (4.2), may be handled analytically. Suppose g^k and g^{k-1}_ψ are independent (the other case involves projecting on one halfspace only). Then one of the following three

cases may arise: $\lambda_k^k = -\alpha_k^k/|g^k|^2$ and $\lambda_\psi^k = 0$ (if $\alpha_k^k\langle g^k, g_\psi^{k-1}\rangle \le \alpha_\psi^k|g^k|^2$); $\lambda_k^k = 0$ and $\lambda_\psi^k = -\alpha_\psi^k/|g_\psi^{k-1}|^2$; or

$$\lambda_k^k = (\langle g^k, g_\psi^{k-1}\rangle\alpha_\psi^k - |g_\psi^{k-1}|^2\alpha_k^k)/(|g^k|^2|g_\psi^{k-1}|^2 - \langle g^k, g_\psi^{k-1}\rangle^2), \tag{4.3a}$$

$$\lambda_\psi^k = (\langle g^k, g_\psi^{k-1}\rangle\alpha_k^k - |g^k|^2\alpha_\psi^k)/(|g^k|^2|g_\psi^{k-1}|^2 - \langle g^k, g_\psi^{k-1}\rangle^2). \tag{4.3b}$$

In particular, if $\alpha_\psi^k = 0$, then either $d^k = -\alpha_k^k g^k/|g^k|^2$ if $\langle g^k, g_\psi^{k-1}\rangle \ge 0$ or

$$-d^k/\lambda_k^k = g^k - \langle g^k, g_\psi^{k-1}\rangle g_\psi^{k-1}/|g_\psi^{k-1}|^2 \quad \text{and} \quad \langle d^k, g_\psi^{k-1}\rangle = 0 \quad \text{if} \quad \langle g^k, g_\psi^{k-1}\rangle < 0, \tag{4.4}$$

where $g_\psi^{k-1} = -d^{k-1}/\lambda_s^{k-1}$ if $k > 1$, so that $-d^k/\lambda_k^k = g^k - \langle g^k, d^{k-1}\rangle d^{k-1}/|d^{k-1}|^2$ and $\langle d^k, d^{k-1}\rangle = 0$ if $\langle g^k, d^{k-1}\rangle > 0$. Hence subgradient aggregation is related to the CS techniques of [CFM75, ShU89]; see §6.

Let us now describe subgradient selection for the dual methods of [Kiw96, §9]. Let λ_j^k, $j \in J^k$, denote a solution to (2.4) using (2.5) for $\hat\phi^k = \hat f^k$. As in the primal case, if $\hat J^k \equiv \{j \in J^k : \lambda_j^k > 0\} \subset J_s^k \subset J^k$, then J_s^k may replace J^k in (2.5) without changing p^k, and we may select $J^{k+1} = J_s^k \cup \{k+1\}$. Again, many QP methods will ensure $|\hat J^k| \le N+1$, and the required storage is of order $(N+2)^2$.

Aggregation is natural in the dual methods, since they produce an aggregate linearization ϕ^k (from $\hat\phi^k$) that determines y^k. Specifically, employing $\hat\phi^k = \max\{\hat f^k, \phi^{k-1}\}$ in (2.4) to find ϕ^k, we may choose $\hat\phi^{k+1} = \max\{\hat f^{k+1}, \phi^k\}$ with J^{k+1} containing $k+1$ and all but one of the elements of J^k to ensure bounded storage.

The following (primal) *pairwise projections* strategy generalizes one in [KKA87]. Having several f^j, $j \in J^k$, let $\phi^k = \max\{f^k, f^{\hat\jmath}\}$ for $\hat\jmath \in J^k$ chosen to maximize the resulting $|y^k - x^k|$ when $\{k, \hat\jmath\}$ replaces J^k in (4.1). For example, use the formula (cf. (4.3))

$$|d^k|^2 = [(|g^k|\alpha_{\hat\jmath}^k)^2 - 2\langle g^k, g^{\hat\jmath}\rangle\alpha_{\hat\jmath}^k\alpha_k^k + (|g^{\hat\jmath}|\alpha_k^k)^2]/(|g^k|^2|g^{\hat\jmath}|^2 - \langle g^k, g^{\hat\jmath}\rangle^2) \quad \text{if} \quad \lambda_{\hat\jmath}^k, \lambda_k^k > 0.$$

Such $\hat\jmath$ may be included in J^{k+1}. Alternatively, if $\hat f^k(y^k) > f_{\text{lev}}^k$, we may replace f^k by the aggregate linearization of f^k and $f^{\hat\jmath}$, pick $\hat\jmath$ such that $f^{\hat\jmath}(y^k) > f_{\text{lev}}^k$, and recompute y^k. Of course, more than two constraints can be used at a time, and projections may continue until y^k becomes almost feasible in (4.1). Moreover, if $N \gg |J^k|$, then maintaining a matrix of inner products between g^j, $j \in J^k$, allows us to compute pairwise projections without additional expensive inner products; cf. (4.3). One may use Lemma 9.4 in [Kiw96] to show that pairwise projections are essentially equivalent to the surrogate method $S2$ of [Oko92] applied to the inequalities $f^j(x) \le f_{\text{lev}}^k$, $j \in J^k$, starting from x^k. (The remaining surrogate methods of [Oko92] are obtained by using triples of inequalities and successive aggregation.)

Remark 4.2. It is worth observing that step 4 may perform *several relaxations* using the accumulated linearizations. Specifically, at step 4(iii), instead of going to step 6, we may return to step 4(i) to choose *any* $\phi^k \in \Phi$ for the next relaxation with ρ_k replaced by $\rho_k + \rho_\phi^k + \rho_S^k$ and x^k by x^{k+1} (the replacement being justified by Remark 2.2); any number of such returns can occur, and all but the final one may skip the projection on S by setting $x^{k+1} = z^k$. For example, suppose J^k is so large that we do not want to solve (4.1). Then, until y^k becomes almost feasible in (4.1), each execution of step 4 may use $\phi^k = \max\{f^{\hat\jmath}, \phi^{k-1}\}$, where $\hat\jmath \in \operatorname{Arg\,max}_{j\in J^k} f^j(x^k)$ and ϕ^{k-1} is

the current aggregate linearization; i.e., it may solve (4.2) with J^k replaced by $\{\hat{\jmath}\}$. Alternatively, $\{\hat{\jmath}\}$ may be replaced by some larger set for which the solution of (4.2) is "cheap"; cf. §5. The dual methods can be used iteratively in the same way. In other words, we may attempt to *accelerate* our algorithm by performing *extra iterations on models* of f to exploit more fully the accumulated information about f and hence to reduce the number of f-evaluations at the cost of more work per iteration.

5. Relaxation with surrogate inequalities. This section introduces "cheap" QP-based implementations by extending the framework of deep surrogate cuts of relaxation methods for linear inequalities [BGT81, GoT82, Tod79].

We need additional notation. For any set $\mathcal{A} \subset \mathbb{R}^N$, $\operatorname{lin}\mathcal{A}$ denotes its linear span and $\operatorname{cone}\mathcal{A} = \{a : a = \sum_{i=1}^n \lambda_i a^i, a^i \in \mathcal{A}, \lambda_i \geq 0, n < \infty\}$ denotes its convex conical hull. We let $\mathcal{A}^- = \{x : \langle x, y\rangle \leq 0 \ \forall y \in \mathcal{A}\}$ and $\mathcal{A}^+ = -\mathcal{A}^-$ denote its negative and positive polar cones, respectively. For a matrix $A \in \mathbb{R}^{n\times n}$, a_{ij} and a^i denote its ijth element and ith column, respectively. Given a set $\mathcal{I} \subset \{1{:}\,n\}$, $A_{\mathcal{I}}$ denotes the matrix with columns a^i, $i \in \mathcal{I}$. For a given $b \in \mathbb{R}^n$, $b_{\mathcal{I}}$ denotes the vector with elements b_i, $i \in \mathcal{I}$. Matrix inequalities hold componentwise. A is called a *Stieltjes* matrix if $a_{ij} = a_{ji} \leq 0 \ \forall i \neq j$, $i, j = 1{:}\,n$, and $A^{-1} \geq 0$.

Given $A \in \mathbb{R}^{N\times m}$ and $b \in \mathbb{R}^m$, consider the system of linear inequalities $\langle a^i, x\rangle \leq b_i$, $i = 1{:}\,m$, having a (possibly empty) solution set $\mathcal{P} = \{x : A^T x \leq b\}$. Suppose $a^i \neq 0$ for $i = 1{:}\,m$. Then each inequality defines a closed halfspace $H_i = \{x : \langle a^i, x\rangle \leq b_i\}$, and $\mathcal{P} = \bigcap_{i=1}^m H_i$ is a convex polyhedron.

Remark 5.1. We are mainly interested in the case where $\mathcal{P} = \mathcal{L}(\hat{f}^k, f^k_{\text{lev}})$, but to compare the convergence results of [Kiw96] with those for relaxation methods [Gof81, Tel82] one may observe the following. If $\mathcal{P} \neq \emptyset$, then $\mathcal{P} = \operatorname{Arg\,min} f$, where $f = \max_{i=1:m}(\langle a^i, \cdot\rangle - b_i)_+$ has a subgradient $g_f(x) = a^i$ if $\langle a^i, x\rangle - b_i = f(x) > 0$, $g_f(x) = 0$ if $f(x) = 0$, satisfying $|g_f(x)| \leq L_f := \max_{i=1:m} |a^i|$. In this case we may let $f^k_{\text{lev}} \equiv f^* = 0$ in Algorithm 2.1 and proceed as if S were $\mathbb{R}^N$, replacing $\operatorname{diam}(S)$ in [Kiw96, Eq. (1.4)] by $|x^* - x^1|$ for any $x^* \in \mathcal{P}$; cf. [Kiw96, §6]. Then the SPA of (1.2) describes the *maximal residual* version of the relaxation method; the *maximal distance* version corresponds to dividing each a^i and b_i by $|a^i|$ initially.

In classical versions of relaxation and ellipsoid methods for finding a point in $\mathcal{P}$, given a current point $\tilde{x} \notin \mathcal{P}$, one finds the next point $\bar{x}$ by projecting $\tilde{x}$ on the halfspace H_i that is furthest from $\tilde{x}$, since for faster convergence one wishes to maximize $|\bar{x} - \tilde{x}|$. By combining inequalities one can sometimes obtain halfspaces that are further from $\tilde{x}$.

If $\lambda \in \mathbb{R}^m_+$, $a^\lambda = A\lambda$, and $b_\lambda = b^T\lambda$, then the *surrogate inequality* $\langle a^\lambda, x\rangle \leq b_\lambda$ is valid ($A^T x \leq b \Rightarrow \lambda^T A^T x \leq \lambda^T b$). The *deepest* surrogate inequality that maximizes the distance $(\langle a^\lambda, \tilde{x}\rangle - b_\lambda)_+/|a^\lambda|$ from $\tilde{x}$ to $H_\lambda = \{x : \langle a^\lambda, x\rangle \leq b_\lambda\}$ corresponds to

$$\tilde{\lambda} \in \operatorname{Arg\,max}\{\, \tilde{s}^T\lambda/|A\lambda| : \lambda \geq 0\,\}, \tag{5.1}$$

where $\tilde{s} := A^T\tilde{x} - b \not\leq 0$ ($\tilde{x} \notin \mathcal{P}$). Clearly, if $\mathcal{P} \neq \emptyset$, then $H_{\tilde{\lambda}}$ is the unique halfspace containing $\mathcal{P}$ that is furthest from $\tilde{x}$, and $H_{\tilde{\lambda}} = \{x : \langle \tilde{d}, x - \tilde{x}\rangle \geq |\tilde{d}|^2\}$, where $\tilde{d} = P_{\mathcal{P}}(\tilde{x}) - \tilde{x}$ (since for any halfspace $H \ni P_{\mathcal{P}}(\tilde{x})$, $d_H(\tilde{x}) < d_{\mathcal{P}}(\tilde{x})$ unless $P_H(\tilde{x}) = P_{\mathcal{P}}(\tilde{x})$). Of course, $\tilde{d}$ solves the QP problem

$$\tilde{d} = \arg\min\{\, |d|^2/2 : A^T d \leq -\tilde{s}\,\}. \tag{5.2}$$

By duality, we may equivalently find its (possibly nonunique) Lagrange multiplier

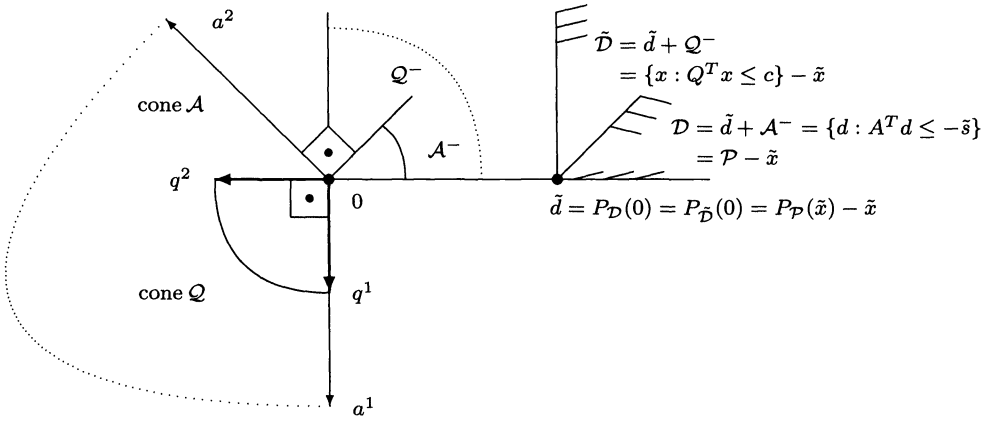


FIG. 5.1. *Illustration of orthogonal surrogates.*

vector

$$\tilde{\lambda} \in \operatorname{Arg\,min}\{\, |A\lambda|^2/2 - \tilde{s}^T\lambda : \lambda \ge 0 \,\}. \tag{5.3}$$

Indeed, by the KKT conditions, $\tilde{d}$ and $\tilde{\lambda}$ satisfy (5.2)–(5.3) iff $\tilde{d} = -A\tilde{\lambda}$, $A^T\tilde{d} \le -\tilde{s}$, $\tilde{\lambda} \ge 0$, and $\tilde{\lambda}^T(A^T\tilde{d} + \tilde{s}) = 0$. Hence $\tilde{s}^T\tilde{\lambda} = |\tilde{d}|^2$ and $\langle a^{\tilde{\lambda}}, \tilde{x}\rangle - |\tilde{d}|^2 = \tilde{\lambda}^T A^T\tilde{x} - \tilde{x}^T A\tilde{\lambda} + b^T\tilde{\lambda} = b_{\tilde{\lambda}}$, so $\langle a^{\tilde{\lambda}}, x\rangle \le b_{\tilde{\lambda}}$ iff $\langle \tilde{d}, x - \tilde{x}\rangle \ge |\tilde{d}|^2$, and $P_{H_{\tilde{\lambda}}}(\tilde{x}) = P_{\mathcal{P}}(\tilde{x})$. (We may add that the optimal values in (5.1) and (5.3) are infinite iff $\mathcal{P} = \emptyset$. Since the objective of (5.1) is positively homogeneous, the deepest cut can also be found by solving $\min\{|A\lambda| : \tilde{s}^T\lambda = 1, \lambda \ge 0\}$, $\max\{\tilde{s}^T\lambda : |A\lambda| = 1, \lambda \ge 0\}$ or $\min\{|A\lambda|/\tilde{s}^T\lambda : \sum_i \lambda_i = 1, \lambda \ge 0\}$. We note that (5.1) is a special case of [Kiw96, Eq. (9.7)], whereas the restricted variant (2.4) does not seem to have been considered explicitly in the context of linear inequalities.)

Of course, finding the deepest surrogate inequality via (5.1)–(5.3) may be too expensive, *except when A is orthonormal and $\tilde{s} \ge 0$*, in which case $\tilde{d} = -A\tilde{\lambda}$ and $\tilde{\lambda} = \tilde{s}$ by (5.3). In the general case, we may project on a surrogate $\tilde{\mathcal{P}} = \{x : Q^T x \le c\}$ of $\mathcal{P}$, where $Q^T x \le c$ is a surrogate of $A^T x \le b$ (so that $\mathcal{P} \subset \tilde{\mathcal{P}}$) and Q is orthonormal. As in [BGT81, GoT82, Tod79], it is convenient to work with a subset of inequalities, indexed by $\mathcal{I} \subset \{1{:}m\}$, say, that satisfy the *obtuse angles condition* $\langle a^i, a^j\rangle \le 0$ $\forall i \ne j$, $i, j \in \mathcal{I}$. Taking $\mathcal{I} = \{1{:}m\}$ first for simplicity, we now show how to construct suitable surrogates via orthogonalization (see Figure 5.1).

LEMMA 5.2. *Let $\mathcal{A} = \{a^i\}_{i=1}^m$, $C = \operatorname{cone}\mathcal{A}$, $\hat{m} = \operatorname{rank} A$, and $\mathcal{G} = A^T A$. If $\langle a^i, a^j\rangle \le 0$ $\forall i \ne j$, $i, j = 1{:}m$, then*

(i) *C contains an orthonormal system $\mathcal{Q} = \{q^i\}_{i=1}^{\hat{m}}$ such that $\operatorname{lin} C = \operatorname{lin}\mathcal{Q}$ and $A = QR$, where $Q \in \mathbb{R}^{N\times\hat{m}}$ is orthonormal and $R \in \mathbb{R}^{\hat{m}\times m}$ is upper triangular, with $r_{ii} \ge 0$ and $r_{ij} \le 0$ for $i = 1{:}\hat{m}$, $j = i+1{:}m$. Q and R can be found via the Gram–Schmidt orthogonalization: set $m_1 = 0$, and for $j = 1{:}m$ set $\tilde{q}^j = a^j - \sum_{i=1}^{m_j}\langle a^j, q^i\rangle q^i$, $r_{jj} = |\tilde{q}^j|$, $r_{ij} = \langle a^j, q^i\rangle$ for $i = 1{:}m_j$, $r_{ij} = 0$ for $i > m_j$, $i \ne j$, $q^j = \tilde{q}^j/|\tilde{q}^j|$ and $m_{j+1} = m_j + 1$ if $\tilde{q}^j \ne 0$, $m_{j+1} = m_j$ otherwise. (The final $m_{m+1} = \hat{m}$.)*

(ii) *$C^+ \cap \operatorname{lin} C \subset \mathcal{Q}^+ \cap \operatorname{lin}\mathcal{Q} = \mathcal{Q}^+ \cap \operatorname{cone}\mathcal{Q} \subset C$.*

(iii) *If $\operatorname{rank} A = m$, then $R^{-1} \ge 0$ and $\mathcal{G}^{-1} \ge 0$; i.e., $\mathcal{G}$ is a Stjelties matrix.*

(iv) *If $\operatorname{rank} A = m$, then R is the unique Cholesky factor of $\mathcal{G}$ having a positive diagonal, $Q = AR^{-1}$ and $\mathcal{P} \subset \{x : Q^T x \le c\}$, where $c = R^{-T}b$. In particular,*

each inequality $\langle q^j, x\rangle \le c_j$ *is a surrogate of the system* $\langle q^i, x\rangle \le c_i$, $i = 1{:}\, j-1$, $\langle a^j, x\rangle \le b_j$, *and hence a surrogate of the (possibly stronger) system* $\langle a^i, x\rangle \le b_i$, $i = 1{:}\, j$.

Proof. (i) If $\mathcal{Q}_j = \{q^i\}_{i=1}^{m_j} \subset C_j = \operatorname{cone}\{a^i\}_{i=1}^{j-1}$, then $a^j \in C_j^-$ yields $\langle a^j, q^i\rangle \le 0$ for $i = 1{:}\, m_j$, so $\tilde{q}^j \in \operatorname{cone}(\{a^j\} \cup \mathcal{Q}_j) \subset C_{j+1}$ and $\mathcal{Q}_{j+1} \subset C_{j+1}$. The rest follows by induction.

(ii) $a \in \mathcal{Q}^+ \cap \operatorname{lin}\mathcal{Q} \iff a = \sum_{i=1}^{\hat{m}} \langle a, q^i\rangle q^i$ with $\langle a, q^i\rangle \ge 0$ for $i = 1{:}\, \hat{m} \iff a \in \operatorname{cone}\mathcal{Q}$, since $a = \sum_{i=1}^{\hat{m}} \lambda_i q^i$ iff $\lambda_i = \langle a, q^i\rangle$ for $i = 1{:}\, \hat{m}$. Clearly, $\mathcal{Q} \subset \operatorname{cone}\mathcal{Q} \subset C$ and $C^+ \subset \mathcal{Q}^+$.

(iii) If $\hat{m} = m$, then $r_{ii} > 0$ for $i = 1{:}\, m$, so R and $\mathcal{G} = R^T R$ are nonsingular. Since $r_{ij} \le 0$ for $i < j$, if $Ru = v \ge 0$, then $u \ge 0$ from $u_i = (v_i - \sum_{j=i+1}^{m} r_{ij}u_j)/r_{ii}$, $i = m, \ldots, 1$. Hence $R^{-1} \ge 0$ and $\mathcal{G}^{-1} = R^{-1}R^{-T} \ge 0$.

(iv) The uniqueness of R is well known. Use $(Q^T, c) = R^{-T}(A^T, b)$ and $(q^j, c_j) = [(a^j, b_j) - \sum_{i=1}^{j-1} r_{ij}(q^i, c_i)]/r_{jj}$, $j = 1, \ldots, m$, with $R^{-T} \ge 0$ and $r_{ij} \le 0$ by (i) and (iii), to get the desired conclusion, noting that $A^T x \le b \Rightarrow R^{-T}A^T x = Q^T x \le R^{-T} b = c$. □

The next result helps in selecting subsets of inequalities for which the projections are "easy." For any $\mathcal{I} \subset \{1{:}\, m\}$, let $\mathcal{A}_\mathcal{I} = \{a^i\}_{i\in\mathcal{I}}$, $\mathcal{P}_\mathcal{I} = \{x : A_\mathcal{I}^T x \le b_\mathcal{I}\}$, $\mathcal{D}_\mathcal{I} = \{d : A_\mathcal{I}^T d \le -\tilde{s}_\mathcal{I}\}$, and $\mathcal{G}_{\mathcal{I}\mathcal{I}} = A_\mathcal{I}^T A_\mathcal{I}$. If $\mathcal{P}_\mathcal{I} \ne \emptyset$, let $\tilde{d}_{(\mathcal{I})} = P_{\mathcal{D}_\mathcal{I}}(0)$, so that $\tilde{d}_{(\mathcal{I})} = P_{\mathcal{P}_\mathcal{I}}(\tilde{x}) - \tilde{x}$ from $\tilde{s}_\mathcal{I} = A_\mathcal{I}^T \tilde{x} - b_\mathcal{I}$.

LEMMA 5.3. *Suppose* $\mathcal{I} \subset \{1{:}\, m\}$, $\langle a^i, a^j\rangle \le 0$ $\forall i \ne j$, $i, j \in \mathcal{I}$, *and* $\tilde{s}_\mathcal{I} \ge 0$. *Then*

(i) *If* $\operatorname{rank} A_\mathcal{I} = |\mathcal{I}|$, *then* $\tilde{d}_{(\mathcal{I})} = -A_\mathcal{I}\tilde{\lambda}_\mathcal{I}$, *where* $\tilde{\lambda}_\mathcal{I} = \mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1}\tilde{s}_\mathcal{I} \ge 0$, *and*

$$\tilde{d}_{(\mathcal{I})} = \arg\min\{\, |d|^2/2 : A_\mathcal{I}^T d = -\tilde{s}_\mathcal{I}\,\}. \tag{5.4}$$

Moreover, for $\hat{\mathcal{I}} = \{i : \tilde{\lambda}_i > 0\}$ *and each* $j \in \mathcal{I}$, $\tilde{\lambda}_j > 0$ *if* $\tilde{s}_j > 0$; *if* $\tilde{s}_j = \tilde{\lambda}_j = 0$, *then* $a^j \perp \mathcal{A}_{\hat{\mathcal{I}}}$; *and* $\tilde{\lambda}_j = 0$ *if* $\tilde{s}_j = 0$ *and* $a^j \perp \mathcal{A}_{\mathcal{I}\setminus\{j\}}$.

(ii) *If* $\operatorname{rank} A_\mathcal{I} = |\mathcal{I}|$, $j \in \{1{:}\, m\} \setminus \mathcal{I}$, $A_\mathcal{I}^T a^j \le 0$, $\mathcal{J} = \mathcal{I} \cup \{j\}$, *and either* $\tilde{s}_j > 0$ *or* $\tilde{s}_j = 0$ *and* $\tilde{s}_\mathcal{I} > 0$, *then* $\operatorname{rank} A_\mathcal{J} = |\mathcal{J}| \iff \mathcal{D}_\mathcal{J} \ne \emptyset \iff \mathcal{P}_\mathcal{J} \ne \emptyset$.

(iii) $\operatorname{rank} A_\mathcal{I} = |\mathcal{I}| \iff A_\mathcal{I}^T \check{d} < 0$ *for some* $\check{d}$.

(iv) *If* $\tilde{s}_\mathcal{I} > 0$ *and* $\mathcal{P}_\mathcal{I} \ne \emptyset$, *then* $\operatorname{rank} A_\mathcal{I} = |\mathcal{I}|$ *and* $\tilde{\lambda}_\mathcal{I} = \mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1}\tilde{s}_\mathcal{I} > 0$.

(v) *If* $\mathcal{G}_{\mathcal{I}\mathcal{I}} = R^T R$ *is nonsingular, where* $R \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|}$ *is upper triangular,* $Q = A_\mathcal{I} R^{-1}$, *and* $\sigma = R^{-T}\tilde{s}_\mathcal{I}$, *then* $\tilde{d}_{(\mathcal{I})} = -Q\sigma$ *and* $\tilde{\lambda}_\mathcal{I} = R^{-1}\sigma$ *in* (i). *Moreover, if* $r_{ii} > 0$ *for* $i = 1{:}\, |\mathcal{I}|$ *and* $\tilde{\mathcal{D}} = \{d : Q^T d \le -\sigma\}$, *then* $\tilde{\mathcal{D}} \supset \mathcal{D}_\mathcal{I}$, $\tilde{d}_{(\mathcal{I})} = P_{\tilde{\mathcal{D}}}(0)$ *and* $Q^T \tilde{d}_{(\mathcal{I})} = -\sigma$.

Proof. (i) Replace $\{1{:}\, m\}$ with $\mathcal{I}$ in Lemma 5.2 to get $\mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1} \ge 0$. Hence $\tilde{\lambda}_\mathcal{I} = \mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1}\tilde{s}_\mathcal{I} \ge 0$, and letting $\bar{d} = -A_\mathcal{I}\tilde{\lambda}_\mathcal{I}$ we have $A_\mathcal{I}^T \bar{d} = -\tilde{s}_\mathcal{I}$, so (5.4) holds with $\bar{d} = \tilde{d}_{(\mathcal{I})}$ by the KKT conditions. Since $\mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1} \ge 0$ is positive definite, it has a positive diagonal, so $\tilde{\lambda}_j > 0$ if $\tilde{s}_j > 0$. Next, suppose $\tilde{s}_j = 0$ and let $\mathcal{J} = \mathcal{I} \setminus \{j\}$. If $\tilde{\lambda}_j = 0$, then $0 = -\tilde{s}_j = \langle a^j, \bar{d}\rangle = -\sum_{i\in\hat{\mathcal{I}}} \tilde{\lambda}_i \langle a^j, a^i\rangle$ with $\tilde{\lambda}_i > 0$ and $\langle a^j, a^i\rangle \le 0$ imply $a^j \perp \mathcal{A}_{\hat{\mathcal{I}}}$. Conversely, if $a^j \perp \mathcal{A}_\mathcal{J}$, then after symmetric permutations we have

$$\mathcal{G}_{\mathcal{I}\mathcal{I}} = \begin{bmatrix} \mathcal{G}_{\mathcal{J}\mathcal{J}} & 0 \\ 0 & |a^j|^2 \end{bmatrix} \text{ and } \mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1} = \begin{bmatrix} \mathcal{G}_{\mathcal{J}\mathcal{J}}^{-1} & 0 \\ 0 & |a^j|^{-2} \end{bmatrix},$$

so $\tilde{\lambda}_j = \tilde{s}_j/|a^j|^2 = 0$.

(ii) If $\operatorname{rank} A_\mathcal{J} = |\mathcal{J}|$, then $-A_\mathcal{J}\mathcal{G}_{\mathcal{J}\mathcal{J}}^{-1}\tilde{s}_\mathcal{J} \in \mathcal{D}_\mathcal{J}$. If $\operatorname{rank} A_\mathcal{J} = |\mathcal{I}|$, then $a^j \in \operatorname{lin}\mathcal{A}_\mathcal{I}$. Since $a^j \in (\operatorname{cone}\mathcal{A}_\mathcal{I})^-$, Lemma 5.2(ii) applied to $\mathcal{A}_\mathcal{I}$ yields $-a^j \in \operatorname{cone}\mathcal{A}_\mathcal{I}$; i.e.,

there exists $\lambda_{\mathcal{I}} \geq 0$ such that $-a^j = A_{\mathcal{I}}\lambda_{\mathcal{I}}$. Thus if $A_{\mathcal{I}}^T d \leq -\tilde{s}_{\mathcal{I}}$ for some d, then $\langle a^j, d\rangle = -\lambda_{\mathcal{I}}^T A_{\mathcal{I}}^T d \geq \lambda_{\mathcal{I}}^T \tilde{s}_{\mathcal{I}} \geq 0$ from $\tilde{s} \geq 0$, so $D_{\mathcal{J}} = \emptyset$ because $\lambda_{\mathcal{I}}^T \tilde{s}_{\mathcal{I}} \geq 0 > -\tilde{s}_j$ if $\tilde{s}_j > 0$, whereas $\lambda_{\mathcal{I}}^T \tilde{s}_{\mathcal{I}} > 0 = -\tilde{s}_j$ if $\tilde{s}_j = 0$ and $\tilde{s}_{\mathcal{I}} > 0$, with $\lambda_{\mathcal{I}} \neq 0$ due to $a^j \neq 0$.

(iii) For "$\Rightarrow$," choose $s_{\mathcal{I}} > 0$ and $\check{d} = -A_{\mathcal{I}}\mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1} s_{\mathcal{I}}$ to get $A_{\mathcal{I}}^T \check{d} = -s_{\mathcal{I}} < 0$. For "$\Leftarrow$," suppose $A_{\mathcal{I}}^T \check{d} < 0$. Replacing $\tilde{x}$ by $\tilde{x} - \check{d}$ if necessary, we may assume $\tilde{s}_{\mathcal{I}} > 0$. Since $t\check{d} \in \mathcal{D}_{\mathcal{I}}$ for large $t > 0$, $\mathcal{D}_{\mathcal{I}} \neq \emptyset$. For any $i \in \mathcal{I}$, $\operatorname{rank} A_{\{i\}} = 1$, since $a^i \neq 0$. If $\tilde{\mathcal{I}} \subset \mathcal{I}$, $\operatorname{rank} A_{\tilde{\mathcal{I}}} = |\tilde{\mathcal{I}}|$, $j \in \mathcal{I} \setminus \tilde{\mathcal{I}}$, and $\tilde{\mathcal{J}} = \tilde{\mathcal{I}} \cup \{j\}$, then $\operatorname{rank} A_{\tilde{\mathcal{J}}} = |\tilde{\mathcal{J}}|$ by (ii) with $\mathcal{I}$ replaced by $\tilde{\mathcal{I}}$, since $\tilde{s}_j > 0$ and $\mathcal{D}_{\tilde{\mathcal{J}}} \supset \mathcal{D}_{\tilde{\mathcal{I}}} \neq \emptyset$. Hence, by induction, $\operatorname{rank} A_{\mathcal{I}} = |\mathcal{I}|$.

(iv) Combine (iii) and (i), noting that $\tilde{\lambda}_{\mathcal{I}} > 0$ if $\tilde{s}_{\mathcal{I}} > 0$.

(v) Clearly, $\tilde{d}_{(\mathcal{I})} = -A_{\mathcal{I}}\tilde{\lambda}_{\mathcal{I}} = -QR(R^TR)^{-1}\tilde{s}_{\mathcal{I}} = -Q\sigma$. Suppose $r_{ii} > 0$ for $i = 1{:}|\mathcal{I}|$. Apply Lemma 5.2(iii,iv) to $A_{\mathcal{I}}$ and $c = R^{-T}b_{\mathcal{I}}$ to get $\mathcal{D}_{\mathcal{I}} \subset \tilde{\mathcal{D}}$ from $\sigma = R^{-T}(A_{\mathcal{I}}^T\tilde{x} - b_{\mathcal{I}}) = Q^T\tilde{x} - c$, with $\sigma = R^{-T}\tilde{s}_{\mathcal{I}} \geq 0$ since $R^{-1} \geq 0$ and $\tilde{s}_{\mathcal{I}} \geq 0$. Hence, replacing $(A_{\mathcal{I}}, \tilde{s}_{\mathcal{I}})$ by (Q, σ) in (i), we have $P_{\tilde{\mathcal{D}}}(0) = -Q(Q^TQ)^{-1}\sigma = -Q\sigma = \tilde{d}_{(\mathcal{I})}$ and $Q^T\tilde{d}_{(\mathcal{I})} = -\sigma$. □

Lemmas 5.2–5.3 extend some results of [Tod79] in a way that is useful for algorithmic developments. For example, consider the following extension of the simultaneous projections method of [Tod79] for solving a possibly inconsistent system $A^Tx \leq b$.

PROCEDURE 5.4 (for finding a point in $\mathcal{P} = \{x : A^Tx \leq b\}$).

Step 0 (*Initialization*). Select $\tilde{x}^1 \in \mathbb{R}^N$, a feasibility tolerance $\epsilon_{\text{tol}} \geq 0$ and $\tilde{D} < \infty$ such that $d_{\mathcal{P}}(\tilde{x}^1) \leq \tilde{D}$ if $\mathcal{P} \neq \emptyset$. Choose $I^0 \subset \{1{:}m\}$ such that $\operatorname{rank} A_{I^0} = |I^0|$ and $\langle a^i, a^j\rangle \leq 0\ \forall i \neq j$, $i, j \in I^0$, e.g., $I^0 = \emptyset$. Set $\tilde{\rho}_1 = 0$ and $n = 1$.

Step 1 (*Constraint evaluation*). Calculate $s^n = A^T\tilde{x}^n - b$ and i_n such that $s_{i_n}^n = \max_i s_i^n$.

Step 2 (*Stopping criterion*). If $s_{i_n}^n \leq \epsilon_{\text{tol}}$, terminate.

Step 3 (*Selection*). Set $\tilde{I}^{n-1} = \{i \in I^{n-1} : \langle a^{i_n}, a^i\rangle \leq 0, s_i^n \geq 0\}$ and $I^n = \tilde{I}^{n-1} \cup \{i_n\}$. If desired, repeat the following for some $i \in \{1{:}m\} \setminus I^n$: if $A_{I^n}^T a^i \leq 0$ and either $s_i^n > 0$ or $s_i^n = 0$, and $A_{I^n}^T a^i \neq 0$ and either $s_{I^n}^n > 0$ or $\operatorname{rank} A_{I^n \cup \{i\}} = |I^n| + 1$, then augment I^n with i.

Step 4 (*Relaxation*). Print "$\mathcal{P} = \emptyset$" and terminate if $\operatorname{rank} A_{I^n} < |I^n|$. Otherwise set $\tilde{y}^n = P_{\mathcal{P}_{I^n}}(\tilde{x}^n) = \tilde{x}^n - A_{I^n}\lambda_{I^n}^n$ with $\lambda_{I^n}^n = \mathcal{G}_{I^nI^n}^{-1} s_{I^n}^n$. Choose a stepsize $\tilde{t}_n \in T$ and set $\tilde{x}^{n+1} = \tilde{x}^n + \tilde{t}_n(\tilde{y}^n - \tilde{x}^n)$ and $\tilde{\rho}_{n+1} = \tilde{\rho}_n + \tilde{t}_n(2 - \tilde{t}_n)|\tilde{y}^n - \tilde{x}^n|^2$.

Step 5 (*Infeasibility detection*). If $\tilde{\rho}_{n+1} > \tilde{D}^2$ or $(\tilde{D} - |\tilde{x}^{n+1} - \tilde{x}^1|)^2 > \tilde{D}^2 - \tilde{\rho}_{n+1}$, print "$\mathcal{P} = \emptyset$" and terminate.

Step 6. Increase n by 1 and go to step 1.

If $\mathcal{P} \neq \emptyset$, we may identify Procedure 5.4 with a version of Algorithm 2.1 that minimizes $f = \max_{i=1:m}(\langle a^i, \cdot\rangle - b_i)_+$ using $f_{\text{lev}}^k \equiv f^* = 0$; cf. [Kiw96, §6] and Remark 5.1. In particular $\mathcal{P} \subset \mathcal{P}_{I^n} \subset \{x : \langle a^{i_n}, x\rangle \leq b_{i_n}\}$ corresponds to $\phi^k \in \Phi_1^k$, and step 4 may be validated by applying Lemma 5.3 inductively at step 3 to get $\operatorname{rank} A_{I^n} = |I^n|$ if $\mathcal{P} \neq \emptyset$. Hence if $\mathcal{P} \neq \emptyset$, then Procedure 5.4 shares the convergence properties of Algorithm 2.1 from [Kiw96, §6], as well as those of classical relaxation methods [Agm54, Gof81, MoS54, Tel82, Tod79], such as linear rate of convergence and possible finite termination. The infeasibility test of step 5 is justified similarly as for Algorithm 2.1; cf. [Kiw96, §4]. Note that step 3 may include in I^n several i with $s_i^n = 0$; e.g., $i \in I^{n-1}$ if $\tilde{t}_{n-1} = 1$ and $s_{I^{n-1}}^n = 0$ from $\tilde{x}^n = \tilde{y}^{n-1}$. It is natural to choose I^n as large as possible, although one need not insist on maximality. Of course, in practice detecting $\operatorname{rank} A_{I^n} < |I^n|$ will require tolerances tuned to the factorization of A_{I^n}.

Remark 5.5. By using the Gram matrix $\mathcal{G} = A^TA$ one may avoid expensive

scalar products in updating s^n without forming $\tilde{x}^n$; cf. [Tod79]. Specifically, let $s^1 = A^T\tilde{x}^1 - b$, $\nu^1 = 0 \in \mathbb{R}^m$, $s^{n+1}_{I^n} = (1-\tilde{t}_n)s^n_{I^n}$, $s^{n+1}_{I^n_c} = s^n_{I^n_c} - \tilde{t}_n \mathcal{G}_{I^n_c I^n}\lambda^n_{I^n}$, and $\nu^{n+1} = \nu^n + \tilde{t}_n\lambda^n$ with $\lambda^n_{I^n_c} = 0$ and $I^n_c = \{1{:}\,m\} \setminus I^n$ for all n, so that $\tilde{x}^n = \tilde{x}^1 - A\nu^n$ and $\nu^n = \sum_{j=1}^n \tilde{t}_j\lambda^j$ for all n.

Remark 5.6. If we compute the Cholesky factorization $A^T_{I^n}A_{I^n} = R^TR$, then λ^n can be found by solving the two systems $R^T\sigma = s^n_{I^n}$ and $R\lambda^n_{I^n} = \sigma$; cf. Lemma 5.3(v). To save work, R and σ may be updated when I^n changes. However, as with normal equations for least-squares problems, one may need to employ iterative refinement to improve accuracy in the presence of rounding errors. Alternatively, one may use any stable method for computing the "skinny" QR-factorization $A_{I^n} = QR$, where Q is orthonormal, so that $A^T_{I^n}A_{I^n} = R^TR$. The classical Gram–Schmidt process may fail due to rounding errors, but reorthogonalization can ensure higher accuracy. Moreover, by Lemma 5.3, $\tilde{d}^n = \tilde{y}^n - \tilde{x}^n$ satisfies

$$\tilde{d}^n = \arg\min\{\, |d|^2/2 : A^T_{I^n}d = -s^n_{I^n}\,\}, \tag{5.5}$$

and this equality QP problem can be solved via many well-known methods. All these aspects are treated in depth in, e.g., [Bjö90, Fle87, GMW91, GVL89].

Remark 5.7. Suppose step 3 of Procedure 5.4 chooses $I^n = \tilde{I}^{n-1} \cup \{i_n\}$ with $s^n_{\tilde{I}^{n-1}} = 0$. (Recall that $s^n_{I^{n-1}} = 0$ if $\tilde{t}_{n-1} = 1$ and $\tilde{x}^n = \tilde{y}^{n-1}$.) Let $\hat{m} = |I^n|$ and let $e^{\hat{m}}$ denote column $\hat{m}$ of the $\hat{m} \times \hat{m}$ identity matrix, so that $s^n_{I^n} = s^n_{i_n}e^{\hat{m}}$. Then, using $R^T\sigma = s^n_{I^n}$ and $R\lambda^n_{I^n} = \sigma$ as in Remark 5.6, we have $\sigma = s^n_{i_n}e^{\hat{m}}/r_{\hat{m}\hat{m}}$ and only the system $R\lambda^n_{I^n} = s^n_{i_n}e^{\hat{m}}/r_{\hat{m}\hat{m}}$ must be solved. This system may be used even if $s^n_{\tilde{I}^{n-1}} \neq 0$. Specifically, decreasing $s^n_{I^n}$ to $\tilde{s}^n_{I^n}$ with $\tilde{s}^n_{\tilde{I}^{n-1}} = 0$ and $\tilde{s}^n_{i_n} = s^n_{i_n}$ corresponds to setting $\tilde{y}^n = P_{\mathcal{P}^n}(\tilde{x}^n)$, where $\mathcal{P}^n = \{x : A^T_{I^n}x \le b_{I^n} + s^n_{I^n} - \tilde{s}^n_{I^n}\}$ satisfies $\mathcal{P}_{I^n} \subset \mathcal{P}^n \subset \{x : \langle a^{i_n}, x\rangle \le b_{i_n}\}$, so the efficiency results remain true. This simplification is used in [Ceg92] when $\tilde{t}^{n-1} < 1$. It may, however, result in slower convergence, since $\mathcal{P}^n$ can be much bigger than $\mathcal{P}_{I^n}$. (The method of [Ceg92] scales the constraints of (5.5) before computing R, but $\tilde{d}^n$ is not affected.)

Remark 5.8. Using $\tilde{d}^n = -A_{I^n}\lambda^n_{I^n}$ and $R\lambda^n_{I^n} = s^n_{i_n}e^{\hat{m}}/r_{\hat{m}\hat{m}}$ as above, for the QR-factorization $A_{I^n} = QR$ we have $\tilde{d}^n = -s^n_{i_n}q^{\hat{m}}/r_{\hat{m}\hat{m}}$, where we may take $r_{\hat{m}\hat{m}} = |q|$ and $q^{\hat{m}} = q/|q|$ for $q = a^{i_n} - \sum_{i=1}^{\hat{m}-1}\langle a^{i_n}, q^i\rangle q^i$ as in Lemma 5.2. Thus only Q could be updated by computing some elements of $R = Q^TA_{I^n}$. However, using Q instead of R would require more storage and work if $\hat{m} \le N$, and could be less accurate without reorthogonalization.

We may add that the idea of using the obtuse angle property to identify cheap projections has wider implications. (A *cheap* projection requires only the solution of one or two linear systems in contrast to combinatorial QP.) For instance, it may be employed to accelerate general projection methods for convex feasibility problems; see [Kiw95].

Let us now show how to employ Procedure 5.4 as a subroutine for implementing step 4 of Algorithm 2.1; cf. §§2 and 3. Suppose Procedure 5.4 is called to find a point in $\mathcal{P} = \mathcal{L}(\hat{f}^k, f^k_{\text{lev}})$, starting from $\tilde{x}^1 = x^k$ (with $\epsilon_{\text{tol}} = 0$). Then it may be exited at any iteration $\bar{n} \ge 1$ also at step 6. Specifically, in view of [Kiw96, §4], we may take $\tilde{D} = r_k = (\bar{D}^2 - \rho_k)^{1/2}$, and step 5 may use the additional test $(\bar{D} - |\tilde{x}^{n+1} - x^{k(l)+1}|)^2 > r_k^2 - \tilde{\rho}_{n+1}$. Upon termination at step j, say, set $z^k = \tilde{x}^{\bar{n}}$, $\rho^k_\phi = \tilde{\rho}_{\bar{n}}$ if $j = 2$, $\rho^k_\phi = \infty$ if $j = 4$ or 5, and $z^k = \tilde{x}^{\bar{n}+1}$ and $\rho^k_\phi = \tilde{\rho}_{\bar{n}+1}$ if $j = 6$. Then z^k and ρ^k_ϕ satisfy (2.1) (cf. [Kiw96, §4]). The easiest way to ensure

that $\rho_\phi^k \geq t_{\min}(2-t_{\max})d^2_{\mathcal{L}(f^k,f^k_{\rm lev})}(x^k)$ consists of taking $i_1 \in \operatorname{Arg\,max}_i s_i^1/|a^i|$ at step 1, since then $|\tilde{y}^1 - \tilde{x}^1| = d_{\mathcal{P}_{I^1}}(\tilde{x}^1) \geq d_{\mathcal{L}(f^k,f^k_{\rm lev})}(x^k)$. (In fact the usual choice yields $|\tilde{y}^1-\tilde{x}^1| \geq (f(x^k)-f^k_{\rm lev})/\max_{j\in J^k}|g^j|$, and this suffices; cf. [Kiw96, Lem. 3.1].) Thus, if desired, *only one* iteration of Procedure 5.4 may be executed, but more iterations will yield better z^k and ρ_ϕ^k for Algorithm 2.1. Note that *if step* 6 *always exits*, then we may set $n=k$ and $\tilde{x}^n = x^k$ at step 0, terminating with $x^{k+1} = \tilde{x}^{k+1}$ at steps 4 or 5 or $z^k = \tilde{x}^{k+1}$ and $y^k = \tilde{y}^k$ at step 6.

Remark 5.9. As in §4, the final $\lambda^{\bar{n}}$ may be used for subgradient selection or aggregation. Note that selective aggregation corresponds to dropping from $A_{I^{\bar{n}}}$ one column aggregated into another, thus retaining the crucial property $\langle a^i, a^j\rangle \leq 0\ \forall i \neq j$ in the new $I^{\bar{n}}$ (in contrast to total aggregation that replaces one column by a convex combination of all columns). Of course, the final $I^{\bar{n}}$ may become the initial I^0 on the next call to Procedure 5.4, and the final matrix factorization should be used in a hot start, e.g., if only $f^k_{\rm lev}$ has changed. We may add that most matrix factorizations can be updated to reflect selective aggregation.

Remark 5.10. The following modification is useful when $\mathcal{P} = \mathcal{L}(\hat{f}^k, f^k_{\rm lev})$. Unless $f^k_{\rm lev} \equiv f^*$ is employed, it suffices to discover that $f^k_{\rm lev} \leq f^*$, since then $f^{k+1}_{\rm low} = f^k_{\rm lev}$ can be used without impairing the preceding efficiency results. To this end, step 3 may choose *any* I^n such that $i_n \in I^n$, $s_i^n \geq 0$, and $\langle a^i, a^j\rangle \leq 0\ \forall i \neq j$, $i,j \in I^n$. Indeed, suppose $f^k_{\rm lev} > f^*$. By [Kiw96, Eq. (2.1)], $\hat{f}^k(x^*) \leq f^*$ for any $x^* \in S^*$, and since $\hat{f}^k = \max_{j\in J^k} f^j$ and $\langle a^i, \cdot\rangle - b_i = f^j(\cdot) - f^k_{\rm lev}$ for suitable i and j, we have $\langle a^i, x^*\rangle - b_i \leq \hat{f}^k(x^*) - f^k_{\rm lev} < 0 \leq s_i^n = \langle a^i, \tilde{x}^n\rangle - b_i\ \forall i \in I^n$, so $A^T_{I^n}(x^* - \tilde{x}^n) < 0$ and $\operatorname{rank} A_{I^n} = |I^n|$ by Lemma 5.3(iii). Therefore, if $\operatorname{rank} A_{I^n} < |I^n|$ is revealed by any factorization then $f^k_{\rm lev} \leq f^*$ and Algorithm 2.1 may set $f^{k+1}_{\rm low} = f^k_{\rm lev}$.

Extending [Shc92], we now describe an *orthogonal surrogate projection* (OSP) version of Procedure 5.4 that sets $\tilde{y}^n = P_{\tilde{\mathcal{P}}_n}(\tilde{x}^n)$ for $\tilde{\mathcal{P}}_n = \{x : \langle q^j, x\rangle \leq c_j, j \in \check{J}^n\}$, where each inequality is a surrogate of $A^Tx \leq b$ (so that $\mathcal{P} \subset \tilde{\mathcal{P}}_n$), the system $\mathcal{Q}_{\check{J}^n} = \{q^j\}_{j\in\check{J}^n}$ is orthonormal, and $\check{J}^n \subset \{1\!:\!n\}$. Here $A^T_{\tilde{I}^{n-1}}x \leq b_{\tilde{I}^{n-1}}$ is replaced by the *accumulated* surrogates $\langle q^j, x\rangle \leq c_j$, $j \in \check{J}^{n-1}$, at step 3 in constructing the new surrogate $\langle q^n, x\rangle \leq c_n$ via orthogonalization as in Lemma 5.2. Specifically, at step 0 set $\check{J}^0 = \emptyset$. At step 3 set $\tilde{J}^{n-1} = \{j \in \check{J}^{n-1} : \langle a^{i_n}, q^j\rangle \leq 0\}$ and $\check{J}^n = \tilde{J}^{n-1} \cup \{n\}$. At step 4 set

$$\text{(5.6a)}\qquad \tilde{q}^n = a^{i_n} - \sum_{j\in\tilde{J}^{n-1}} \langle a^{i_n}, q^j\rangle q^j = 0 \quad\text{and}\quad q^n = \tilde{q}^n/|\tilde{q}^n| \quad\text{if}\quad \tilde{q}^n \neq 0,$$

print "$\mathcal{P} = \emptyset$," and terminate if $\tilde{q}^n = 0$; otherwise set $\sigma^n_{\tilde{J}^{n-1}} = (1-\tilde{t}_{n-1})\sigma^{n-1}_{\tilde{J}^{n-1}}$,

$$\text{(5.6b)}\qquad (c_n, \sigma_n^n) = \left[(b_{i_n}, s^n_{i_n}) - \sum_{j\in\tilde{J}^{n-1}} \langle a^{i_n}, q^j\rangle (c_j, \sigma_j^n)\right]/|\tilde{q}^n|,$$

$\tilde{d}^n = -Q_{\check{J}^n}\sigma^n_{\check{J}^n}$, and $\tilde{y}^n = \tilde{x}^n + \tilde{d}^n$, and choose $\tilde{t}_n \leq 1$. Here $Q_{\check{J}^n} = [Q_{\tilde{J}^{n-1}}, q^n]$, where $Q_{\tilde{J}^{n-1}}$ is the $N \times |\tilde{J}^{n-1}|$ orthonormal matrix corresponding to $\mathcal{Q}_{\tilde{J}^{n-1}}$.

To validate this modification, suppose $\mathcal{P} \neq \emptyset$, $Q_{\tilde{J}^{n-1}}$ is orthonormal, $\sigma^n_{\tilde{J}^{n-1}} = Q^T_{\tilde{J}^{n-1}}\tilde{x}^n - c_{\tilde{J}^{n-1}} \geq 0$, and $Q^T_{\tilde{J}^{n-1}}x \leq c_{\tilde{J}^{n-1}}$ is a surrogate of $A^Tx \leq b$, so that $\mathcal{P} \subset \mathcal{P}_n = \tilde{x}^n + \mathcal{D}_n$, where $\mathcal{D}_n = \{d : Q^T_{\tilde{J}^{n-1}}d \leq -\sigma^n_{\tilde{J}^{n-1}}, \langle a^{i_n}, d\rangle \leq -s^n_{i_n}\}$. Also $s^n_{i_n} > 0$ and $Q^T_{\tilde{J}^{n-1}}a^{i_n} \leq 0$ by construction. Hence, replacing $\mathcal{D}_{\mathcal{I}}$ by $\mathcal{D}_n \neq \emptyset$ in Lemma

5.3, we deduce that $\operatorname{rank}[Q_{\check{J}^{n-1}}, a^{i_n}] = \hat{m} := |\check{J}^n|$, so $\tilde{q}^n \neq 0$ and by construction $[Q_{\check{J}^{n-1}}, a^{i_n}] = Q_{\check{J}^n} R$, where $Q_{\check{J}^n}$ is orthonormal, $r_{jj} = 1$ and $r_{j\hat{m}} = \langle a^{i_n}, q^j \rangle \leq 0$, $j = 1{:}\hat{m}-1$, $r_{\hat{m}\hat{m}} = |\tilde{q}^n|$, and the remaining $r_{ij} = 0$. Then by (5.6), $\langle q^n, x \rangle \leq c_n$ is a surrogate of $Q_{\check{J}^{n-1}}^T x \leq c_{\check{J}^{n-1}}$, $\langle a^{i_n}, x \rangle \leq b_{i_n}$ (and hence of $A^T x \leq b$) with $\sigma_n^n = \langle q^n, \tilde{x}^n \rangle - c_n > 0$, and $R^{-T}[(\sigma^n_{\check{J}^{n-1}})^T, s^n_{i_n}]^T = \sigma^n_{\check{J}^n}$. Therefore, since $\tilde{d}^n = -Q_{\check{J}^n}\sigma^n_{\check{J}^n}$, we deduce from Lemma 5.3(v) applied to $\mathcal{D}_n$ that $Q^T_{\check{J}^n}\tilde{d}^n = -\sigma^n_{\check{J}^n}$,

$$\mathcal{D}_n \subset \tilde{\mathcal{D}}_n := \{\, d : Q^T_{\check{J}^n} d \leq -\sigma^n_{\check{J}^n} \,\} \quad \text{and} \quad \tilde{d}^n = P_{\mathcal{D}_n}(\tilde{x}^n) = P_{\tilde{\mathcal{D}}_n}(\tilde{x}^n). \tag{5.7}$$

Thus $\tilde{y}^n = P_{\mathcal{P}_n}(\tilde{x}^n) = P_{\tilde{\mathcal{P}}_n}(\tilde{x}^n)$, where $\mathcal{P}_n \subset \{x : \langle a^{i_n}, x \rangle \leq b_{i_n}\}$ and $\tilde{\mathcal{P}}_n = \tilde{x}^n + \tilde{\mathcal{D}}_n$. Using $Q^T_{\check{J}^n}\tilde{d}^n = -\sigma^n_{\check{J}^n}$, $\tilde{x}^{n+1} = \tilde{x}^n + \tilde{t}_n\tilde{d}^n$, and $\tilde{t}_n \leq 1$ gives $\sigma^{n+1}_{\check{J}^n} = Q^T_{\check{J}^n}\tilde{x}^{n+1} - c_{\check{J}^n} = \sigma^n_{\check{J}^n} + \tilde{t}_n Q^T_{\check{J}^n}\tilde{d}^n = (1-\tilde{t}_n)\sigma^n_{\check{J}^n} \geq 0$. Also $Q^T_{\check{J}^n} x \leq c_{\check{J}^n}$ is a surrogate of $A^T x \leq b$. Hence one may use induction to show that this OSP version shares the convergence properties of the original one.

Note that if $\tilde{t}_{n-1} = 1$, then $\sigma^n_{\check{J}^{n-1}} = 0$ and $\tilde{d}^n = -\sigma^n_n q^n = -s^n_{i_n} q^n/|\tilde{q}^n|$ as in Remark 5.8. Thus $\tilde{y}^n$ is the projection of $\tilde{x}^n$ on the "orthogonalized" surrogate $\langle q^n, x \rangle \leq c_n$. Also the preceding validation would go through if, to save work, we increased $c_{\check{J}^{n-1}}$ by $\sigma^n_{\check{J}^{n-1}} \neq 0$, i.e., enlarged $\mathcal{P}_n$ to get $\sigma^n_{\check{J}^{n-1}} = 0$ and $\tilde{d}^n = -s^n_{i_n} q^n/|\tilde{q}^n|$ as in Remark 5.7.

Step 3 could construct more than one surrogate. Specifically, if $s^n_i > 0$ and $Q^T_{\check{J}^n} a^i \leq 0$ for some i, then step 3 could append to $Q^T_{\check{J}^n} x \leq c_{\check{J}^n}$ another surrogate derived from $\langle a^i, x \rangle \leq b_i$ and $Q^T_{\check{J}^n} x \leq c_{\check{J}^n}$ as in (5.6), and this may be repeated for other violated constraints. Again, Lemma 5.3 validates this extension.

To improve accuracy, *iterative refinement* of the form $\tilde{y}^n \leftarrow \tilde{y}^n + Q_{\check{J}^n}(c_{\check{J}^n} - Q^T_{\check{J}^n}\tilde{y}^n)$ or $\tilde{d}^n \leftarrow \tilde{d}^n + Q_{\check{J}^n}(\sigma^n_{\check{J}^n} - Q^T_{\check{J}^n}\tilde{d}^n)$ may be used at some iterations, and q^n should be *reorthogonalized* with respect to $Q_{\check{J}^{n-1}}$; see, e.g., [Bjö90, DGKS76].

When the OSP procedure is called repeatedly with $\mathcal{P} = \mathcal{L}(\hat{f}^k, f^k_{\text{lev}})$, we may generate a vector $\check{c}_{\check{J}^n} > 0$ recursively via $\check{c}_n = (1 - \sum_{j \in \check{J}^{n-1}} \langle a^{i_n}, q^j \rangle \check{c}_j)/|\tilde{q}^n|$. Then, by induction on the surrogate construction (5.6), there exist $\nu_j \geq 0$ such that

$$(q^n, c_n) = \sum_{j=1}^n \nu_j (a^{i_j}, b_{i_j}) \quad \text{and} \quad \check{c}_n = \sum_{j=1}^n \nu_j \geq \nu_n = 1/|\tilde{q}^n| > 0, \tag{5.8}$$

so dividing by $\check{c}_n$ yields $(\langle q^n, \cdot \rangle - c_n)/\check{c}_n \in \operatorname{co}\{\langle a^i, \cdot \rangle - b_i\}_{i=1}^m \in \operatorname{co}\{f^j\}_{j \in J^k} - f^k_{\text{lev}}$, and hence $\operatorname{co}\{(q^j, c_j)/\check{c}_j\}_{j \in \check{J}^n} \in \operatorname{co}\{(a^i, b_i)\}_{i=1}^m$. Thus, when f^k_{lev} changes to f^{k+1}_{lev} on the next call then $(f^{k+1}_{\text{lev}} - f^k_{\text{lev}})\check{c}_{\check{J}^n}$ should be added to $c_{\check{J}^n}$ and $\sigma^n_{\check{J}^n}$. If $x^{k+1} \neq \tilde{x}^n$ (e.g., due to $x^{k+1} = P_S(\tilde{x}^n)$), then $\sigma^n_{\check{J}^n} = Q^T_{\check{J}^n}\tilde{x}^1 - c_{\check{J}^n}$ must be recomputed for $\tilde{x}^1 = x^{k+1}$. Alternatively, using $Q^T_{\check{J}^n} x - c_{\check{J}^n} = Q^T_{\check{J}^n}(x - \tilde{x}^n) + \sigma_{\check{J}^n}$, we may update $\sigma^n_{\check{J}^n} \leftarrow \sigma^n_{\check{J}^n} + Q^T_{\check{J}^n}(x^{k+1} - \tilde{x}^n)$ (then $c_{\check{J}^n}$ is not required). Next, dropping j with $\sigma^n_j < 0$ from $\check{J}^n$, if any, a hot start can proceed as if $\check{J}^0$ were $\check{J}^n$. Note that Remark 5.10 holds for the OSP version with A_{I^n} replaced by $[Q_{\check{J}^{n-1}}, a^{i_n}]$ (since $A^T x^* < b \Rightarrow Q^T_{\check{J}^{n-1}} x^* - c_{\check{J}^{n-1}} < 0 \leq \sigma^n_{\check{J}^{n-1}} = Q^T_{\check{J}^{n-1}}\tilde{x}^n - c_{\check{J}^{n-1}}$ by (5.8)).

Remark 5.11. By the preceding argument, the *surrogate linearizations*

$$\tilde{f}^j(\cdot) = (\langle q^j, \cdot \rangle - c_j)/\check{c}_j + f^k_{\text{lev}} = (\langle q^j, \cdot - \tilde{x}^n \rangle + \sigma^n_j)/\check{c}_j + f^k_{\text{lev}}, \quad j \in \check{J}^n, \tag{5.9}$$

satisfy $\tilde{f}^j \in \operatorname{co}\{f^j\}_{j \in J^k} \subset \Phi$. Hence they may be used as any other linearizations of f. For instance, no additional storage is required if, at step 6, $\tilde{f}^n$ replaces the

f^j corresponding to $\langle a^{i_n}, \cdot \rangle - b_{i_n}$. Also $\lambda_j^n = \check{c}_j \sigma_j^n$ are Lagrange multipliers for $\tilde{f}^j$, $j \in \check{J}^n$, since $\tilde{d}^n = P_{\check{\mathcal{D}}_n}(\tilde{x}^n)$ with $\check{\mathcal{D}}_n = \{d : \langle q^j/\check{c}_j, d \rangle \le -\sigma_j^n/\check{c}_j, j \in \check{J}^n\}$, whereas σ_j^n are Lagrange multipliers for $\tilde{d}^n = P_{\tilde{\mathcal{D}}_n}(\tilde{x}^n)$ in (5.7). Thus $\lambda_{\check{J}^{\bar{n}}}^{\bar{n}}$ can be used for selective aggregation as in Remark 5.9, and normalization of the aggregated column ensures orthonormality of the new $Q_{\check{J}^{\bar{n}}}$.

Remark 5.12. Consider the simplest case where step 6 always exits, so that we may let $n = k$ and $\tilde{x}^n = x^k$. Suppose $\mathcal{P} = \mathcal{L}(\phi^k, f_{\text{lev}}^k)$ with $\phi = \max\{f^k, \psi^{k-1}\}$ as for (4.2), where ψ^{k-1} is the previous aggregate satisfying $\psi^{k-1}(\tilde{x}^n) = f_{\text{lev}}^k$, so that a hot start occurs from $\check{J}^{n-1} = \{n-1\}$ using $\tilde{q}^{n-1} = g_\psi^{k-1}$, $q^{n-1} = \tilde{q}^{n-1}/|\tilde{q}^{n-1}|$, $c_{n-1} = 1/|\tilde{q}^{n-1}|$ and $\sigma_{n-1}^n = 0$. Assume $\langle g^k, g_\psi^{k-1} \rangle < 0$. By simple calculation, either infeasibility is detected at steps 4 or 5, or step 6 terminates with $\tilde{d}^n = -(f(x^k) - f_{\text{lev}}^k)\tilde{q}^n/|\tilde{q}^n|^2$, where $\tilde{q}^n = g^k - \langle g^k, g_\psi^{k-1} \rangle g_\psi^{k-1}/|g_\psi^{k-1}|^2$ has $|\tilde{q}^n|^2 = |g^k|^2 - \langle g^k, g_\psi^{k-1} \rangle^2/|g_\psi^{k-1}|^2$ (use $\tilde{d}^n = -\sigma_n^n q^n$, $\sigma_n^n = s_{i_n}^n/|\tilde{q}^n|$ and $s_{i_n}^n = f(x^k) - f_{\text{lev}}^k$). Also, since $\sigma_{n-1}^n = 0$, the aggregate ψ^k of f^k and ψ^{k-1} coincides with $\tilde{f}^n$. Note that for $\langle g^k, g_\psi^{k-1} \rangle \ge 0$ we would get $\tilde{q}^n = g^k$ (as if $\phi^k = f^k$), and that if we had $\psi^{k-1}(\tilde{x}^n) > f_{\text{lev}}^k$, then the same formulae would hold if we set $c_{n-1}^n = 0$. It is not surprising that *the same* $\tilde{d}^n$ and ψ^k would be produced via the original version of Procedure 5.4 restarted from $A_{I^{n-1}} \cdot - b_{I^{n-1}} \equiv \psi^{k-1} - f_{\text{lev}}^k$. (Use $\alpha_\psi^k = 0$ in (4.3)–(4.4) to get $\lambda_k^k = (f(x^k) - f_{\text{lev}}^k)/|\tilde{q}^n|^2$ and $d^k = -\lambda_k^k \tilde{q}^n = \tilde{d}^n$.) We add that $\psi^k(x^{k+1}) \ge f_{\text{lev}}^{k+1}$ if $t_k \le 1$ (cf. Lemma 6.1(v)).

We may add that the method of [Shc79, Shc92] corresponds to a version of Algorithm 2.1 that attempts to solve the inequality $f(x) \le 0$. It sets $f_{\text{lev}}^k = f_{\text{low}}^k = 0$ and finds $x^{k+1} = \tilde{x}^2$ via one iteration of a simplified OSP version of Procedure 5.4 that starts with $\tilde{x}^1 = x^k$ and $f^k(x) \le 0$ appended to the accumulated surrogates and exits at step 6 unless infeasibility is detected earlier, in which case it terminates. First, it sets $\tilde{t}_n = 1$ for all n, whereas our OSP version allows smaller stepsizes that may be useful at initial iterations. Second, it expresses $\tilde{d}^n$ as $\tilde{d}^n = -|\tilde{a}^n|^2 q^n/\langle \tilde{a}^n, q^n \rangle$, where $\tilde{a}^n = s_{i_n}^n a^{i_n}/|a^{i_n}|^2$, so $\tilde{d}^n = -s_{i_n}^n q^n/|\tilde{q}^n|$ from $\langle a^{i_n}, \tilde{q}^n \rangle = |\tilde{q}^n|^2$ by orthogonality. (In fact it replaces a^{i_n} by $\tilde{a}^n$ in calculating q^n, but this does not affect $Q_{\check{J}^n}$.) Third, it does not compute $c_{\check{J}^n}$ and $\check{c}_{\check{J}^n}$, thus preventing iterative refinement and hot starts that would be necessary for handling the additional constraint $x \in S$ (except when S is a flat [Shc87]). Fourth, both methods should cope with the instability of the Gram–Schmidt process. Periodic resets to $\check{J}^n = \{n\}$ recommended in [Shc87, Shc92] slow down convergence. It seems better to employ iterative refinement and reorthogonalization in computing q^n. To sum up, our method appears competitive with that of [Shc92]. Finally, note that such methods project on sets $\mathcal{P}_n$ that may or may not be larger than $\mathcal{P}_{I^n}$ with $I^n = \{i_j\}_{j \in \check{J}^n}$. Thus it is not clear whether our OSP version could compete with, e.g., a Cholesky-based implementation of Procedure 5.4. We note that encouraging numerical results were obtained in [Ceg92] by a method that combines greatly simplified versions of Algorithm 2.1 and Procedure 5.4 for solving a consistent inequality $f(x) \le 0$.

6. Conjugate subgradient techniques. In this section we use the dual framework of [Kiw96, §9] for extending some CS techniques; see, e.g., [Brä93, CFM75, KiA90, SaK87, SKR87, ShU89, Sho79].

First, we identify surrogate linearizations of f that may be generated via CS methods.

LEMMA 6.1. *Let* $0 < \mu \le 1$. *Suppose iteration* $k-1$ *provides an affine model* ψ^{k-1} *of* f_S *of the form* $\psi^{k-1}(\cdot) = \psi^{k-1}(x^k) + \langle g_\psi^{k-1}, \cdot - x^k \rangle$ *with* $\psi^{k-1}(x^k) \ge f_{\text{lev}}^k$ *such that*

$\psi^{k-1} \in \Phi$ *if* $f^k_{\text{lev}} \geq f^*$. *Let* $\check{\psi}^{k-1} = \psi^{k-1} + f^k_{\text{lev}} - \psi^{k-1}(x^k)$ *denote a shifted version of* ψ^{k-1} *such that* $\check{\psi}^{k-1}(x^k) = f^k_{\text{lev}}$. *The corresponding current models of* f_S *are given by* $\hat{\phi}^k = \max\{f^k, \psi^{k-1}\}$ *and* $\check{\phi}^k = \max\{f^k, \check{\psi}^{k-1}\}$. *Next, let* $\check{\phi}^k = f(x^k) + \langle \check{g}^k_\phi, \cdot - x^k \rangle = f^k + \beta_k \langle g^{k-1}_\psi, \cdot - x^k \rangle$ *be the current CS model of* f_S, *where* $\check{g}^k_\phi = g^k + \beta_k g^{k-1}_\psi$ *for* $\beta_k \geq 0$ *such that* $|\check{g}^k_\phi| \leq |g^k|/\mu$, *and let* $\tilde{\phi}^k = (f^k + \beta_k \psi^{k-1})/(1+\beta_k)$ *denote another CS-like model of* f_S *that is a convex combination of* f^k *and* ψ^{k-1}. *Finally, let* $\check{\beta}_k = \arg\min_{\beta \geq 0} |g^k + \beta g^{k-1}_\psi|$. *If* $f^k_{\text{lev}} \geq f^*$, *then*

(i) $\psi^{k-1}, \check{\psi}^{k-1} \in \Phi$, $\hat{\phi}^k, \check{\phi}^k \in \Phi^k_1$ *and* $\check{\phi}^k, \tilde{\phi}^k \in \Phi^k_\mu$ (*cf.* (2.3)). *In particular,*

$$d_{\mathcal{L}(\hat{\phi}^k, f^k_{\text{lev}})}(x^k) \geq d_{\mathcal{L}(\tilde{\phi}^k, f^k_{\text{lev}})}(x^k) = [f(x^k) - f^k_{\text{lev}} + \beta_k(\psi^{k-1}(x^k) - f^k_{\text{lev}})]/|\check{g}^k_\phi|$$
$$\geq d_{\mathcal{L}(\check{\phi}^k, f^k_{\text{lev}})}(x^k) = (f(x^k) - f^k_{\text{lev}})/|\check{g}^k_\phi| \geq \mu\kappa\Delta^k/L_f$$

and $d_{\mathcal{L}(\hat{\phi}^k, f^k_{\text{lev}})}(x^k) \geq d_{\mathcal{L}(\check{\phi}^k, f^k_{\text{lev}})}(x^k) \geq d_{\mathcal{L}(f^k, f^k_{\text{lev}})}(x^k)$.

(ii) $\check{\beta}_k = \langle g^k, -g^{k-1}_\psi \rangle_+ / |g^{k-1}_\psi|^2$. *Moreover,* $|\check{g}^k_\phi|^2 = |g^k|^2 - \langle g^k, g^{k-1}_\psi \rangle^2 / |g^{k-1}_\psi|^2$ *if* $\beta_k = \check{\beta}_k$ *and* $\langle g^k, g^{k-1}_\psi \rangle < 0$, $|\check{g}^k_\phi| \leq |g^k|$ *if* $\beta_k \leq 2\check{\beta}_k$, *and* $|\check{g}^k_\phi| \leq 2|g^k|$ *if* $\beta_k = |g^k|/|g^{k-1}_\psi|$.

(iii) $\beta_k \leq 2\check{\beta}_k \Rightarrow d_{\mathcal{L}(\check{\phi}^k, f^k_{\text{lev}})}(x^k) \geq d_{\mathcal{L}(f^k, f^k_{\text{lev}})}(x^k)$; $\beta_k < 2\check{\beta}_k \Rightarrow d_{\mathcal{L}(\check{\phi}^k, f^k_{\text{lev}})}(x^k) > d_{\mathcal{L}(f^k, f^k_{\text{lev}})}(x^k)$.

(iv) *If* $\langle g^k, g^{k-1}_\psi \rangle < 0$ *and* $\beta_k = \check{\beta}_k$, *then* $P_{\mathcal{L}(\check{\phi}^k, f^k_{\text{lev}})}(x^k) = P_{\mathcal{L}(\tilde{\phi}^k, f^k_{\text{lev}})}(x^k)$.

(v) $\phi^k(x^{k+1}) \geq f^k_{\text{lev}}$ *if* $t_k \leq 1$ *and* $\phi^k(x^k) > f^k_{\text{lev}}$; *e.g.,* $\phi^k = \check{\phi}^k$, $\tilde{\phi}^k$, $\hat{\phi}^k$, *or* $\check{\phi}^k$ *at step* 4.

Conversely, if $\phi^k = \check{\phi}^k$, $\tilde{\phi}^k$, $\hat{\phi}^k$, *or* $\check{\phi}^k$ *and* $\mathcal{L}(\phi^k, f^k_{\text{lev}}) = \emptyset$, *then* $f^k_{\text{lev}} < f^*$.

Proof. (i) Let $x^* \in S^*$. Since $\psi^{k-1}(x^k) + \langle g^{k-1}_\psi, x^* - x^k \rangle = \psi^{k-1}(x^*) \leq f^*$ and $f^k(x^*) \leq f^*$, we have $\langle g^{k-1}_\psi, x^* - x^k \rangle \leq f^* - \psi^{k-1}(x^k) \leq 0$ and $\check{\phi}^k(x^*) = f^k(x^*) + \beta_k \langle g^{k-1}_\psi, x^* - x^k \rangle \leq f^*$. Next, $d_{\mathcal{L}(\check{\phi}^k, f^k_{\text{lev}})}(x^k) = (f(x^k) - f^k_{\text{lev}})/|\check{g}^k_\phi| \geq \kappa\Delta^k/(|g^k|/\mu) \geq \mu\kappa\Delta^k/L_f$, while $\tilde{\phi}^k(x) = \tilde{\phi}^k(x^k) + \langle \tilde{g}^k_\phi, x - x^k \rangle$ with $\tilde{\phi}^k(x^k) - f^k_{\text{lev}} = [f^k(x^k) - f^k_{\text{lev}} + \beta_k(\psi^{k-1}(x^k) - f^k_{\text{lev}})/(1+\beta_k)$ and $\tilde{g}^k_\phi = (g^k + \beta_k g^{k-1}_\psi)/(1+\beta_k)$ yield $d_{\mathcal{L}(\tilde{\phi}^k, f^k_{\text{lev}})}(x^k) = [f(x^k) - f^k_{\text{lev}} + \beta_k(f(x^k) - f^k_{\text{lev}})]/|\check{g}^k_\phi|$, so the conclusion follows from $\check{\psi}^{k-1} \leq \psi^{k-1}$, $f^k \leq \check{\phi}^k \leq \hat{\phi}^k$, and $\tilde{\phi}^k \leq \check{\phi}^k$.

(ii) Solve $\min_{\beta \geq 0} |g^k + \beta g^{k-1}_\psi|^2$, and use $|\check{g}^k| \leq |g^k| + \beta_k |g^{k-1}_\psi|$.

(iii) Invoke (i) with $|\check{g}^k| \leq |g^k|$ and $d_{\mathcal{L}(f^k, f^k_{\text{lev}})}(x^k) = (f(x^k) - f^k_{\text{lev}})/|g^k|$.

(iv) Use $|\check{g}^k_\phi|^2 = |g^k|^2 - \langle g^k, g^{k-1}_\psi \rangle^2 / |g^{k-1}_\psi|^2$ and $\alpha^k_\psi = 0$ in (4.3)–(4.4) to get $\lambda^k_k = (f(x^k) - f^k_{\text{lev}})/|\check{g}^k_\phi|^2$ and $d^k = -\lambda^k_k \check{g}^k_\phi = P_{\mathcal{L}(\check{\phi}^k, f^k_{\text{lev}})}(x^k) - x^k$ with $J^k = \{k\}$ in (4.2).

(v) Using (1.4), $x^{k+1} = P_S(z^k)$, $x^k \in S$ and $t_k \leq 1$, we get $|x^{k+1} - x^k| \leq |z^k - x^k| = t_k |y^k - x^k| \leq |y^k - x^k|$. But $y^k = P_{\mathcal{L}(\phi^k, f^k_{\text{lev}})}(x^k)$ with $\phi^k(x^k) > f^k_{\text{lev}}$, so $f^k_{\text{lev}} = \phi^k(y^k) \leq \phi^k(x^{k+1})$. □

Lemma 6.1 suggests the following *CS implementation* of Algorithm 2.1. Let $0 < \mu \leq 1$ and $\phi^1 = f^1$. At iteration $k \geq 2$, let $\psi^{k-1} = \phi^{k-1}$ and $\phi^k = \check{\phi}^k$ with $\beta_k \geq 0$ such that $|\check{g}^k_\phi| \leq |g^k|/\mu$ if $\psi^{k-1}(x^k) \geq f^k_{\text{lev}}$, and $\beta_k = 0$ ($\phi^k = f^k$) otherwise. Then by induction, as in [Kiw96, §9], we see that only the first terms of the constants in all the efficiency estimates and the right side of (7.3) in [Kiw96] need be divided by μ^2, with $\mu = 1$ if $\beta_k \leq 2\check{\beta}_k$ $\forall k$; of course, Δ^k_{lev} replaces $\kappa\Delta^k$ in Lemma 6.1 for the frozen level gaps of [Kiw96, §7].

Note that by construction, $d^k = -(f(x^k) - f^k_{\rm lev})\breve{g}^k_\phi/|\breve{g}^k_\phi|^2$, and if $d^{k-1} \neq 0$, then $d^{k-1}/|d^{k-1}| = -g^{k-1}_\psi/|g^{k-1}_\psi|$, so if $\beta_k = \breve{\beta}_k$ and $\langle g^k, d^{k-1}\rangle > 0$, then $\breve{g}^k_\phi = g^k - \langle g^k, d^{k-1}\rangle d^{k-1}/|d^{k-1}|^2$ and $\langle d^k, d^{k-1}\rangle = 0$. These CS relations correspond to those of the methods in [Brä93, CFM75, KiA90, ShU89, Sho79], which set $f^k_{\rm lev} = f^*$ and $t_k \leq 1$. Incidentally, when $t_k \leq 1$ and $f^{k+1}_{\rm lev} \leq f^k_{\rm lev}$, then $\phi^k(x^{k+1}) \geq f^{k+1}_{\rm lev}$ by Lemma 6.1(v), so such methods can skip computing $\psi^{k-1}(x^k)$ $(\geq f^*)$; this is the main reason for choosing $t_k \leq 1$. Usually $\beta_k \leq 2\breve{\beta}_k$ is advocated; with the choice of $\beta_k = |g^k|/|g^{k-1}_\psi|$ from [ShU89], the direction d^k simply bisects the angle between $-g^k$ and d^{k-1} and in this sense is an *average direction*. Since $|y - y^k|^2 \leq |y - x^k|^2 - d^2_{\mathcal{L}(\phi^k, f^k_{\rm lev})}(x^k)$ if $y \in S^*$, $y^k = P_{\mathcal{L}(\phi^k, f^k_{\rm lev})}(x^k)$, and $f^k_{\rm lev} \geq f^*$, Lemma 6.1(iii) augments the usual angle-based motivation for using $\breve{\phi}^k$ instead of f^k, while Lemma 6.1(iv) complements results in [KiA90]. In particular, the CS implementation with $\beta_k = \breve{\beta}_k$ corresponds to the simplest OSP implementation of Remark 5.12 with ψ^{k-1} replaced by $\breve{\psi}^{k-1}$ to zero its σ^n_{n-1} and $\breve{g}^k_\phi = \tilde{q}^n$ obtained by orthogonalizing g^k and g^{k-1}_ψ.

Lemma 6.1 says that we may easily improve classical CS techniques by taking $\phi^k = \breve{\phi}^k$, $\tilde{\phi}^k$, or $\hat{\phi}^k$ instead of $\phi^k = \breve{\phi}^k$ to increase $d_{\mathcal{L}(\phi^k, f^k_{\rm lev})}(x^k)$; cf. [Kiw96, Rem. 3.9]. In particular, $\tilde{\phi}^k$ is a convex combination of f^k and ψ^{k-1}, and other such combinations could be developed as in [Kiw96, §9]. It seems, however, that $\hat{\phi}^k = \max\{f^k, \psi^{k-1}\}$ is preferable anyway. First, Lemma 6.1 and [Kiw96, Lem. 9.4] show that $\hat{\phi}^k$ is best in terms of efficiency estimates. Second, the resulting choice of $\psi^{k-1} = \phi^{k-1}$ and $\hat{\phi}^k = \max\{f^k, \phi^{k-1}\}$ corresponds to the aggregate subgradient implementation of §4, which, in contrast to the other CS choices, does not require $t_k \leq 1$ and does not need to resort to the poorest model $\phi^k = f^k$ when $\phi^{k-1}(x^k) < f^k_{\rm lev}$ or $\langle g^k, g^{k-1}_\psi\rangle \geq 0$. Third, it involves little additional work (cf. (4.3)). Last, but not least, it is simpler conceptually. Incidentally, the choices of $\phi^k = \breve{\phi}^k$ and $\phi^k = \hat{\phi}^k$ may be compared in dual terms by noting that $\breve{\beta}_k = \arg\max_{\beta \geq 0}(f(x^k) - f^k_{\rm lev})/|g^k + \beta g^{k-1}_\psi|$ and $(\lambda^k_k, \lambda^k_\psi) \in \operatorname{Arg\,max}\{[\lambda_k(f(x^k) - f^k_{\rm lev}) + \lambda_\psi(\psi^{k-1}(x^k) - f^k_{\rm lev})]/|\lambda_k g^k + \lambda_\psi g^{k-1}_\psi| : \lambda_k \geq 0, \lambda_\psi \geq 0, \lambda_k + \lambda_\psi = 1\}$, where the second maximum $(= d_{\mathcal{L}(\hat{\phi}^k, f^k_{\rm lev})}(x^k))$ can be much greater.

An obvious extension of the CS techniques is to take $\phi^k = \max\{\hat{f}^k, \breve{\phi}^k\}$ or $\max\{\hat{f}^k, \tilde{\phi}^k\}$ to increase $d_{\mathcal{L}(\phi^k, f^k_{\rm lev})}(x^k)$. Further, more than one CS step can be made as in Remark 4.2.

7. Constraint modelling. Since Algorithm 2.1 minimizes f on S, ϕ^k should be chosen to model the extended objective $f_S = f + \delta_S$ and not just f alone, where $\delta_S(x) = 0$ if $x \in S$, $\delta_S(x) = \infty$ if $x \notin S$. Failure to do so may result in severe deficiencies, as shown in the following simple example.

Example 7.1. Let $N = 2$, $S = [0,1] \times [0,1]$, and $f(x) = \epsilon x_1 + x_2$, where $0 < \epsilon < 1$ is a small parameter. Then $S^* = \{(0,0)\}$, $f^* = 0$, $\operatorname{diam}(S) = \sqrt{2}$, and $L_f = \sqrt{1+\epsilon^2}$. Let $x^1 = (1,0)$ and $t_{\min} = t_{\max} = 1$. The following facts are easy to verify by induction. The SPA (1.2) generates $x^k = ((1+\epsilon^2)^{1-k}, 0)$ and $f(x^k) = \epsilon(1+\epsilon^2)^{1-k}$ $\forall k$; i.e., its convergence is linear but very slow for small ϵ. The situation is even slightly worse for the SPLA (2.2) with $f^1_{\rm low} = f^*$, which yields $x^k = ([1 - \kappa\epsilon^2]/(1+\epsilon^2)^{k-1}, 0)$ $\forall k$. In contrast, Algorithm 2.1 with $f^1_{\rm low} = f^*$, $\bar{D} \geq \sqrt{2}$ and $\phi^k \equiv f^k + \delta_S$ gives $x^k = ((1-\kappa)^{k-1}, 0)$ $\forall k$; i.e., it is much faster for typical κ, independent of ϵ. In fact for $\kappa = 1$ (cf. [Kiw96, Thm 6.1]) it terminates with $x^2 \in S^*$, being equivalent to the iteration (1.3).

Example 7.1 and [Kiw96, Rem. 3.9] suggest that the following modification of

(2.2),

$$x^{k+1} = \arg\min\{\,|x - x^k|^2/2 : f(x^k) + \langle g_f(x^k), x - x^k\rangle \le f^k_{\text{lev}}, x \in S\,\}, \tag{7.1}$$

should be more efficient in practice. Supposing S is a box of the form $[x^{\text{low}}, x^{\text{up}}]$, let $x(\nu) = \arg\min_{x\in S}\{|x - x^k|^2/2 + \nu\langle g^k, x\rangle\}\ \forall \nu \ge 0$. Then $x^{k+1} = x(\hat\nu)$, where $\hat\nu \ge 0$ solves the equation $h(\nu) \equiv f(x^k) + \langle g^k, x(\nu) - x^k\rangle - f^k_{\text{lev}} = 0$. Since $x_i(\nu) = \max\{\min[x_i^{\text{low}}, x_i^k - \nu g_i^k], x_i^{\text{up}}\}$, $i = 1{:}N$, and h is nonincreasing and piecewise linear, $\hat\nu$ is easy to compute.

Of course, projecting on S is easy only if S is simple enough, e.g., a Cartesian product of boxes, simplices, balls, ellipsoids, cylinders, etc. Additional linear constraints may complicate the projections; e.g., for $\phi^k = \hat f^k + \delta_S$ we must find

$$y^k = \arg\min\{\,|x - x^k|^2/2 : f^j(x) \le f^k_{\text{lev}}, j \in J^k, x \in S\,\}. \tag{7.2}$$

Fortunately, accurate projections are not really necessary. For instance, (7.2) can be implemented approximately by projecting *cyclically* on $\mathcal{L}(\hat f^k, f^k_{\text{lev}})$ and S as in Remark 4.2, possibly with inexact projections on $\mathcal{L}(\hat f^k, f^k_{\text{lev}})$ being performed via the methods of §§4, 5 and 6. Also if S is polyhedral then (7.2) is just (4.1) augmented with the inequalities of S, so it can be solved approximately by several steps of these methods.

It is crucial to observe that even if S is not polyhedral, it may still be *linearized* via inequalities generated in the course of calculations. First, such inequalities may be recovered *geometrically* by noting that $S \subset \mathcal{H}^k := \{x : \langle z^{k-1} - x^k, x - x^k\rangle \le 0\}$ from $x^k = P_S(z^{k-1})$. Hence we may replace S in (7.2) by $S^k = \cap_{j\in J^k_S}\mathcal{H}^j$ with $J^k_S \subset \{2{:}k\}$. Second, similar inequalities may be generated *analytically* if $S = \{x : F(x) \le 0\}$ say, where $F : \mathbb{R}^N \to \mathbb{R}$ is convex, and we can find its linearization $\bar F(\cdot; x) = F(x) + \langle g_F(x), \cdot - x\rangle$ with $g_F(x) \in \partial F(x)$ for any x. Then we may use $\mathcal{H}^k = \{x : \bar F(x; x^k) \le 0\}$ as above. In other words, we may accumulate $\mathcal{H}^j$ for the model δ_{S^k} of δ_S in the same way as we use f^j in the model $\hat f^k$ of f, so that $f + \delta_S$ is approximated by $\phi^k = \hat f^k + \delta_{S^k}$, with $\delta_{S_k} \le \delta_S$ from $S \subset S^k$. To save storage, some of the inequalities defining S^k may be aggregated as in §4.

The following observation is useful for the dual models of §2. If we take $\hat\phi^k = \hat f^k + \delta_S$ for a polyhedral $S = \{x : \langle a^i, x\rangle \le b_i, i = 1{:}m\}$, say, then $\partial_{\epsilon_k}\hat\phi^k(x^k) = \{\partial_{\epsilon'}\hat f^k(x^k) + \partial_{\epsilon''}\delta_S(x^k) : \epsilon' + \epsilon'' \le \epsilon_k, \epsilon', \epsilon'' \ge 0\}$ with

$$\partial_\epsilon\delta_S(x^k) = \left\{\sum_{i=1}^m \nu_i a^i : \sum_{i=1}^m \nu_i(b_i - \langle a^i, x^k\rangle) \le \epsilon, \nu_i \ge 0, i = 1{:}m\right\},$$

so $p^k = \arg\min\{|p|^2/2 : p \in \partial_{\epsilon_k}\hat\phi^k(x^k)\}$ of (2.4) can again be found via QP using (2.5). Next, letting $I^k = \{i : \langle a^i, x^k\rangle = b_i\}$ and $\hat S^k = \{x : \langle a^i, x\rangle \le b_i, i \in I^k\}$, consider the simpler model $\hat\phi^k = f^k + \delta_{\hat S^k}$. Clearly, $\hat\phi^k \le f_S$, $\hat\phi^k(x^k) = f(x^k)$ and $\partial_{\epsilon_k}\hat\phi^k(x^k) = g^k + \partial\delta_{\hat S^k}(x^k)$ for *any* $\epsilon_k \ge 0$; i.e., p^k does not depend on ϵ_k. Hence by [Kiw96, Lem. 9.4], $\epsilon_k = 0$ gives the "optimal" dual method with $d^k = y^k - x^k = -(f(x^k) - f^k_{\text{lev}})p^k/|p^k|^2 = P_{\mathcal{L}(\hat\phi^k, f^k_{\text{lev}})}(x^k) - x^k$ if $p^k \ne 0$; otherwise $f^k_{\text{lev}} < f^*$ by [Kiw96, Lem. 9.2]. (Thus it is not suprising that this dual method behaves like the primal one in Example 7.1.) Additional insight may be gained as follows. The cones $C = \text{cone}\{a^i\}_{i\in I^k}$ and $C^- = \{x : \langle a^i, x\rangle \le 0, i \in I^k\}$ provide the classical orthogonal decomposition

$$g = P_C(g) + P_{C^-}(g), \quad P_C(g) \perp P_{C^-}(g) \quad \forall g \in \mathbb{R}^N,$$

since $P_{C^-}(g) = \arg\min\{|x-g|^2/2 : A_{I^k}^T x \le 0\}$ has multipliers

$$\lambda_{I^k} \in \operatorname{Arg\,min}\{|A_{I^k}\lambda_{I^k} - g|^2/2 : \lambda_{I^k} \ge 0\}$$

satisfying $P_C(g) = A_{I^k}\lambda_{I^k}$ and $\lambda_{I^k}^T A_{I^k}^T P_{C^-}(g) = 0$ by the KKT conditions, so for $g = -g^k$,

$$-p^k = -g^k - P_{\mathcal{N}_S(x^k)}(-g^k) = P_{\mathcal{T}_S(x^k)}(-g^k), \tag{7.3}$$

where $\mathcal{N}_S(x^k) = \partial\delta_S(x^k) = C$ and $\mathcal{T}_S(x^k) = C^-$ are the normal and tangent cones to S at x^k, respectively. Hence $-p^k$ and d^k are feasible directions for S at x^k. Thus if t_k is sufficiently small, then $z^k = x^k + t_k d^k \in S$, so that one may take $x^{k+1} = z^k$, skipping its projection on S. This motivates a similar technique in [KiU89] (with $f_{\text{lev}}^k = f^*$), but small stepsizes may yield slow convergence.

Of course, if S is not polyhedral, then the preceding construction may employ its accumulated approximation S^k. A simple but useful example is given in the following lemma.

LEMMA 7.2. *Suppose $\mathcal{H}^k = \{x : \langle a^k, x - x^k\rangle \le 0\}$ is a nontrivial outer approximation to S at x^k; e.g., $a^k = z^{k-1} - x^k \ne 0$. Let $\hat\gamma_k = \arg\min_{\gamma\ge 0}|g^k + \gamma a^k|$, choose $0 \le \gamma_k \le 2\hat\gamma_k$, and let $g_S^k = g^k + \gamma_k a^k$. Then $f_S^k(\cdot) = f^k(\cdot) + \gamma_k\langle a^k, \cdot - x^k\rangle = f(x^k) + \langle g_S^k, \cdot - x^k\rangle$ is a valid linearization of f_S, i.e., $f_S^k \in \Phi_1^k$, satisfying $f_S^k(x^*) \le f^* \ \forall x^* \in S^*$, and $d_{\mathcal{L}(f_S^k, f_{\text{lev}}^k)}(x^k) \ge d_{\mathcal{L}(f^k, f_{\text{lev}}^k)}(x^k)$, with strict inequality if $0 < \gamma_k < 2\hat\gamma_k$. In particular, if $\langle g^k, a^k\rangle < 0$ and $\gamma_k = \hat\gamma_k = -\langle g^k, a^k\rangle/|a^k|^2$, then $g_S^k = g^k - \langle g^k, a^k\rangle a^k/|a^k|^2 = -P_{\mathcal{H}^k}(-g^k)$.*

Proof. Clearly, $f_S^k(x^k) = f(x^k)$ and $f^k(x) \ge f_S^k(x)\ \forall x \in S \subset \mathcal{H}^k$, so $f_S^k(x^*) \le f^*$ if $x^* \in S^*$. As for the rest, solve $\min_{\gamma\ge 0}|g^k + \gamma a^k|^2$ and use $d_{\mathcal{L}(f_S^k, f_{\text{lev}}^k)}(x^k) = (f(x^k) - f_{\text{lev}}^k)/|g_S^k|$. □

In view of Lemma 7.2, we may replace g^k with $g_S^k \in \partial f_S(x^k)$, a conditional subgradient of f on S. In general, f_S^k is a worse model of f_S than $f^k + \delta_{\mathcal{H}^k}$, but it may be easier to handle.

We now extend the *average direction strategy* (ADS) of [ShU89].

LEMMA 7.3. *Suppose $t_k \le 1$ and $\phi^k = \breve\phi^k \in \Phi_\mu^k$, where $\breve\phi^k = f(x^k) + \langle \breve g_\phi^k, \cdot - x^k\rangle$ is the CS model of Lemma 6.1. Let $\psi^k = \breve\phi^k(x^{k+1}) + \langle g_\psi^k, \cdot - x^{k+1}\rangle$, where $g_\psi^k = \breve g_\phi^k + \gamma_k(z^k - x^{k+1})$ for some $\gamma_k \ge 0$. Then $\psi^k(x^{k+1}) \ge f_{\text{lev}}^k$ and $\psi^k \in \Phi$ if $f_{\text{lev}}^k \ge f^*$. Moreover, if $\gamma_k = \breve\gamma_k := |\breve g_\phi^k|^2/(\breve\phi^k(x^k) - f_{\text{lev}}^k)t_k$, then $\breve d^k = -g_\psi^k/\gamma_k$, where $\breve d^k = x^{k+1} - x^k$ is the actual direction of motion which includes the effect of the projection operation. In particular, if $\breve d^k \ne 0$ and ψ^k is used to define the next $\phi^{k+1} = \breve\phi^{k+1}$ with $\beta_{k+1} = |g^{k+1}|/|g_\psi^{k+1}|$, then $\breve g_\phi^{k+1} = g^{k+1} - |g^{k+1}|\breve d^k/|\breve d^k|$; i.e., the move $y^{k+1} = x^{k+1} - (f(x^{k+1}) - f_{\text{lev}}^{k+1})\breve g_\phi^{k+1}/|\breve g_\phi^{k+1}|^2$ occurs along the average direction of $-g^{k+1}$ and $\breve d^k$.*

Proof. If $f_{\text{lev}}^k \ge f^*$ and $x^* \in S^*$, then $\breve\phi^k(x^{k+1}) + \langle \breve g_\phi^k, x^* - x^{k+1}\rangle = \breve\phi^k(x^*) \le f^*$ by (2.3) and $\langle z^k - x^{k+1}, x^* - x^{k+1}\rangle \le 0$ because $x^{k+1} = P_S(z^k)$ and $x^* \in S$, so $\psi^k(x^*) \le f^*$, while $\psi^k(x^{k+1}) = \phi^k(x^{k+1}) \ge f_{\text{lev}}^k$ by Lemma 6.1(v). Next, suppose $\gamma_k = \breve\gamma_k$. Then $z^k - x^k = -t_k(\breve\phi^k(x^k) - f_{\text{lev}}^k)\breve g_\phi^k/|\breve g_\phi^k|^2$ yields $-g_\psi^k/\gamma_k = z^k - x^k - (z^k - x^{k+1}) = \breve d^k$. The rest follows by construction. □

Lemma 7.3 suggests the following *ADS version* of the CS implementation from §6. Let $0 < \mu \le 1$. At iteration k, let $\phi^1 = f^1$ if $k = 1$, otherwise use ψ^{k-1} to find $\phi^k = \breve\phi^k$ with $\beta_k \ge 0$ such that $|\breve g_\phi^k| \le |g^k|/\mu$ if $\psi^{k-1}(x^k) \ge f_{\text{lev}}^k$, and $\beta_k = 0$

otherwise; in both cases choose $\gamma_k \geq 0$ to construct ψ^k as in Lemma 7.3. In other words, instead of using $\psi^{k-1} = \phi^{k-1}$, the ADS version modifies ψ^{k-1} to include the effect of the projection operation. Clearly, the ADS version shares the efficiency estimates of the CS one's. On the other hand, the ADS version of a similar method in [ShU89] (with $\gamma_k \equiv \check{\gamma}_k$) performed better than the standard CS version of [CFM75] (corresponding to $\beta_k = \check{\beta}_k$ in Lemma 6.1). (By the way, the ADS version was not validated theoretically in [ShU89].) We note, however, that the arguments of §6 that favor $\hat{\phi}^k = \max\{f^k, \psi^{k-1}\}$ versus $\check{\phi}^k$ hold also for this modified form of ψ^{k-1}.

8. Conclusions. We have extended various acceleration techniques for subgradient methods, such as surrogate constraints, deepest surrogate cuts, simultaneous projections, orthogonal surrogate projections, CSs, and projected (conditional) subgradients. We have also proposed to use subgradient aggregation and parallel projection methods for implementing our methods in the large-scale case.

Of course, some of our ideas have been inspired by other popular approaches [AHKS87, BaS81, HWC74, KKA87, SeS86, ShM88, ShU89] and may in turn be used to modify the methods given in these papers. For example, the concept of relying on subgradient aggregation to provide some "conjugacy" (cf. §§4, 6, and 7) would enable the method of [ShU89] to use "deeper cuts," thus enabling faster convergence. We hope, therefore, that this paper will contribute to the development of other subgradient methods.

Acknowledgments. This work was started during my six-month stay at INRIA, Rocquencourt, in 1992, owing to the kind invitation by C. Lemaréchal and the French Ministry for Research and Technology, whose financial support is gratefully acknowledged. I would also like to thank the two anonymous referees for their valuable comments.

REFERENCES

[Agm54] S. AGMON, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 382–392.

[AhC89] R. AHARONI AND Y. CENSOR, *Block-iterative projection methods for parallel computation of solutions to convex feasibility problems*, Linear Algebra Appl., 120 (1989), pp. 165–175.

[AHKS87] E. ALLEN, R. HELGASON, J. KENNINGTON, AND B. SHETTY, *A generalization of Polyak's convergence result for subgradient optimization*, Math. Programming, 37 (1987), pp. 309–317.

[BaG79] M. S. BAZARAA AND J. J. GOODE, *A survey of various tactics for generating Lagrangian multipliers in the context of Lagrangian duality*, European J. Oper. Res., 3 (1979), pp. 322–338.

[BaS81] M. S. BAZARAA AND H. D. SHERALI, *On the choice of step size in subgradient optimization*, European J. Oper. Res., 7 (1981), pp. 380–388.

[BeT89] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[BGT81] R. B. BLAND, D. GOLDFARB, AND M. J. TODD, *The ellipsoid method: A survey*, Oper. Res., 29 (1981), pp. 1039–1091.

[Bjö90] Å. BJÖRCK, *Least squares methods*, in Finite Difference Methods (Part 1) — Solution of Equations in R^n (Part 1), Handbook of Numerical Analysis, vol. I, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 1990, pp. 465–652.

[Brä93] U. BRÄNNLUND, *On Relaxation Methods for Nonsmooth Convex Optimization*, Ph.D. thesis, Department of Mathematics, Royal Institute of Technology, Stockholm, 1993.

[Ceg92] A. CEGIELSKI, *On a relaxation algorithm in convex optimization problems*, tech. report, Institute of Mathematics, Higher College of Engineering, Zielona Góra, Poland, 1992.

[CFM75] P. M. CAMERINI, L. FRATTA, AND F. MAFFIOLI, *On improving relaxation methods by modified gradient techniques*, Math. Programming Stud., 3 (1975), pp. 26–34.

[DGKS76] J. W. DANIEL, W. B. GRAGG, L. C. KAUFMAN, AND G. W. STEWART, *Reorthogonalization and stable algorithms for updating the Gram-Schmidt QR factorization*, Math. Comp., 30 (1976), pp. 772–795.

[Fle87] R. FLETCHER, *Practical Methods of Optimization*, 2nd ed., Wiley, Chichester, 1987.

[GMW91] P. E. GILL, W. MURRAY, AND M. H. WRIGHT, *Numerical Linear Algebra and Optimization*, Addison-Wesley, Redwood City, CA, 1991.

[Gof78] J.-L. GOFFIN, *Nondifferentiable optimization and the relaxation method*, in Nonsmooth Optimization, C. Lemaréchal and R. Mifflin, eds., Pergamon Press, Oxford, 1978, pp. 31–49.

[Gof81] ———, *Convergence results in a class of variable metric subgradient methods*, in Nonlinear Programming 4, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1981, pp. 283–326.

[GoT82] D. GOLDFARB AND M. J. TODD, *Modifications and implementation of the ellipsoid algorithm for linear programming*, Math. Programming, 23 (1982), pp. 1–19.

[GVL89] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, Maryland, 1989.

[HUL93] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

[HWC74] M. HELD, P. WOLFE, AND H. P. CROWDER, *Validation of subgradient optimization*, Math. Programming, 6 (1974), pp. 62–88.

[IDP91] A. N. IUSEM AND A. R. DE PIERRO, *On the convergence of Han's method for convex programming with quadratic objective*, Math. Programming, 52 (1991), pp. 265–284.

[KAC91] S. KIM, H. AHN, AND S.-C. CHO, *Variable target value subgradient method*, Math. Programming, 49 (1991), pp. 359–369.

[KiA90] S. KIM AND H. AHN, *Convergence properties of the modified subgradient method of Camerini et al.*, Naval Res. Logist. Quart., 37 (1990), pp. 961–966.

[KiU89] S. KIM AND B.-S. UM, *Polyak's subgradient method with simplified projection for nondifferentiable optimization with linear constraints*, Optimization, 20 (1989), pp. 451–456.

[KiU93] ———, *An improved subgradient method for constrained nondifferentiable optimization*, Oper. Res. Lett., 14 (1993), pp. 61–64.

[Kiw85] K. C. KIWIEL, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Mathematics 1133, Springer-Verlag, Berlin, 1985.

[Kiw89] ———, *A survey of bundle methods for nondifferentiable optimization*, in Mathematical Programming: Recent Developments and Applications, M. Iri and K. Tanabe, eds., KTT/Kluwer, Tokyo, 1989, pp. 263–282.

[Kiw90] ———, *Proximity control in bundle methods for convex nondifferentiable minimization*, Math. Programming, 46 (1990), pp. 105–122.

[Kiw93] ———, *The efficiency of subgradient projection methods for convex nondifferentiable optimization*, Research report 1845, INRIA, Rocquencourt, 1993.

[Kiw95] ———, *Block-iterative surrogate projection methods for convex feasibility problems*, Linear Algebra Appl., 215 (1995), pp. 225–259.

[Kiw96] ———, *The efficiency of subgradient projection methods for convex optimization, part I: General level methods*, SIAM J. Control Optim., 34 (1996), pp. 660–676.

[KKA87] S. KIM, S. KOH, AND H. AHN, *Two-direction subgradient method for non-differentiable optimization problems*, Oper. Res. Lett., 6 (1987), pp. 43–46.

[KuF90] A. N. KULIKOV AND V. R. FAZILOV, *Convex optimization with prescribed accuracy*, Zh. Vychisl. Mat. i Mat. Fiz., 30 (1990), pp. 663–671. (In Russian.)

[Lem89] C. LEMARÉCHAL, *Nondifferentiable optimization*, in Optimization, Handbooks in Operations Research and Management Science, vol. 1, G. L. Nemhauser, A. H. G. Rinnooy-Kan, and M. J. Todd, eds., North-Holland, Amsterdam, 1989, pp. 529–572.

[LNN95] C. LEMARÉCHAL, A. S. NEMIROVSKII, AND YU. E. NESTEROV, *New variants of bundle methods*, Math. Programming, 69 (1995), pp. 111–147.

[LoH88] G. LOU AND S.-P. HAN, *A parallel projection method for solving generalized linear least-squares problems*, Numer. Math., 53 (1988), pp. 255–264.

[MoS54] T. MOTZKIN AND I. J. SCHOENBERG, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 393–404.

[Oko92] S. O. OKO, *Surrogate methods for linear inequalities*, J. Optim. Theory Appl., 72

(1992), pp. 247–268.

[Pol69] B. T. POLYAK, *Minimization of unsmooth functionals*, Zh. Vychisl. Mat. i Mat. Fiz., 9 (1969), pp. 509–521. (In Russian.) English transl. in U.S.S.R. Comput. Math. and Math. Phys. 9 (1969), pp. 14–29.

[SaK87] S. SARIN AND M. H. KARWAN, *Computational evaluation of two subgradient search methods*, Comput. Oper. Res., 14 (1987), pp. 241–247.

[ScZ92] H. SCHRAMM AND J. ZOWE, *A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results*, SIAM J. Optim., 2 (1992), pp. 121–152.

[SeS86] S. SEN AND H. D. SHERALI, *A class of convergent primal-dual subgradient algorithms for decomposable convex programs*, Math. Programming, 35 (1986), pp. 279–297.

[Shc79] M. B. SHCHEPAKIN, *On a modification of a class of algorithms for mathematical programming*, Zh. Vychisl. Mat. i Mat. Fiz., 19 (1979), pp. 1387–1395. (In Russian.)

[Shc87] ———, *On the method of orthogonal descent*, Kibernetika, (1987), pp. 58–62. (In Russian.)

[Shc92] ———, *An algorithm of orthogonal descent for searching for the zero of a convex function, identification of unsolvability of the problem and the questions of the speed of convergence*, Kibernetika, (1992), pp. 87–96. (In Russian.)

[ShM88] H. D. SHERALI AND D. C. MYERS, *Dual formulations and subgradient optimization strategies for linear programming relaxations of mixed-integer programs*, Discrete Appl. Math., 20 (1988), pp. 51–68.

[Sho79] N. Z. SHOR, *Minimization Methods for Non-Differentiable Functions*, Naukova Dumka, Kiev, 1979. (In Russian.) English transl., Springer-Verlag, Berlin, 1985.

[ShU89] H. D. SHERALI AND O. ULULAR, *A primal-dual conjugate subgradient algorithm for specially structured linear and convex programming problems*, Appl. Math. Optim., 20 (1989), pp. 193–221.

[SKR87] S. SARIN, M. H. KARWAN, AND R. L. RADIN, *A new surrogate-dual multiplier search procedure*, Naval Res. Logist. Quart., 34 (1987), pp. 431–450.

[Spi87] J. E. SPINGARN, *A projection method for least-squares solutions to overdetermined systems of linear inequalities*, Linear Algebra Appl., 86 (1987), pp. 211–236.

[Tel82] J. TELGEN, *On relaxation methods for systems of linear inequalities*, European J. Oper. Res., 9 (1982), pp. 184–189.

[Tod79] M. J. TODD, *Some remarks on the relaxation method for linear inequalities*, Tech. report 419, Cornell Univ., Ithaca, 1979.

[Tse90a] P. TSENG, *Dual ascent methods for problems with strictly convex costs and linear constraints: A unified approach*, SIAM J. Control Optim., 28 (1990), pp. 214–242.

[Tse90b] ———, *Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming*, Math. Programming, 48 (1990), pp. 249–263.

[YaM92] K. YANG AND K. G. MURTY, *New iterative methods for linear inequalities*, J. Optim. Theory Appl., 72 (1992), pp. 163–185.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 698–711, March 1996

016

GENERAL OPTIMALITY CONDITIONS FOR CONSTRAINED CONVEX CONTROL PROBLEMS*

MAÏTINE BERGOUNIOUX† AND DAN TIBA‡

Abstract. In this paper we investigate some optimal convex control problems, with mixed constraints on the state and the control. We give a general condition which allows us to set optimality conditions for nonqualified problems (in the Slater sense). Then we give some applications and examples involving generalized bang-bang results.

Key words. optimal control with mixed constraints, qualification condition, optimality conditions

AMS subject classification. 49B22

1. Formulation of the problem. Let $V \subset H \subset V'$ compactly and densely be Hilbert spaces; $A(t) : V \to V'$, $B : U \to V'$ be linear bounded operators (U is another nontrivial Hilbert space); and $L : L^2(0,T;H \times U) \to \mathbb{R}$, $l : H \to \mathbb{R}$ be convex, continuous mappings.

We consider the following optimal control problem:

$$\text{(P)} \qquad \text{Min } \{ L(x,u) + l(x(T)) \}$$

subject to

$$x'(t) + A(t)x(t) = Bu(t) + f(t) \qquad \text{a.e. in }]0,T[, \tag{1.1}$$

$$[x,u] \in D \subset \mathcal{X} \times L^2(0,T;U), \text{ closed convex subset}, \tag{1.2}$$

where

- $\mathcal{X} = L^2(0,T;V) \cap W^{1,2}(0,T;V')$,
- $f \in L^2(0,T;V')$,
- L is coercive in the sense

$$\begin{gathered} \exists c \text{ (a generic constant)} > 0 \text{ such that} \\ \forall [y,u] \in L^2(0,T;H \times U) \qquad L(y,u) > c\, \|u\|^2_{L^2(0,T;U)} - c, \end{gathered} \tag{1.3}$$

- $\forall z \in V \quad t \mapsto A(t)z$ is V'-measurable on]0,T[, and

$$\forall z \in V \quad \|A(t)z\|_{V'} \leq c\|z\|_V, \tag{1.4}$$

$$\exists \alpha\ ,\ \exists \beta > 0\ ,\ \forall z \in V \quad \langle A(t)z, z\rangle_{V \times V'} + \alpha\|z\|^2_H \geq \beta\|z\|^2_V\ . \tag{1.5}$$

* Received by the editors January 26, 1994; accepted for publication (in revised form) December 12, 1994.

† Département de Mathématiques, URA CNRS 1803, Université d'Orléans, B.P. 6759, 45067 Orléans Cedex 2, France

‡ Institute for Mathematics, Romanian Academy of Sciences, P.O. Box 1-764, 70700 Bucuresti, Romania. Part of this research was performed while this author was visiting Technische Universität München under the support of the Alexander von Humboldt Foundation.

We assume the initial condition $x(0) = x^o \in H$, and the possible final restrictions are included in the definition of D. The evolution equation (1.1) has a unique solution $x \in \mathcal{X}$ for any $u \in L^2(0,T;U)$ by theorem 4.5 in Barbu and Precupanu [3, Chap. 1].

Moreover, by condition (1.3), it is a standard argument to show that (P) has at least one optimal pair (denoted $[x^*, u^*]$) in D if some admissibility assumption is fulfilled:

$$\exists\, [x,u] \in D \text{ such that } T(x,u) = 0 \tag{1.6}$$

with

$$\forall [x,u] \in \mathcal{X} \times L^2(0,T;U) \qquad T(x,u) = x' + A(t)x - Bu - f.$$

The problem (P) is a generalization of the Bolza optimal control problem studied by Barbu and Precupanu [3, Chap. 4] both with respect to the cost functional and with respect to the form of the mixed constraints. The continuity hypothesis on L and l is quite restrictive, but as we keep the constraints separate (i.e., we do not include them into the cost via the indicator function of D), then the class of examples is very large.

For state-constrained control problems, one usually assumes a Slater-type interiority condition. In the general setting of (1.2), it takes the form

$$(\mathcal{S}) \quad \exists\, [\bar{x}, \bar{u}] \text{ feasible for (P) such that } \bar{x} \in \operatorname{int} \{\, y \in \mathcal{X} \mid [y, \bar{u}] \in D\} \text{ in } \mathcal{C}(0,T;H),$$

and it has very severe implications for the set of possible applications.

It is our main concern to weaken this classical qualification constraint. Namely, instead of $(\mathcal{S})$, we shall suppose that

$$(\mathcal{H}) \qquad \begin{array}{c} \exists \mathcal{M} \subset D \text{ bounded in } \mathcal{C}(0,T;H) \times L^2(0,T;U) \\ \text{such that } 0 \in \operatorname{int} T(\mathcal{M}) \text{ in } L^2(0,T;V'). \end{array}$$

Let us first notice that 0 appears naturally in $(\mathcal{H})$ since the problem constraint is expressed as $T(x,u) = 0$. Moreover, the elements (pairs) of the set $\mathcal{M}$ need not be feasible for (P).

We first compare the two conditions $(\mathcal{S})$ and $(\mathcal{H})$; the following proposition proves that $(\mathcal{S})$ is always stronger than $(\mathcal{H})$.

PROPOSITION 1.1. $(\mathcal{S}) \Rightarrow (\mathcal{H})$.

Proof. Thanks to $(\mathcal{S})$, (1.1), and (1.2) we have

$$\bar{x}' + A(t)\bar{x} = B\bar{u} + f, \ \bar{x}(0) = x^o \ \text{ a.e. in }]0,T[. \tag{1.7}$$

Let $\rho > 0$ and $\xi \in L^2(0,T;V')$ such that $\|\xi\|_{L^2(0,T;V')} = 1$. We denote by x_ξ the solution of

$$x_\xi' + A(t)x_\xi = B\bar{u} + f + \rho\xi, \ x_\xi(0) = x^o \ \text{ a.e. in }]0,T[. \tag{1.8}$$

Taking the difference between (1.7) and (1.8), we get that

$$\|\bar{x} - x_\xi\|_{\mathcal{X}} \le k\rho$$

with k independent of ξ. Then, if ρ is small enough, $(\mathcal{S})$ gives $[x_\xi, \bar{u}] \in D$ for all $\xi \in L^2(0,T;V')$ such that $\|\xi\|_{L^2(0,T;V')} = 1$. Here we set

$$\mathcal{M} = \operatorname{conv}(\{\, [x_\xi, \bar{u}] \mid \xi \in L^2(0,T;V'),\ \|\xi\|_{L^2(0,T;V')} = 1 \,\}),$$

where conv(E) is the convex hull of the set E, and the proof is complete. □

Remark 1.1. Assume that $U \subset V'$ continuously and $B : U \to V'$ is the canonical injection. Then one makes an interiority assumption with respect to the control of the type

$$(\mathcal{I}) \qquad \begin{array}{c} \exists\ [\tilde{x}, \tilde{u}] \text{ feasible for (P) such that} \\ \text{Int } \{\ u \in U \mid [\tilde{x}, u] \in D\ \} \text{ is nonempty in } L^2(0,T;V'). \end{array}$$

This again implies $(\mathcal{H})$ by an argument as above. In this case the Slater condition need not be fulfilled; that is to say, the condition $(\mathcal{H})$ is strictly weaker than $(\mathcal{S})$.

Condition $(\mathcal{H})$ or its weaker variant $(\mathcal{H}')$ from §3 may be mainly compared with Zowe and Kurcyusz's [18] condition in the mathematical programming theory. This was previously used in abstract control problems by Tröltzsch [16], [17], combined with interiority-type assumptions at the level of applications. In the examples of §3, we show that the interior of the constraint set may be empty even in the uniform topology, but the argument still applies.

In the recent work of Barbu and Pavel [2] another case of empty interior constraints is discussed for optimal control problems governed by periodic evolution equations and by a different method. Our approach is based on the penalization of the only state system rather than of both the state system and the constraints (as in Bonnans and Casas [8]); the constraints are kept explicit throughout the proof. This is quite a classical philosophy in connection to Lagrange multipliers techniques (see, for instance, the monograph of Tikhomirov [15]). In the setting of partial differential equations, it has been extensively exploited in books by Lions [12] and Tiba [14, Chap. 2] in connection to nonlinear singular control problems. Recently Bergounioux [5, 6] has applied this method to control problems with state constraints governed by elliptic systems, and Bergounioux, Männikkö, and Tiba [7] have studied some examples of parabolic control problems. Applications of the obtained optimality system to augmented Lagrangian algorithms were also indicated.

Finally, we point out that the technique used in the next section makes evident with full accuracy the relationship between the operator T and the set D of constraints.

2. The optimality system. We define the penalized problem as follows:

$$(\mathrm{P}_\varepsilon) \qquad \begin{aligned} \text{Min} \Big\{ & L_\varepsilon(x,u) + l_\varepsilon(x(T)) + \frac{1}{2}\int_0^T \|u - u^*\|_U^2\ dt \\ & + \frac{1}{2\varepsilon}\int_0^T \|x' + Ax - Bu - f\|_{V'}^2\ dt \Big\} \end{aligned}$$

over all $[x,u] \in D$. It should be noted that the first integral is an "adapted" penalization term according to Barbu [1], while L_ε and l_ε are the Moreau–Yosida regularization of the convex mappings L and l.

Remark 2.1. Let us briefly recall the Moreau–Yosida regularization. Let f be a proper, convex mapping on a Banach space X. For any $\varepsilon > 0$, the Moreau–Yosida regularization of f is

$$\forall x \in X \qquad f_\varepsilon(x) = \inf \left\{ \frac{1}{2\varepsilon}\|x - y\|_X^2 + f(y) \mid y \in X \right\}.$$

A thorough study of the properties of L_ε and l_ε may be found in Barbu and Precupanu [3, Chap. 2]. We shall recall some of them when needed in the text.

The existence of a unique optimal pair $[x_\varepsilon, u_\varepsilon]$ is obvious. We also denote

$$r_\varepsilon = \frac{1}{\varepsilon} J^{-1}(x'_\varepsilon + Ax_\varepsilon - Bu_\varepsilon - f) \in L^2(0,T;V),$$

where $J : V \to V'$ is the canonical isomorphism.

PROPOSITION 2.1. *We have*

$$x_\varepsilon \to x^* \quad \textit{strongly in } \mathcal{X}, \tag{2.1}$$

$$u_\varepsilon \to u^* \quad \textit{strongly in } L^2(0,T;U), \tag{2.2}$$

$$\varepsilon^{\frac{1}{2}} r_\varepsilon \quad \textit{is bounded in } L^2(0,T;V). \tag{2.3}$$

Proof. The optimality of the pair $[x^*, u^*]$ and the properties of the convex regularized mappings give

$$\begin{aligned} L_\varepsilon(x_\varepsilon, u_\varepsilon) + l_\varepsilon(x_\varepsilon(T)) + \frac{1}{2}\int_0^T \|u_\varepsilon - u^*\|_U^2 \, dt + \frac{1}{2\varepsilon}\int_0^T \|x'_\varepsilon + A(t)x_\varepsilon - Bu_\varepsilon - f\|_{V'}^2 dt \\ \leq L_\varepsilon(x^*, u^*) + l_\varepsilon(x^*(T)) \leq L(x^*, u^*) + l(x^*(T)). \end{aligned} \tag{2.4}$$

With the coercivity assumption (1.3), the relation (2.4) gives

$$l_\varepsilon(x_\varepsilon(T)) + \frac{1}{2}\int_0^T \|u_\varepsilon - u^*\|_U^2 \, dt + \frac{1}{2\varepsilon}\int_0^T \|x'_\varepsilon + A(t)x_\varepsilon - Bu_\varepsilon - f\|_{V'}^2 dt \leq c.$$

Moreover l_ε is lower bounded by an affine mapping uniformly with respect to $\varepsilon > 0$ so that

$$\frac{1}{2}\int_0^T \|u_\varepsilon - u^*\|_U^2 \, dt + \frac{1}{2\varepsilon}\int_0^T \|x'_\varepsilon + A(t)x_\varepsilon - Bu_\varepsilon - f\|_{V'}^2 dt \leq c + c\,\|x_\varepsilon(T)\|_H.$$

As the initial condition is contained in D and the dependence from the right-hand side as defined by (1.1) is sublinear, then we see that (u_ε) is bounded in $L^2(0,T;U)$, x_ε is bounded in $\mathcal{X}$, and $\varepsilon^{\frac{1}{2}} r_\varepsilon$ is bounded in $L^2(0,T;V)$.

We denote by $[\hat{x}, \hat{u}]$ their weak limit on a subsequence. Since

$$x'_\varepsilon + A(t)x_\varepsilon = Bu_\varepsilon + f + \varepsilon J(r_\varepsilon)\ ,$$

we can pass to the limit and $[\hat{x}, \hat{u}]$ is an feasible pair for (P). We have

$$L_\varepsilon(x_\varepsilon, u_\varepsilon) = L((I + \varepsilon\partial L)^{-1}(x_\varepsilon, u_\varepsilon)) + \frac{1}{2\varepsilon}\|[x_\varepsilon, u_\varepsilon] - (I + \varepsilon\partial L)^{-1}(x_\varepsilon, u_\varepsilon)\|_{H\times U}^2,$$

where I is the identity in $H \times U$. Coming back to (2.4), we get easily that $L_\varepsilon(x_\varepsilon, u_\varepsilon)$ is bounded, so $(I + \varepsilon\partial L)^{-1}(x_\varepsilon, u_\varepsilon) \rightharpoonup [\hat{x}, \hat{u}]$ weakly in $L^2(0,T;H \times U)$ by the above formula. Taking into account the weak lower semicontinuity of L and l we can pass to the limit in (2.4):

$$L(\hat{x}, \hat{u}) + l(\hat{x}(T)) + \frac{1}{2}\int_0^T \|\hat{u} - u^*\|_U^2 \, dt \leq L(x^*, u^*) + l(x^*(T)).$$

Then $\hat{u} = u^*$, $\hat{x} = x^*$, and we have (2.2) and (2.1) by a strong convergence criterion in Hilbert spaces. □

PROPOSITION 2.2. *We have the following first-order optimality condition:*

$$\begin{aligned}(2.5)\qquad &\langle \nabla L_\varepsilon(x_\varepsilon, u_\varepsilon), [x_\varepsilon, u_\varepsilon] - [x, u]\rangle_{L^2(0,T;H\times U)} + \langle \nabla l_\varepsilon(x_\varepsilon(T)), x_\varepsilon(T) - x(T)\rangle_H \\ &+ \int_0^T \langle u_\varepsilon - u, u_\varepsilon - u^*\rangle_U \, dt - \int_0^T \langle x' + Ax - Bu - f, J(r_\varepsilon)\rangle_{V'} \, dt \le 0\end{aligned}$$

for any $[x, u]$ in D. (Here ∇ denotes the Gâteaux derivative.)

This is a standard result in the optimization of convex differentiable functionals (see, for instance, Lions [11, Chap. 1]) and $J(r_\varepsilon)$ plays the role of a Lagrange multiplier.

Proof. We make feasible variations in x and u:

$$\begin{aligned}&L_\varepsilon(x_\varepsilon, u_\varepsilon) + l_\varepsilon(x_\varepsilon(T)) + \frac{1}{2}\int_0^T \|u_\varepsilon - u^*\|_U^2 \, dt + \frac{1}{2\varepsilon}\int_0^T \|x'_\varepsilon + Ax_\varepsilon - Bu_\varepsilon - f\|_{V'}^2 \, dt \\ &\le L_\varepsilon(x_s, u_s) + l_\varepsilon(x_s(T)) + \frac{1}{2}\int_0^T \|u_s - u^*\|_U^2 \, dt + \frac{1}{2\varepsilon}\int_0^T \|x'_s + Ax_s - Bu_s - f\|_{V'}^2 \, dt,\end{aligned}$$

where $x_s = x_\varepsilon + s(x - x_\varepsilon)$, $u_s = u_\varepsilon + s(u - u_\varepsilon)$, $s \in]0, 1]$, and $[x, u] \in D$ arbitrary.

Moving all the terms to the left-hand side, dividing by $s > 0$, and letting s tend to 0, we obtain

$$\begin{aligned}(2.6)\qquad &\langle \nabla L_\varepsilon(x_\varepsilon, u_\varepsilon), [x_\varepsilon, u_\varepsilon] - [x, u]\rangle_{L^2(0,T;H\times U)} + \langle \nabla l_\varepsilon(x_\varepsilon(T)), x_\varepsilon(T) - x(T)\rangle_H \\ &+ \int_0^T \langle u_\varepsilon - u, u_\varepsilon - u^*\rangle_U \, dt - \int_0^T \langle x' + Ax - Bu - f - \varepsilon J(r_\varepsilon), J(r_\varepsilon)\rangle_{V'} \, dt \le 0\end{aligned}$$

for any $[x, u]$ in D. Then (2.5) follows from (2.6) since $\varepsilon\|J(r_\varepsilon)\|^2_{L^2(0,T;V')} \ge 0$. (We have also used the properties of $J(r_\varepsilon)$.) □

Remark 2.2. The condition (2.6) is also sufficient for optimality in (P). We can reexpress (2.5) as

$$\begin{aligned}(2.7)\qquad &\langle \nabla L_\varepsilon(x_\varepsilon, u_\varepsilon), [x_\varepsilon, u_\varepsilon] - [x, u]\rangle_{L^2(0,T;H\times U)} + \langle \nabla l_\varepsilon(x_\varepsilon(T)), x_\varepsilon(T) - x(T)\rangle_H \\ &+ \int_0^T \langle u_\varepsilon - u, u_\varepsilon - u^*\rangle_U \, dt - \int_0^T \langle x' + Ax - Bu - f, r_\varepsilon\rangle_{V'\times V} \, dt \le 0\end{aligned}$$

for any $[x, u]$ in D.

Now we define the simplified adjoint system

$$(2.8)\qquad -p'_\varepsilon + A^*(t)p_\varepsilon = \nabla_x L_\varepsilon(x_\varepsilon, u_\varepsilon),$$

$$(2.9)\qquad p_\varepsilon(T) = \nabla l_\varepsilon(x_\varepsilon(T)),$$

where A^* denotes the adjoint operator of A.

Multiply (2.8) by $x_\varepsilon - x$ for any x in $\mathcal{X}$ such that $x(0) = x^o$, and integrating by parts we get

$$\begin{aligned}(2.10)\qquad \langle \nabla_x L_\varepsilon(x_\varepsilon, u_\varepsilon), x_\varepsilon - x\rangle_{L^2(0,T;H)} &= \int_0^T \langle -p'_\varepsilon + A^*p_\varepsilon, x_\varepsilon - x\rangle_H \, dt \\ &= \int_0^T \langle x'_\varepsilon - x' + A(x_\varepsilon - x), p_\varepsilon\rangle_{V'\times V} \, dt \\ &\quad - \langle \nabla l_\varepsilon(x_\varepsilon(T)), x_\varepsilon(T) - x(T)\rangle_H .\end{aligned}$$

Replacing (2.10) in (2.6), we obtain the equivalent form

$$\begin{aligned}(2.11)\qquad &\langle \nabla_u L_\varepsilon(x_\varepsilon,u_\varepsilon),u_\varepsilon-u\rangle_{L^2(0,T;U)}+\int_0^T \langle x_\varepsilon'-x'+A(x_\varepsilon-x),p_\varepsilon\rangle_{V'\times V}\,dt\\ &+\int_0^T \langle u_\varepsilon-u,u_\varepsilon-u^*\rangle_U\,dt-\int_0^T \langle x'+Ax-Bu-f,J(r_\varepsilon)\rangle_{V'}\,dt\\ &\qquad\le -\varepsilon\|J(r_\varepsilon)\|^2_{L^2(0,T;V')}\le 0\end{aligned}$$

for any $[x,u]$ in D.

Taking in turn $u=u_\varepsilon$, $x=x_\varepsilon$, a short computation provides the following decoupled system:

$$(2.12)\qquad \int_0^T \langle x_\varepsilon'-x'+Ax_\varepsilon-Ax,p_\varepsilon+r_\varepsilon\rangle_{V'\times V}\,dt\le 0,$$

$$\begin{aligned}(2.13)\qquad &\int_0^T \langle u_\varepsilon-u,u_\varepsilon-u^*\rangle_U\,dt+\langle \nabla_u L_\varepsilon(x_\varepsilon,u_\varepsilon),u_\varepsilon-u\rangle_{L^2(0,T;U)}\\ &\qquad-\int_0^T \langle u_\varepsilon-u,B^*J(r_\varepsilon)\rangle_U\,dt\le 0\end{aligned}$$

for any $[x,u]$ such that $[x,u_\varepsilon]\in D$ and $[x_\varepsilon,u]\in D$.

Remark 2.3. The relations (2.8)–(2.9) and (2.12)–(2.13) give the optimality conditions for the problem (P_ε) in a more usual form. In particular, if $D=K\times U_{\mathrm{ad}}$ (i.e., the constraints are separate) and if $\mathcal{N}(u_\varepsilon)$ denotes the normal cone to the control constraints set at u_ε, that is,

$$\mathcal{N}(u_\varepsilon)=\{\,z\in L^2(0,T;U)\mid \langle z,u_\varepsilon-u\rangle_{L^2(0,T;U)}\ge 0\ \forall u\in U_{\mathrm{ad}}\,\},$$

then (2.13) becomes

$$\nabla_u L_\varepsilon(x_\varepsilon,u_\varepsilon)+u_\varepsilon-u^*+\mathcal{N}(u_\varepsilon)\ni B^*J(r_\varepsilon),$$

which is a standard form of the Pontryagin maximum principle (Barbu and Precupanu [3, Chap. IV] and Tiba [14, Chap. II]).

PROPOSITION 2.3. *On a subsequence, we have*

$$(2.14)\qquad \nabla l_\varepsilon(x_\varepsilon(T))\rightharpoonup w\in\partial l(x^*(T))\ \textit{weakly in } H,$$

$$(2.15)\qquad \nabla L_\varepsilon(x_\varepsilon,u_\varepsilon)\rightharpoonup (w_1,w_2)\in\partial L(x^*,u^*)\ \textit{weakly in } L^2(0,T;H\times U),$$

$$(2.16)\qquad p_\varepsilon\to p^*\ \textit{strongly in } C(0,T;H),$$

where p^* is the solution of the simplified adjoint system

$$(2.17)\qquad -\frac{dp^*}{dt}+A^*p^*=w_1,$$

$$(2.18)\qquad p^*(T)=w.$$

Proof. We have $\nabla L_\varepsilon(x_\varepsilon, u_\varepsilon) \in \partial L((I+\varepsilon\partial L)^{-1})(x_\varepsilon, u_\varepsilon)$. With an argument similar to the one in the proof of Proposition 2.1, it yields that $(I+\varepsilon\partial L)^{-1})(x_\varepsilon, u_\varepsilon)$ strongly converges to $[x^*, u^*]$ in $L^2(0,T; H \times U)$. As L is continuous, it is everywhere subdifferentiable and ∂L is locally bounded. Then $\nabla L_\varepsilon(x_\varepsilon, u_\varepsilon)$ is bounded in $L^2(0,T; H \times U)$ and (2.15) is a consequence of the demiclosedness of maximal monotone operators.

The argument is the same for relation (2.14), and (2.16)–(2.18) may be obtained by taking the limit in (2.8)–(2.9). □

PROPOSITION 2.4. *Under hypothesis* $(\mathcal{H})$, (r_ε) *is bounded in* $L^2(0,T;V)$ *and* $r_\varepsilon \rightharpoonup r^*$ *on a subsequence weakly in* $L^2(0,T;V)$.

Proof. We use the relation (2.7) and take test functions $[x_\xi, u_\xi] \in \mathcal{M} \subset D$ such that

$$T(x_\xi, u_\xi) = \rho\, \xi$$

for any $\xi \in L^2(0,T;V')$ such that $\|\xi\|_{L^2(0,T;V')} = 1$ and for some $\rho > 0$. The boundedness of $\mathcal{M}$ and Propositions 2.1 and 2.3 allow us to infer

$$\rho \int_0^T \langle \xi, r_\varepsilon \rangle_{V' \times V}\, dt \;\leq\; c,$$

where c is an absolute constant independent of $\varepsilon > 0$ and ξ. □

We finally have the following theorem.

THEOREM 2.1. *If the pair* $[x^*, u^*]$ *is optimal for the problem* (P), *then*

$$\int_0^T \langle x^{*\prime} - x' + Ax^* - Ax, p^* + r^* \rangle_{V' \times V}\; dt \;\leq 0, \tag{2.19}$$

$$\langle w_2, u^* - u \rangle_{L^2(0,T;U)} - \int_0^T \langle u^* - u, B^* J^{-1}(r^*) \rangle_U\; dt \;\leq 0 \tag{2.20}$$

for any $[x,u]$ *such that* $[x,u^*] \in D$ *and* $[x^*,u] \in D$.

Moreover the inequality summing (2.19) *and* (2.20) *is valid for any* $[x,u] \in D$; *it is also sufficient for the optimality of* $[x^*,u^*]$.

Proof. The necessity has been established with the previous sequence of propositions. Let us prove the sufficiency of the condition. Let $[x,u]$ be any feasible pair for (P) and add (2.19) and (2.20):

$$\int_0^T \langle x^{*\prime} - x' + Ax^* - Ax, p^* + r^* \rangle_{V' \times V}\, dt + \langle w_2, u^* - u \rangle_{L^2(0,T;U)} - \int_0^T \langle Bu^* - Bu, r^* \rangle_{V' \times V}\, dt \leq 0.$$

As $[x,u]$ is feasible, we get

$$\langle w_2, u^* - u \rangle_{L^2(0,T;U)} + \int_0^T \langle x^{*\prime} - x' + Ax^* - Ax, p^* \rangle_{V' \times V}\; dt \;\leq 0\,.$$

Integrating by parts and taking in account the adjoint equation, we obtain

$$\langle w_2, u^* - u \rangle_{L^2(0,T;U)} + \langle w_1, x^* - x \rangle_{L^2(0,T;H)} + \langle w, x^*(T) - x(T) \rangle_H \leq 0\,.$$

The definition of the subdifferential achieves the proof. □

Remark 2.4. To get a better insight into the relation (2.19), let us assume that $D = K \times U_{ad}$ (closed convex subsets in appropriate spaces). Let r^* be in $\mathcal{X}$ (regularity). Then (2.19) can be written as follows:

$$\int_0^T \langle \partial_x L(x^*, u^*) - r^{*\prime} + A^* r^*, x^* - x \rangle_{V \times V'} \, dt + \langle \partial l(x^*(T)) + r^*(T), x^*(T) - x(T) \rangle_H \leq 0 \tag{2.21}$$

by partial integration and for any x in K. If we consider the evolution system, which gives the adjoint equation of Barbu and Precupanu [3],

$$-\frac{dr^*}{dt} + A^*(t) r^* + \partial 1_K(x^*) \ni -\partial_x L(x^*, u^*) \quad \text{a.e. in }]0, T[, \tag{2.22}$$

$$r^*(T) \in -\partial l(x^*(T)), \tag{2.23}$$

(where 1_K is the indicatrix function of K), then (2.21) is as a weak variant of (2.22)–(2.23). In particular, when no state constraints are imposed, one may easily infer that $p^* = -r^*$. We see that condition $(\mathcal{H})$ yields the existence of a Lagrange multiplier, while $(\mathcal{S})$ ensures better regularity properties for it (Barbu and Precupanu [3]).

Remark 2.5. The form of the optimality conditions may also be compared with the works of Bonnans and Casas [8, 9]. Basically, we decouple the influence of the state constraints from the adjoint equation and put it as an independent inequality (2.19). The remaining simplified adjoint system (2.17)–(2.18) just performs the necessary integrations by parts in order to reexpress the gradient of the cost functional, and it is identical to the case without any state constraints (Lions [11]). This also avoids the delicate analysis of adjoint systems with measures as data, which is necessary when the classical approach is used (Casas [10]).

3. Some applications. Let $W \subset L^2(0, T; V')$ continuously, densely be a Banach space. We replace $(\mathcal{H})$ with the weaker variant

$$\begin{aligned} &\exists \mathcal{M} \subset D \text{ bounded in } \mathcal{C}(0, T; H) \times L^2(0, T; U) \text{ such that} \\ &0 \in \text{Int } T(\mathcal{M}) \text{ in the } W \text{ topology.} \end{aligned} \tag{$\mathcal{H}'$}$$

Since condition $(\mathcal{H})$ is used only in the proof of Proposition 2.4, then Propositions 2.1–2.3 remain valid. We also ask the following pairing compatibility condition, which is automatically fulfilled in many examples:

$$\langle v, w \rangle_{W \times W'} = \int_0^T \langle v, w \rangle_{V \times V'} \tag{3.1}$$

when both terms have sense. We keep the notations of the proof of Proposition 2.4. Condition $(\mathcal{H}')$ yields that

$$\rho \ \langle \xi, r_\varepsilon \rangle_{W \times W'} < \ c; \tag{3.2}$$

that is, (r_ε) is bounded in the "larger" space W' instead of $L^2(0, T; V)$. Let r^* denote a weak * cluster point for this set.

THEOREM 3.1. *The pair* $[x^*, u^*]$ *is an optimal pair for* (P) *if and only if*

$$\langle w_2, u^* - u \rangle_{L^2(0,T;U)} + \int_0^T \langle x^{*\prime} - x' + Ax^* - Ax, p^* \rangle_{V' \times V} \, dt - \langle x' + Ax - Bu - f, r^* \rangle_{W \times W'} \leq 0 \tag{3.3}$$

for any $[x,u]$ *such that* $T(x,u) \in W$.

Proof. Necessity is a direct consequence of (2.7) and (3.2) since one may pass to the limit in all the terms if $T(x,u) \in W$.

For the sufficiency we notice that any feasible pair for (P) satisfies $T(x,u) = 0 \in W$. Then (3.3) may be used, and only the first two terms will remain. The proof is completed as in Theorem 2.1. □

Remark 3.1. In applications, $T(x,u) \in W$ may be valid for any pair $[x,u] \in D$, or this may be equivalent to a regularity condition which is possible to include in the definition of the state and control spaces. See [9] for the details of this technique in a different setting.

3.1. A first example: Empty interior constraints. We analyse in some detail the following example of optimal control problem governed by a parabolic partial differential equation

$$\text{(PP)} \qquad \text{Min} \left\{ \frac{1}{2} \int_Q (y - z_d)^2 \, dx \, dt + \frac{N}{2} \int_Q u^2 \, dx \, dt \right\}$$

subject to

$$\frac{\partial y}{\partial t} - \Delta y = f + u \quad \text{in } Q = \Omega \times]0,T[, \tag{3.4}$$

$$y(x,t) = 0 \quad \text{on } \Sigma = \partial\Omega \times]0,T[, \tag{3.5}$$

$$y(x,0) = y_o(x) \quad \text{in } \Omega \tag{3.6}$$

and the constraints

$$e(x,t) \le y(x,t) \le g(x,t) \quad \text{a.e. in } Q, \tag{3.7}$$

$$a(x,t) \le u(x,t) \le b(x,t) \quad \text{a.e. in } Q. \tag{3.8}$$

Here Ω is a smooth, open, and bounded domain of $\mathbb{R}^n$; $z_d \in L^2(Q)$; $N \ge 0$; $y_o \in L^2(\Omega)$; f, a, and b are in $L^\infty(Q)$; and e and g in $C(\bar{Q})$. We denote

$$K = \{ \, y \in L^2(0,T; H_o^1(\Omega)) \cap W^{1,2}(0,T; H^{-1}(\Omega)) \mid e \le y \le g, \; y(.,0) = y_o \, \},$$

$$U_{\text{ad}} = \{ \, u \in L^2(Q) \mid a \le u \le b \;\; \text{a.e. in } Q \, \},$$

$$D = K \times U_{\text{ad}},$$

which are closed convex sets. One has to assume the compatibility condition

$$e \le y^o \le g \qquad \text{in } \Omega$$

and some admissibility hypothesis. We ask

$$(\mathcal{E}) \qquad \exists \alpha > 0, \; \exists \tilde{u} \in U_{\text{ad}} \quad \text{such that} \quad e \le Y(\tilde{u} - \alpha) \le Y(\tilde{u} + \alpha) \le g \qquad \text{in } Q,$$

where Y is the solution operator $u \mapsto y$ defined by (3.4)–(3.6). By comparison, $(\mathcal{E})$ implies that the pair $[\tilde{u}, Y(\tilde{u})]$ is feasible for (PP). However, this is not an interiority assumption since e may be equal to g in some points. For instance, we may allow $e(x,t) = g(x,t) = 0$ on the border of the domain. Moreover, the mappings $\tilde{u} + \alpha, \tilde{u} - \alpha$ need not belong to U_{ad}, which may also have a void interior, i.e., $a = b$ on some subset.

Remark 3.2. A stronger variant of $(\mathcal{E})$ is that there exist two controls $\tilde{u}, \hat{u}$ feasible for (PP), which can be "strictly separated." It means that $(\mathcal{E})$ is stronger than the standard admissibility assumption but weaker than the hypothesis of existence of two feasible pairs (with this separation property). As $(\mathcal{E})$, in turn, yields $(\mathcal{H}')$, this gives a hint on the generality of the hypothesis $(\mathcal{H}')$. Moreover $(\mathcal{H}')$ requires that the problem (P) is nontrivial; that is, the set of admissible pairs is "rich."

In order to apply the abstract theory, we take the spaces $V = H_o^1(\Omega)$, $H = U = L^2(\Omega)$; the operators $A(t) : V \to V'$, $A(t)z = -\Delta z$, $B : H \to V'$, $B = i$, the canonical injection; and the mappings $l = 0$ and

$$L(y,u) = \frac{1}{2}\int_Q (y - z_d)^2 \, dx \, dt + \frac{N}{2}\int_Q u^2 \, dx \, dt.$$

The hypothesis $(\mathcal{H}')$ is a clear consequence of $(\mathcal{E})$ with $W = L^\infty(Q)$ by fixing

$$y_\xi = Y(u_\xi) = Y(\tilde{u} + \alpha\xi), \quad \xi \in W, \ \|\xi\|_W = 1.$$

Again a comparison argument shows that $[y_\xi, u_\xi] \in D$ for any ξ as above, and we can choose in $(\mathcal{H}')$ the bounded set

$$\mathcal{M} = \text{ conv } \{ \, [y_\xi, u_\xi] \mid \xi \in W, \ \|\xi\|_W = 1 \, \}.$$

Then relation (3.3) may be rewritten as

$$\begin{aligned} \int_Q Nu^*(u^* - u) \, dx \, dt + \int_Q (y^{*\prime} - y' - \Delta(y^* - y))p^* \, dx \, dt \\ - \langle y' - \Delta y - u - f, r^* \rangle_{W \times W'} \leq 0 \end{aligned} \tag{3.9}$$

for any y in K, u in U_{ad} such that $T(y,u) \in W = L^\infty(Q)$.

Note also that, since $f \in W, U_{\text{ad}} \subset W$, the last condition $(T(y,u) \in W = L^\infty(Q))$ is equivalent to a regularity condition on y : $y' - \Delta y \in L^\infty(Q)$, which is satisfied by y^*.

Here $[y^*, u^*]$ is the optimal pair of (PP), and p^* satisfies

$$\begin{aligned} -p^{*\prime} - \Delta p^* &= y^* - z_d \quad \text{in } Q, \\ p^* &= 0 \quad \text{on } \Sigma, \\ p^*(T) &= 0 \quad \text{in } \Omega. \end{aligned}$$

Choosing in turn $u = u^*$ and $y = y^*$ in (3.9), we get

(3.10)
$$\forall y \in K \text{ such that } y' - \Delta y \in L^\infty(Q) \qquad \langle y^{*\prime} - y' - \Delta(y^* - y), p^* + r^* \rangle_{W \times W'} \leq 0,$$

$$\forall u \in U_{\text{ad}} \qquad \langle Nu^* - r^*, u^* - u \rangle_{W \times W'} \leq 0. \tag{3.11}$$

The relations (3.10) and (3.11) represent the decoupled form of the optimality condition (3.9) and may be compared with (2.19) and (2.20).

Now we are going to use these above relations to get some more precise information about the optimal pair. Let us define the sets

$$Q_e = \{ \, (x,t) \in \bar{Q} \mid y^*(x,t) = e(x,t) \, \}, \qquad Q_g = \{ \, (x,t) \in \bar{Q} \mid y^*(x,t) = g(x,t) \, \},$$

$$Q^o = Q - (Q_e \cup Q_g) \ .$$

Thanks to (3.4)–(3.6) and the Sobolev embedding theorem, we infer that $y^* \in C(\bar{Q})$ if y_o is regular. Then Q_e and Q_g are closed sets and Q^o is an open subset of Q. Let $d \in \mathcal{D}(Q)$ be a test function with compact support supp $d \subset Q^o$. By the continuity of y^*, e, g and the compactness of supp d, one can find $\rho > 0$ such that $y^* + \rho d$ and $y^* - \rho d$ remain in K. Obviously, they are also regular, and we can use them in (3.10) as test functions to infer

$$\left\langle p^* + r^*, \frac{\partial d}{\partial t} - \Delta d \right\rangle_{W \times W'} = 0$$

for any $d \in \mathcal{D}(Q)$ with compact support in Q^o. Taking in account this relation and the equation satisfied by p^*, we see that there exists a distribution $j \in \mathcal{D}'(Q)$ with support included in $Q - Q^o$ such that

$$\frac{\partial r^*}{\partial t} - \Delta r^* + j = y^* - z_d \qquad \text{in } \mathcal{D}'(Q). \tag{3.12}$$

The previous equation is another familiar form of the adjoint system for state constrained control problems. In particular, it shows that $r^* \in W^{2,1,p}_{loc}(Q^o)$ for $p > 1$ if z_d belongs to $L^p(Q)$. Then $r^* \in C(Q^o)$ by the Sobolev theorem if p is big enough.

We are now prepared to give a result on the structure of the optimal pair of (PP), which may be termed a generalized bang-bang result (Tröltzsch [16]). We suppose from here that $N = 0$.

COROLLARY 3.1. *We have*

$$\begin{aligned} Q^o \subseteq & \{ \ (x,t) \mid y^*(x,t) = z_d(x,t) \ \} \\ & \cup \{ \ (x,t) \mid u^*(x,t) = a(x,t) \ \} \cup \{ \ (x,t) \mid u^*(x,t) = b(x,t) \ \}. \end{aligned}$$

Proof. Choose $u = u^*$ in $Q - Q^o$ so that (3.11) yields

$$\forall u \in U_{\text{ad}} \qquad \int_{Q^o} r^*(u^* - u) \, dx \, dt \ \geq 0. \tag{3.13}$$

Since $r^* \in C(Q^o)$, obviously

$$\begin{aligned} Q^o = & \{ \ (x,t) \in Q^o \mid r^*(x,t) > 0 \ \} \\ & \cup \{ \ (x,t) \in Q^o \mid r^*(x,t) < 0 \ \} \cup \{ \ (x,t) \in Q^o \mid r^*(x,t) = 0 \ \} \end{aligned}$$

and (3.13) shows that $u^* = b$ on the first set and $u^* = a$ on the second set. If the last set has positive measure, then (3.12) and the maximal regularity of r^* on Q^o give that $y^* = z_d$ on this subset. □

Remark 3.3. Taking into account the definition of Q^o, we see that at least one from y^* and u^* equals the extremal values e, g, a, b, or z_d in any point of Q. A similar analysis may be pursued when $N > 0$, but the structure of the optimal pair will be more complicated.

3.2. A second example: "Bottleneck" problems. We examine "bottleneck"-type problems, which were introduced by Bellman [4] in connection with some models for industrial production processes. They were discussed by a different approach in the work of Miricà [13].

We assume that the state equation and the cost functional are the same as in the previous example, with $f \in L^\infty(Q)$, $y_o \in W_o^{1,\infty}(\Omega) \cap W^{2,\infty}(\Omega)$, $y_o \geq 0$ a.e. in Ω, but the constraint has the form

$$|y| \leq u \quad \text{a.e. in } Q. \tag{3.14}$$

This is equivalent to $-u \leq y \leq u$, so the set D defined by (3.14) is convex.

Remark 3.4. If $f \geq 0$, the maximum principle gives $y \geq 0$ (with (3.4)) and the constraint (3.14) is equivalent to

$$0 \leq y \leq u \quad \text{a.e. in } Q, \tag{3.14$'$}$$

which is the original form considered by Bellman [4]. We also emphasize that the boundary condition (3.5) shows that the feasible pairs *are not* in the interior of D even in the $L^\infty(Q)$-topology; that is, Slater-type conditions cannot be valid in (3.14′).

Now take $\tilde{u} = \alpha e^t$, $\alpha > 0$, and $\tilde{y}$, the solution of (3.4)–(3.6) associated with $\tilde{u}$. If α is great enough, then $f + \tilde{u} \geq 0$ a.e. in Q and $\tilde{y} \geq 0$ a.e. in Q by the maximum principle. Let $w = \alpha e^t - \alpha \geq 0$ a.e. in Q. We notice that w satisfies

$$\begin{array}{rcll} \dfrac{\partial w}{\partial t} - \Delta w & = & \alpha e^t & \text{in } Q, \\ w & = & (\alpha e^t - \alpha)_{|\Sigma} \quad \geq 0 & \text{on } \Sigma, \\ w(0) & = & 0 & \text{in } \Omega. \end{array} \tag{3.15}$$

Let us denote by y^v the solution of (3.4)–(3.6) associated with v (so that y^o is associated with $v = 0$). By comparison, it yields that $w \geq \tilde{y} - y^o$; that is, $\tilde{u} - \alpha + y^o \geq \tilde{y} \geq 0$. There is some constant m such that $-m \leq y^o \leq m$ a.e. in Q. Then if $\alpha > 2m$ is large enough, we have

$$0 \leq \tilde{y} \leq \tilde{u} - m \qquad \text{a.e. in } Q. \tag{3.16}$$

The pair $[\tilde{y}, \tilde{u}]$ is feasible and (3.16) shows that hypothesis $(\mathcal{H}')$ is satisfied. Indeed we take

$$\mathcal{M} = \{\ [y^v, \tilde{u}] \mid v \in B_{L^p(Q)}(\tilde{u}, \lambda)\ \}, \qquad p > \frac{n+2}{2},$$

where $B_{L^p(Q)}(\tilde{u}, \lambda)$ is the L^p-ball centered in $\tilde{u}$ with radius λ and $\lambda > 0$ is small. By the continuity with respect of the right-hand side and the Sobolev embedding theorem, we can choose λ such that

$$\|\tilde{y} - y_v\|_{C(\bar{Q})} \leq m,$$

that is, $\mathcal{M} \subset D$. Moreover $T(\mathcal{M}) = B_{L^p(Q)}(0, \lambda)$ and $(\mathcal{H}')$ is fulfilled in $L^p(Q)$. If $n = 1$, we may take $p = 2$ and even condition $(\mathcal{H})$ is fulfilled.

Remark 3.5. We also notice that for the "linear" constraint (3.14′), a simpler argument may be used. For instance, we may define the new control function

$$\varphi = y - u,$$

and D may be reexpressed in the "decoupled" form

$$y \geq 0, \quad \varphi \leq 0 \text{ a.e. in } Q.$$

The same substitution may be performed in (3.4)–(3.6) and in the cost functional so that the penalization method from §2 may be used directly.

Finally, we prove that a generalized bang-bang result also remains valid in this case under some regularity assumptions.

COROLLARY 3.2. *Let N be equal to 0, and assume that y^*, u^* exist and are continuous functions on $\bar{Q}$. Then*

$$(3.17)\quad Q = \{\ (x,t) \in Q \mid |y^*(x,t)| = u^*(x,t)\ \} \cup \{\ (x,t) \in Q \mid y^*(x,t) = z_d(x,t)\ \}.$$

(The two sets above need not be disjoint. The first one corresponds to the case where the constraint is active.)

Proof. As hypothesis $(\mathcal{H}')$ is fulfilled, Theorem (3.1) gives the existence of $r^* \in L^q(Q)$ where $\frac{1}{p} + \frac{1}{q} = 1$ and

$$(3.18)\quad \int_Q (y^{*\prime} - y' - \Delta(y^* - y))p^*\ dx\ dt - \int_Q (y' - \Delta y - u - f)r^*\ dx\ dt \le 0$$

for any $[y,u] \in D$ such that $T(y,u) \in L^p(Q)$.

Here we used that $N = 0$ and p^* is given by

$$(3.19)\quad \begin{array}{rcll} -p^{*\prime} - \Delta p^* & = & y^* - z_d & \text{in } Q, \\ p^* & = & 0 & \text{on } \Sigma, \\ p^*(T) & = & 0 & \text{in } \Omega. \end{array}$$

Let $Q^* \subset Q$ be the open set defined as follows:

$$Q^* = \{(x,t) \in Q \mid |y^*(x,t)| < u^*(x,t)\},$$

where the constraint is inactive. First we take test pairs of the type

$$[y,u] = [y^* \pm \lambda_d d, u^*] \in D,$$

where $d \in \mathcal{D}(Q)$, supp $d \subset Q^*$, and $\lambda_d > 0$ is a small constant given by the Weierstrass theorem applied to the continuous functions $|y^*|$, u^* on the compact set supp d such that

$$|y^*(x,t) \pm \lambda_d d(x,t)| \le |y^*(x,t)| + \lambda_d |d(x,t)| \le u^*(x,t).$$

Thanks to (3.18), we obtain after a short calculation that

$$(3.20)\quad \int_Q (d' - \Delta d)(p^* + r^*)\ dx\ dt = 0$$

for every $d \in \mathcal{D}(Q)$ with compact support in Q^*. Then, we may find a distribution $j \in \mathcal{D}'(Q)$ with support in $Q - Q^*$ (active constraints set) such that

$$(3.21)\quad \frac{\partial r^*}{\partial t} + \Delta r^* + j = y^* - z_d \quad \text{in } \mathcal{D}'(Q).$$

This follows from (3.19) and (3.20) and implies a local regularity property for the Lagrange multiplier r^* : $r^* \in W^{2,1,2}_{\text{loc}}(Q^*)$ since $z_d \in L^2(Q)$.

Now let us take the test pairs

$$[y,u] = [y^*, u^* \pm \lambda_d d\ ] \in D,$$

where d, λ_d are as above. Again, by (3.18) we get

$$\int_Q d \quad r^* \, dx \, dt = 0.$$

Multiplying (3.21) by d, we infer

$$\int_Q d(y^* - z_d) \, dx \, dt = \langle r^{*\prime} + \Delta r^*, d \rangle_{\mathcal{D}'(Q) \times \mathcal{D}(Q)} = - \int_Q r^*(d' - \Delta d) \, dx \, dt = 0$$

for any $d \in \mathcal{D}(Q)$ with support in Q^*. This proves that $y^* = z_d$ in Q^*, and the proof is complete. □

Acknowledgment. The second author thanks the Alexander von Humboldt Foundation for their support.

REFERENCES

[1] V. BARBU, *Optimal Control of Variational Inequalities*, Research Notes in Mathematics 100, Pitman, London, 1984.

[2] V. BARBU AND N. PAVEL, *Optimal control problems with two point boundary condition*, J. Optim. Theory Appl., 77 (1993), pp. 51–78.

[3] V. BARBU AND TH. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Sijthoff and Noordhoff, Leyden, 1978.

[4] R. BELLMAN, *Dynamic Programming*, Princeton University Press, 1957.

[5] M. BERGOUNIOUX, *Contrôle optimal de problèmes elliptiques avec contraintes sur l'Etat*, C. R. Acad. Sci. Paris sér I, 310 (1990), pp. 391–396.

[6] ———, *A penalization method for optimal control of elliptic problems with state constraints*, SIAM J. Control Optim., 30 (1992), pp. 305–323.

[7] M. BERGOUNIOUX, T. MÄNNIKKÖ, AND D. TIBA, *On Nonqualified Parabolic Control Problems*, Preprint 148, University of Jyväskylä, Finland, July 1992.

[8] J. F. BONNANS AND E. CASAS, *Optimal control of semilinear multistate systems with state constraints*, SIAM J. Control Optim., 27 (1989), pp. 446–455.

[9] ———, *On the choice of the function spaces for some state-constrained control problems*, Numer. Funct. Anal. Optim., 7 (1984–85), pp. 333–348.

[10] E. CASAS, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.

[11] J. L. LIONS, *Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[12] ———, *Contrôle des systèmes distribués singuliers*, Gauthier-Villars, Paris, 1983.

[13] S. MIRICÀ, *Optimal feedback control for a class of bottleneck problems*, J. Math. Anal. Appl., 112 (1985), pp. 221–235.

[14] D. TIBA, *Optimal Control of Nonsmooth Distributed Parameter Systems*, Lecture Notes in Math. 1459, Springer-Verlag, Berlin, 1990.

[15] V. TIKHOMIROV, *Fundamental Principles of the Theory of Extremal Problems*, John Wiley, Chichester, 1986.

[16] F. TRÖLTZSCH, *Optimality Conditions for Parabolic Control Problems and Applications*, Teubner Texte, Leipzig, 1984.

[17] ———, *A modification of the Zowe and Kurcyusz regularity condition with application to the optimal control of Noether operator equations with constraints on the control and the state*, Math. Operationforsch. Statist., Ser. Optimization, 14 (1983), pp. 245–253.

[18] J. ZOWE AND S. KURCYUSZ, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49–62.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 712–756, March 1996

017

STOCHASTIC APPROXIMATION METHODS FOR SYSTEMS OVER AN INFINITE HORIZON*

HAROLD J. KUSHNER† AND FELISA J. VÁZQUEZ-ABAD‡

Abstract. The paper develops efficient and general stochastic approximation (SA) methods for improving the operation of parametrized systems of either the continuous- or discrete-event dynamical systems types and which are of interest over a long time period. For example, one might wish to optimize or improve the stationary (or average cost per unit time) performance by adjusting the systems parameters. The number of applications and the associated literature are increasing at a rapid rate. This is partly due to the increasing activity in computing pathwise derivatives and adapting them to the average-cost problem. Although the original motivation and the examples come from an interest in the infinite-horizon problem, the techniques and results are of general applicability in SA. We present an updating and review of powerful ordinary differential equation–type methods, in a fairly general context, and based on weak convergence ideas. The results and proof techniques are applicable to a wide variety of applications. Exploiting the full potential of these ideas can greatly simplify and extend much current work. Their breadth as well as the relative ease of using the basic ideas are illustrated in detail via typical examples drawn from discrete-event dynamical systems, piecewise deterministic dynamical systems, and a stochastic differential equations model. In these particular illustrations, we use either infinitesimal perturbation analysis–type estimators, mean square derivative-type estimators, or finite-difference type estimators. Markov and non-Markov models are discussed. The algorithms for distributed/asynchronous updating as well as the fully synchronous schemes are developed.

Key words. stochastic approximation, ordinary differential equation method, weak convergence, recursive optimization, Monte Carlo optimization, discrete-event dynamical systems, piecewise deterministic dynamical systems, stationary cost problems

AMS subject classifications. 62L20, 93C40, 93E25, 90B25

1. Introduction.

The paper is concerned with efficient and general stochastic approximation (SA) methods for parametrized systems of either continuous or discrete event dynamical systems that are of interest over a long time period. For example, one might wish to optimize or improve the stationary (or average cost per unit time) performance by adjusting the systems parameters. The number of applications and the associated literature are increasing at a rapid rate. Although the motivation and examples come from an interest in this infinite-horizon problem, the techniques and results are of general applicability in SA. Basic techniques for such problems have appeared in [2, 22, 27]. These techniques are still fundamental for applications to to the general problems of current interest. Exploiting their full potential can greatly simplify and extend much current work. We present a full development of the basic ideas in [22, 27] and related works in a more general context, with the particular goal of illustrating their breadth as well as the relative ease of using them in particular applications.

To fix ideas, let θ denote an adjustable parameter of a dynamical system and $x(\cdot,\theta)$ the associated system state process. For a cost rate $c(\theta,x)$, define $C_T(\theta,x(0)) = E\int_0^T c(\theta,x(t,\theta))dt/T$ and $C(\theta,x(0)) = \lim C_T(\theta,x(0))$. We wish to minimize $C(\theta,x(0))$

* Received by the editors February 23, 1994; accepted for publication (in revised form) May 12, 1995.

† Applied Mathematics Department, Brown University, Providence, RI 02912. The research of this author was supported by NSF grant ECS-9302137, AFOSR contract F49620-92-J-0081, and ARO contract DAAL03-92-G-01157.

‡ Département d'Informatique et Recherche Opérationelle, Université de Montréal, Montréal, PQ, H3C 3J7, Canada. The research of this author was supported by NSERC grant WFAO139015.

by the dynamic adjustment of the parameter θ, using estimates of the derivatives made from measurements of the sample path. Indeed, much of the recent interest in SA methods has been motivated by the increasing availability of good estimators of the derivatives of objects such as $C_T(\theta, x(0))$, say, of the infinitesimal perturbation analysis (IPA) or related types [13, 14, 18, 34, 42, 45] or of the mean square derivative type [5]. With ϵ_n a step-size parameter and θ_n the nth estimate of the parameter, the basic SA algorithm is $\theta_{n+1} = \theta_n + \epsilon_n Y_n$, where Y_n is the measurement used for the current update. One is concerned with the asymptotic properties of the sequence θ_n. The ordinary differential equations (ODE) method shows that the asymptotic properties can be characterized in terms of the limit properties of the solution to an ODE $\dot{\theta} = g(\theta)$, where, loosely speaking, $g(\theta)$ is the stationary mean value of Y_n given that the parameter value is always fixed at θ. Thus the individual Y_n themselves need not be (asymptotically) unbiased estimators of the gradient at the current parameter values. The fact that the estimators are taken over a finite time interval but one actually wishes to use them effectively for the infinite-time problem has led to various ad hoc approaches, often driven by the proof technique. One technique was to let the successive estimation intervals go to infinity. It will appear from the results in §§3–5 (a direct consequence of the results in [22, 27]) that to get the desired limit result one generally need not reinitialize the estimator periodically nor let the intervals go to infinity. One basically does what is more natural: keep the successive updating time intervals bounded and appropriately update the estimator without "reinitializing" it. The proofs of such results are the essence of the "local averaging" intuition in the ODE method, initiated by Ljung [33], although the techniques used here are quite different.

The paper is not concerned with optimization per se but rather with getting the appropriate ODE for the SA algorithm of interest and in showing the great flexibility in the algorithms that one can use and analyze. For the optimization problem, one generally needs to show that the solution of the ODE converges to the desired point, and this requires a closer look at the right-hand side of the ODE. In some cases, this involves showing that the right side of the ODE is the negative of the gradient of a desired cost function with a particular structure. Indeed, in §§7 and 8, we show that the right side is indeed the negative of the gradient of the desired ergodic cost. But in any application, one needs first to characterize the correct ODE and then to analyze the limits of its solutions. The latter job is highly problem dependent.

One can try to prove that the convergence either is with probability one (w.p.1) or is in a weak (or generalized distributional) sense. Our framework for getting the asymptotic properties is that of weak convergence. This allows the use of what might be the simplest mathematical techniques and conditions. For example, for the SA with decreasing step sizes $0 < \epsilon_n \to 0$ satisfying $\sum \epsilon_n = \infty$, no additional conditions need be imposed on the ϵ_n. Conditions of the often used type [2] $\sum \epsilon_n^{1+\alpha} < \infty$ for some $\alpha > 0$ are not needed. The sequence of estimators need only be uniformly integrable, and no additional moment conditions are needed. The weak convergence technique correctly identifies the places where the process spends either almost all or all of its (asymptotic) time, and gives us a fairly complete stability structure of the algorithm.

For the decreasing step-size algorithms, the difference between the probability-one and the weak convergence results is not as great as what one might at first suppose. Indeed, known results show that under quite weak additional conditions, probability one convergence follows directly from the weak convergence results, and we

now comment loosely on this. Suppose that the ODE is locally asymptotically stable about a point θ^* with open domain of attraction Γ. The ODE method associated with the weak convergence approach quite generally allows us to show that some such set Γ is entered infinitely often. Then, under very weak conditions, one can appeal to existing applications of large deviations methods to SA's to get probability one convergence. This idea is fully developed in [9]. Among other things, it is shown in this reference that one gets probability one convergence with the only additional requirement on the step-size sequence ϵ_n is that it satisfy

$$\sum_n e^{-\delta/\epsilon_n} < \infty$$

for each $\delta > 0$. That is, we need only that $\epsilon_n < c_n/\log n$, where $c_n \to 0$. The conditions on the noise process in [9] are satisfied by the usual processes that are not too "heavily tailed," and for such processes these probability one convergence results might be about the best now available. The main point is that once the weak convergence results are available, probability one results follow directly from existing works under broad conditions, and the basic weak convergence techniques are very much simpler than those required for probability-one convergence.

It is worth noting that in applications, probability-one results might be of illusory advantage over weak convergence results. The algorithms generally have stopping rules and, when these are applied, one generally has only probabilistic or distributional information about the last iterate.

We are also concerned with constant step size cases where one can only use weak (and not w.p.1) convergence ideas. Indeed, in problems of tracking time-varying systems one must use constant step sizes. In adaptive problems in communication theory and signal processing, constant step sizes are the common practice. Even if the problem is such that decreasing step sizes can be used, one often lets them be constant due to robustness considerations. Indeed, in practice one often prefers algorithms which get to a neighborhood of the desired point quickly, and this argues for a constant step size.

The development in the paper requires only some of the elementary concepts from the theory of weak convergence. These are reviewed in §2. Perhaps the only required nonelementary fact concerns the use in Theorem 3.1 of random variables which are measure valued. Our application of this concept is straightforward, since for our purposes the important facts concerning such random variables are determined by their mean values and will be implied by the conditions imposed on the "noise" terms. The concept of measure-valued random variables allows us to deal more easily than in the past with unbounded noise.

The basic result of the paper is Theorem 3.1. It is basic in that it lays out the fundamental ideas of the averaging method, and most subsequent results can be derived by mild modifications of the technique of that theorem. The theorem is for the constant-step-size case. But, as seen in §4, the case where $\epsilon_n \to 0$ differs only in the way certain terms are grouped in the proof. In Theorem 3.1, we have tried to use conditions that are fairly general. Since one's imagination in constructing algorithms is endless, no set of conditions is "completely general." But it will be seen that the conditions used are quite minimal, and allow the few basic ideas to be exposed. The first basic idea is to repose the problem as a "martingale problem," which allows us to replace the noise terms by appropriate conditional expectations given the past, and greatly facilitates the averaging. Then we are confronted by the fact that the

noise at step n can depend on the values of the state at that time as well as at previous times. In Theorem 3.1, this is handled in a convenient way (coming originally from [27]) by the use of a Markov model for the joint (noise, state) process, and imposing appropriate weak continuity conditions on the transition function. (Non-Markov models are treated in Appendix 1, but the Markov assumption in Theorem 3.1 is quite powerful, since the state space can be a complete separable metric space, thus allowing convenient "Markovianizations.") In doing the local averaging to get the appropriate ODE, these weak continuity assumptions allow us to average as though the state did not change. They facilitate the use of the appropriate (mean) ergodic theorems for the noise processes which, for the purposes of averaging, can be assumed to evolve as though the state did not change. These few basic and powerful ideas underlie all the results and are widely adaptable. The averaging idea of Theorem 3.1 is like that in [27] but is somewhat more general, particularly in the treatment of unbounded noise.

Section 3.2 concerns the asymptotic points of the algorithm, and in Theorem 3.2 they are identified with the two-sided invariant set (in the sense of dynamical systems theory) of the ODE. The other parts of §3 concern the simplifications when the basic observation has an "additive" character or the problem has a regenerative structure and one wishes to update at the regeneration times. This "additivity" property is common to numerous applications, as seen in §§6–9. In general, updating at regenerative intervals, even if the process has a regenerative structure, is not needed and might not even be a good idea. More will be said about this later. It is certainly inadvisable when the regenerative periods are very long. In §4, we make the few necessary changes when the step sizes ϵ_n go to zero.

Section 5 gives the simple alterations when the iterate is to be confined to some constraint set. It was noted in [25] and elsewhere subsequently that the ODE for the constrained problem follows directly from that for the unconstrained problem by use of a simple decomposition of the iterate into the sum of the unprojected value plus an "error." The "error" is easy to treat since it is what brings an infeasible point back to the constraint set. The unprojected values are treated as for the unconstrained algorithm. So, under appropriate conditions on the constraint set, the constrained problems are easy extensions of the unconstrained problem.

Section 6 formally introduces the application of Theorem 3.1 and its extensions for use on systems whose performance function involves a stationary average. The basic heuristic illustration is for a system where an IPA- or mean square derivative–type estimator might be used and we wish to minimize a stationary cost. The right side of the limit ODE is the negative of the derivative of the stationary cost with respect to the adjustable parameter. All of this is a consequence of the basic theorem. Many authors [6, 7, 16, 32, 34, 45] consider finite-horizon gradient estimators. They reset the estimation (reset the accumulator, to use current jargon) at the start of each observation interval, whose length becomes large as $n \to \infty$. It will be seen that quite often one does not need to let the observation intervals become large nor to reset the estimator. Indeed, these latter techniques are frequently adopted just because it is under those conditions that the authors have proved their convergence results.

To illustrate the basic simplicity and power of the approach, in §§7 and 8 we have chosen examples of current importance and on which much work has been done. Each example is typical of a large class of great current interest and illustrates the application of the methods to that class. The problem of §7 concerns the optimization of a single queue with respect to a service time parameter. This problem has been well

studied and is typical of the use of IPA in many discrete event dynamical systems. The problem in §8 concerns the optimization of an "unreliable" manufacturing system via the choice of suitable production rates and thresholds and is typical of many applications to piecewise deterministic systems. In both cases, the general techniques discussed here are relatively quick to apply and yield good results for many forms of the algorithms and under conditions which are weaker than those generally used. The power of the approach allows much flexibility in the SA algorithm.

In §9 we apply the ideas of §3 to a stochastic differential equation (SDE) model, where the sample derivative is obtained from the equation for the adjoint or mean square derivative. This is just the SDE analogue of the IPA-type estimator and has been in use for a long time. In such examples one often has the problem of proving stability of the derivative estimators, and there is no regenerative structure to help. When stability can be proved, the results are exactly as for the discrete-event and piecewise deterministic dynamical systems cases; one need not restart the estimator nor let the estimation periods increase with time, each of which might not be good practice. In the limit one gets the basic ODE, whose right side is the negative of the gradient of the stationary cost with respect to the parameter. When stability of the mean square derivative process cannot be proved, one can use various forms of finite differences. For example, one might use one continuous run either with the parameter being perturbed over successive intervals of, say, fixed length, or with the use of independent samples for the positive and negative perturbations. In the former (one sample) case, it is noteworthy that we can often get something close to the desired limit ODE. Either finite difference method can be employed when the functions in the cost or dynamical equation are not smooth (say, the cost involves the indicator function of some event) or when we do not know the model well enough to even try to compute a pathwise derivative.

The appendices contain various extensions. Appendix 1 uses a perturbed test function type method (of the type used in [23]) to avoid the Markov assumption. In Appendix 2, we illustrate the use of a method with which one can sometimes avoid the use of occupation measures in the argument of Theorem 3.1 and which is adaptable to many uses. Appendix 3 contains the few additional details when one wishes to work within a regenerative context but possibly update at rather arbitrary random times during the interval as well as at its end. Appendix 4 contains the essential ideas for dealing with a decentralized algorithm, where the different processors update on their own (asynchronous) schedule, with possible delays in communication. Using simple time-change arguments (extensions of the type first used in [30]), we show that the proof and the end results are essentially as for the basic synchronized case, except for some notational changes. This approach generalizes the results in [40]. Thus, the described approach efficiently encompasses a very diverse group of algorithms and applications.

Although the essential ideas are all in Theorem 3.1, the paper is long because we wish to show the great flexibility of the ideas and how to extend them effectively in the many possible (not entirely obvious) directions which are of increasing current interest and to properly illustrate their practical use via concrete applications to important problems.

We note that the convergence can generally be accelerated using the iterate averaging methods initiated by Polyak and discussed in [29, 36, 46, 47].

2. Some background on weak convergence. The methods of the theory of weak convergence are powerful and widely used tools for problems concerning approx-

imations and limit theorems for random processes [3, 10, 22]. They do for random processes what the central limit theorem and the law of large numbers do for sequences of vector-valued random variables. Because they are averaging methods for random processes evolving on different time scales, they are natural methods for SA and have been widely used. Only the basic definitions will be given, since the ideas will be used in a simple way. Further information for those interested can be found in the references.

Let $\{X_n, n < \infty\}$ be a sequence of random variables with values in a complete and separable metric space (CSMS) S. In this paper, S will generally be either some Euclidean space R^k, a space of functions representing the paths of the SA process, or a set of probability measures, as specified below. We say that $\{X_n, n < \infty\}$ *converges weakly* to a random variable X and write $X_n \Rightarrow X$, if for each continuous and bounded real-valued function $f(\cdot)$ on S we have $Ef(X_n) \to Ef(X)$. Thus, weak convergence is an extension of the concept of convergence in distribution of a sequence of real valued random variables to more general spaces. If P_n and P, resp., are the measures of X_n and X, resp., we also say that $P_n \Rightarrow P$. The sequence $\{X_n, n < \infty\}$ is said to be *tight* if for each $\delta > 0$ there is a compact set $K_\delta \subset S$ such that $P\{X_n \notin K_\delta\} \leq \delta$ for all n. Equivalently, a set of measures $\{P_n, n < \infty\}$ on the Borel sets of S is said to be tight if $P_n\{S - K_\delta\} \leq \delta$ for all n. Tightness implies the existence of a weakly convergent subsequence [10, p. 104].

The Skorohod representation. Since weak convergence is a generalized distributional convergence, it does not depend on the actual probability space that is used. It is often more convenient in the analysis to work with w.p.1 convergence rather than with weak convergence directly. The *Skorohod representation* [10] guarantees that we can choose the probability space so that w.p.1 convergence holds if weak convergence does, as follows. Suppose that $X_n \Rightarrow X$ weakly. Then we can find a probability space with random variables $\{\tilde{X}_n, n < \infty\}, \tilde{X}$ defined on it, where $\tilde{X}_n$ (resp., $\tilde{X}$) has the same measure as X_n (resp., X) and on which $\tilde{X}_n \to \tilde{X}$ w.p.1 in the topology of S [10, p. 102]. We will use the Skorohod representation where convenient.

The path spaces. Define $D^r[0, \infty)$ or $D^r(-\infty, \infty)$, where $D^r(I)$ is the space of R^r-valued functions on the interval I which are right continuous and have left-hand limits (and are continuous at $t = 0$ in the case of $D^r[0, \infty)$). The topology will be that of uniform convergence on finite intervals, making both spaces into CSMSs.

Notation on interpolated processes. In the SA algorithms that we study, we will have recursions of the form $X^\epsilon_{n+1} = X^\epsilon_n + \epsilon\chi^\epsilon_n$, where χ^ϵ_n is a sequence of R^r-valued random variables. We are interested in studying the limit behavior as $\epsilon \to 0$ and $n \to \infty$. The piecewise constant interpolation of X^ϵ_n on $(-\infty, \infty)$ is defined as $X^\epsilon(t) = X^\epsilon_n$ for $t \in [\epsilon n, \epsilon(n+1))$ and $X^\epsilon(t) = X^\epsilon_0$ for $t < 0$. For $t \geq 0$, let $[t]$ denote the integer part of t. Then,

$$X^\epsilon(t) = X(0) + \epsilon \sum_{i=0}^{[t/\epsilon]-1} \chi^\epsilon_i, \qquad t \geq 0. \tag{2.1}$$

$X^\epsilon(\cdot)$ is a random process with paths in $D^r(-\infty, \infty)$. We also view it as a random variable with values in $D^r(-\infty, \infty)$. We will also use shifted processes, as follows. Let q_ϵ be a sequence of integers such that $\epsilon q_\epsilon \to \infty$ as $\epsilon \to 0$. Then $X^\epsilon(\epsilon q_\epsilon + \cdot)$ will also be of interest, since the "tail" of the original process is now in the "vicinity of the origin" for large ϵq_ϵ. We will be interested in the limits of the $X^\epsilon(\cdot)$ as $\epsilon \to 0$. For notational simplicity, we henceforth write $[t/\epsilon]$ simply as t/ϵ in the limits of the sums.

Next, let ϵ_j be a sequence of positive numbers which goes to zero and such that $\sum_j \epsilon_j = \infty$. Let χ_i be a sequence of R^r-valued random variables. Define $t_n = \sum_{i=0}^{n-1} \epsilon_i$, $m(t) = \max\{i : t_i \leq t\}$. With $X(0)$ given, define the interpolation $X^0(\cdot)$ by $X^0(t) = X(0)$ for $t \leq 0$, and for $t \geq 0$,

$$X^0(t) = X(0) + \sum_{i=0}^{m(t)-1} \epsilon_i \chi_i. \tag{2.2}$$

The shifted processes $X^n(t) = X^0(t+t_n)$ will play an important role since they bring the tail of $X^0(\cdot)$ to the forefront. The following result will be used. It is not hard to prove directly and follows from [3, Thm. 15.2].

LEMMA 2.1. *Suppose that* $\{\chi_n^\epsilon, \epsilon > 0, n < \infty\}$ *and* $\{\chi_n, n < \infty\}$ *are uniformly integrable and* $\{X_n\}$, $\{X_n^\epsilon\}$ *are tight. Then* $\{X^\epsilon(\cdot), \epsilon > 0\}$, $\{X^\epsilon(\epsilon q_\epsilon + \cdot), \epsilon > 0\}$, *and* $\{X^n(\cdot), n < \infty\}$ *are tight and any weak limit has Lipschitz continuous paths w.p.1.*

Random measures. The treatment of the unbounded noise case (which is generally the situation in the problems of interest here) will be simplified and extended (over that in [22, 27]) by the use of random variables which are measure valued. The concept will be used in a rather simple way and all that we need to know will now be stated.

Let $\mathcal{P}(S)$ denote the set of probability measures over the Borel subsets of S. The Prohorov metric [10] will be used on this space. An important point is that under this metric $\mathcal{P}(S)$ is a CSMS, since S is [10, p. 101]. Convergence $P_n \to P$ in this topology is equivalent to weak convergence of $\{P_n, n < \infty\}$ to P [10, p. 108].

Now let $\{R_n, n < \infty\}$ be a sequence of *random variables* whose values are points in $\mathcal{P}(S)$. By definition, $\{R_n, n < \infty\}$ converges weakly (as a sequence of random variables) to the measure-valued random variable R if $EF(R_n) \to EF(R)$ for each bounded and continuous real-valued function $F(\cdot)$ on $\mathcal{P}(S)$. The function $\bar{R}_n$ defined by $\bar{R}_n = ER_n$ is a measure in $\mathcal{P}(S)$. We will need the following important fact

LEMMA 2.2 (see [24, pp. 14–15]). *The set* $\{R_n, n < \infty\}$ *has a weakly convergent subsequence if* $\{\bar{R}_n, n < \infty\}$ *is tight.*

This characterization in terms of the mean values is of great help, since the mean values $\{\bar{R}_n, n < \infty\}$ are much easier to deal with. Recall that the sequence $\{\bar{R}_n, n < \infty\}$ is tight if the associated sequence of random variables is tight. Let $f(\cdot)$ be a bounded, continuous, and real-valued function on S. The function defined by $F_f(P) = \int f(x)P(dx)$ is real valued, bounded, and continuous on $\mathcal{P}(S)$. Thus, if $R_n \Rightarrow R$ then $F_f(R_n) \to F_f(R)$ in distribution for each $f(\cdot)$. If the Skorohod representation is used, then we can say that w.p.1 for each such $f(\cdot)$

$$F_f(R_n) \to F_f(R). \tag{2.3}$$

3. The basic SA algorithms. The section contains several parts. Section 3.1 gives the main convergence theorem from which all others will be derived. Section 3.2 concerns the limit points of the ODEs which characterize the asymptotics of the SA. In many cases, the observation has a certain decomposition property which simplifies the verification of the assumptions, and this is exploited in §3.3. A simplified result for regenerative type processes, where we update at the end of the regeneration intervals, is in §3.4.

3.1. The canonical algorithm. Let $\epsilon > 0$. We will develop the basic ideas for the algorithm

$$\theta_{n+1}^{\epsilon} = \theta_n^{\epsilon} + \epsilon Y_n^{\epsilon}, \qquad n \geq 0,\ \theta_n^{\epsilon} \in R^r, \tag{3.1}$$

where Y_n^{ϵ} is a sequence of R^r-valued random variables. The proofs of subsequent results for other SA forms will be more or less simple variations of the proof for (3.1). We next state the conditions which will be needed. The conditions seem to be nearly minimal and will be illustrated in the examples in §§7–9.

Let $\mathcal{B}_n^{\epsilon}$ be a sequence of nondecreasing sequence of sigma-algebras where $\mathcal{B}_n^{\epsilon}$ measures at least $\{\theta_0^{\epsilon}, Y_i^{\epsilon}, i < n\}$ and E_n^{ϵ} be the expectation conditioned on $\mathcal{B}_n^{\epsilon}$. Write $E_n^{\epsilon} Y_n^{\epsilon} = \bar{Y}_n^{\epsilon}$. Then the δY_n^{ϵ} defined by $Y_n^{\epsilon} = \bar{Y}_n^{\epsilon} + \delta Y_n^{\epsilon}$ are $\mathcal{B}_n^{\epsilon}$-martingale differences. Generally, $\mathcal{B}_n^{\epsilon}$ will measure "all the information" which is used to get the $\{Y_i^{\epsilon}, i < n\}$. Suppose that

$$\{Y_n^{\epsilon}, n < \infty, \epsilon > 0\} \text{ is uniformly integrable.} \tag{3.2}$$

Suppose that there is a process $\{\xi_n^{\epsilon}, n < \infty\}$ which takes values in some CSMS and measurable functions $G_n^{\epsilon}(\cdot)$ such that we can write

$$\bar{Y}_n^{\epsilon} = G_n^{\epsilon}(\theta_n^{\epsilon}, \xi_n^{\epsilon}). \tag{3.3}$$

Assume that

$$\text{the set } \{\xi_n^{\epsilon}, \theta_n^{\epsilon}, \epsilon > 0, n < \infty\} \text{ is tight.} \tag{3.4}$$

Tightness of $\{\theta_n^{\epsilon}, \epsilon > 0, n < \infty\}$ holds if a projection algorithm is used (§5). Otherwise a stability argument might need to be used. Suppose that for each ϵ, θ, n there is a transition function $P_n^{\epsilon}(\cdot, \cdot|\theta)$ such that $P_n^{\epsilon}(\cdot, A|\cdot)$ is measurable for each Borel set A in the range space of ξ and

$$P\{\xi_{n+1}^{\epsilon} \in \cdot | \xi_i^{\epsilon}, \theta_i^{\epsilon}, i \leq n\} = P_n^{\epsilon}(\xi_n^{\epsilon}, \cdot | \theta_n^{\epsilon}). \tag{3.5}$$

By this Markov assumption, E_n^{ϵ} is the expectation conditioned on $(\theta_n^{\epsilon}, \xi_n^{\epsilon})$. For each fixed θ, let there be a transition function $P(\xi, \cdot|\theta)$ such that

$$P_n^{\epsilon}(\xi, \cdot|\theta) \Rightarrow P(\xi, \cdot|\theta) \text{ as } n \to \infty, \epsilon \to 0, \tag{3.6}$$

where the limit is uniform on each compact (θ, ξ) set; i.e., for each bounded and continuous real-valued function $f(\cdot)$,

$$\int f(\tilde{\xi}) P_n^{\epsilon}(\xi, d\tilde{\xi}|\theta) \to \int f(\tilde{\xi}) P(\xi, d\tilde{\xi}|\theta)$$

uniformly on each compact (θ, ξ) set. Assume

$$P(\xi, \cdot|\theta) \text{ is weakly continuous in } (\theta, \xi). \tag{3.7}$$

For each fixed θ the transition function $P(\cdot, \cdot|\theta)$ determines a Markov chain and we let $\{\xi_n(\theta)\}$ denote the associated random variables. Let $\mu(\cdot|\theta)$ denote the invariant measures under the transition function $P(\xi, \cdot|\theta)$. Suppose that

$$\{\mu(\cdot|\theta), \theta \in \Theta\} \text{ is tight for each compact } \Theta. \tag{3.8}$$

Henceforth, let q_ϵ be a sequence of integers such that

$$\text{either} \quad q_\epsilon \equiv 0 \qquad \text{or} \quad \epsilon q_\epsilon \to \infty. \tag{3.8'}$$

Suppose that there is a continuous function $G(\cdot)$ such that for each $\delta > 0$

$$\lim_{\epsilon} \limsup_{n} P\{|G_n^\epsilon(\theta_n^\epsilon, \xi_n^\epsilon) - G(\theta_n^\epsilon, \xi_n^\epsilon)| \geq \delta\} = 0 \tag{3.9}$$

and that for each compact θ-set Θ there is $K_0(\Theta) < \infty$ such that for all stationary processes $\{\xi_n(\theta)\}$

$$\sup_{\theta \in \Theta} E|G(\theta, \xi_j(\theta))| < K_0(\Theta). \tag{3.10}$$

Finally, we assume either (3.11a) or (3.11b):

$$\text{For each } \theta, \mu(\cdot|\theta) \text{ is unique.} \tag{3.11a}$$

There is a continuous $g(\cdot)$ such that for each θ and initial condition $\xi_0(\theta)$

$$\lim_N \frac{1}{N} \sum_{n=0}^{N-1} EG(\theta, \xi_n(\theta)) = g(\theta). \tag{3.11b}$$

Under (3.11a), define

$$g(\theta) = \int G(\theta, \xi)\mu(d\xi|\theta).$$

Define the continuous parameter interpolation $\theta^\epsilon(\cdot)$ by $\theta^\epsilon(t) = \theta_n^\epsilon$ for $t \in [n\epsilon, n\epsilon + \epsilon), n \geq 0$. For $t < 0$, set $\theta^\epsilon(t) = \theta_0^\epsilon$.

THEOREM 3.1. *Assume the conditions* (3.2)–(3.11). *Each subsequence of* $\{\theta^\epsilon(q_\epsilon\epsilon + \cdot), \epsilon > 0\}$ *has a further subsequence which converges weakly to a bounded solution* $\theta(\cdot)$ *of*

$$\dot{\theta} = g(\theta) \tag{3.12}$$

on $[0, \infty)$ *if* $q_\epsilon = 0$ *and on* $(-\infty, \infty)$ *if* $\epsilon q_\epsilon \to \infty$. *Also,* $g(\cdot)$ *is a continuous function of* θ.

Remark. We note that in current applications it is often the case that the P_n^ϵ and the G_n^ϵ do not depend on either ϵ or n. See the examples in §§6–8. A way of avoiding the Markovianization is described in Appendix 1. Condition (3.11b) is often much easier to check than is uniqueness of the invariant measure. In typical examples where one uses some sort of weak sense derivative or an IPA-type estimator, it is equivalent to the asymptotic consistency of the estimator under fixed θ, as will be seen in the examples in §§6–8. This is a minimal condition. The ability to use such a condition is basically a consequence of the "martingale problem" formulation used in the proof. It is exploited in the use of conditional expectations in the expressions from (3.17) on.

The basic idea in the proof is to first replace the Y_n^ϵ by its conditional expectation, given the past. Then use a piecewise constant approximation to the state process, and finally exploit this last approximation via an ergodic condition. The type of continuity

and uniform integrability conditions required seem rather weak and have their roots in the basic references [22, 27].

Remark. If Y_n^ϵ can be represented as $g(\theta_n^\epsilon)$ plus a "martingale difference" plus a term which goes to zero in mean[1] as $\epsilon \to \infty$ and/or $n \to \infty$, then the proof becomes nearly trivial since no averaging needs to be done. The difficulties arise when the conditional expectation (given past data) of Y_n^ϵ depends on the past, and this holds true in many important cases. The basic structure and motivation of the proof are analogous to those of [22, 27], but many of the details are different. Here there is a smoother development of the unbounded noise case under weaker conditions. The proof also provides a simpler way of characterizing the limit points (see Theorem 3.2) and dealing with the other extensions. In order to simplify the notation, we use $q_\epsilon = 0$ in the proof. The details are exactly the same for the general case.

Proof. Part 1. *A continuity result.* Until the last part of the proof, assume (3.11a). Let $f(\cdot)$ be bounded, continuous, and real valued. Given $\theta_0 \in R^r$, let θ_n be a deterministic sequence tending to θ_0. We have

$$\int f(\xi)\mu(d\xi|\theta_n) = \int \left[\int f(\tilde{\xi})P(\xi, d\tilde{\xi}|\theta_n)\right] \mu(d\xi|\theta_n).$$

Now as $n \to \infty$ $P(\xi, \cdot|\theta_n)$ converges weakly to $P(\xi, \cdot|\theta_0)$ uniformly on each compact (θ, ξ) set by (3.7). Using (3.8), extract a weakly convergent subsequence of $\{\mu(\cdot|\theta_n), n < \infty\}$ and denote the limit by $\tilde{\mu}(\cdot)$. Then

$$\int f(\xi)\tilde{\mu}(d\xi) = \int \left[\int f(\tilde{\xi})P(\xi, d\tilde{\xi}|\theta_0)\right] \tilde{\mu}(d\xi),$$

which implies, via uniqueness, that $\tilde{\mu}(\cdot) = \mu(\cdot|\theta_0)$. This argument yields the continuity of $\int f(\xi)\mu(d\xi|\theta)$.

Part 2. *A martingale problem representation.* By (3.1) and (3.3)

$$\theta^\epsilon(t) = \theta_0^\epsilon + \epsilon \sum_{i=0}^{t/\epsilon-1} G_i^\epsilon(\theta_i^\epsilon, \xi_i^\epsilon) + \epsilon \sum_{i=0}^{t/\epsilon-1} \delta Y_i^\epsilon. \tag{3.13}$$

First, we show that the martingale term (the one on the right) goes to zero as $\epsilon \to 0$. This would be easy if the uniform integrability in (3.2) were replaced by square integrability, since then the martingale would be square integrable and its variance at t would be bounded by $\epsilon^2(t/\epsilon)\sup_{\epsilon,n} \operatorname{var}(\delta Y_n^\epsilon) = O(\epsilon t)$. Hence the term would have the zero process as a weak limit. We get the same result by a truncation argument, as follows. For large positive B, let $I_{n,B}^\epsilon$ be the indicator function of the event that Y_n^ϵ does not exceed B in absolute magnitude. Then use $Y_n^\epsilon I_{n,B}^\epsilon$ in lieu of Y_n^ϵ, as follows. Define $\delta Y_{n,B}^\epsilon$ and $\beta_{n,B}^\epsilon$ by

$$Y_n^\epsilon I_{n,B}^\epsilon = E_n^\epsilon Y_n^\epsilon I_{n,B}^\epsilon + \delta Y_{n,B}^\epsilon, \qquad Y_n^\epsilon = Y_n^\epsilon I_{n,B}^\epsilon + \beta_{n,B}^\epsilon.$$

We have $\sup_{\epsilon,n} E|\beta_{n,B}^\epsilon| \to 0$ as $B \to \infty$ by the uniform integrability. Since $\{\delta Y_{n,B}^\epsilon\}$ are bounded, for each $B < \infty$ the martingale term $\epsilon \sum_{i=0}^{t/\epsilon-1} \delta Y_{i,B}^\epsilon$ contributes nothing to the limit by the "square integrability" theory. Now the uniform integrability (3.2) yields

$$\limsup_B \sup_{n,\epsilon} E|E_n^\epsilon Y_n^\epsilon I_{n,B}^\epsilon - G_n^\epsilon(\theta_n^\epsilon, \xi_n^\epsilon)| = 0.$$

[1] If, for a sequence $Z_n, E|Z_n| \to 0$, we say that it converges in mean to zero.

These results imply that (3.13) can be written as

$$\theta^\epsilon(t) = \theta^\epsilon(0) + \epsilon \sum_{j=0}^{t/\epsilon-1} G_j^\epsilon(\theta_j^\epsilon, \xi_j^\epsilon) + \rho^\epsilon(t), \tag{3.14}$$

where $|\rho^\epsilon(t)| \to 0$ in the mean uniformly on each bounded t-interval. Now the uniform integrability (3.2) and the form (3.1) imply that $\{\theta^\epsilon(\cdot), \epsilon > 0\}$ is tight and that any weak limit has Lipschitz-continuous paths w.p.1 (see Lemma 2.1).

The conditions seem to be weakest if we work with a "martingale problem" formulation, and we proceed to do so. Now, with a slight abuse of notation, let ϵ index a weakly convergent subsequence of $\{\theta^\epsilon(\cdot), \epsilon > 0\}$ with limit process denoted by $\theta(\cdot)$. Let t, τ be arbitrary positive numbers; q be an integer; $s_i, i \le q$, be nonnegative numbers no larger than t; and $h(\cdot)$ be a bounded, continuous, and real-valued function of its arguments. As is common in weak convergence–type arguments, we will show that

$$Eh(\theta(s_i), i \le q)\left[\theta(t+\tau) - \theta(t) - \int_t^{t+\tau} g(\theta(u))du\right] = 0. \tag{3.15}$$

By the arbitrariness of the $h(\cdot), q, t, \tau, s_i$, (3.15) implies that $\theta(t) - \theta(0) - \int_0^t g(\theta(u))du$ is a martingale (with respect to the filtration which it generates). Since $E|\rho^\epsilon(\cdot)| \to 0$ and $\{G_n^\epsilon(\theta_n^\epsilon, \xi_n^\epsilon), \epsilon > 0, n < \infty\}$ is uniformly integrable, the form (3.14) implies that the martingale has zero quadratic variation; hence it is constant. Since it takes the value zero at $t = 0$, it is identically zero w.p.1. Thus, the theorem will be proved once (3.15) is proved.

Part 3. *Approximating the* $G_j^\epsilon(\cdot)$. By the properties of $\rho^\epsilon(\cdot)$, we can write

$$Eh(\theta^\epsilon(s_i), i \le q)\left[\theta^\epsilon(t+\tau) - \theta^\epsilon(t) - \epsilon \sum_{j=t/\epsilon}^{(t+\tau)/\epsilon-1} G_j^\epsilon(\theta_j^\epsilon, \xi_j^\epsilon)\right] \to 0 \tag{3.16}$$

as $\epsilon \to 0$. We proceed to rearrange the terms in (3.16) so that efficient averaging methods can be used. Let $n_\epsilon \to \infty$ be a sequence of integers such that $\delta_\epsilon = \epsilon n_\epsilon \to 0$ and τ is an integral multiple of δ_ϵ. Without loss of generality and for notational simplicity, suppose that t is also an integral multiple of δ_ϵ. By collecting terms in groups of size n_ϵ and using the freedom that we have with taking the conditional expectations given "past data" inside the brackets in (3.16), we can write the left side of (3.16) as

$$Eh(\theta^\epsilon(s_i), i \le q)\left\{\theta^\epsilon(t+\tau) - \theta^\epsilon(t) - \sum_{l:l\delta_\epsilon=t}^{t+\tau-\delta_\epsilon} \delta_\epsilon\left[\frac{1}{n_\epsilon}\sum_{j=ln_\epsilon}^{ln_\epsilon+n_\epsilon-1} E_{ln_\epsilon}^\epsilon G_j^\epsilon(\theta_j^\epsilon, \xi_j^\epsilon)\right]\right\}. \tag{3.17}$$

For a real-valued function $f(\cdot)$ and $\infty > B > 0$, define $f_B(\cdot)$ by $f_B(x) = \min[f(x), B]$ for $f(x) \ge 0$ and by $f(x) = \max[f(x), -B]$ otherwise.

By the uniform integrability (3.2), given any $\rho > 0$ there is $B < \infty$ such that we can use $G_{n,B}^\epsilon(\cdot)$ while changing the expectations of the absolute values of the summands in the brackets of (3.17) by at most ρ. Continuing in the bracketed term, first replace $G_j^\epsilon(\theta_j^\epsilon, \xi_j^\epsilon)$ with $G_{j,B}^\epsilon(\theta_j^\epsilon, \xi_j^\epsilon)$ plus a small error term. Then use (3.9) (which also holds for the B-truncated functions) to replace $G_{j,B}^\epsilon(\theta_j^\epsilon, \xi_j^\epsilon)$ with $G_B(\theta_j^\epsilon, \xi_j^\epsilon)$ plus

a small error. Finally, use the (uniform in compact (θ, ξ) sets) continuity of $G_B(\cdot, \xi)$, (3.4) and the fact that (in probability, uniformly in l)

$$\sup_{j \le n_\epsilon} |\theta^\epsilon_{ln_\epsilon + j} - \theta^\epsilon_{ln_\epsilon}| \to 0 \quad \text{as } \epsilon \to 0 \tag{3.18}$$

to justify replacing θ^ϵ_j by $\theta^\epsilon_{ln_\epsilon}$ (plus a small error term), yielding that (3.17) equals (3.19)

$$Eh(\theta^\epsilon(s_i), i \le q) \left\{ \theta^\epsilon(t+\tau) - \theta^\epsilon(t) - \sum_{l: l\delta_\epsilon = t}^{t+\tau-\delta_\epsilon} \delta_\epsilon \left[\frac{1}{n_\epsilon} \sum_{j=ln_\epsilon}^{ln_\epsilon + n_\epsilon - 1} E^\epsilon_{ln_\epsilon} G_B(\theta^\epsilon_{ln_\epsilon}, \xi^\epsilon_j) \right] \right\}$$

modulo an error ρ' which can be made as small as desired in mean value by choosing B large enough and then ϵ small enough. The sum in (3.19) can be written as $\int_t^{t+\tau} \tilde{G}^\epsilon_B(s)ds$, with the obvious definition of $\tilde{G}^\epsilon_B(\cdot)$ as the process which is constant on intervals $[l\delta_\epsilon, (l+1)\delta_\epsilon)$, as defined[2] by the bracketed term.

The weak convergence arguments in the next parts will show that

$$Eh(\theta^\epsilon(s_i), i \le q)\tilde{G}^\epsilon_B(s) \to Eh(\theta(s_i), i \le q) g_B(\theta(s)), \tag{3.20}$$

where $g_B(\theta) = \int G_B(\theta, \xi)\mu(d\xi|\theta)$. This will imply that the outer sum in (3.19) can be replaced by $\int_t^{t+\tau} g_B(\theta(s))ds$ in the limit as $\epsilon \to 0$. These results will yield that

$$\dot{\theta} = g_B(\theta) + \rho_B, \tag{3.21}$$

where $E\int_0^t |\rho_B(s)|ds \to 0$ as $B \to \infty$. The proof under (3.11a) will then be completed in part 5 by showing that we can let $B = \infty$. Part 6 will deal with (3.11b).

Part 4. *Averaging out the* ξ^ϵ_j. To complete our program, we need to average out the ξ^ϵ_j terms in (3.19). Define the measure-valued random variable (an average of conditional probabilities)

$$R(l, \epsilon, \cdot) = \frac{1}{n_\epsilon} \sum_{j=ln_\epsilon}^{ln_\epsilon + n_\epsilon - 1} P\{\xi^\epsilon_j \in \cdot | \theta^\epsilon_{ln_\epsilon}, \xi^\epsilon_{ln_\epsilon}\}, \tag{3.22}$$

and recall that the $E^\epsilon_{ln_\epsilon}$ is the expectation conditioned on $(\theta^\epsilon_{ln_\epsilon}, \xi^\epsilon_{ln_\epsilon})$ by the Markov assumption. The inner square bracketed term in (3.19) can now be written

$$\int R(l, \epsilon, d\xi) G_B(\theta^\epsilon_{ln_\epsilon}, \xi). \tag{3.23}$$

The set of measure-valued random variables $\{R(l, \epsilon, \cdot), l < \infty, \epsilon > 0\}$ is tight, since the mean values are just

$$\bar{R}(l, \epsilon, \cdot) = \frac{1}{n_\epsilon} \sum_{j=ln_\epsilon}^{ln_\epsilon + n_\epsilon - 1} P\{\xi^\epsilon_j \in \cdot\}$$

[2] We note here that the major problem in averaging the inner square bracket is in showing that the ξ^ϵ_j in the inner sum in (3.19) can be replaced by $\xi_j(\theta^\epsilon_{ln_\epsilon})$, all of which have the same value of θ as an argument, so that some sort of ergodic theorem or averaging principle can be used to get the ultimate averaged limit. Recall that $\xi_i(\theta)$ is the Markov chain with fixed parameter θ. This idea is basic to all of the averaging methods. The continuity (3.7) of the transition function is the basic property that is used. Various alternatives will appear in the appendices.

and the tightness of $\{\bar{R}(l,\epsilon,\cdot), l \geq 0, \epsilon > 0\}$ is just the tightness of $\{\xi_n^\epsilon, \epsilon > 0, n < \infty\}$, and this latter sequence is tight by assumption (3.4) (see Lemma 2.2). We will characterize the limits of weakly convergent subsequences of $\{R(l,\epsilon,\cdot), l < \infty, \epsilon > 0\}$ as measures whose values are (w.p.1) just the invariant measure $\mu(\cdot|\theta)$ with appropriate values of θ. Now we follow the ideas in the development in [22, p. 110], except for the use of the random measures in place of their pointwise values. (We note that the use of random measures here greatly simplifies the treatment of the unbounded noise case over that in the references.)

Fix $s > 0$, and let l_ϵ be such that $s \in [l_\epsilon\delta_\epsilon, l_\epsilon\delta_\epsilon + \delta_\epsilon)$ for all ϵ. Let $f(\cdot)$ be a bounded and continuous real-valued function of ξ. Using (3.8), extract a weakly convergent subsequence of $\{R(l_\epsilon,\epsilon,\cdot), \theta^\epsilon(\cdot), \epsilon > 0\}$, and index it by $\epsilon(p), p \to \infty$. The proof will show that this further subsequence is irrelevant, due to the uniqueness of the $\mu(\cdot|\theta)$, and that we can let $\epsilon(p) = \epsilon$. We also suppose that the Skorohod representation is used so that the convergences are w.p.1 in the appropriate topologies. Denote the limit by $(\tilde{R}(\cdot), \theta(\cdot))$.[3] Define $m_{\epsilon(p)} = l_{\epsilon(p)} n_{\epsilon(p)}$. Note that $m_\epsilon(p) \to \infty$ as $p \to \infty$. We can write

$$\begin{aligned}\int \tilde{R}(d\xi) f(\xi) &= \lim_{p\to\infty} \int R(l_{\epsilon(p)}, \epsilon(p), d\xi) f(\xi) \\ &= \lim_{p\to\infty} \frac{1}{n_{\epsilon(p)}} \sum_{j=m_{\epsilon(p)}}^{m_{\epsilon(p)}+n_{\epsilon(p)}-1} \int P\{\xi_j^{\epsilon(p)} \in d\xi | \theta^{\epsilon(p)}_{m_{\epsilon(p)}}, \xi^{\epsilon(p)}_{m_{\epsilon(p)}}\} f(\xi).\end{aligned} \tag{3.24}$$

The first equality follows from the definition of the weak limit. The second follows from definition of $R(l_\epsilon,\epsilon,\cdot)$. Continuing, we use the one-step transition function $P_j^{\epsilon(p)}(\cdot)$ to rewrite the right side of (3.24) as (minus the first term of the sum)

(3.25)

$$\lim_p \frac{1}{n_{\epsilon(p)}} \sum_{j=m_{\epsilon(p)}+1}^{m_{\epsilon(p)}+n_{\epsilon(p)}-1} \int\int P\{\theta_{j-1}^{\epsilon(p)} \in d\tilde{\theta}, \xi_{j-1}^{\epsilon(p)} \in d\tilde{\xi} | \theta^{\epsilon(p)}_{m_{\epsilon(p)}}, \xi^{\epsilon(p)}_{m_{\epsilon(p)}}\} P_{j-1}^{\epsilon(p)}(\tilde{\xi}, d\xi|\tilde{\theta}) f(\xi).$$

Condition (3.6) yields

$$\lim_{p,n} \int P_n^{\epsilon(p)}(\tilde{\xi}, d\xi|\tilde{\theta}) f(\xi) = \int P(\tilde{\xi}, d\xi|\tilde{\theta}) f(\xi) \equiv \bar{f}(\tilde{\theta}, \tilde{\xi}),$$

and the limit is uniform on each compact $(\tilde{\theta}, \tilde{\xi})$ set. By (3.7), the right-hand side is continuous. By using these facts and the fact that the limit $\theta(\cdot)$ is continuous, we can concentrate the measure of $\tilde{\theta}$ in (3.25) at $\theta^{\epsilon(p)}_{m_{\epsilon(p)}}$ without affecting the limit. With this replacement, (3.25) can be written as

$$\lim_p \int R(l_{\epsilon(p)}, \epsilon(p), d\xi) \bar{f}(\theta^{\epsilon(p)}_{m_{\epsilon(p)}}, \xi),$$

which by the use of the weak convergence of $\{R^{\epsilon(p)}(l_{\epsilon(p)}, \epsilon(p), \cdot), \theta^{\epsilon(p)}_{m_{\epsilon(p)}}\}$ to $(\tilde{R}(\cdot), \theta(s))$ as $p \to \infty$ equals

$$\int \tilde{R}(d\xi) \bar{f}(\theta(s), \xi) = \int\int \tilde{R}(d\tilde{\xi}) P(\tilde{\xi}, d\xi|\theta(s)) f(\xi). \tag{3.26}$$

[3] Strictly speaking, when taking limits of $\{\tilde{R}(l_{\epsilon(p)}, \epsilon(p), \cdot), \theta^{\epsilon(p)}(\cdot), p < \infty\}$ and using Skorohod representation, the probability space might be different from what was used when we got the original weakly convergent subsequence with limit $\theta(\cdot)$ in part 2. But since $\{\theta^{\epsilon(p)}(\cdot), p < \infty\}$ is a subsequence of $\{\theta^\epsilon(\cdot), \epsilon > 0\}$ and all that matters are the distributions of the resulting limits anyway, we write $\theta(\cdot)$ for the limit as $p \to \infty$ for notational simplicity and without loss of generality.

Let ω denote the canonical probability space variable. Equating the right side of (3.26) with the left-hand side of (3.24) yields that (w.p.1) each sample value $\tilde{R}(\cdot,\omega)$ must be an invariant measure for the transition probability $P(\xi,\cdot|\theta(s,\omega))$. By the uniqueness of the invariant measure, $\tilde{R}(\cdot)=\mu(\cdot|\theta(s))$ w.p.1, and the subsequence of $\{R(l_\epsilon,\epsilon,\cdot),\epsilon>0\}$ which is used is irrelevant. This implies that the limit in (3.23) is $g_B(\theta(s))$ w.p.1, which yields (3.20).

Part 5. *Replacing* $g_B(\cdot)$ *by* $g(\cdot)$. The results of the previous parts imply that we can replace the square bracketed term of (3.20) by (3.19) by $g_B(\theta(s))$ and that (3.20), (3.21) hold. We need only show that we can let $B\to\infty$ in (3.21). We can let $B\to\infty$ and replace $g_B(\cdot)$ by $g(\cdot)$ if $G(\theta,\cdot)$ is $\mu(\cdot|\theta)$ integrable for each θ, the integral is bounded on each compact θ set, and

$$\int G_B(\theta,\xi)\mu(d\xi|\theta)\to\int G(\theta,\xi)\mu(d\xi|\theta) \tag{3.27}$$

uniformly on each compact θ set. But this follows from (3.10) and the monotone convergence theorem.

Part 6. *Using* (3.11b). If we drop the uniqueness condition, then the subsequence $\epsilon(p)$ might be important. However, note that the above proof established that, whether or not there is uniqueness,

$$\dot{\theta}(s)=\int G(\theta(s),\xi)\mu(d\xi|\theta(s))$$

for some invariant measure, which might depend on (ω,s). But (3.11b) implies that the right-hand side equals $g(\theta(s))$ for the function $g(\cdot)$ defined there. □

3.2. Limit points and nonunique invariant measures: Limit points of (3.12). We use some elementary facts from the theory of differential equations. Given an ODE $\dot{x}=f(x)$, $x\in R^r$, with continuous $f(\cdot)$ and a bounded solution $x(\cdot)$ on $[0,\infty)$, let L denote the set of limit points of the path $x(\cdot)$. Define an *invariant set* M for the ODE as follows. For each $y\in M$, there is a solution $y(\cdot)$ to the ODE on $(-\infty,\infty)$ such that $y(t)\in M$ for all t and $y(0)=y$. Then [15] L is a compact invariant set. We can now state the following result.

THEOREM 3.2. *Assume the conditions of Theorem* 3.1. *The limit points of* (3.12) *are contained in the largest bounded invariant set* M *of* $\dot{x}=g(x)$. *Now let* $\epsilon q_\epsilon\to\infty$ *and* $\theta^\epsilon(\epsilon q_\epsilon+\cdot)\Rightarrow\theta(\cdot)$ *as* $\epsilon\to 0$. *Then, w.p.1 for each* $t\in(-\infty,\infty)$, $\theta(t)\in M$.

Remark. The last assertion holds since the solution is defined on the doubly infinite time interval. If the ODE has a unique stationary point $\hat{\theta}$, then the last assertion implies that $\theta(t)=\hat{\theta}$ for all t. Appropriate perturbation schemes will guarantee that the iterate won't get stuck at a maximum or at a saddle point. Reference [1] shows that the set of limit points are confined to the set of chain recurrent points, which might be smaller than the largest bounded invariant set, but the conditions are stronger.

Nonunique invariant measure. Suppose that $\mu(\cdot|\theta)$ is not unique and (3.11b) cannot be verified. We might still be able to get a useful result. Let $V(\theta)$ denote the (convex) set of invariant measures under θ. The proof of Theorem 3.1 can be easily modified to get the following theorem.

THEOREM 3.3. *Assume the conditions of Theorem* 3.1 *except for* (3.11). *Then w.p.1 and for almost all* t, *the theorem holds with* (3.12) *replaced by*

$$\dot{\theta}(t)\in\left\{\int G(\theta(t),\xi)\mu(d\xi|\theta(t)),\ \mu(\cdot|\theta(t))\in V(\theta(t))\right\}. \tag{3.28}$$

3.3. "Atomic" increments. Return to the basic algorithm (3.1). In many applications the Y-variables have an additivity property which can simplify the verification of the conditions and which we now explain and exploit. Define $q_0^\epsilon = 0$ and suppose that we update at the increasing random times $q_n^\epsilon, n = 1, \ldots$. Let $k^\epsilon(n)$ be the last time of updating before and including time n. By additivity, we mean that the observations can be divided up such that the algorithm can be written

$$\theta_{n+1}^\epsilon = \theta_n^\epsilon + \epsilon \sum_{i=q_n^\epsilon}^{q_{n+1}^\epsilon - 1} Y_i^\epsilon, \tag{3.29}$$

where the Y_n^ϵ obey the conditions of Theorem 3.1. At each instant i in the interval $[q_n^\epsilon, q_{n+1}^\epsilon)$ the value θ_n^ϵ is used to get Y_i^ϵ. From the point of view of the convergence theory, one can just as well update in "real time" and use the modified algorithm

$$\theta_{n+1}^\epsilon = \theta_n^\epsilon + \epsilon Y_n^\epsilon, \tag{3.30}$$

where $\theta_{k^\epsilon(n)}^\epsilon$ is used to get Y_n^ϵ. Suppose that $q_{n+1}^\epsilon - q_n^\epsilon$ is bounded by some constant independently of ϵ, n. Then the conditions of Theorem 3.1 guarantee that its conclusions hold for the SA (3.30) and similarly for its extensions. If the intervals $q_{n+1}^\epsilon - q_n^\epsilon$ are unbounded, then in order for the two time scales of (3.29) and (3.30) to be compatible, we need in addition that the conditions of Theorem 3.1 hold for the original algorithm, in particular that $[\theta_{n+1}^\epsilon - \theta_n^\epsilon]/\epsilon$ be uniformly (in n, ϵ) integrable.

The advantage of this "atomic" decomposition is that it makes it easier to verify the conditions on the Markov chain ξ_n^ϵ for (3.30) than for (3.39), since the transitions are viewed "more locally."

3.4. Updating at regeneration times. Suppose that the problem has a structure that allows $g(\theta)$ to be estimated regeneratively. When $g(\theta)$ is the derivative of $C(\theta)$, a continuously differentiable, "stationary cost" function of a regenerative process, an excellent treatment of the regenerative estimation of the derivative is in [14]. We might wish to use SA to minimize $C(\theta)$. In particular, suppose that θ_n^ϵ will be updated at the end of each new k regeneration intervals and that there are Y_n^ϵ which are "nearly" unbiased estimators of $g(\theta_n^\epsilon)$ and which depend on data in regeneration intervals $[nk+1, nk+k]$ only. Let (3.1) be used. Define

$$G_n^\epsilon(\theta) = E[Y_n^\epsilon | \theta_n^\epsilon = \theta, \theta_i^\epsilon, i < n] = E[Y_n^\epsilon | \theta_n^\epsilon = \theta].$$

Assume that $\{\theta_n^\epsilon, \epsilon > 0, n < \infty\}$ is tight, $\{Y_n^\epsilon, \epsilon > 0, n < \infty\}$ is uniformly integrable, and, for each $\delta > 0$,

$$\lim_\epsilon \limsup_n P\{|G_n^\epsilon(\theta_n^\epsilon) - g(\theta_n^\epsilon)| \geq \delta\} = 0. \tag{3.31}$$

Then a simpler proof than that of Theorem 3.1 says that the conclusions of Theorem 3.1 hold (and similarly for the other theorems). The proof is simpler since there is no need to introduce ξ_n^ϵ or the averaging measures. We note that sometimes the minimization of an average cost per unit time can be reduced to an SA iteration where we update at the end of each k intervals for some integer k. See, for example, §2.2 of [32]. The results of Theorem 3.1, as expanded in Appendices 3 and 4, show that it is not necessary to use regeneration intervals as the basis for updating. Indeed, updating only at the ends of these intervals might be a poor idea in practice in general, despite the fact that the proofs are simplified. For network problems, a regeneration

interval-based approach would be a handicap, since the intervals would generally be very long. The situation is even worse if the processing is distributed (Appendix 4). The same point was made by [35].

4. Time-varying gains $\epsilon_n \to 0$. Suppose that the positive real numbers ϵ_n go to zero such that

$$\sum_{n=0}^{\infty} \epsilon_n = \infty \tag{4.1a}$$

and

$$\text{either } \sum_n |\epsilon_{n+1} - \epsilon_n| < \infty \text{ or } \epsilon_n/\epsilon_{n+1} \to 1. \tag{4.1b}$$

These ϵ_j could actually be random if they are nonanticipative and satisfy (4.1). The SA algorithm is

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n. \tag{4.2}$$

Let $\mathcal{B}_n$ be a sequence of nondecreasing sigma-algebras measuring at least $\{\theta_0, Y_i, i < n\}$. Write E_n for the conditional expectation given $\mathcal{B}_n$. The δY_n defined by $Y_n = E_n Y_n + \delta Y_n$ are $\mathcal{B}_n$-martingale differences. As for Theorem 3.1, suppose that there is a process $\{\xi_n, n < \infty\}$ taking values in a CSMS and functions $G_n(\cdot)$ such that $E_n Y_n = G_n(\theta_n, \xi_n)$. Define $t_n = \sum_{i=0}^{n-1} \epsilon_i$. Following the definition (2.2) for $t \geq 0$, define $\theta^0(t) = \theta_n$ on $[t_n, t_{n+1})$ and set $\theta^0(t) = \theta_0$ on $(-\infty, 0]$. Define $\theta^n(t) = \theta^0(t_n + t)$. Thus $\theta^n(0) = \theta_n$.

Theorems 3.1–3.3 readily lead to the following theorem.

THEOREM 4.1. *Assume the conditions of Theorem* 3.1 *with* (4.1), (4.2), *and the following replacements. Equations* (3.2)–(3.6) *and* (3.10) *hold with the superscript* ϵ *dropped. Use* $\limsup_n$ *in* (3.9). *Then* $\{\theta^n(\cdot), n < \infty\}$ *is tight and the limit of any weakly convergent subsequence satisfies* (3.12) *on* $(-\infty, \infty)$ *w.p.*1. *Also, w.p.*1., *for all* t, $\theta(t) \in M$, *the largest bounded invariant set of* (3.12). *If* (3.11) *is dropped, the the conclusions of Theorem* 3.3 *still hold. The obvious analogue of the results for the "atomic" increments formulation also hold.*

Remarks on the proof. Again, the general structure is similar to that used in [28] but with differing details. The proof is essentially the same as those of Theorems 3.1 and 3.3. The only difference concerns the way the terms are grouped, i.e., the analogy to the arrangement in (3.17). For simplicity, let $t \geq 0, \tau > 0$. Define $m_n(t) = \max\{j \geq n : t_j - t_n \leq t\}$. The following expression replaces (3.16):

$$Eh(\theta^n(s_i), i \leq q)\left[\theta^n(t+\tau) - \theta^n(t) - \sum_{j=m_n(t)}^{m_n(t+\tau)-1} \epsilon_j (G_j(\theta_j, \xi_j) + \delta Y_j)\right] = 0. \tag{4.3}$$

The martingale term

$$\sum_{j=m_n(t)}^{m_n(t+\tau)-1} \epsilon_j \delta Y_j$$

goes to zero as $n \to \infty$ by a bounding argument of the type used in Theorem 3.1 and the fact that $\epsilon_n \to 0$. Note in particular that no condition of the form $\sum \epsilon_j^{1+\delta} < \infty$

for $\delta > 0$ is needed. It is only required that $\epsilon_n \to 0$. Indeed, if the $\{Y_n, n < \infty\}$ were uniformly square integrable, then the variance of the martingale term would be $O(1)\sum_{m_n(t)}^{m_n(t+\tau)-1} \epsilon_j^2$ and this goes to zero if $\epsilon_j \to 0$, since $\sum_{j=m_n(t)}^{m_n(t+\tau)-1} \epsilon_j \approx \tau$. In general, one uses uniform integrability to the same end as in part 2 of the proof of Theorem 3.1.

We now comment briefly on the appropriate grouping of the terms. The δ_ϵ used for the grouping in (3.17) is replaced with a sequence of positive numbers $\delta_n \to 0$ which satisfy $\lim_n \sup\{\epsilon_j/\delta_n : j \geq n\} = 0$. Define $m(n,0) = n$. For each n, define an increasing sequence of integers $m(n,l), l = 1, \ldots,$ by

$$m(n,l) = \min\left\{j : \sum_{i=n}^{j-1} \epsilon_i \geq l\delta_n\right\}.$$

Thus

$$\sum_{j=m(n,l)}^{m(n,l+1)-1} \epsilon_j \approx \delta_n.$$

For each n, we will arrange the terms in groups of successive sizes $m(n,l+1) - m(n,l)$ as follows. Suppose, for notational simplicity, that both t and $t+\tau$ are integral multiples of δ_n. The changes for the general case should be obvious. Now, analogously to what was done in part 3 of the proof of Theorem 3.1, replace the sum of the $\epsilon_j G_j$ in (4.3) with

$$\sum_{l:l\delta_n=t}^{(t+\tau)-\delta_n} \delta_n \left[\frac{1}{\delta_n} E_{m(n,l)} \sum_{j=m(n,l)}^{m(n,l+1)-1} \epsilon_j G_j(\theta_j, \xi_j)\right]. \tag{4.4}$$

Then argue that the $G_j(\theta_j, \xi_j)$ in the sum on the right can be replaced by $G_B(\theta_{m(n,l)}, \xi_j)$ plus an arbirarily small (in the mean) error for large B, n. Define the measure-valued random variables $R(l,n,\cdot)$ by

$$R(l,n,\cdot) = \frac{1}{\delta_n} \sum_{j=m(n,l)}^{m(n,l+1)-1} \epsilon_j P\{\xi_j \in \cdot | \theta_{m(n,l)}, \xi_{m(n,l)}\}. \tag{4.5}$$

Thus, as in part 4 of the proof of Theorem 3.1, we approximate the bracketed term in (4.4) by

$$\int G_B(\theta_{m(n,l)}, \xi) R(l,n,d\xi).$$

Next consider the analogue of the factorization taking the sum on the right side of (3.24) to that in (3.25). The analogue of (3.25) (before extracting the convergent subsequence and without the limit) is

$$\frac{1}{\delta_n} \sum_{j=m(n,l)+1}^{m(n,l+1)-1} \int\int \epsilon_j P\{\theta_{j-1} \in d\tilde{\theta}, \xi_{j-1} \in d\tilde{\xi} | \theta_{m(n,l)}, \xi_{m(n,l)}\} P_{j-1}(\tilde{\xi}, d\xi | \tilde{\theta}) f(\xi).$$

Analogously to part 4 of the proof of Theorem 3.1, we can fix $\tilde{\theta}$ at $\theta_{m(n,l)}$ and replace $P_j(\cdot)$ with $P(\cdot)$ to get the representation

$$\int R(l,n,d\tilde{\xi}) \int f(\xi) P(\tilde{\xi}, d\xi | \theta_{m(n,l)}),$$

where ϵ_{j-1} replaces ϵ_j in the definition of $R(\cdot)$. By either option in (4.1b), this replacement does not affect the limit.

5. A constrained algorithm. Let H be a closed set in R^r. Let $\Pi_H(x)$ denote the closest point in H to x. The following development is a slight extension of [22, pp. 111–114]. Define the projected form of (3.1) as

$$\theta_{n+1}^{\epsilon} = \Pi_H(\theta_n^{\epsilon} + \epsilon Y_n^{\epsilon}). \tag{5.1}$$

Rewrite (5.1) as

$$\theta_{n+1}^{\epsilon} = \theta_n^{\epsilon} + \epsilon Y_n^{\epsilon} + \epsilon z_n^{\epsilon},$$

where z_n^{ϵ} is the "correction term." The decomposition (5.1) is the key to the analysis and first appeared in [25]. Under (3.2), the sequence $\{z_n^{\epsilon}, \epsilon > 0, n < \infty\}$ is uniformly integrable. Let q_ϵ be a sequence of integers satisfying (3.8′). The proof of Theorem 3.1 yields immediately that the limit of any convergent subsequence of $\{\theta^{\epsilon}(q_\epsilon \epsilon + \cdot), \epsilon > 0\}$ as $\epsilon \to 0$ has the form

$$\dot{\theta} = g(\theta) + z, \tag{5.2}$$

where $\int_0^t z(s)ds$ is the limit of the process with values $\epsilon \sum_{n=q_\epsilon}^{q_\epsilon + t/\epsilon - 1} z_n^{\epsilon}$ for $t \geq 0$, and with the obvious change for $t < 0$. Thus the only problem concerns the characterization of $z(\cdot)$. By the uniform integrability, $(z(\cdot), \theta(\cdot))$ are Lipschitz continuous w.p.1 (Lemma 2.1). Clearly, $z(u) = 0$ on any interval $(t, t+\tau)$ in which $\theta(u) \in H^0$, the interior of H. To proceed, we need to specify H more fully, and we assume either of I or II below.

I. Let $q_i(\cdot), i = 1, \cdots, p$, be continuously differentiable real-valued functions on R^r, with gradients $q_{i,x}(\cdot)$. Without loss of generality, let $q_{i,x}(x) \neq 0$ if $q_i(x) = 0$. Define $H = \{x : q_i(x) \leq 0, i = 1, \ldots, p\}$ and assume that it is nonempty. Define $A(x)$, the set of active constraints at x, by $A(x) = \{i : q_i(x) = 0\}$. Define $C(x)$ to be the closed convex cone generated by $\{y : y = q_{i,x}(x), i \in A(x)\}$. Suppose that for each x with nonempty $A(x)$, the set $\{q_{i,x}(x), i \in A(x)\}$ is linearly independent.

II. H is an R^{r-1}-dimensional connected surface with a continuously differentiable outer normal. In this case, define $C(x), x \in H$, to be just the linear span of the outer normal at x.

THEOREM 5.1. *Assume the conditions above and the conditions of Theorem* 3.1. *Then the conclusions of Theorem* 3.1 *and Theorem* 3.2 *hold with* (5.2) *replacing* (3.12) *and* $z \in -C(\theta(t))$, *where the limit points of* (5.2) *replace* M. *If* (3.11) *is dropped, then the the conclusions of Theorem* 3.3 *still hold. The same conclusions hold for the constrained form of Theorem* 4.1.

Proof. The basic proof is a straightforward extension of that of Theorem 3.1. To characterize $z(t)$ we use the fact that if for any (t, x), $\theta^{\epsilon}(t) = x$, then $z_{t/\epsilon}^{\epsilon}$ is in a small neighborhood of $-C(y)$ for some y near x when ϵ is small. Then use the fact that $C(x)$ is upper semicontinuous in the sense that if $N_\delta(x)$ is a $\delta-$neighborhood of x, then

$$\cap_{\delta > 0} \cup_{y \in N_\delta(x)} C(y) \subset C(x);$$

i.e., the set of active constraints at x contains that for points very close to it. □

Note. If, under I, there is only one active constraint (say, i) at t, and $g(\theta(t))$ points out of H, then the right-hand side of (5.2) is just the projection of $g(\theta(t))$ onto the boundary surface.

6. Applications of Theorem 3.1: Introduction. As they are stated, the results in §§3–5 do not explicitly deal with the optimization of an average cost over an infinite interval. In the examples in the remaining sections, we show that they are very powerful tools for proving convergence for just such problems. Three canonical examples of optimization will be described in detail. All use some approximation to a gradient search procedure. We will use constant step sizes as in Theorem 3.1, but the extensions to the decreasing step-size case will follow immediately from Theorem 4.1. Note that the constant step-size case $\epsilon_n = \epsilon$ has applications in tracking and adaptive control also. The examples concern the minimization of a stationary average cost associated with the path of a dynamical system. This section deals with a general discussion of the issues. Section 7 concerns a discrete event dynamical system example and an IPA-type estimator [13, 18]. Section 8 concerns a "piecewise deterministic" example, also using an IPA-type estimator and involving a problem in manufacturing. The third example involves a stochastic differential equations model. The examples are illustrative of many others using various methods of estimating derivatives.

Let us consider a canonical continuous time model in a rather informal way, since we wish only to illustrate the basic ideas in an unincumbered way. The general considerations hold also for discrete-time models, as will be seen in the next two sections. Among the points to be clarified is the so-called resetting of the IPA "accumulator." It will be seen that it is often neither necessary nor desirable. The basic ideas are in [22, 27], but their full potential has not been realized in the literature.

Suppose that for fixed parameter θ, $x(\cdot,\theta)$ represents the dynamical state process of the system. In order to fix ideas, let $x(\cdot,\theta)$ be defined by the SDE

$$dx(t,\theta) = b(x(t,\theta),\theta))dt + dw.$$

For the sake of simple notation, let both x and θ be real valued and the function $b(\cdot)$ be smooth enough so that the following calculations make sense. We return to this example in a more thorough way in §9. For initial condition $x(0,\theta) = x(0)$, fixed parameter θ, and cost rate $c(\theta, x(s,\theta))$, define the average cost per unit time on $[0,T]$ by

$$C_T(\theta, x(0)) = E\frac{1}{T}\int_0^T c(\theta, x(s,\theta))ds.$$

Suppose that $C_T(\theta, x(0))$ is continuously differentiable with respect to θ with gradient $C_{T,\theta}(\theta, x(0))$. Suppose that for each θ the limit $C(\theta) = \lim_T C_T(\theta, x(0))$ exists and does not depend on $x(0)$. Suppose that the pointwise limit of $C_{T,\theta}(\theta, x(0))$ exists and is denoted by $\hat{C}_\theta(\theta)$. Then $\hat{C}_\theta(\theta) = C_\theta(\theta)$.

We wish to use SA to minimize $C(\theta)$. A common procedure for updating θ via gradient seach is based on the consistency of $C_{T,\theta}(\theta, x(0))$; i.e., it is a good estimator of $C_\theta(\theta)$ if T is large. Pursuing this idea, let us update the parameter at times $nT, n = 1, 2, \ldots$, as follows. Letting θ_n^ϵ denote the nth choice of the parameter, use it on $[nT, nT+T)$ to get an estimator $\hat{Y}_n^\epsilon$ of $-C_{T,\theta}(\theta_n^\epsilon, x(\theta_n^\epsilon, nT))$. Then use

$$\theta_{n+1}^\epsilon = \theta_n^\epsilon + \epsilon\hat{Y}_n^\epsilon. \tag{6.1}$$

Let $x^\epsilon(\cdot)$ (with $x^\epsilon(0) = x(0)$) denote the actual physical state process with the time varying θ_n^ϵ used, i.e., on $[nT, nT+T)$ $x^\epsilon(t) = x(t, \theta_n^\epsilon)$ with the "initial condition" of $x(\cdot, \theta_n^\epsilon)$ at time nT being

$$x(nT, \theta_n^\epsilon) = x^\epsilon(nT). \tag{6.2}$$

Continuing, suppose that $\hat{Y}_n^\epsilon$ is an unbiased estimator of $-C_{T,\theta}(\theta_n^\epsilon, x^\epsilon(nT))$. This is equivalent to "restarting" the estimation procedure anew at each nT with initial condition $x^\epsilon(nT)$. To see what the limit ODE for (6.1) might be, proceed purely formally, let $\xi_n^\epsilon = x^\epsilon(nT)$ and apply Theorem 3.1 to get

$$\dot{\theta} = -\int C_{T,\theta}(\theta, \xi)\mu(d\xi|\theta). \tag{6.3}$$

The right side of (6.3) would not be close to $-C_\theta(\theta)$ unless (at least) T is large. For this reason, it is often suggested that T depend on either or both ϵ, n and go to infinity as one or both of these quantities goes to its limit. In [41], there are conditions for the convergence of the right side of (6.3) to $-C_\theta(\theta)$ for Markov chain models.

Before showing how to improve (6.3), let us look at a typical procedure more closely. In order to get a (pathwise) gradient estimator one generally introduces an auxiliary process $y(\cdot, \theta)$. For IPA estimators [13, 18, 32], this would be the pathwise derivative of $x(\cdot, \theta)$ with respect to θ; for likelihood ratio estimators [32, 37, 38] this would be the score function which keeps the information on the derivative of the measure. Other methods such as smoothed perturbation analysis and rare perturbation analysis [4] use auxiliary information that represents the difference between the path $x(\cdot, \theta)$ and a perturbed one. See also the discussion of mean square derivatives and finite differences in §9.

For the model used in our illustrative example, the appropriate $y(\cdot, \theta)$ process is the mean square derivative defined by $x_\theta(t, \theta) = y(t, \theta)$:

$$\dot{y}(t, \theta) = b_x(x(t, \theta), \theta) + b_\theta(x(t, \theta), \theta)$$

with initial condition $y(0, \theta) = 0$. Define $z(\cdot, \theta) = (x(\cdot, \theta), y(\cdot, \theta))$ The estimator of $C_{T,\theta}(\theta, x(0))$ has the form

$$\frac{1}{T}\int_0^T \lambda(\theta, z(s, \theta))ds,$$

where

$$\lambda(\theta, z(s, \theta)) = c_x(\theta, x(s, \theta))y(s, \theta) + c_\theta(\theta, x(s, \theta)).$$

Let $z^\epsilon(\cdot) = (x^\epsilon(\cdot), y^\epsilon(\cdot))$ be the actual process with the time-varying parameter used. Then the formal procedure leading to (6.3) would use

$$\hat{Y}_n^\epsilon = -\frac{1}{T}\int_{nT}^{nT+T} \lambda(\theta, z^\epsilon(s))ds, \tag{6.4}$$

where on $[nT, nT + T)$ we have $y^\epsilon(t) = y(t, \theta_n^\epsilon)$ with initial condition $y^\epsilon(nT) = 0$.

Now consider an alternative where $x^\epsilon(\cdot)$ is as above but $y^\epsilon(\cdot)$ is not reset to zero at times nT. Use $y^\epsilon(0) = 0$ and on $[nT, nT + T)$ use $y^\epsilon(\cdot) = y(\cdot, \theta_n^\epsilon)$ with initial condition at nT defined recursively by

$$y^\epsilon(nT) = y(nT, \theta_n^\epsilon). \tag{6.5}$$

Then with the new definition of $z^\epsilon(\cdot)$, use the estimator

$$Y_n^\epsilon = -\frac{1}{T}\int_{nT}^{nT+T} \lambda(\theta_n^\epsilon, z^\epsilon(s))ds. \tag{6.6}$$

Note that in general (6.6) would not be an unbiased estimator of $-C_{T,\theta}(\theta_n^\epsilon, x^\epsilon(nT))$ due to the "memory " in its "initial conditions."

Now define the process $\xi_n^\epsilon = z^\epsilon(nT)$, and assume the conditions of Theorem 3.1. Then the ODE which characterizes the limit behavior is (3.12), where

$$G(\theta, \xi) = -E[Y_n^\epsilon | \xi_n^\epsilon = \xi, \theta_n^\epsilon = \theta].$$

By the definition of the invariant measure, we then have

$$g(\theta) \equiv \int G(\theta, \xi)\mu(d\xi|\theta) = \lim_n \frac{1}{n}\sum_{i=1}^{n} EG(\theta, \xi_i(\theta)), \tag{6.7}$$

where $\xi_i(\theta)$ is the stationary process under θ. Under either (3.11a) or (3.11b), the limit on the right side is the same if we used the process $\xi_n(\theta)$ with initial condition $\xi_0(\theta) = (x(0), 0)$. Thus, with this new initial condition (6.7) equals

$$-\lim \frac{1}{nT} E \int_0^{nT} \lambda(\theta, z(s, \theta))ds = -\lim_T C_{T,\theta}(\theta, x(0)) = -C_\theta(\theta). \tag{6.8}$$

This is what we want since it yields the gradient descent ODE

$$\dot{\theta} = -C_\theta(\theta) \tag{6.9}$$

in lieu of the "biased" (6.3). In the parlance of the literature (e.g., [32]), (6.9) results when we do not reset the "accumulator." While there has been some discussion of this preferable alternative, proofs and a clear understanding were lacking. In the next three sections, the details are filled in for three classes of applications.

7. A discrete example: A GI/G/1 queue. We consider the problem treated in [7, 11, 31, 32]. The model is a single-server queue with a renewal arrival process and general service time distribution, which is parametrized by $\theta > 0$. For notational simplicity, we suppose that θ is real valued, but the development and results are the same in general. For fixed θ let $X_i(\theta)$ denote the sojourn time of the ith customer and $K(\theta)$ be a bounded real-valued function with a continuous and bounded gradient. The cost of interest is

$$C(\theta) = \lim_N \frac{1}{N}\sum_{i=1}^{N} EX_i(\theta) + K(\theta) \equiv \hat{C}(\theta) + K(\theta), \tag{7.1}$$

and we wish to use SA to get the minimizing θ. Again, we suppose that the parameter θ is bounded. Indeed, the parameter might have to be restrained to some particular interval $[\theta_-, \theta_+]$ in order for the assumptions below to hold, and we assume that this is done. The example is widely studied, but the conditions used here are about as simple as one can expect. The structure of the problem is similar (from the point of view of SA) to those arising in other applications to single queues (and even for some network problems). For example, consider the multiclass problem [34], admission control [42], flow control in a closed network [43], routing in an open network [40], and routing in a closed network [17]. Appendix 4 discusses the decentralized case that is of interest in network models.

Fixed θ-process: Application of IPA. We proceed to make the usual assumptions to assure that $dX_i(\theta)/d\theta$ exists and can be estimated via IPA. Define the

parametrized service time distribution $F(\cdot|\theta)$, and suppose that it is weakly continuous in θ. Define the inverse function $F^{-1}(\cdot|\theta)$ by

$$F^{-1}(\chi|\theta) = \min\{\zeta : F(\zeta|\theta) \geq \chi\}, \qquad \chi \in [0,1],$$

and assume that it is differentiable in θ for each χ, with a bounded and continuous (uniformly in χ) derivative denoted by $F_\theta^{-1}(\chi|\theta)$. For fixed θ, let $\{\zeta_i(\theta), i < \infty\}$ denote the sequence of service times and define $\chi_n(\theta) = F(\zeta_n(\theta)|\theta)$ and the derivative $Z_n(\theta) = F_\theta^{-1}(\chi_n(\theta)|\theta)$. Let $Q_i(\theta)$ denote the queue length and $\tau_i(\theta)$ the (residual) time until the next arrival, all taken just after the departure of the ith customer. Then $x_n(\theta) = (Q_n(\theta), \tau_n(\theta))$ is a Markov process. Define the cost for the first N customers, initialized at an arbitrary initial condition, as

$$C_N(\theta, x_0) = \hat{C}_N(\theta, x_0) + K(\theta) = \frac{1}{N}\sum_{i=1}^{N} EX_i(\theta) + K(\theta). \tag{7.2}$$

Suppose that the busy periods have finite mean length for each fixed θ. Let $\bar{Z}_n(\theta)$ denote the sum of the $|Z_j(\theta)|$ in the nth busy period. Suppose that

$$\sup_\theta E|\bar{Z}_n(\theta)| < \infty. \tag{7.3}$$

Remark on (7.3). In many cases where IPA can be applied, the θ is a scale parameter of the service distribution. Then we have the form $F(\zeta|\theta) = F(\theta\zeta|1)$, $\zeta_n(\theta) = \theta F^{-1}(\chi_n(\theta)|1)$, and $Z_n(\theta) = \zeta_n(\theta)/\theta$, where the $\chi_n(\theta)$. Letting $N(\theta)$ denote the number of services in a busy period, Wald's identity yields $E|\bar{Z}_n(\theta)| = E\bar{Z}_n(\theta) = EN(\theta)EZ_j(\theta)$. If the system is stable, then $EN(\theta) < \infty$ and (7.3) holds.

Continuing, consider the estimator

$$\hat{Z}_m(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=v_i(\theta)}^{i} Z_j(\theta), \tag{7.4}$$

where $v_i(\theta)$ is the index of the first arrival in the busy period in which customer i arrives. If $Q_0 = 0$, then (7.4) is an unbiased estimator of the derivative of $\hat{C}_m(\theta, x_0)$, where x_0 is the state at time zero. It is an asymptotically consistent estimator in that [13] for each initial condition

$$E\hat{Z}_m(\theta) \to \hat{C}_\theta(\theta). \tag{7.5}$$

Henceforth, just to simplify notation and not have to worry about the possibly separate indices for arrivals and departures, we suppose that the queue starts empty. The conditions and results are the same in general.

The estimator for the SA. We wish to use SA to minimize the cost (7.1) via use of the IPA estimator. Suppose that the parameter is updated after the departure of each successive group of N customers. We use the "customer number" instead of real time. For fixed θ, the estimator used on the nth interval (the departures $[nN+1, nN+N]$) is to be

$$\hat{Y}_n(\theta) = \frac{1}{N}\sum_{i=nN+1}^{nN+N}\sum_{j=v_i(\theta)}^{i} Z_j(\theta).$$

Recall that $v_{nN}(\theta)$ is the index of the first arrival in the busy period in which arrival (equivalently, departure) nN occurs. Let $D_n(\theta)$ denote the number of departures from the $(nN+1)$st to the end of the busy period in which departure nN occurs. It is zero if the nNth departure ends a busy period. Now, to separate the "past" from the contributions over $[nN+1, nN+N)$, we split $\hat{Y}_n(\theta)$ by defining

$$\psi_n(\theta) = \sum_{j=v_{nN}(\theta)}^{nN} Z_j(\theta), \text{ "past"}$$

$$A_n(\theta) = \frac{1}{N} \sum_{i=nN+1}^{nN+N} \sum_{j=v_i(\theta)\vee(nN+1)-1}^{i} Z_j(\ t), \text{ "future,"}$$

$$B_n(\theta) = \frac{1}{N} \sum_{i=nN+1}^{nN+N} \sum_{j=v_i(\theta)}^{v_i(\theta)\vee(nN+1)-1} Z_j(\theta),$$

Then

$$\hat{Y}_n(\theta) = A_n(\theta) + B_n(\theta).$$

We can write

$$B_n(\theta) = D_n(\theta)\psi_n(\theta)/N.$$

Following the basic framework of §3, define the Markov chain $\xi_n(\theta) = (Q_n(\theta), \tau_{nN}(\theta), \psi_n(\theta))$. Define $G_i(\cdot)$ by

$$G_0(\theta, \xi_n(\theta)) = E_n A_n(\theta), \qquad G_1(\theta, \xi_n(\theta)) = E_n D_n(\theta)/N,$$

where E_n denotes the expectation conditioned on all the systems data up to and including the time of the nNth departure. It is equivalent to conditioning on $\xi_n(\theta)$.

The functions $G_i(\cdot, \xi)$ are continuous in θ, uniformly in each compact (θ, ξ) set by the continuity assumptions made on the distribution of the service interval and the derivative of its inverse. In preparation for the conclusion of the SA argument, note that (and define $G(\cdot)$ by)

$$E_n \hat{Y}_n(\theta) = G_0(\theta, \xi_n(\theta)) + G_1(\theta, \xi_n(\theta))\psi_n(\theta) \equiv G(\theta, \xi_n(\theta))$$

and that

$$\lim_n \frac{1}{n} \sum_{i=1}^{n} EG(\theta, \xi_n(\theta)) = \lim_n \frac{1}{nN} E \sum_{i=1}^{nN} \sum_{j=v_i(\theta)}^{i} Z_j(\theta) = \hat{C}_\theta(\theta) \tag{7.6}$$

for each initial condition.

The SA problem. For the actual physical system with the time-varying parameter, let ζ_n^ϵ denote the actual service time of the nth customer and Z_n^ϵ the derivative of the inverse function, using the parameter θ_n^ϵ for $nN+1 \le j \le nN+N$. Let v_i^ϵ be the index of the first arrival in the busy period in which customer i departs. Let

τ_i^ϵ be the residual time to the next arrival and Q_i^ϵ the queue length, all taken at the time of the ith departure. To update θ_n^ϵ we use the estimator

$$\hat{Y}_n^\epsilon = \frac{1}{N} \sum_{i=nN+1}^{nN+N} \sum_{j=v_i^\epsilon}^{i} Z_j^\epsilon. \tag{7.7}$$

The SA algorithm is

$$\theta_{n+1}^\epsilon = \left[\theta_n^\epsilon - \epsilon \hat{Y}_n^\epsilon - \epsilon K_\theta(\theta_n^\epsilon)\right]_{\theta_-}^{\theta_+}. \tag{7.8}$$

Now define

$$\psi_n^\epsilon = \sum_{j=v_{nN}^\epsilon}^{nN} Z_j^\epsilon,$$

and define $A_n^\epsilon, B_n^\epsilon$ analogously to what was done for the fixed θ case. Then $\hat{Y}_n^\epsilon = A_n^\epsilon + B_n^\epsilon \psi_n^\epsilon$ and $\xi_n^\epsilon = (Q_{nN}^\epsilon, \tau_{nN}^\epsilon, \psi_n^\epsilon)$ is a Markov chain. We can write

$$E_n^\epsilon \hat{Y}_n^\epsilon = G_0(\theta_n^\epsilon, \xi_n^\epsilon) + G_1(\theta_n^\epsilon, \xi_n^\epsilon)\psi_n^\epsilon = G(\theta_n^\epsilon, \xi_n^\epsilon). \tag{7.9}$$

In this example, the P_n^ϵ in (3.5) does not depend on ϵ, n. Thus, (3.6) holds. Assumption (3.7) follows from the assumptions on the service time distribution. Also, the G_n^ϵ in (3.9) does not depend on ϵ, n. Assumption (3.11b) follows from (7.6).

We need conditions which guarantee (3.2), (3.4), and (3.8). Define $\bar{Z}_n^\epsilon$ to be the sums of $|Z_j^\epsilon|$ over the nth busy period. Suppose that

$$\{\bar{Z}_n^\epsilon, \epsilon > 0, n < \infty\} \text{ is uniformly integrable,} \tag{7.10}$$

$$\sup_\theta E\zeta(\theta) < E[\text{ interarrival time}]. \tag{7.11}$$

By (7.10), (3.2) holds for $\hat{Y}_n^\epsilon$. The θ_n^ϵ are bounded and tightness of $\{\xi_n^\epsilon, \epsilon > 0, n < \infty\}$ follows from (7.3) and (7.11). Condition (3.8) follows from (7.3) and (7.11), and (3.10) is a consequence of (7.3). Now the convergence to the ODE

$$\dot{\theta} = -C_\theta(\theta) - K_\theta(\theta) \tag{7.12}$$

projected onto $[\theta_-, \theta_+]$ follows from Theorem 3.1. The case $\epsilon_n \to 0$ follows from Theorem 4.1.

Remark. We note that the updates need not be at regular intervals. If the interupdate times are bounded, then the end result is the same. Unbounded interupdate times might conceivably be of future interest, but at the moment it is hard to imagine an application in which it would be allowed, but with suitable conditions on the increments between updates we get the same result. If the estimator were reset at each nN, then the $\hat{Y}_n^\epsilon$ in (7.8) is replaced by A_n^ϵ and the $C_\theta(\theta)$ in (7.12) would be replaced by the biased quantity $\int \hat{C}_{N,\theta}(\theta, x)\mu_0(dx|\theta)$, where $\mu_0(\cdot|\theta)$ is the invariant measure of the $x_n(\theta)$ process defined above (7.2).

8. An example from manufacturing: A piecewise deterministic problem. We now consider an interesting example from [44]. The reference considers a manufacturing system with two unreliable tandem machines and is typical of many applications to production rate scheduling problems. Let $\alpha_i(t)$ denote the indicator function that machine i is working, and assume that these processes are independent renewal processes. The production rates $u_i(\cdot), i = 1, 2$, of the machines can be controlled, subject to the machines' working and to upper bounds $\bar{u}_i$ on the rates. Machine 1 feeds into machine 2 via a buffer for surplus inventory, and the demand rate for the output of machine 2 is fixed at d. The dynamical state is the current inventory level $x(\cdot) = (x_1(\cdot), x_2(\cdot))$. The inventory of machine 2 can be negative (backlog). The reference assumes that the inventory process defined below satisfies a Harris recurrence condition, but we will not need to suppose that. The dynamical state equation is $x(0) = 0$ and

$$\dot{x}_1(t) = u_1(t) - u_2(t), \qquad \dot{x}_2(t) = u_2(t) - d.$$

The control problem is actually more conveniently formulated in terms of the surplus variables, defined by

$$s_1(t) = x_1(t) + x_2(t) = \int_0^t u_1(s)ds, \qquad s_2(t) = x_2(t).$$

They consider control strategies of a threshold type on the $s_i(\cdot)$. (Their thresholds B_1, B_2 are the θ_1, θ_2 here.) In order to illustrate our method, we shall focus on one of their strategies, called the surplus control. Loosely speaking, for this control the production rate is held at the maximum value if the surplus is less than the threshold and tries to stay at the threshold if it ever reaches it. If (during the transient initial period) the surplus on some machine is higher than the threshold, then production on that machine is zero. For notational simplicity, we suppose that the initial surpluses do not exceed the thresholds. The surplus process $s_i(\cdot)$ evolves with deterministic slopes ($\bar{u}_i$, 0, or $-d$) which correspond to maximum production rate minus demand, production rate equals demand, and no production, and these change at random times which depend on the values of the renewal $\alpha(\cdot)$ and state processes $s(\cdot)$. With minor exceptions, the assumptions used here are implied by those in [44]. The $\alpha_i(\cdot)$ processes do not depend on the thresholds. It is assumed that the maximum production rates satisfy $\bar{u}_1 > \bar{u}_2 > d$. Also, it is supposed that $\theta_1 > \theta_2 \geq 0$, with each threshold subject to an upper bound, and that the SA algorithm is constructed to guarantee this.

DEFINITIONS (the fixed θ processes). Henceforth, for the fixed θ processes, we write the state and surplus variables as $x(\cdot, \theta), s(\cdot, \theta)$, resp. For fixed thresholds θ, the control is such that the surplus processes are defined by

$$\dot{s}_1(t,\theta) = \begin{cases} (\bar{u}_1 - d)I_{\{s_1(t,\theta) < \theta_1\}} - dI_{\{s_1(t,\theta) > \theta_1\}} & \text{if } \alpha_1(t) = 1, \\ -d & \text{otherwise,} \end{cases} \tag{8.1}$$

$$\begin{aligned} \dot{s}_2(t,\theta) &= (\bar{u}_2 - d)I_{\{s_2(t,\theta) < \theta_2\}} - dI_{\{s_2(t,\theta) > \theta_2\}} \\ &\qquad \text{if } \alpha_2(t) = 1 \text{ and } (s_1(t,\theta) \geq s_2(t,\theta) \text{ or } \alpha_1(t) = 1), \\ \dot{s}_2(t,\theta) &= -d \quad \text{otherwise.} \end{aligned} \tag{8.2}$$

The dynamics are such that $s_1(t,\theta) - s_2(t,\theta) \geq 0$ if this condition holds at time zero, which we suppose. (The condition will eventually hold in any case.)

The system just described is an example of a piecewise deterministic control system [8, 16], where $x(\cdot,\theta)$ is piecewise linear, with the intervals being random. In [44], the interval distributions are exponential so that the state process is Markovian. For $c_1 > 0, c_2^{\pm} > 0$, the cost rate of concern is

$$c(s) = c_1 s_1 + c_2^+ s_2^+ + c_2^- s_2^-.$$

Actually, the reference starts with $\tilde{c}(x) = c_1 x_1 + c_2^+ x_2^+ + c_2^- x_2^-$, but in the derivative calculations switches to $c(\cdot)$. The two cost rates are equivalent, with appropriate definitions of the coefficients.

The thresholds are adjusted via an SA algorithm using IPA-type derivatives of the cost with respect to the θ_i, with the aim of minimizing the cost function

$$C(\theta) = \lim_M \frac{1}{M} \int_0^M c(s(t,\theta))dt, \tag{8.3}$$

where we suppose that the limit exists for each θ in the desired range. The reference [44] derives auxiliary processes $y_j^{(i)}(t,\theta)$, which is the pathwise derivative of $s_j(t,\theta)$ with respect to θ_i. We now describe these derivative processes.

The (IPA) derivative processes. The following results are taken from the reference, with only the terminology changed to bring it into line with our own. Define the random time $\tau(\theta) = \inf\{t > 0 : s_1(t,\theta) = \theta_1\}$. Then

$$y_1^{(1)}(t,\theta) = I_{\{t \geq \tau(\theta)\}}.$$

Define the random times $\tau_i^k(\theta)$ recursively by $\tau_2^0(\theta) = \tau(\theta)$ and

$$\tau_1^k(\theta) = \min\{t \geq \tau_2^{k-1}(\theta) \colon s_1(t,\theta) = s_2(t,\theta)\},$$

$$\tau_2^k(\theta) = \min\{t \geq \tau_1^k(\theta) \colon s_2(t,\theta) = \theta_2\}.$$

Then the pathwise derivative of the surplus process at machine 2 with respect to θ_1 is

$$y_2^{(1)}(t,\theta) = \sum_{k=1}^{\infty} I_{\{\tau_1^k(\theta) \leq t \leq \tau_2^k(\theta)\}}.$$

Note that at most one of the indicator functions in the sum can be positive at a time. Clearly, the expression for $\dot{s}_1(t,\theta)$ implies that $y_1^{(2)}(t) = 0$. Define the additional random times $\gamma_i^k(\theta)$ recursively by $\gamma_2^0(\theta) = 0$ and

$$\gamma_1^k(\theta) = \min\{t \geq \gamma_2^{k-1}(\theta) \colon s_2(t,\theta) = \theta_2\},$$

$$\gamma_2^k(\theta) = \min\{t \geq \gamma_1^k(\theta) \colon s_1(t,\theta) = s_2(t,\theta)\}.$$

Then the pathwise derivative with respect to θ_2 of the surplus process at machine 2 is

$$y_2^{(2)}(t,\theta) = \sum_{k=1}^{\infty} I_{\{\gamma_1^k(\theta) \leq t \leq \gamma_2^k(\theta)\}}.$$

Again, at most one of the indicator functions can be positive at a time.

Let $z(\cdot,\theta) = (x(\cdot,\theta), y(\cdot,\theta))$. Define $Y_n(\theta) = (Y_n^1(\theta), Y_n^2(\theta))$, where

$$Y_n^i(\theta) = -\frac{1}{T}\int_{nT}^{nT+T} \lambda_i(\theta, z(t,\theta))dt, \tag{8.4}$$

where

$$\lambda_1(\theta, z(s,\theta)) = c_1 y_1^{(1)}(t,\theta) + c_2^+ y_2^{(1)}(t,\theta) I_{\{s_2(t,\theta)\geq 0\}} - c_2^- y_2^{(1)}(t,\theta) I_{\{s_2(t,\theta)<0\}}, \tag{8.5a}$$

$$\lambda_2(\theta, z(s,\theta)) = c_2^+ y_2^{(2)}(t,\theta) I_{\{s_2(t,\theta)\geq 0\}} - c_2^- y_2^{(2)}(t,\theta) I_{\{s_2(t,\theta)<0\}}. \tag{8.5b}$$

Then $Y_n^i(\theta)$ is an unbiased estimator for $-C_{T,\theta_i}(\theta, x(0))$.

The SA updates will be at times $nT, n = 1, \ldots$, for some $T > 0$. Below, we will be concerned with the set $(s(nT,\theta), y(nT,\theta), \alpha(nT))$. For a general renewal process, this set is not a Markov process. To Markovianize, we augment it by adding the residual times until the next change of values of the $\alpha_1(\cdot), \alpha_2(\cdot)$ after nT. Let $\xi_n(\theta)$ denote the consequent quadruple.

To minimize work, we suppose that for each initial condition the limit

$$-\lim_M \frac{1}{M}\int_0^M E\lambda(\theta, z(\theta,t))dt \tag{8.6}$$

exists for each θ value of interest and is continuous in θ. Define it as $g(\theta)$. These conditions are weaker than those in the reference. Then $C_\theta(\theta) = -g(\theta)$. In any case, these conditions amount to nothing more than asymptotic consistency, and are a minimal condition for the convergence. Define $G(\xi,\theta) = E[Y_n(\theta)|\xi_n(\theta) = \xi]$. Then $G(\cdot)$ is continuous and bounded. The limit (8.6) is the same as

$$\lim_n \frac{1}{n}\sum_{i=1}^n EG(\xi_n(\theta),\theta).$$

In order to prove the tightness of $\{s^\epsilon(t)\}$, we need the following conditions. Let p_i be the stationary probability that machine i is working. We suppose that

$$\bar{u}_2 p_2 > \bar{u}_1 p_1 > d; \tag{8.7}$$

i.e., the average maximum possible production rate for machine 2 is greater than that of machine 1, which is greater than the demand rate. Also, suppose that, where E_t is the expectation given $\{\alpha_i(v), v \leq t, i = 1,2\}$,

$$\int_t^\infty E_t[\alpha_i(v) - p_i]dv = O(1), \tag{8.8}$$

where $O(1)$ means that the term is bounded uniformly in all variables. Loosely speaking, (8.8) is equivalent to the expectation of the time to the next change in the $\alpha_i(\cdot)$ being uniformly bounded, conditioned on the current data. This is certainly not a strong condition.

THE SA ALGORITHM. Fix $T > 0$, the time interval between parameter updates. Let $s^\epsilon(\cdot)$ denote the actual surplus process with the time-varying parameter. Define $y^\epsilon(\cdot)$ as the derivative process with the random times determined by the actual time of

the associated events in the true physical process. It is not reset at each nT. Then set $z^\epsilon(\cdot) = (x^\epsilon(\cdot), z^\epsilon(\cdot))$. Let Y_n^ϵ be (8.4) with $z^\epsilon(\cdot)$ used in lieu of $z(\cdot, \theta)$. The algorithm is now (3.1).

Convergence of the SA. Condition (3.2) holds since the Y_n^ϵ are uniformly bounded. Tightness of all the components of the Markov chain (as needed for (3.4)) follows once tightness of $\{s^\epsilon(nT), \epsilon > 0, n < \infty\}$ is shown. This will be discussed at the end of the section. The $P(\cdot)$ and $G(\cdot)$ do not depend on n or ϵ. The weak continuity of the transition probability is a consequence of the basic structure of the problem. In particular, of the continuous effects of the threshold variations and the monotone nature of the evolution of the residual times. Finally, (3.11b) holds by assumption (8.6). Thus, the conclusions of Theorem 3.1 hold, and the extensions of Theorem 3.1 can also be readily handled. Theorem 3.1 asserts that the limit ODE is $\dot{\theta} = g(\theta)$ for the function $g(\theta)$ defined above. The reference [44] presents numerical data which implies that the cost function has a unique minimum and that their SA converges nicely.

The requirements are generally much weaker than those in the reference, and we do not need to restart the estimator periodically or let $T \to \infty$ as $\epsilon \to 0$. Some other references concerned with the use of SA in related manufacturing problems are [6, 39, 16, 45]. In [16], another interesting work on the same subject, they use an SA with gains $\epsilon_n \to 0$ and an IPA-type estimator where the estimation intervals go to infinity as $n \to \infty$. They do not "reset the accumulator." The conditions used here are simpler whether or not the step size is constant. The paper [19] was one of the early works which attempted to improve the operation of a production line subject to random breakdowns using IPA-type estimates, and dealt with a production line in an automobile factory. Some of the background analytical work is in [20].

Tightness of $\{s^\epsilon(t)$, small $\epsilon > 0, t < \infty\}$. Let $\mathcal{B}_t$ denote the minimal sigma-algebra measuring $\{\alpha_i(v), v \le t, i = 1, 2\}$ and E_t the associated conditional expectation. We define a differential operator $\tilde{A}$ and its domain.

The real-valued functions $f(\cdot), g(\cdot)$ of (t, ω) will be measurable with $f(t), g(t)$ being $\mathcal{B}_t$-measurable. Suppose that for each $T < \infty$,

$$\sup_{t \le T} E|g(t)| < \infty, \ \limsup_{\delta \downarrow 0} \sup_{t \le T} E\left|\frac{E_t f(t+\delta) - f(t)}{\delta}\right| < \infty,$$

$$\lim_{\delta \downarrow 0} E\left|\frac{E_t f(t+\delta) - f(t)}{\delta} - g(t)\right| = 0.$$

Then we say that $\tilde{A}f(\cdot) = g(\cdot)$. The process $f(t) - \int_0^t g(v)dv$ is a martingale [21], [22, §3.2.2].

The $s_i^\epsilon(t)$ are bounded above by the upper bounds to the thresholds. Thus the tightness problem concerns the probability of large negative excursions. We will work with altered processes, which provide the appropriate bounds from below. First we work with $s_1^\epsilon(t)$. To get a lower bound, we can suppose that $\theta_{i,n}^\epsilon = 0$. Let $q_1(t)$ be the process (8.1) with $\theta_1 = 0$. Then $\tilde{A}q_1^2(t)^2/2 = q_1(t)\dot{q}_1(t)$, which is $q_1(t)[\bar{u}_1\alpha_1(t) - d]$. To help in averaging the term with the $\alpha_1(t)$, define

$$\tilde{q}_1(t) = q_1(t)\bar{u}_1 \int_t^\infty E_t[\alpha_1(v) - p_1]dv.$$

We have

$$\tilde{A}\tilde{q}_1(t) = -q_1(t)\bar{u}_1[\alpha_1(t) - p_1] + O(1).$$

Define $Q_1(t) = q_1^2(t)/2 + \tilde{q}_1(t)$. Then

$$\tilde{A}\tilde{Q}_1(t) = q_1(t)[\bar{u}_1 p_1 - d] + O(1). \tag{8.9}$$

By (8.7), $\bar{u}_1 p_1 - d > 0$. Thus there are $k_i > 0$ such that for $q_1 \leq -k_1$, we have the right side of (8.9) less than $-k_2$. This implies that when $q_1 \leq -k_1$, $Q_1(\cdot)$ has the supermartingale property (until it hits the interval $[-k_1, 0]$). These considerations and the quadratic dependence of $Q_1(t)$ on $q_1(t)$ imply the tightness of $\{Q_1(t), t < \infty\}$. The tightness of $\{q_1(t), t < \infty\}$ (hence of $\{s_1^\epsilon(t), \text{small } \epsilon > 0, t < \infty\}$ follows from this tightness and the quadratic dependence of $Q_1(t)$ on $q_1(t)$.

The tightness of the $\{s_2^\epsilon(t)\}$ is proved in the same way. By the above results, it is sufficient to prove tightness for $s_1^\epsilon(t) - s_2^\epsilon(t)$ instead. Again, this can be done by a bounding argument. We have $s_1^\epsilon(t) - s_2^\epsilon(t) \geq 0$. Thus, we need to be concerned with large positive excursions of this difference. We start by fixing the thresholds at the upper bound for s_1^ϵ and the lower bound for s_2^ϵ. Once these thresholds are fixed, their actual values do not affect the result, so we can set them equal to zero without loss of generality. Let $q_i(\cdot)$ denote the new processes with the thresholds fixed at zero. One starts the argument by using a tentative Liapunov function (for the variables with the thresholds fixed at zero): $(q_1(t) - q_2(t))^2/2$. One bounds the derivative from above. Then introduces a function $\tilde{q}_2(\cdot)$ whose purpose is analogous to that of $\tilde{q}_1(\cdot)$ above. We omit the rest of the details due to lack of space. But by an argument similar to what was done for $q_1(\cdot)$ above, we get the tightness under the conditions (8.7), (8.8).

9. A continuous time SDE example: The system. We continue the discussion of the SDE model of §6 but with more detail and a more general system. We start by using the mean square derivatives and then discuss finite-difference forms. The finite-difference forms can be advantageous. One can use them without knowing the exact model and for more general cost functions. They can also be used for discrete-event systems in the same way. Let θ be real valued (for notational simplicity only) and $x \in R^k$. Let $b(\cdot)$ be a R^k-valued and continuously differentiable function of (x, θ) with bounded x and θ first derivatives, $\sigma(\cdot)$ a continuously differentiable matrix-valued function of x with bounded first derivatives, and let the fixed θ state process satisfy the SDE

$$dx(t,\theta) = b(x(t,\theta),\theta)dt + \sigma(x(t,\theta))dw(t), \tag{9.1}$$

where $w(t)$ is a standard vector valued Wiener process. Define the auxiliary process $y(t,\theta)$ by

$$dy(t,\theta) = b_x(x(t,\theta),\theta)y(t,\theta)dt + b_\theta(x(t,\theta),\theta)dt + (\sigma, y)(t,\theta)dw(t), \tag{9.2}$$

where the vector $(\sigma, y)(t,\theta)dw(t)$ is defined by its components

$$\sum_{j,p} \frac{\partial \sigma_{ij}(x(t,\theta))}{\partial x_p} y_p(t,\theta)dw_j(t), \qquad i = 1, \ldots, k,$$

The $y(t,\theta)$ is the pathwise (mean square) derivative of $x(t,\theta)$ with respect to θ. This "pathwise derivative" for the SDE was in use [12] long before its analogue for the

discrete case was developed. Define $z(\cdot,\theta) = (x(\cdot,\theta), y(\cdot,\theta))$. Let $c(\cdot,\cdot)$ be a bounded, real-valued, continuously differentiable function of (θ, x) with bounded derivatives, and define $C_T(\theta) = \int_0^T c(\theta, x(s,\theta))ds/T$ as in §6.

The SA procedure. Use the method of §6, where we update at intervals $nT, n = 1, \ldots,$ with θ_n^ϵ being the parameter value used on $[nT, nT+T)$. Use (3.1) with $x^\epsilon(\cdot)$ again defined as the state process with the time-varying parameter used. Define $y^\epsilon(\cdot)$ as above (6.5) (i.e., it is never reset), and define $z^\epsilon(\cdot) = (x^\epsilon(\cdot), y(\cdot))$. Define

$$Y_n^\epsilon = -\frac{1}{T}\int_{nT}^{nT+T}\left[\sum_j c_{x_j}(\theta_n^\epsilon, x^\epsilon(s))y_j^\epsilon(s) + c_\theta(\theta_n^\epsilon, x^\epsilon(s))\right]ds. \tag{9.3}$$

We assume the following conditions.

(9.4) The process $\{z(nT,\theta)\}$ has a unique invariant measure for each θ.

$$\{z^\epsilon(nT), \theta_n^\epsilon, \epsilon > 0, n < \infty\} \text{ is tight} \tag{9.5}$$

$$\{Y_n^\epsilon, \epsilon > 0, n < \infty\} \text{ is uniformly integrable.} \tag{9.6}$$

$$\{z(0,\theta) : \theta \in \Theta \text{ compact }, z(\cdot,\theta) \text{ stationary }\} \text{ is tight.} \tag{9.7}$$

Condition (9.4) implies that the limit $C(\theta)$ of $C_T(\theta, x(0)))$ exists and does not depend on $x(0)$. Under these conditions, Theorem 3.1 and its extensions hold. Thus (6.9) holds for algorithm (3.1). An SA procedure using mean square derivatives was used to good practical effect in [5, 26]. There is an analagous result under (3.11b).

Finite-difference methods. The main difficulties in applications concern the verification of the various conditions on the y processes. This was an unresolved issue in [5]. These difficulties can be alleviated by using a finite-difference method rather than the derivative $y^\epsilon(\cdot)$ process. We will discuss two forms of the finite-difference method. The first is the more traditional, using separate runs for the different components of the difference. The second combines these runs into one "concatenated difference" and provides a useful alternative since it can be used on line. There is an obvious analogue for discrete-event systems.

A finite-difference alternative: Simultaneous runs. Tightness and uniqueness of the appropriate invariant measure are often much easier to prove if a finite-difference method is used in lieu of the estimator (9.3), since then the troublesome $y^\epsilon(\cdot)$ process does not appear. We retain the conditions of the last part, with the exception of those concerning the y process. We also let $c(\cdot)$ be simply bounded and continuous. Given a finite-difference interval $\delta\theta$, replace the integrand in (9.3) with

$$\frac{c(\theta_n^\epsilon + \delta\theta, x(s, \theta_n^\epsilon + \delta\theta)) - c(\theta_n^\epsilon - \delta\theta, x(s, \theta_n^\epsilon - \delta\ t))}{2(\delta\theta)}. \tag{9.8}$$

Here we use two separate simulations, one for $\{\theta_n^\epsilon + \delta\theta\}$ and one for $\{\theta_n^\epsilon - \delta\theta\}$. We thus run two processes $x^{\epsilon,\pm}(\cdot)$ defined by $x^{\epsilon,\pm}(0) = x(0)$, and on $[nT, nT+T)$ set $x^{\epsilon,\pm}(\cdot) = x(\cdot, \theta_n^\epsilon \pm \delta\theta)$ with initial condition at nT defined recursively by $x(nT, \theta_n^\epsilon \pm \delta\theta) = x^{\epsilon,\pm}(nT)$. Generally, one would want to use the same Wiener process to drive

the two processes. This (common random variables) form often yields essentially the same path properties as does the use of the derivative process.

Under the given conditions, Theorem 3.1 yields that the limit ODE is

$$\dot{\theta} = -\frac{1}{2\delta\theta}\int [c(\theta+\delta\theta,\xi)\mu(d\xi|\theta+\delta\theta) - c(\theta-\delta\theta,\xi)\mu(d\xi|\theta-\delta\theta)]\,, \tag{9.9}$$

where $\mu(\cdot|\theta)$ is the invariant measure of $\{x(nT,\theta)\}$, and with the analogous formula for the multidimensional θ case. Due to the additive way that the two terms appear in (9.8), we do not need to have a unique invariant measure of the pair $\{x(nT,\theta+\delta\theta), x(nT,\theta-\delta\theta)\}$ for each θ but only of $\{x(nT,\theta)\}$ for each θ.

The finite-difference approach can be either easier or harder than the pathwise derivative approach. The order of the SDEs to be solved in each case is the same. If $\sigma(x)$ actually depended on x, then the pathwise derivative procedure cannot be conducted "on line," since we need to know the Wiener process to get $y(\cdot,\theta)$. If $\sigma(x)$ does not depend on x, then the equation for $y(\cdot,\theta)$ or $y^\epsilon(\cdot)$ is linear in the y variable (but with time varying coefficients) and it is simpler to solve. The procedure can then be done "on line," at least in principle. An additional point to be kept in mind is that any simulation can only approximate the solution to (9.1) and (9.2). Thus, there is the additional question concerning the relations between the estimators for the approximations and those of the original model. See [26] for some results on this important problem. Finally, the finite-difference method can be used for cases where the $c(\cdot), b(\cdot)$ are not smooth, e.g., where $c(\cdot)$ is an indicator function of a event of interest.

Finite differences with only one run. Alternatively to the traditional simultaneous run method discussed above, a single run can be used to get a good estimate of the desired quantity and will be useful when the optimization must be done "on line," where simultaneous runs might not be possible. Let $T > 0$ and $\delta\theta > 0$ be given. For the "one run" method, we use $\theta_n^\epsilon + \delta\theta$ on the interval $[2nT, 2nT+T)$ and then $\theta_n^\epsilon - \delta\theta$ on $[2nT+T, 2nT+2T)$. Let $x^\epsilon(\cdot)$ denote the actual process with the $\theta_n^\epsilon \pm \delta\theta$ being used on the appropriate alternating time intervals. The appropriate fixed θ process, which we call $\hat{x}(\cdot,\theta)$, uses parameter value $\theta+\delta\theta$ on $[0,T)$ and then alternates between $\theta-\delta\theta$ and $\theta+\delta\theta$ on successive intervals of width T. We use

$$Y_n^\epsilon = -\frac{1}{2T\delta\theta}\int_0^T [c(\theta_n^\epsilon+\delta\theta, x^\epsilon(nT+s)) - c(\theta_n^\epsilon-\delta\theta, x^\epsilon(nT+T+s))]\,ds.$$

The analysis follows the lines of Theorem 3.1, but the limit form will be slightly different from that above. It is worth commenting on the differences between the simultaneous and single run cases since they are of practical importance and of interest in related algorithms. The main additional problem is due to the fact that the transition function for the fixed θ process depends periodically on time.

Let $\xi_n^+(\theta) = \hat{x}(2nT,\theta)$ and $\xi_n^-(\theta) = \hat{x}(2nT+T,\theta)$. Suppose that the stationary processes exist and are unique, with invariant measures $\mu^+(\cdot|\theta)$ and $\mu^-(\cdot|\theta)$, resp. Define

$$G^+(\theta,\xi) = \frac{1}{2T\delta\theta}\int_0^T E\,[c(\theta+\delta\theta, x(s,\theta+\delta\theta))|x(0)=\xi]\,ds,$$

$$G^-(\theta,\xi) = \frac{1}{2T\delta\theta}\int_0^T E\,[c(\theta-\delta\theta, x(s,\theta-\delta\theta))|x(0)=\xi]\,ds.$$

The right side of the limit ODE is

$$g(\theta) = -\int \left[G^+(\theta,\xi)\mu^+(d\xi|\theta) - G^-(\theta,\xi)\mu^-(d\xi|\theta)\right]. \tag{9.10}$$

Let $P_T(\xi,\cdot|\theta+\delta\theta)$ denote the transition function for the process $x(nT,\theta+\delta\theta)$. Note that

$$\mu^-(d\xi|\theta) = \int \mu^+(d\tilde{\xi}|\theta)P_T(\tilde{\xi},d\xi|\theta+\delta\theta). \tag{9.11}$$

Thus, as $\delta\theta \to 0$, the $\mu^{\pm}(\cdot|\theta)$ converge weakly to $\mu(\cdot|\theta)$, and so do the $\mu(\cdot|\theta\pm\delta\theta)$. Thus the $\mu^{\pm}(\cdot|\theta)$ become closer to the $\mu(\cdot|\theta\pm\delta\theta)$, which are the measures in the right side of (9.9). This line of reasoning suggests that the one sample procedure might be quite reasonable. The obvious form of (3.11b) can replace the assumption of uniqueness of the invariant measures.

To better understand the above "one-run" procedure, one needs to compare it to an alternative one-run procedure, say where we restart the process each T units of time at some fixed initial value, still using the $\theta \pm \delta\theta$ on the alternate intervals (assuming that such restarts were possible in the application). This would yield a right side of the form (9.10), where the $\mu^{\pm}$ are replaced by the measures concentrated on the fixed initial values. We expect that this "restarted method" would be much inferior to the original procedure, since the $\mu^{\pm}(\cdot|\theta)$ defined above would be much closer to the desired values $\mu(\cdot|\theta\pm\delta\theta)$, particularly for large T. The situation would be a little more complicated if θ were vector valued, but the general idea is the same. Analogous remarks can be made on the use of finite differences for discrete-event systems.

Appendix 1. Non-Markov models. Consider the algorithm (3.1). Suppose that due to the nature of the correlations, there is no convenient Markov chain $\{\xi_n^\epsilon, n < \infty\}$ for each ϵ. For example, the service or interarrival intervals in a queue might be correlated in a "non-Markovian way." The first-order perturbed test function methods of [22] are often very helpful in such circumstances, and we will outline the general idea in the context of Theorem 3.1.

For each $\epsilon > 0$, $\{Y_n^\epsilon, n < \infty\}$ denotes the observation sequence, and the uniform integrability (3.2) is assumed. The θ_n will be assumed to be in a compact set to make the development simpler. For fixed parameter θ and each integer m we define the fixed θ process $\{Y_j^m(\theta), j \geq m\}$, and define $Y_j^m(\theta) = Y_j^\epsilon$ for $j \leq m$ by supposing that after time m the sequence evolves as though the parameter were held fixed at θ. This process is the analogue of the fixed θ Markov chain of §3. The key to the development is to work with an appropriately chosen "perturbed" θ_n^ϵ, which differs only slightly from θ_n^ϵ and for which the theorem can be proved. Suppose that there is a continuous function $g(\cdot)$ such that for each large $T_1 < T_2 < \infty$ and $m < T_1/\epsilon$, the sum defined by

$$\delta f_m^\epsilon(\theta) = \sum_{j=m}^{T_2/\epsilon} \epsilon E_m^\epsilon \left[Y_j^m(\theta) - g(\theta)\right] \tag{A1.1}$$

goes to zero in mean, uniformly in $m \leq T_1/\epsilon$ as $\epsilon \to 0$. The convergence of (A1.1) is a condition on the "mixing rate" of the noise process.

Define $\delta f_n^\epsilon = \delta f_n^\epsilon(\theta_n^\epsilon)$. In the analysis, $\tilde{\theta}_n^\epsilon = \theta_n^\epsilon + \delta f_n^\epsilon(\theta_n^\epsilon)$ replaces θ_n^ϵ. We also need the continuity condition that

$$\text{(A1.2)} \qquad \sum_{j=m}^{T_2/\epsilon} E_m^\epsilon \left[Y_j^m(\theta+\delta\theta) - g(\theta+\delta\theta)\right] - \sum_{j=m}^{T_2/\epsilon} E_m^\epsilon \left[Y_j^m(\theta) - g(\theta)\right] \to 0$$

in the mean, uniformly in $m \le T_1/\epsilon$ as $\delta\theta \to 0$.

We have the following theorem.

THEOREM A1.1. *Let q_ϵ satisfy* (3.8′). *Then, assuming* (3.2) *and the conditions concerning* (A1.1) *and* (A1.2), $\{\theta^\epsilon(\epsilon q_\epsilon + \cdot), \epsilon > 0\}$ *is tight, and the limit of any weakly convergent subsequence satisfies* (3.12). *If $\epsilon q_\epsilon \to \infty$, then the conclusions of Theorem* 3.2 *hold.*

Proof. The proof is much simpler than that of Theorem 3.1. Again, for simplicity, we let $q_\epsilon = 0$. Let $\tilde{\theta}^\epsilon(\cdot)$ denote the continuous parameter interpolation (interval ϵ) of the $\tilde{\theta}_n^\epsilon$ sequence defined above (A1.2). $\{\theta^\epsilon(\cdot)\}$ is tight. For notational simplicity, let ϵ index a weakly convergent subsequence. Let $h(\cdot), s_i, t, \tau$ be as in Theorem 3.1 with $t+\tau \le T_1$, and suppose for notational simplicity that ϵ indexes a weakly convergent subsequence. By the definition of conditional expectation,

$$\text{(A1.3)} \qquad Eh(\tilde{\theta}^\epsilon(s_i), i \le q)\left[\tilde{\theta}^\epsilon(t+\tau) - \tilde{\theta}^\epsilon(t) - \sum_{m=t/\epsilon}^{(t+\tau)/\epsilon-1} E_m^\epsilon(\tilde{\theta}_{m+1}^\epsilon - \tilde{\theta}_m^\epsilon)\right] = 0.$$

We have

$$\text{(A1.4)} \qquad E_m^\epsilon\left(\tilde{\theta}_{m+1}^\epsilon - \tilde{\theta}_m^\epsilon\right) = \epsilon E_m^\epsilon Y_m^\epsilon + E_m^\epsilon\left[\delta f_{m+1}^\epsilon - \delta f_m^\epsilon\right].$$

The last term on the right equals

$$\epsilon\left[g(\theta_m^\epsilon) - E_m^\epsilon Y_m^m(\theta_m^\epsilon)\right] + \epsilon W_m^\epsilon,$$

where

$$W_m^\epsilon = \sum_{j=m+1}^{T_2/\epsilon} E_m^\epsilon\left[Y_j^{m+1}(\theta_{m+1}^\epsilon) - g(\theta_{m+1}^\epsilon)\right] - \sum_{j=m+1}^{T_2/\epsilon} E_m^\epsilon\left[Y_j^m(\theta_m^\epsilon) - g(\theta_m^\epsilon)\right].$$

Hence, we can write (A1.4) as $\epsilon g(\theta_m^\epsilon) + \epsilon W_m^\epsilon$. By definition,

$$Y_j^{m+1}(\theta_{m+1}^\epsilon) = Y_j^m(\theta_{m+1}^\epsilon), \qquad j \ge m+1.$$

Therefore, if the Y_n^ϵ were bounded (so that $|\theta_{m+1}^\epsilon - \theta_m^\epsilon| \to 0$ as $\epsilon \to 0$ uniformly in (m,ω)), we could use (A1.2) to get that $E|W_m^\epsilon| \to 0$ uniformly in $m : m\epsilon \le T_1$ as $\epsilon \to 0$. Then, (A1.3) would imply that

$$\text{(A1.5)} \qquad \lim_\epsilon Eh(\theta^\epsilon(s_i), i \le q)\left[\theta^\epsilon(t+\tau) - \theta^\epsilon(t) - \epsilon \sum_{m=t/\epsilon}^{(t+\tau)/\epsilon-1} g(\theta_m^\epsilon)\right] = 0.$$

The theorem would follow from this last equality, analogously to the situation in Theorem 3.1. If the Y_n^ϵ are not bounded, use the uniform integrability (3.2) to bound them for the purposes of the proof. For $B > 0$ define $Y_{n,,B}^\epsilon$ to be Y_n^ϵ but with

the components truncated at $\pm B$. Define $\hat{\theta}^\epsilon_m$ as follows. Let $\hat{\theta}^\epsilon_{t/\epsilon} = \theta^\epsilon_{t/\epsilon}$. Then, for $(t+\tau)/\epsilon \geq m \geq t/\epsilon$, set $\hat{\theta}^\epsilon_{m+1} = \hat{\theta}^\epsilon_m + \epsilon Y^\epsilon_{m,B}$. Now proceed with $\hat{\theta}^\epsilon_n$ replacing θ^ϵ_n, but continuing to use the original definition of W^ϵ_n. The result is (A1.5) plus an error which goes to zero as $B \to \infty$. □

An interpretation. Refer to the example in §7. Fix the parameter at θ. Suppose that $Y_n(\theta)$ is the IPA estimator on the interval $[nN, nN+N)$ without "resetting the accumulator." Then

$$Y_0(\theta) + \cdots + Y_{n-1}(\theta)$$

is an unbiased estimator of the derivative of the cost on $[0, nN]$. Suppose that for each fixed θ, the system is stationary. The condition (A1.1) is close to the assumption that

$$\frac{1}{n}\sum_{i=0}^{n-1} EY_i(\theta) \to g(\theta) \tag{A1.6}$$

for each initial condition, the only difference being in the conditioning data. Suppose that the mean cost per unit time on $[0,T]$ converges as time goes to infinity. This convergence and the convergence of the mean value of the left side of (A1.6) to $g(\theta)$ imply (the closed graph theorem) that $g(\theta)$ is the derivative of the mean ergodic cost at θ. Analogous comments apply to the example of §8.

The extensions of Theorem A1.1 are handled analogously to the way that the extensions of Theorem 3.1 were handled; e.g., for the analog of Theorem 3.3, replace the T_2/ϵ in (A1.1) by $m_n(T_2)$ and ϵ by ϵ_m. The general scheme is very flexible and allows many variations. More background and examples satisfying the conditions is in [22].

Appendix 2. An alternative averaging method. Return to (3.17) and the problem of replacing $(\xi^\epsilon_j, \theta^\epsilon_j)$ by $(\xi_j(\theta^\epsilon_{ln_\epsilon}), \theta^\epsilon_{ln_\epsilon})$ in

$$\frac{1}{n_\epsilon}\sum_{j=ln_\epsilon}^{ln_\epsilon+n_\epsilon-1} E^\epsilon_{ln_\epsilon} G^\epsilon_j(\theta^\epsilon_j, \xi^\epsilon_j) = \frac{1}{n_\epsilon}\sum_{j=0}^{n_\epsilon-1} E^\epsilon_{ln_\epsilon} G^\epsilon_{ln_\epsilon+j}(\theta^\epsilon_{ln_\epsilon+j}, \xi^\epsilon_{ln_\epsilon+j}). \tag{A2.1}$$

We present an alternative approach which avoids the use of the occupation measure $R(l,\epsilon,\cdot)$ but involves some other conditions. The method relies more heavily on continuity properties. Only a brief outline will be given, but the main idea should be clear. The sequence of integers n_ϵ might be different here than in Theorem 3.1. Continue to assume (3.2)–(3.5), (3.8), and (3.10). Suppose that $P^\epsilon_n(\xi, \cdot|\theta)$ is weakly continuous in (θ,ξ), uniformly in (ϵ, n) and in each compact (θ,ξ) set. Let $G^\epsilon_n(\theta,\xi)$ be continuous in (θ,ξ), uniformly in (ϵ,n) and in each compact (θ,ξ) set.

Due to the tightness (3.4), the uniform integrability (3.2) and the assumed uniform θ continuity of the $G^\epsilon_j(\cdot)$, we can suppose (as in Theorem 3.1) that the Y^ϵ_n are truncated and that in the interval $[t, t+\tau]$ of concern (see proof of Theorem 3.1) the θ^ϵ_n take values in some compact set. Thus we can suppose that $|\theta^\epsilon_{j+1} - \theta^\epsilon_j| = O(\epsilon)$ for all j of interest. For notational simplicity, we will not use the truncation notation.

Define

$$P^\epsilon_n\{d\xi, d\theta|\xi^\epsilon_n, \theta^\epsilon_n\} = P^\epsilon_n\{\xi^\epsilon_{n+1} \in d\xi, \theta^\epsilon_{n+1} \in d\theta|\xi^\epsilon_n, \theta^\epsilon_n\}.$$

Writing out the conditional expectation $E^\epsilon_{ln_\epsilon} G^\epsilon_{ln_\epsilon+j}(\theta^\epsilon_{ln_\epsilon+j}, \xi^\epsilon_{ln_\epsilon+j})$ in (A2.1), we have

$$\begin{aligned}\int \cdots \quad & \int P^\epsilon_{ln_\epsilon}\{d\xi_1, d\theta_1 | \xi^\epsilon_{ln_\epsilon}, \theta^\epsilon_{ln_\epsilon}\} \cdots \\ & P^\epsilon_{ln_\epsilon+j-1}\{d\xi_j, d\theta_j | \xi_{j-1}, \theta_{j-1}\} G^\epsilon_{ln_\epsilon+j}(\theta_j, \xi_j).\end{aligned}$$

Let $\epsilon n_\epsilon \to 0$. Now, using the uniform weak continuity of the P^ϵ_k, the uniform continuity of the G^ϵ_k, the tightness (3.4), and the fact that

$$\sup_{0 \le j \le n_\epsilon} |\theta^\epsilon_{ln_\epsilon+j} - \theta^\epsilon_{ln_\epsilon}| \to 0,$$

we can work backwards in the above equation, successively concentrating the measure of $\theta^\epsilon_{ln_\epsilon+j}$ at $\theta^\epsilon_{ln_\epsilon}$ and ultimately yielding the representation

$$\begin{aligned}\int \cdots \quad & \int P^\epsilon_{ln_\epsilon}(\xi^\epsilon_{ln_\epsilon}, d\xi_1 | \theta^\epsilon_{ln_\epsilon}) \cdots \\ & P^\epsilon_{ln_\epsilon+j-1}(\xi_{j-1}, d\xi_j | \theta^\epsilon_{ln_\epsilon}) G^\epsilon_{ln_\epsilon+j}(\theta^\epsilon_{ln_\epsilon}, \xi_j) + \rho^\epsilon_{ln_\epsilon}(\epsilon, j).\end{aligned}$$

The error term satisfies $|\rho^\epsilon_{ln_\epsilon}(\epsilon, j)| \le \rho(\epsilon, n_\epsilon)$, where $\rho(\epsilon, n_\epsilon)$ depends on the moduli of continuity and

$$\rho(\epsilon, n_\epsilon) \to 0 \tag{A2.2}$$

as $\epsilon \to 0$ for each constant $n_\epsilon = m$. Consequently, there are $n_\epsilon \to \infty$ such that (A2.2) holds for this sequence. The above discussion and the proof of Theorem 3.1 imply that the conclusions of Theorem 3.1 will hold under the additional condition that there is a function $g(\cdot)$ (which must be continuous by the above arguments) such that

$$\frac{1}{n_\epsilon} \sum_{j=ln_\epsilon}^{ln_\epsilon+n_\epsilon-1} E^\epsilon_{ln_\epsilon} G^\epsilon_j(\theta, \xi_j(\theta)) \to g(\theta) \tag{A2.3}$$

in mean for each θ, as $l \to \infty$, $\epsilon \to 0$, and $n_\epsilon \to \infty$.

An advantage of this averaging approach is that it can be used for grouping terms when the dependence of G^ϵ_j on ϵ, j does not vanish for large j and small ϵ.

Appendix 3. Arbitrary updatings within a regeneration period. This appendix illustrates the possibilities when updates are made after "partial" observations. It is intended to be suggestive and is a little vague. To ensure that the "partial" observations fit together properly, additional conditions are needed. Recall the example of §7, where we updated after each N departures. Owing to the regeneration stucture of the problem, one could have updated at the end of each regeneration period if the conditions of §3.4 held. These two approaches yield two different time scales in which to get the limit results. The $g(\cdot)$ functions would be different in the two cases but are related by the constant, which is the mean length of the renewal period. The results are equivalent since the two ODEs have the same asymptotic behavior. As seen in §3 and in the examples, there is no need in general to update at the end of regeneration periods. Indeed, even if the problem admits of a regeneration model, for general problems the intervals might be excessively long. If the problem has the "atomic increment" property of §3.3, then the regenerative structure does allow a rather arbitrary method of updating, within the intervals. By a regeneration process, we mean that for each fixed θ the process is regenerative and that for the physical process with the varying θ the conditional distribution of functionals of the

intervals $n, n+1, \ldots$ given the past depends only on the parameter value at the start of the nth interval. We will work within the regeneration setup but wish to update at arbitrary intervals (random times). This falls easily and naturally into our framework, as will now be shown. Only a brief outline will be given.

The basic algorithm is still (3.1). The updating times within the regeneration intervals can be chosen rather arbitrarily, subject to the mild conditions below. But we always update at the end of each regeneration interval. This last condition is not necessary but does simplify the discussion. Otherwise the groupings of the terms would be more involved. Let N_n^ϵ denote the number of updatings in the nth regeneration interval, $n = 1, 2, \ldots$. Define $M_0^\epsilon = 0$ and $M_n^\epsilon = \sum_{i=1}^{n-1} N_i^\epsilon$, $n \geq 1$. We now state the basic redefinitions and assumptions. They are essentially copies of those of Theorem 3.1. But since the estimation process begins anew at the start of each regeneration interval, the assumptions concern what happens within the intervals.

Let $N_n^\epsilon < \infty$ w.p.1 for all $\epsilon > 0, n < \infty$. Let $Y_{n,j}^\epsilon, j = 0, \ldots, N_n^\epsilon - 1$ denote the observations in the nth regeneration interval. We update after each observation. Hence there will be N_n^ϵ updates in the nth interval. Due to the assumption of a regenerative structure, the $\{Y_{m,i}^\epsilon, m \geq n, i \geq 0\}$ are conditionally independent of $\{Y_{m,i}^\epsilon, m < n, i \geq 0\}$ given $\theta_{M_n^\epsilon}^\epsilon$, the parameter value at the start of the nth interval. For $j \geq N_n^\epsilon$, set $Y_{n,j}^\epsilon = 0$. For each $\epsilon > 0, n \geq 1$, let $\mathcal{B}_{n,j}^\epsilon$ be a nondecreasing sequence of sigma-algebras measuring at least $\{\theta_{M_n^\epsilon}^\epsilon, Y_{n,i}^\epsilon, i < j\}$, with $E_{n,j}^\epsilon$ denoting the associated conditional expectation. Assume

(A3.1a) $$\left\{ \sum_{j=0}^{K} |Y_{n,j}^\epsilon|; n, \epsilon \right\} \text{ is uniformly integrable for each } K,$$

(A3.1b) $$E \sum_{N_n^\epsilon \wedge K}^{N_n^\epsilon} |Y_{n,j}^\epsilon| \to 0 \text{ as } K \to \infty.$$

Remark on (A3.1) *for the example of* §7. Return to the physical problem of §7. Suppose that there is an integer M such that we update at least after each new M departures but otherwise use the updating model of this section. Let R_n^ϵ denote the number of customers in the nth regeneration interval, and set $Q_n^\epsilon = \sum_{i=1}^{n-1} R_i^\epsilon$. We have

$$\sum_{i=0}^{N_n^\epsilon - 1} |Y_{n,i}^\epsilon| \leq \sum_{i=1}^{R_n^\epsilon} \sum_{l=1}^{i} |Z_{Q_n^\epsilon + l}^\epsilon|.$$

Condition (A3.1a) holds if $\{Z_l^\epsilon, \epsilon, l\}$ is uniformly integrable. Condition (A3.1b) holds if

$$\lim_{K \to \infty} E \sum_{i=K \wedge R_n^\epsilon}^{R_n^\epsilon} \sum_{l=1}^{i} |Z_{Q_n^\epsilon + l}^\epsilon| = 0,$$

where the limit is taken on uniformly in (n, ϵ).

The SA algorithm and interpolation. Now define $\theta_{n,j}^\epsilon = \theta_{M_n^\epsilon + j}^\epsilon$. The algorithm within the nth interval is

(A3.2) $$\theta_{n,j+1}^\epsilon = \theta_{n,j}^\epsilon + \epsilon Y_{n,j}^\epsilon, \qquad j < N_n^\epsilon.$$

Define the interpolated process $\theta^\epsilon(\cdot)$ by $\theta^\epsilon(\cdot) = \theta^\epsilon_{M^\epsilon_n}$ on the interval $[\epsilon n, \epsilon n + \epsilon)$, $\theta^\epsilon(t) = \theta^\epsilon_0, t \leq 0$. Thus, we update the parameter at arbitrary times but define the interpolation $\theta^\epsilon(\cdot)$ by the values of the parameter at the end of the regeneration intervals only. This makes the scaling easier, allows a nicer representation of the limit ODE, and yields the desired limit points of the algorithm. Assume that

$$\{\theta^\epsilon_{M^\epsilon_n}; \epsilon > 0, n < \infty\} \text{ is tight.} \tag{A3.3}$$

Analogous to the situation in Theorem 3.1, suppose that there are random variables $\{\xi^\epsilon_{n,j}; \epsilon > 0, j < \infty\}$ and measurable $G^\epsilon_{n,j}(\cdot)$ such that for $j < N^\epsilon_n$

$$E^\epsilon_{n,j} Y^\epsilon_{n,j} = G^\epsilon_{n,j}(\theta^\epsilon_{n,j}, \xi^\epsilon_{n,j}).$$

The values of $\xi^\epsilon_{n,j}$ for $j \geq N^\epsilon_n$ are irrelevant, and one can use any convenient one. Let $G^\epsilon_{n,j}(\cdot)$ be continuous, uniformly in ϵ, n, j and on each compact (θ, ξ) set. Assume

$$\{\xi^\epsilon_{n,j}; \epsilon > 0, n < \infty, j < \infty\} \text{ is tight.} \tag{A3.4}$$

Suppose that there are transition functions $P^\epsilon_{n,j}(\cdot)$ such that $P^\epsilon_{n,j}(\cdot, A|\cdot)$ is measurable for each Borel set A and that

$$P^\epsilon_{n,j}(\xi^\epsilon_{n,j}, \xi^\epsilon_{n,j+1} \in \cdot|\theta^\epsilon_{n,j}) = P\{\xi^\epsilon_{n,j+1} \in \cdot|\theta^\epsilon_{M^\epsilon_n}, Y^\epsilon_{n,k}, k \leq j\}. \tag{A3.5}$$

Let $P^\epsilon_{n,j}(\xi, d\tilde{\xi}|\theta)$ be weakly continuous in (θ, ξ), uniformly in ϵ, n, j and in each compact (θ, ξ) set. Now for each n, ϵ and θ, $P^\epsilon_{n,j}(\cdot|\theta), j \geq 0$, defines a nonhomogeneous "fixed θ" Markov chain. Let $\{\xi^\epsilon_{n,j}(\theta), j = 0, 1, \ldots\}$ denote the random variables of this chain.

Assume that there are continuous functions $g^\epsilon_n(\cdot)$ such that

$$g^\epsilon_n(\theta) = E \sum_{j=0}^{N^\epsilon_n - 1} G^\epsilon_{n,j}(\theta, \xi^\epsilon_{n,j}(\theta)). \tag{A3.6}$$

Define $\hat{\theta}^\epsilon_n = \theta^\epsilon_{M^\epsilon_n}$. Let there be a continuous function $g(\cdot)$ such that for each $\delta > 0$

$$\lim_\epsilon \limsup_n P\{|g^\epsilon_{n+1}(\hat{\theta}^\epsilon_n) - g(\hat{\theta}^\epsilon_n)| \geq \delta\} = 0. \tag{A3.7}$$

THEOREM A3.1. *Assume the conditions of this section. Then the conclusions of Theorem* 3.1 *continue to hold for* $\hat{\theta}^\epsilon(\cdot)$ *and* $\hat{\theta}^\epsilon(\epsilon q_\epsilon + \cdot)$ *and similarly for Theorems* 3.2, 3.3, 4.1, *and* 5.1.

Proof. Only a few basic remarks will be made. The assumption (A3.1) allows the set of observations in an interval to be truncated to and well approximated by some finite number K and guarantees uniform integrability of the set of those truncations. By (A3.1) and following the scheme in Theorem 3.1, for $n > 0$ we can write

$$\hat{\theta}^\epsilon_n = \theta^\epsilon_0 + \epsilon \sum_{i=1}^{n} E^\epsilon_{i,0} \sum_{j=0}^{N^\epsilon_i - 1} G^\epsilon_{i,j}(\theta^\epsilon_{i,j}, \xi^\epsilon_{i,j}) + \rho^\epsilon_n, \tag{A3.8}$$

where, for each t, $\sup_{\epsilon, n:\epsilon n \leq t} E|\rho^\epsilon_n| \to 0$ as $\epsilon \to 0$. Indeed, (A3.1) implies the tightness of $\{\hat{\theta}^\epsilon(\cdot), \hat{\theta}^\epsilon(\epsilon q_\epsilon + \cdot), \epsilon > 0\}$ and the fact that any weak limit will have Lipschitz

continuous paths w.p.1. The assumed uniform θ continuity of $G^\epsilon_{n,j}(\cdot)$, (A3.1), and the fact that

$$\sup_{j<N^\epsilon_i} |\theta^\epsilon_{i,j} - \hat{\theta}^\epsilon_{i-1}| \to 0$$

(in probability, uniform in i) as $\epsilon \to 0$ imply that we can replace the $\theta^\epsilon_{i,j}$ in (A3.8) with $\hat{\theta}^\epsilon_{i-1}$ without affecting the limit.

The only remaining problem concerns the fact that the distribution of the $\xi^\epsilon_{i,j}$ in (A3.8) depends on all the $\theta^\epsilon_{i,k}, k < j$. But the representation (A3.5), condition (A3.1), and the asserted uniform continuity of the $P^\epsilon_{i,j}, G^\epsilon_{i,j}$ can be used to show that $\xi^\epsilon_{i,j}$ can be replaced with $\xi^\epsilon_{i,j}(\hat{\theta}^\epsilon_{i-1})$ without changing the limits. Finally, (A3.6) and (A3.7) are used to complete the proof of the analogue of Theorem 3.1. The analogues of the other theorems will then follow. □

Appendix 4. Distributed/asynchronous updating: A network example. We will discuss a useful canonical form for a SA procedure that operates in a decentralized way and where different components of the iterate might be updated at different (random) times. Some components might be updated much more frequently than others. This is typical of a growing number of applications. One example is given below. Owing to this asynchronous behavior between the components, one needs to work with interpolated processes in an appropriate *real time* scale. Heretofore, the interpolations were based on the iterate number. But now, due to the possibly different times and frequencies of updating of the different components, one needs to use a common time scale for all the components, and this will be an appropriate "real" time scale. The general idea of the proof is just that of Theorem 3.1. The main added feature concerns the difference in the time scaling of the interpolations. Working directly with the iterates can lead to a notational nightmare. We avoid the need to deal directly with the possibly different and random interpolation intervals by using appropriate rescaling. This puts the problem into a framework where the previous results can be directly applied. The result is a simplification and extension of the results in [30, 40], where the ideas of time scaling first appeared. The central idea of the rescaling is easier to see if we start with a centralized and synchronized updating and interpolate in real time. This will be done in the subsection below. The general result for the decentralized problem is in §A4.2.

A4.1. A synchronized updating: Real time scale. Assume the conditions of Theorem 3.1. The algorithm is

$$\theta^\epsilon_{n+1} = \theta^\epsilon_n + \epsilon Y^\epsilon_n.$$

Then Theorem 3.1 holds for the interpolations $\{\theta^\epsilon(\cdot), \theta^\epsilon(\epsilon q_\epsilon + \cdot)\}$.

Now let us rewrite the interpolation in real time. Let $\delta\tau^\epsilon_n$ denote the time interval between the nth and $(n+1)$st updating. Let $\mathcal{B}^\epsilon_n$ be a nondecreasing sequence of sigma-algebras such that $\mathcal{B}^\epsilon_n$ measures at least $\{\theta^\epsilon_0, Y^\epsilon_i, \delta\tau^\epsilon_i, i < n\}$. Let E^ϵ_n denote the associated conditional expectation. We keep the same framework as in §3. Suppose that there is a Markov chain ξ^ϵ_n and a continuous and strictly positive function $u(\cdot)$ such that $E^\epsilon_n \delta\tau^\epsilon_n = u(\theta^\epsilon_n, \xi^\epsilon_n)$. Assume the tightness condition (3.4), the uniform integrability condition (3.2), and

(A4.1) $\{\delta\tau^\epsilon_n, n < \infty\}$ is uniformly integrable.

Also suppose that there is a continuous and bounded function $\hat{u}(\cdot)$ such that

$$\frac{1}{n}\sum_{i=0}^{n-1} Eu(\theta, \xi_i(\theta)) \to \hat{u}(\theta) \tag{A4.2}$$

for each θ and initial condition.

Define

$$\tau_n^\epsilon = \epsilon \sum_{i=0}^{n-1} \delta\tau_i^\epsilon,$$

$$N^\epsilon(t) = \epsilon[\text{ number of updatings by time } t/\epsilon].$$

Let $\tau^\epsilon(\cdot)$ be the interpolation of $\{\tau_n^\epsilon, n < \infty\}$ defined by $\tau^\epsilon(t) = \tau_n^\epsilon$ on $[\epsilon n, \epsilon(n+1))$. Note that $\tau^\epsilon(\cdot)$ is the inverse of $N^\epsilon(\cdot)$ in the sense that $N^\epsilon(\tau^\epsilon(t)) = n\epsilon$ for $t \in [n\epsilon, n\epsilon + \epsilon)$ and

$$\tau^\epsilon(t) = \inf\{s : N^\epsilon(s) \geq t\}. \tag{A4.3}$$

Define $\hat{\theta}^\epsilon(t) = \theta^\epsilon(N^\epsilon(t))$. This is the interpolation in the real time scale, not the iterate time scale. The weak convergence and characterization of the ODE for the $\hat{\theta}^\epsilon(\cdot)$ are now easily done. In all cases we suppose that the original sequence indexed by ϵ converges weakly. Otherwise take appropriate subsequences. By the above conditions and the proof of Theorem 3.1, $(\tau^\epsilon(\cdot), \hat{\theta}^\epsilon(\cdot), N^\epsilon(\cdot), \theta^\epsilon(\cdot)) \Rightarrow (\tau(\cdot), \hat{\theta}(\cdot), N(\cdot), \theta(\cdot))$, where

$$\hat{\theta}(t) = \theta(N(t)) \quad \text{and} \quad \tau(t) = \int_0^t \hat{u}(\theta(s))ds. \tag{A4.4}$$

From the positivity of $\hat{u}(\theta)$ and the "inverse" definitions of $N^\epsilon(\cdot)$ and $\tau^\epsilon(\cdot)$, it follows that $N^\epsilon(\cdot) \Rightarrow N(\cdot)$, where $N(\tau(t)) = t$. Taking derivatives, we get $\dot{N}(\tau(t))\dot{\tau}(t) = 1$. Call $s = \tau(t)$. Then using (A4.4), the slope of $N(s)$ is $\dot{N}(s) = 1/\hat{u}(\theta(\tau^{-1}(s))) = 1/\hat{u}(\hat{\theta}(s))$. Therefore

$$N(t) = \int_0^t \frac{ds}{\hat{u}(\hat{\theta}(s))}. \tag{A4.5}$$

By Theorem 3.1, $\theta(\cdot)$ satisfies $\dot{\theta} = g(\theta)$. Recall that

$$\hat{\theta}^\epsilon(\cdot) = \theta^\epsilon(N^\epsilon(\cdot)) \Rightarrow \theta(N(\cdot)) \equiv \hat{\theta}(\cdot). \tag{A4.6}$$

Thus, using the fact that $N(\tau(t)) = t$, we can write

$$\dot{\hat{\theta}}(t) = [\dot{\theta}(N(t))]\dot{N}(t) = g(\theta(N(t)))/\hat{u}(\hat{\theta}(t)) = g(\hat{\theta}(t))/\hat{u}(\hat{\theta}(t)). \tag{A4.7}$$

Thus, the proof is just Theorem 3.1 plus a time change argument. The purpose of the time change argument is to avoid dealing with random interpolation intervals and the interaction of the Y_n^ϵ and the $\delta\tau_n^\epsilon$. It exploits the convergence of both the "time" processes and of the original interpolation $\theta^\epsilon(\cdot)$.

Remark. The above argument is for processes that start at time zero with limits defined on the interval $[0, \infty)$. Suppose that we wish to get the limit on $(-\infty, \infty)$ of

$\hat{\theta}^\epsilon(T^\epsilon + \cdot)$, where T^ϵ is a sequence of real numbers tending to infinity. The T^ϵ is simply the replacement for the ϵq_ϵ in Theorem 3.1. Then the analysis is the same as above, except that the initial condition of the interpolation is

$$\hat{\theta}^\epsilon(T^\epsilon) = \theta^\epsilon(N^\epsilon(T^\epsilon)) = \theta^\epsilon_{N^\epsilon(T^\epsilon)/\epsilon},$$

the values of the parameter at increasingly large iterate numbers, and the uniform integrability and tightness conditions must reflect this change. In particular, we need tightness of

$$\{\xi^\epsilon_{N^\epsilon(t)/\epsilon+n}, \theta^\epsilon_{N^\epsilon(t)/\epsilon+n}; \epsilon, n, t\} \tag{A4.8}$$

and uniform integrability of

$$\{Y^\epsilon_{N^\epsilon(t)/\epsilon+n}, \delta\tau^\epsilon_{N^\epsilon(t)/\epsilon+n}; \epsilon, n, t\}. \tag{A4.9}$$

Since all sorts of dependencies among the two sequences Y^ϵ_n and $\delta\tau^\epsilon_n$ can be constructed, little can be said without further assumptions. But a casual examination of some simple cases suggests that (A4.8) and (A4.9) are not very restrictive.

A distributed and decentralized network model. We work with one canonical model in order to illustrate some of the possibilities and minimize notation. To simplify the notation, some of the conditions will be less general than can be handled by the introduced technique. The basic work is in setting up the notation for the various time scales. Basically the general method uses the idea of the above subsection separately on different parts of the problem, as will now be seen.

Let $\theta = (\theta_1, \ldots, \theta_K)$, where the θ_α are the scalar components of θ. Consider a system with K controllers, each of which is responsible for the updating of one component. We wish to minimize a function $F(\cdot)$ which takes the form $F(\theta) = \sum_{\beta=1}^K F^\beta(\theta)$ for real-valued and continuously differentiable $F^\beta(\cdot)$. Let $F^\beta_\alpha(\theta) = \partial F^\beta(\theta)/\partial\theta_\alpha$. In our model, for each α subsystem β produces a sequence of estimates $Y^{\beta,\epsilon}_{\alpha,n}, n = 0, \ldots,$ which it sends to node α for help in estimating F^β_α at whatever the current value of θ is. It also sends the current values of its own component θ_β.

Example. An important class of examples that provides a guide to the development are the problems of optimal routing in queueing networks. Let the network have K nodes, with θ the K vector of routing parameters, where θ_α is the component associated with the αth node. Let $F^\beta(\theta)$ denote the stationary average queue length at node β under parameter value θ. We wish to minimize the stationary average number of customers in the network $F(\theta) = \sum F^\beta(\theta)$. The problem arises in control of telecommunication networks and has been treated in [42, 40]. The controller at node α updates the component θ_α of θ, and it does so based on both its own observations and relevant data sent from other nodes. In one useful approach, called the *surrogate estimation method* in the above references, each node β estimates the sensitivity of the mean length of its own queue to variations in *external* inputs to that node. Then one uses the mean systems flow equations to get acceptable estimates of the $F^\beta_\alpha(\theta)$. These estimates are transmitted to node α for use in estimating the derivative of $F(\theta)$ with respect to θ_α at the current value of θ and then updating the value of θ_α. After each transmission, new estimates are taken and the process is repeated. The method gave good results in simulations.

The times required for the estimation intervals can depend heavily and randomly on the node. They might be functions of the number of service completions or simply

deterministic time intervals. The nodes would transmit their estimates in an asynchronous way. Thus the SA is both decentralized and unsynchronized. In general, θ_α would be a vector of routing probabilities. For simplicity of notation, we shall consider only scalar components. The extensions to the vector case are straightforward. In a typical application of SA, each time a new estimate of $F_\alpha^\beta(\theta)$ (at the current value of θ) is received at node α, that estimate is multiplied by a step-size parameter and subtracted from the current value of state component θ_α. This "additive" structure allows us to represent the algorithm in a useful decomposed way by writing the current value of the component θ_α as the sum of the initial value plus K terms. The βth such term is the product of an appropriate step-size times the sum of the past transmissions from node β to node α of the estimates of $F_\alpha^\beta(\theta)$ at whatever the operating values of the parameter were when the estimates were made. In the development below, this decomposition is formalized and provides a useful simplification.

We shall now return to our general model. The time for transmission of information can have bounded delays, and these delays cause no problems in the analysis. But only to simplify notation, we work under the assumption that there are no delays and that the parameters are updated as soon as new information is available. The reader can fill in the few additional details for the delayed case. We are reluctant to try a very general development since the entire field of decentralized/asynchronous optimization is in its infancy, and one expects many new models and methods for estimation to appear in the next few years. But the methods employed would be fundamental to any extensions.

Notation. Let $\delta\tau_{\alpha,n}^{\beta,\epsilon}$ denote the interval between the nth and $(n+1)$st transmissions from β to α. Define

$$\tau_{\alpha,n}^{\beta,\epsilon} = \epsilon \sum_{i=0}^{n-1} \delta\tau_{\alpha,i}^{\beta,\epsilon},$$

ϵ times the real time required by the first n transmissions from β to α. Define

$$N_\alpha^{\beta,\epsilon}(t) = \epsilon \text{ [number of transmissions from } \beta \text{ to } \alpha \text{ to reach real time } t/\epsilon].$$

Let $\tau_\alpha^{\beta,\epsilon}(t)$ be the interpolation of the $\tau_{\alpha,n}^{\beta,\epsilon}$ with interpolation intervals ϵ and initial condition zero. Analogously to the situation in the last subsection, $N_\alpha^{\beta,\epsilon}(\cdot)$ and $\tau_\alpha^{\beta,\epsilon}(\cdot)$ are inverses of one another.

The SA algorithm. The notation is a little complex but very natural. It enables us to carry over the results of Theorem 3.1 to a much more complex situation via several time-change arguments and thus saves a great deal of work over a direct analysis. Let $\hat{\theta}^\epsilon(\cdot) = \{\hat{\theta}_i^\epsilon(\cdot), i \le K\}$ denote the interpolation in the real time (times ϵ) scale. As mentioned in the discussion of the example, it is convenient to separate $\hat{\theta}_\alpha^\epsilon(\cdot)$ into components which come from the same node. This suggests the following decomposed representation for the SA algorithm. For each α, β, let $c_\alpha^\beta(\cdot)$ be a continuous and bounded real-valued function and define the sequence $\theta_{\alpha,n}^{\beta,\epsilon}$ by

$$\theta_{\alpha,n+1}^{\beta,\epsilon} = \theta_{\alpha,n}^{\beta,\epsilon} + \epsilon c_a^\beta(\hat{\theta}^\epsilon(\tau_{\alpha,n}^{\beta,\epsilon}))Y_{\alpha,n}^{\beta,\epsilon}. \tag{A4.10}$$

The role of the $c_\alpha^\beta(\cdot)$ functions is to partially compensate for the fact that the frequency of the intervals between updates might depend on θ, α, β, and will be further commented upon at the end of the section. In many cases, we would use $c_\alpha^\beta(\theta) \equiv 1$. Note

that by the definitions $\hat{\theta}^{\epsilon}(\tau_{\alpha,n}^{\beta,\epsilon})$ is the state value at the time of the nth transmission from node β to node α. Indeed, we can write

$$\begin{aligned} &\hat{\theta}_{\alpha}^{\beta,\epsilon}(t) = \theta_{\alpha,N_{\alpha}^{\beta,\epsilon}(t)/\epsilon}^{\beta,\epsilon} = \theta_{\alpha}^{\beta,\epsilon}(N_{\alpha}^{\beta,\epsilon}(t)), \\ &\hat{\theta}_{\alpha}^{\beta,\epsilon}(\tau_{\alpha}^{\beta,\epsilon}(t)) = \theta_{\alpha}^{\beta,\epsilon}(t), \text{ where } \theta_{\alpha}^{\beta,\epsilon}(t) = \theta_{\alpha,n}^{\beta,\epsilon} \text{ for } t \in [n\epsilon, (n+1)\epsilon). \end{aligned} \tag{A4.11}$$

We can now define the actual interpolated iterate in the appropriate real time scale in terms of the components as

$$\hat{\theta}_{\alpha}^{\epsilon}(t) = \hat{\theta}_{\alpha}^{\epsilon}(0) + \sum_{\beta=1}^{K} \hat{\theta}_{\alpha}^{\beta,\epsilon}(t), \ \hat{\theta}_{\alpha}^{\beta,\epsilon}(0) = 0. \tag{A4.12}$$

It will be shown that the proofs are just adaptations of the argument in the last subsection to the vector case. It will be seen from the argument that all sorts of groupings and variations can be added to the format.

Assumptions. Let $\mathcal{B}_t^{\epsilon}$ be a nondecreasing sequence of sigma-algebras which measure at least the initial conditions and all the data transmitted by all the nodes up to real time t. Let $E_{\alpha,n}^{\beta,\epsilon}$ equal the expectation conditioned on $\mathcal{B}_{\tau_{\alpha,n}^{\beta,\epsilon}/\epsilon}^{\epsilon}$. Thus, $E_{\alpha,n}^{\beta,\epsilon}$ can be interpreted as the expectation conditioned on the information which is available at the time $\tau_{\alpha,n}^{\beta,\epsilon}$ of the nth transmission from β to α. We next give the conditions on the interupdate intervals. We suppose that for each α, β, ϵ there is a Markov chain $\{\xi_{\alpha,n}^{\beta,\epsilon}, n < \infty\}$ whose transition functions satisfy the obvious analogue of (3.5) and continuous functions $u_{\alpha}^{\beta}(\cdot)$ such that

$$\{\delta\tau_{\alpha,n}^{\beta,\epsilon}; \alpha, \beta, \epsilon, n\} \text{ is uniformly integrable,} \tag{A4.13}$$

$$E_{\alpha,n}^{\beta,\epsilon}\delta\tau_{\alpha,n}^{\beta,\epsilon} = u_{\alpha}^{\beta}(\hat{\theta}^{\epsilon}(\tau_{\alpha,n}^{\beta,\epsilon}), \xi_{\alpha,n}^{\beta,\epsilon}). \tag{A4.14}$$

Let there be fixed θ Markov chains $\{\xi_{\alpha,n}^{\beta}(\theta)\}$ with transition probabilities satisfying the analogues of (3.6)–(3.8). Let there be continuous functions $\hat{u}_{\alpha}^{\beta}(\cdot)$ such that for each initial condition and each θ

$$\lim_{N} \frac{1}{N} \sum_{i=0}^{N-1} E u_{\alpha}^{\beta}(\theta, \xi_{\alpha,n}^{\beta}(\theta)) = \hat{u}_{\alpha}^{\beta}(\theta). \tag{A4.15}$$

We also need that

$$\inf_{\theta,\xi} u_{\alpha}^{\beta}(\theta, \xi) > 0. \tag{A4.16}$$

Let there be uniform integrability of

$$\{Y_{\alpha,n}^{\beta,\epsilon}; \alpha, \beta, \epsilon, n\} \tag{A4.17}$$

and tightness of

$$\{\xi_{\alpha,n}^{\beta,\epsilon}, \theta_{\alpha,n}^{\beta,\epsilon}; \alpha, \beta, \epsilon, n\}. \tag{A4.18}$$

Define $G_{\alpha}^{\beta}(\cdot)$ in the usual way:

$$E_{\alpha,n}^{\beta,\epsilon} Y_{\alpha,n}^{\beta,\epsilon} = G_{\alpha}^{\beta}(\hat{\theta}(\tau_{\alpha,n}^{\beta,\epsilon}), \xi_{\alpha,n}^{\beta,\epsilon}) + \text{ small error,}$$

where the small error goes to zero in the mean as $\epsilon \to 0$ and $n \to \infty$. Suppose that there are continuous functions $g_\alpha^\beta(\theta)$ such that for each initial condition and each θ

$$\lim_N \frac{1}{N} \sum_{n=1}^{N} EG_\alpha^\beta(\theta, \xi_{\alpha,n}^\beta(\theta)) = g_\alpha^\beta(\theta). \tag{A4.19}$$

The above conditions are for starting at time zero. If we wish to work with $\hat{\theta}^\epsilon(T^\epsilon + \cdot)$ as in the first subsection, then we need to shift the indices analogously to what was done there.

Remark. In the case of the example mentioned above, the assumptions on the interupdate intervals are obviously satisfied if the nodes compute the estimators at constant intervals but also in many cases where (for example) either a fixed number of service completions or perhaps a "local" regenerative approach is used for the local estimation. Note that the chains are "local" in the sense that they can depend on the pair α, β. Thus, we work with each pair separately, which can give a simpler chain than what would appear if we treated all the pairs simultaneously. Equation (A4.16) is used to guarantee that $\{N_\alpha^{\beta,\epsilon}(\cdot), \epsilon > 0\}$ is tight and has continuous limits.

THEOREM A4.1. *Every subsequence of $\hat{\theta}(\cdot)$ has a further subsequence that converges weakly to a solution of the ODE:*

$$\dot{\hat{\theta}}_\alpha = \sum_{\beta=1}^{K} \frac{c_\alpha^\beta(\hat{\theta})}{\hat{u}_\alpha^\beta(\hat{\theta})} g_\alpha^\beta(\hat{\theta}).$$

Comments. Now we see the use of the $c_\alpha^\beta(\cdot)$ functions as a way of dealing with the variable $\hat{u}_\alpha^\beta(\theta)$. For the example, the $g_\alpha^\beta(\theta)$ are supposed to be approximations to the $F_\alpha^\beta(\theta)$. The proof of the theorem uses the ideas of the previous subsection. The $\{\theta_\alpha^{\beta,\epsilon}(\cdot), \tau_\alpha^{\beta,\epsilon}(\cdot)\}$ is tight, and all weak limits have Lipschitz-continuous paths w.p.1. Also, $\{N_\alpha^{\beta,\epsilon}(\cdot)\}$ is tight, and all limits have Lipschitz-continuous paths. Thus, $\{\hat{\theta}_\alpha^{\beta,\epsilon}(\cdot), \hat{\theta}^\epsilon(\cdot)\}$ is tight and has Lipschitz-continuous limits. Let $(\theta_\alpha^\beta(\cdot), \theta(\cdot), \hat{\theta}(\cdot), \tau_\alpha^\beta(\cdot), N_\alpha^\beta(\cdot))$ denote the limit processes. We have $\tau_\alpha^{\beta,\epsilon}(\cdot) \Rightarrow \tau_\alpha^\beta(\cdot)$, where

$$\tau_\alpha^\beta(t) = \int_0^t \hat{u}_\alpha^\beta(\hat{\theta}(\tau_\alpha^\beta(s)))ds.$$

In the centralized case of the previous subsection, the argument of the $\hat{u}$ reduces to just $\theta(s)$. Also $N_\alpha^{\beta,\epsilon}(\cdot) \Rightarrow N_\alpha^\beta(\cdot)$, where

$$N_\alpha^\beta(t) = \int_0^t \frac{ds}{\hat{u}_\alpha^\beta(\hat{\theta}(s))}.$$

The form of the algorithm (A4.10), the weak convergence of $(\hat{\theta}^\epsilon(\cdot), \tau_\alpha^{\beta,\epsilon}(\cdot))$ to $(\hat{\theta}(\cdot), \tau_\alpha^\beta(\cdot))$, and Theorem 3.1 yield that

$$\dot{\theta}_\alpha^\beta(t) = c_\alpha^\beta(\hat{\theta}(\tau_\alpha^\beta(t)))g_\alpha^\beta(\hat{\theta}(\tau_\alpha^\beta(t))).$$

The theorem follows by writing the expression for $\dot{\hat{\theta}}_\alpha^\beta(t)$ and using the fact that $N_\alpha^\beta(\tau_\alpha^\beta(t)) = t$.

Remark. Note the great advantage in using the rescaling idea. It allows us to separate the intervals form the values of the updates in the analysis and permits a result under quite weak conditions with minimal new work. It is a technique of considerable utility. The analogues of Theorems 3.2, 4.1, and 5.1 also hold.

REFERENCES

[1] M. BENAIM, *A dynamical systems approach to stochastic approximations*, Univ. of California, Berkeley, CA, 1993, preprint.

[2] A. BENVENISTE, M. METIVIER, AND P. PRIOURET, *Adaptive Algorithms and Stochastic Approximation*, Springer-Verlag, New York, Berlin, 1990.

[3] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[4] P. BRÉMAUD AND F. J. VÁZQUEZ-ABAD, *On the pathwise computation of derivatives with respect to the rate of a point process: The phantom RPA method*, Queueing Systems Theory Appl., 10 (1992), pp. 249–270.

[5] F. CAMPILLO AND E. PARDOUX, *Numerical methods in ergodic optimal stochastic control*, in Applied Stochastic Analysis, I. Karatzas and D. Ocone, eds., Springer-Verlag, Berlin, 1991, pp. 59–73.

[6] M. CARAMANIS AND G. LIBEROPOULOS, *Perturbation analysis for the design of flexible manufacturing system flow controllers*, Oper. Res., 40 (1992), pp. 1107–1125.

[7] E. K. P. CHONG AND P. J. RAMADGE, *Optimization of queues using an infinitesimal perturbation analysis-based stochastic algorithm with general update times*, SIAM J. Control Optim., 31 (1993), pp. 698–732.

[8] M. H. A. DAVIS, *Markov Models and Optimization*, Chapman-Hall, New York, London, 1993.

[9] P. DUPUIS AND H. J. KUSHNER, *Stochastic approximation and large deviations: Upper bounds and w.p.1 convergence*, SIAM J. Control Optim., 27 (1989), pp. 1108–1135.

[10] S. N. ETHIER AND T. G. KURTZ, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.

[11] M. C. FU, *Convergence of the stochastic approximation algorithm for the GI/G/1 queue using infinitesimal perturbation analysis*, J. Optim. Theory Appl., 65 (1990), pp. 149–160.

[12] I. I. GIHMAN AND A. V. SKOROHOD, *Stochastic Differential Equations*, Springer-Verlag, Berlin, 1972.

[13] P. GLASSERMAN, *Gradient Estimation via Perturbation Analysis*, Kluwer, Amsterdam, 1990.

[14] ———, *Filtered Monte Carlo*, Math. Oper. Res., 18 (1993), pp. 610–634.

[15] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurctions of Vector Fields*, Springer-Verlag, Berlin, 1983.

[16] A. HAURIE, P. L'ECUYER, AND CH. VAN DELFT, *Convergence of stochastic approximation coupled with perturbation analysis in a class of manufacturing flow control models*, J. Discrete Event Dynamical Systems, 4 (1994), pp. 87–111.

[17] Y.-C. HO AND X.-R CAO, *Performance sensitivity to routing changes in queueing networks and flexible manufacturing systems using perturbation analysis*, IEEE J. on Robotics and Automation, RA-1 (1985), pp. 165–172.

[18] ———, *Perturbation Analysis of Discrete Event Dynamical Systems*, Kluwer, Boston, 1991.

[19] Y. C. HO, M. A. EYLER, AND T. T. CHIEN, *A gradient technique for general buffer storage design in a production line*, Internat. J. of Production Research, 17 (1979), pp. 577–580.

[20] ———, *A new approach to determine parameter sensitivities of transfer lines*, Management Science, 29 (1983), pp. 700–714.

[21] T. G. KURTZ, *Semigroups of conditional shifts and approximation of Markov processes*, Ann. Probab., 4 (1975), pp. 618–642.

[22] H. J. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic System Theory*, MIT Press, Cambridge, MA, 1984.

[23] ———, *Robustness and approximation of escape times and large deviations estimates for systems with small noise effects*, SIAM J. Appl. Math., 44 (1984), pp. 160–182.

[24] ———, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Systems and Control, Vol. 3, Birkhäuser, Boston, 1990.

[25] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation for Constrained and Unconstrained Systems*, Springer-Verlag, Berlin, New York, 1978.

[26] H. J. KUSHNER AND J. YANG, *A Monte Carlo method for sensitivity analysis and parametric optimization of nonlinear stochastic systems: The ergodic case*, SIAM J. Control Optim., 30 (1992), pp. 440–464.

[27] H. J. KUSHNER AND A. SHWARTZ, *An invariant measure approach to the convergence of stochastic approximations with state dependent noise*, SIAM J. Control Optim., 22 (1984), pp. 13–27.

[28] ———, *Weak convergence and asymptotic properties of adaptive filters with constant gains*, IEEE Trans. Inform. Theory, IT-30 (1984), pp. 177–182.

[29] H. J. KUSHNER AND J. YANG, *Stochastic approximation with averaging: Optimal asymptotic rates of convergence for general processes*, SIAM J. Control Optim., 31 (1993), pp. 1045–

1062.

[30] H. J. KUSHNER AND G. YIN, *Stochastic approximation algorithms for parallel and distributed processing*, Stochastics, 22 (1987), pp. 219–250.

[31] P. L'ECUYER, N. GIROUX, AND P. W. GLYNN, *Stochastic optimization by simulation: Numerical experiments for a simple queue in steady state*, Management Sci., 40 (1994), pp. 1245–1261.

[32] P. L'ECUYER AND P. W. GLYNN, *Stochastic optimization by simulation: Convergence proofs for the GI/G/1 queue*, Management Sci., 40 (1994), pp. 1562–1578.

[33] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 551–575.

[34] J. PAN AND C. G. CASSANDRAS, *Flow control of bursty traffic using infinitesimal perturbation analysis*, J. Discrete Event Systems, 4 (1994), pp. 325–358.

[35] G. CH. PFLUG, *On line optimization of simulated Markov processes*, Math. Oper. Res., 15 (1990), pp. 381–395.

[36] B. T. POLYAK AND A. B. JUDITSKY, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.

[37] M. I. REIMAN AND A. WEISS, *Sensitivity analysis for simulation via likelihood ratios*, Oper. Res., 37 (1989), pp. 930–844.

[38] R. Y. RUBINSTEIN, *Sensitivity analysis and performance extrapolation for computer simulation models*, Oper. Res., 37 (1989), pp. 72–81.

[39] R. SURI AND B. R. FU, *On using continuous flow lines for performance estimation of discrete production lines*, in Proc. 1991 Winter Simulation Conference, 1991, B.L. Nelson, W. D. Kelton, and G. M. Clark, eds., pp. 968–977.

[40] F. J. VÁZQUEZ-ABAD, *Stochastic Recursive Algorithms for Optimal Routing in Queueing Networks*, Ph.D. thesis, Brown University, 1989.

[41] F. J. VÁZQUEZ-ABAD AND H. J. KUSHNER, *Estimation of the derivative of a stationary measure with respect to a control parameter*, J. Appl. Probab., 29 (1992), pp. 343–352.

[42] ———, *The surrogate estimation approach for sensitivity analysis in queueing networks*, in G. W. Evans, M. Mollaghasemi, E. C. Russel, and W. E. Biles, eds., Proc. Winter Simulation Conference, 1993, pp. 347–355.

[43] F. J. VÁZQUEZ-ABAD AND L. MASON, *Adaptive decentralized control under non uniqueness of the optimal control*, J. Discrete Event Dyn. Syst., in press.

[44] H. YAN, G. YIN, AND S. X. C. LOU, *Using stochastic approximation to determine threshold values for control of unreliable manufacturing systems*, J. Optim. Theory Appl., 83 (1994), pp. 511–539.

[45] H. YAN, X.Y. ZHOU, AND G. YIN, *Optimal numbers of circulating kanbans in a two machine flowshop with uncertainty*, Oper. Res., to appear.

[46] G. YIN, *On extensions of Polyak's averaging approach to stochastic approximation*, Stochastics, 36 (1992), pp. 245–264.

[47] ———, *Stochastic approximation via averaging: Polyak's approach revisited*, in Simulation and Optimization: Proceedings of the International Workshops on Computationally Intensive Methods in Simulation and Optimization held at the International Institute for Applied Systems Analysis, Laxenburg, Austria, August 23–25, 1990, Lecture Notes in Economics and Mathematical Systems 374, G. Pflug and U. Dieter, eds., Springer-Verlag, Berlin, 1992, pp. 119–134.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 757-812, May 1996

INFINITE-DIMENSIONAL CONTINUOUS-TIME LINEAR SYSTEMS: STABILITY AND STRUCTURE ANALYSIS*

RAIMUND J. OBER† AND YUANYIN WU†

Abstract. The question of exponential and asymptotic stability of infinite-dimensional continuous-time state-space systems is investigated. It is shown that every (par)balanced realization is asymptotically stable. Conditions are given for (par)balanced, input-normal, or output-normal realizations to be asymptotically and/or exponentially stable. The boundedness of the system operators is also studied. Examples of delay systems are given to illustrate the theory.

Key words. linear infinite-dimensional systems, balanced realizations, stability, Hankel operators, semigroups of operators

AMS subject classifications. 93B15, 93B20, 93B28, 93D20

1. Introduction.

For a finite-dimensional linear system with transfer function G, there are standard ways to obtain a minimal, i.e., reachable and observable, state-space realization:

$$\begin{cases} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t). \end{cases}$$

This realization is unique in the sense that every other minimal realization is equivalent to it. The spectrum of the state propagation operator A is precisely the set of poles of the transfer function $G(s) = C(sI - A)^{-1}B + D$, which is proper rational. Hence the realization is exponentially stable if and only if the poles of G are all in the open left half plane. Furthermore, exponential stability of the system is equivalent to asymptotic stability.

This paper is concerned with the question of stability for infinite-dimensional systems. If the transfer function G is not rational, then we have an infinite-dimensional system of the above form, where the system operators A, B, and C are usually unbounded operators. In general, it is no longer true that all observable and reachable realizations are equivalent. The correspondence between the spectrum of the realization and the singularities of the transfer function does not necessarily hold. In general the exponential stability of a system cannot be determined by the location of the singularities of the transfer function (see, e.g., [18]). Also asymptotically stable systems are typically not exponentially stable.

There have been attempts to extend the results for finite-dimensional systems mentioned above to the infinite-dimensional case by restricting the transfer functions to a certain class. For example, Curtain [4], Yamamoto [29], and several other authors considered the equivalence between input/output stability and internal stability. We refer to [4] and [29] and the reference therein for the work in this direction. Inevitably, the stronger the results are, the smaller the class of transfer functions is.

Here we present another approach. Instead of putting too stringent restrictions on the class of transfer functions to be studied, we restrict the class of realizations to (par)balanced realizations and the closely related input-normal and output-normal realizations. These types of realizations have been advocated by several authors [20], [13], [14], [30]. They were introduced in the finite-dimensional case as a means to perform model reduction in an easy fashion

*Received by the editors March 10, 1993; accepted for publication (in revised form) November 9, 1994. This research was supported in part by NSF grant DMS-9304696. The research of the second author was supported in part by Texas Advanced Research Program grant 00974103.

†Center for Engineering Mathematics, University of Texas at Dallas, Richardson, TX 75083-0688.

[20]. Glover, Curtain, and Partington [14] derived infinite-dimensional continuous-time balanced realizations for a class of transfer functions with nuclear Hankel operators. Young [30] developed a general realization theory of balanced realizations of infinite-dimensional discrete-time systems. The results were generalized to the continuous-time case by Ober and Montgomery-Smith [23]. The results by Young were also used by the authors to conduct an analysis of the stability and structural properties for infinite-dimensional discrete-time systems in [24].

In this paper we extend our analysis in [24] to the continuous-time case. The exponential and asymptotic stability properties of parbalanced, input-normal, and output-normal realizations are studied in detail. It is shown that all parbalanced realizations are asymptotically stable. For a subclass of transfer functions—namely, strictly noncyclic functions—results that are reminiscent of the finite-dimensional case are obtained. For this class of transfer functions the location of the singularities of the transfer function determines the exponential stability properties of parbalanced systems. The stability properties of parbalanced realizations are studied without the explicit presentation of the realizations. Structural properties of the realizations are also analyzed. In particular the boundedness of the system operators of the input- and output-normal realizations is investigated.

Most of the results presented in this paper are in terms of the properties of the transfer functions and the Hankel operators with the transfer functions as symbols. This may therefore be regarded as expressing the internal properties of a system in terms of input/output properties. Related topics can be found in Dewilde [7], where systems with strictly noncyclic transfer functions are studied from an input/output point of view. We also refer to Baras, Brockett, and Fuhrmann [2], [3], [11]. For realization theory of nonrational transfer functions, Fuhrmann [11] and Helton [16] reference for transfer provide general references.

Our main tool is a bilinear map that maps discrete-time systems to continuous-time systems. This bilinear map is routinely used for finite-dimensional systems to translate discrete-time results to continuous-time results and vice versa. In [23] properties of this bilinear map were studied for infinite-dimensional systems (see also [11]). Some continuous-time questions, however, such as exponential stability, cannot be directly answered by simply applying the bilinear transform to a discrete-time result. In such cases a more detailed study of the problem is necessary.

The contents of the paper can be summarized as follows. In §2 we review the settings of infinite-dimensional continuous-time systems we will deal with. We restrict ourselves to so-called admissible systems. We relate continuous-time systems to discrete-time systems in §3, using the above-mentioned bilinear map. As Hankel operators play an important role in our approach, we discuss Hankel operators in §4 in both the discrete- and the continuous-time case. Concrete constructions of the continuous time restricted and *-restricted shift realizations are given in §5. They respectively represent the classes of input-normal and output-normal realizations and are intimately related to Hankel operators and translation semigroups. In §6 we establish the asymptotic stability of all parbalanced continuous-time realizations. Conditions for input-normal or output-normal realizations to be asymptotically stable are also given in terms of the cyclicity of the transfer functions. The topic of §7 is exponential stability. Necessary and sufficient conditions are given for the input- and output-normal realizations to be exponentially stable. These conditions are based on the spectral properties of the transfer functions. They also hold for parbalanced realizations as long as the transfer functions are strictly noncyclic. In §8 we investigate when the system operators are bounded, and finally some examples are given in §9.

The following symbols are used:

$\mathbb{D}$	the open unit disk,
$\partial\mathbb{D}$	the unit circle,
$\mathbb{D}_e$	the complement of $(\partial\mathbb{D}) \cup \mathbb{D}$,
$D_X^{U,Y}$	admissible discrete-time systems (§3),
$C_X^{U,Y}$	admissible continuous-time systems (§3),
$D(A) \subseteq X$	the domain of an operator A on X,
$(D(A), \|\cdot\|_A)$	the space $D(A)$ equipped with norm $\|x\|_A^2 = \|x\|^2 + \|Ax\|^2$,
$(D(A)^{(\prime)}, \|\cdot\|')$	$\{f \mid f : (D(A), \|\cdot\|_A) \to \mathbb{C}$, antilinear, bounded$\}$,
$G_d^{\perp}(z)$	$\frac{1}{z}[G_d(\frac{1}{z}) - G_d(\infty)]$, $z \in \mathbb{D}$, for $G_d \in TLD^{U,Y}$,
$G_c(+\infty)$	$\lim_{\substack{r\in\mathbb{R} \\ r\to+\infty}} G_c(r)$,
H_K	the Hankel operator with symbol K,
$H^{\infty}_{L(U,Y)}(W)$	$\{F \mid F : W \to L(U,Y)$ analytic, $\sup_{z\in W} \|F(z)\| < \infty\}$; $W = \mathbb{D}$ or RHP,
$H_Y^2(\mathbb{D})$	$\{f \mid f : \mathbb{D} \to Y$ analytic on $\mathbb{D}$ and $\sup_{0<r<1} \int_0^{2\pi} \|f(re^{it})\|^2 dt < \infty\}$,
$H_Y^2(RHP)$	$\{f \mid f : RHP \to Y$ analytic on RHP and $\sup_{x>0} \int_{-\infty}^{\infty} \|f(x+iy)\|^2 dy < \infty\}$,
$\tilde{K}(z)$	$(K(\bar{z}))^*$,
$\mathcal{L}$	the Laplace transform (§3),
$L(U,Y)$	$\{A \mid A : U \to Y$ a bounded operator$\}$,
$L_Y^2(\Delta)$	$\{f \mid f : L \to Y$ square integrable on $\Delta\}$, $\Delta = \partial\mathbb{D}$ or $i\mathbb{R}$,
LHP	the open left half plane: $\{s \in \mathbb{C} : Re(s) < 0\}$,
P_+	The orthogonal projection of $L_Y^2(\Delta)$ onto $H_Y^2(W)$; $\Delta = \partial\mathbb{D}$, $W = \mathbb{D}$, or $\Delta = i\mathbb{R}$, $W = RHP$,
P_X	the orthogonal projection of $H_Y^2(W)$ onto $X \subseteq H_Y^2(W)$; $W = \mathbb{D}$ or RHP,
RHP	the open right half plane: $\{s \in \mathbb{C} : Re(s) > 0\}$,
S	the forward shift: $(Sf)(z) = zf(z)$ for $f \in H_Y^2(\mathbb{D})$,
S^*	the backward shift: $(S^*f)(z) = z^{-1}[f(z) - f(0)]$ for $f \in H_Y^2(\mathbb{D})$,
$S(Q)$	$P_X S\|_X$, the compression of S to X, where $X = H_Y^2(\mathbb{D}) \ominus (QH_Y^2(\mathbb{D}))$,
$S(Q)^*$	$S^*\|_{H_Y^2(\mathbb{D})\ominus(QH_Y^2(\mathbb{D}))}$, the restriction of S^* to $H_Y^2(\mathbb{D}) \ominus (QH_Y^2(\mathbb{D}))$,
$\sigma(A)$	the spectrum of an operator A,
$\sigma_p(A)$	the point spectrum of an operator A,
$\sigma(Q)$	the spectrum of an inner function $Q \in H_Y^{\infty}(W)$ (Lemma 7.3),
$\sigma_s(G)$	the set of points in $\mathbb{C}$ where G has no analytic continuation (§7),
$TLD^{U,Y}$	$\{G_d \mid G_d : \mathbb{D}_e \to L(U,Y)$ has a reachable and observable admissible realization$\}$,
$TLC^{U,Y}$	$\{G_c \mid G_c : RHP \to L(U,Y)$ has a reachable and observable admissible realization$\}$,
$X \vee Y$	closed linear span of subsets X and Y of a Hilbert space,
$(F,G)_L = I_Y$	F and G are weakly left coprime (§3),
$(F,G)_R = I_U$	F and G are weakly right coprime (§3).

2. Admissible continuous-time state-space systems. The main aim of this section is to briefly set out the notation and introduce the most important system theoretic concepts for this paper. More details can be found in [11], [23], [27], and [6]. In the first subsection, admissible continuous-time systems are discussed. Input-normal, output-normal, and parbalanced

realizations are defined in the second subsection. It is these classes of systems that are being analyzed in detail in later sections. What is meant by system equivalence for infinite-dimensional systems is defined in the third subsection.

2.1. Admissible continuous-time systems. It is well known that if A is the generator of a strongly continuous semigroup of operators $(e^{tA})_{t\geq 0}$ with domain of definition $D(A)$, then $D(A)$ is a Hilbert space with inner product induced by the graph norm

$$\|x\|_A^2 := \|x\|_X^2 + \|Ax\|_X^2, \quad x \in D(A).$$

Since $\|x\|_A \geq \|x\|$ for $x \in D(A)$, we can embed X in $D(A)^{(\prime)}$, the set of antilinear continuous functionals on $(D(A), \|\cdot\|_A)$, by

$$\begin{aligned} E: \quad & X \to D(A)^{(\prime)}, \\ & x \longmapsto (y \longmapsto \langle x, y\rangle). \end{aligned}$$

Note that $D(A)^{(\prime)}$ is a Hilbert space with norm $\|f\|' := \sup_{\|x\|_A\leq 1} |f(x)|$. Since $\langle\cdot,\cdot\rangle$ is linear in the first component, the embedding E is linear. By the above, we have the rigged structure

$$D(A) \subseteq X \subseteq D(A)^{(\prime)}.$$

If $(A, D(A))$ is the generator of a strongly continuous semigroup of contractions $(e^{tA})_{t\geq 0}$ on a Hilbert space, then the adjoint $(A^*, D(A^*))$ of $(A, D(A))$ is the generator of the adjoint semigroup $(e^{tA})_{t\geq 0}^*$ (see [26]). Hence, we have similarly that

$$D(A^*) \subseteq X \subseteq D(A^*)^{(\prime)}.$$

We are now in a position to define admissible continuous-time systems.

DEFINITION 2.1. *A quadruple of operators (A_c, B_c, C_c, D_c) is called an* admissible continuous-time system *with state space X, input space U, and output space Y, where X, U, and Y are separable Hilbert spaces, if*

1. *$(A_c, D(A_c))$ is the generator of a strongly continuous semigroup of contractions on X;*
2. *$B_c : U \to (D(A_c^*)^{(\prime)}, \|\cdot\|')$ is a bounded linear operator;*
3. *$C_c : D(C_c) \to Y$ is linear with $D(C_c) = D(A_c) + (I - A_c)^{-1}B_cU$ and $C_{c|D(A_c)} : (D(A_c), \|\cdot\|_{A_c}) \to Y$ is bounded;*
4. *$C_c(I - A_c)^{-1}B_c \in L(U, Y)$;*
5. *A_c, B_c, and C_c are such that $\lim_{\substack{s\in\mathbb{R}\\ s\to+\infty}} C_c(sI - A_c)^{-1}B_c = 0$ in the norm topology;*
6. *$D_c \in L(U, Y)$.*

We write $C_X^{U,Y}$ for the set of admissible continuous-time systems with input space U, output space Y, and state space X. □

By the resolvent identity, part 4 of the definition implies that $G_c(s) := C_c(sI - A_c)^{-1}B_c \in L(U, Y)$ for all $s \in RHP$ and G_c is analytic on the RHP. The function G_c is called the transfer function of the system, and (A_c, B_c, C_c, D_c) is called a realization of G_c.

2.2. Duality, observability, reachability, and parbalanced realizations. In order to define observability and reachability for continuous-time systems we need to introduce the notion of the dual system of an admissible continuous-time system.

DEFINITION 2.2. *Let $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$. Then the* dual system $(\tilde{A_c}, \tilde{B_c}, \tilde{C_c}, \tilde{D_c})$ *of (A_c, B_c, C_c, D_c) is given by*

1. *$(\tilde{A_c}, D(\tilde{A_c})) = (A_c^*, D(A_c^*))$, the adjoint operator of $(A_c, D(A_c))$;*
2. *$\tilde{B_c} : Y \to D(A_c)^{(\prime)}; y \longmapsto \tilde{B_c}(y)[\cdot] := \langle y, C_c(\cdot)\rangle$;*

3. $\tilde{C}_c : D(\tilde{C}_c) \to U$, $D(\tilde{C}_c) = D(\tilde{A}_c) + (I - \tilde{A}_c)^{-1}\tilde{B}_c Y$, *where* $\tilde{C}_c x_0$ *is defined by*

$$\begin{cases} \langle u, \tilde{C}_c x_0\rangle = B_c(u)[x_0], & x_0 \in D(A_c^*), u \in U, \\ \langle \tilde{C}_c x_0, u\rangle = \langle y_0, C_c(I - A_c)^{-1}B_c u\rangle, & x_0 = (I - \tilde{A}_c)^{-1}\tilde{B}_c y_0,\ y_0 \in Y, u \in U; \end{cases}$$

4. $\tilde{D}_c := D_c^* : Y \to U$. □

It can be directly verified that the dual system $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ of an admissible continuous-time system (A_c, B_c, C_c, D_c) is admissible. If the continuous-time transfer function $G(s) : RHP \to L(U, Y)$ has an admissible realization (A_c, B_c, C_c, D_c), then the dual system $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ is a realization of the transfer function $\tilde{G}(s) := (G(\bar{s}))^*$, $s \in RHP$, i.e., for all $s \in RHP$,

$$\tilde{G}(s) = (G(\bar{s}))^* = \tilde{C}_c(sI - \tilde{A}_c)^{-1}\tilde{B}_c + \tilde{D}_c.$$

The definition of observability and reachability of admissible continuous-time systems is now given.

DEFINITION 2.3. *Let* $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$; *then the operator*

$$\begin{aligned} \mathcal{O}_c : D(\mathcal{O}_c) &\to L_Y^2([0, \infty)), \\ x &\longmapsto (C_c e^{tA_c} x)_{t \geq 0} \end{aligned}$$

is called the observability operator *of the system* (A_c, B_c, C_c, D_c), *where*

$$D(\mathcal{O}_c) = \{x \in X \mid C_c e^{tA_c} x \text{ exists for almost all } t \in [0, \infty), \text{ and } C_c e^{tA_c} x \in L_Y^2([0, \infty))\}.$$

We say that (A_c, B_c, C_c, D_c) *has a bounded observability operator if* $D(A_c) \subseteq D(\mathcal{O}_c)$ *and* $\mathcal{O}_c$ *extends to a bounded operator on* X. *This extension will also be denoted by* $\mathcal{O}_c$.

If (A_c, B_c, C_c, D_c) *has a bounded observability operator* $\mathcal{O}_c$ *such that* $\operatorname{Ker}(\mathcal{O}_c) = \{0\}$, *then the system* (A_c, B_c, C_c, D_c) *is called* observable.

Let $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ *be the dual system of* (A_c, B_c, C_c, D_c). *If the observability operator* $\tilde{\mathcal{O}}_c$ *of* $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ *is a bounded operator on* X, *the adjoint of* $\tilde{\mathcal{O}}_c$ *is called the* reachability operator, *denoted by* $\mathcal{R}_c$, *of* (A_c, B_c, C_c, D_c), *i.e.,*

$$\mathcal{R}_c := \tilde{\mathcal{O}}_c^*.$$

If $\mathcal{R}_c$ *exists and* range$(\mathcal{R}_c)$ *is dense in* X, *the system* (A_c, B_c, C_c, D_c) *is said to be* reachable. □

The set of all reachable and observable continuous-time systems with input space U, output space Y, and state space X is denoted by $LC_X^{U,Y}$. We mainly deal with this set of systems.

The *reachability Gramian* $\mathcal{W}_c$ and the *observability Gramian* $\mathcal{M}_c$ of a continuous-time system with bounded reachability operator $\mathcal{R}_c$ and bounded observability operator $\mathcal{O}_c$ are defined to be

$$\begin{aligned} \mathcal{W}_c &:= \mathcal{R}_c\mathcal{R}_c^* : X \to X, \\ \mathcal{M}_c &:= \mathcal{O}_c^*\mathcal{O}_c : X \to X. \end{aligned}$$

When $\mathcal{W}_c = \mathcal{M}_c$ and the admissible system is observable and reachable, we say that the system is *parbalanced.* A reachable and observable admissible system is said to be *balanced* if $\mathcal{W}_c = \mathcal{M}_c$ and $\mathcal{W}_c$ has a diagonal representation with respect to an orthonormal basis of the state space. If $\mathcal{W}_c = I$, then a reachable and observable admissible system is called *input-normal.* If $\mathcal{M}_c = I$, then a reachable and observable admissible system is called *output-normal.*

2.3. System equivalence. The concept of an equivalent state-space transformation of an admissible continuous-time system is slightly more complicated than in the discrete time case as the system operators are in general unbounded.

Two systems $(A_c^i, B_c^i, C_c^i, D_c^i) \in C_{X_i}^{U,Y}$, $i = 1, 2$, are called *equivalent* if there exists a boundedly invertible operator $V \in L(X_1, X_2)$ such that

$$\begin{aligned}&\left((A_c^2, D(A_c^2)), B_c^2, (C_c^2, D(C_c^2)), D_c^2\right) =\\ &\left((VA_c^1V^{-1}, VD(A_c^1)), VB_c^1, (C_c^1V^{-1}, VD(C_c^1)), D_c^1\right),\end{aligned}$$

where

$$B_c^2 = (VB_c^1) : U \to \left(\left(D(A_c^{2*})\right)^{(\prime)}, \|\cdot\|'\right)$$

is given by

$$[B_c^2(u)](x) = (VB_c^1)(u)[x] := B_c^1(u)[V^*x], \quad u \in U, \ x \in D(A_c^{2*}) = (V^*)^{-1}D(A_c^{1*}).$$

If V is a unitary operator, then the two systems are said to be *unitarily equivalent.*

We have the following results concerning equivalent systems.

PROPOSITION 2.4. *Let* $(A_c^i, B_c^i, C_c^i, D_c^i) \in C_{X_i}^{U,Y}$, $i = 1, 2$, *be two equivalent systems such that*

$$\begin{aligned}&\left((A_c^2, D(A_c^2)), B_c^2, (C_c^2, D(C_c^2)), D_c^2\right) =\\ &\left((VA_c^1V^{-1}, VD(A_c^1)), VB_c^1, (C_c^1V^{-1}, VD(C_c^1)), D_c^1\right)\end{aligned}$$

with $V \in L(X_1, X_2)$ *a boundedly invertible operator. Then*

1. *both* $(A_c^1, B_c^1, C_c^1, D_c^1)$ *and* $(A_c^2, B_c^2, C_c^2, D_c^2)$ *realize the same transfer function.*
2. *if* $(A_c^1, B_c^1, C_c^1, D_c^1) \in C_{X_1}^{U,Y}$ *has observability operator* $\mathcal{O}$ *and reachability operator* $\mathcal{R}$, *then the observability and reachability operators of* $(A_c^2, B_c^2, C_c^2, D_c^2) \in C_{X_2}^{U,Y}$ *are respectively*

$$\mathcal{O}V^{-1} \quad \text{and} \quad V\mathcal{R}.$$

Proof. The proof is straightforward. □

Thus equivalent systems have the same transfer function as well as the same observability and reachability properties. Moreover, it can be seen that unitary equivalent systems have the same Gramians. Hence unitary equivalence preserves parbalancing.

We point out that for an admissible system $((A_c^1, D(A_c^1)), B_c^1, C_c^1, D_c^1) \in C_{X_1}^{U,Y}$ and a unitary operator $V : X_1 \to X_2$, the system

$$\begin{aligned}&\left((A_c^2, D(A_c^2)), B_c^2, (C_c^2, D(C_c^2)), D_c^2\right) =\\ &\left((VA_c^1V^{-1}, VD(A_c^1)), VB_c^1, (C_c^1V^{-1}, VD(C_c^1)), D_c^1\right)\end{aligned}$$

is also admissible, where VB_c^1 is defined as above. Therefore $((A_c^1, D(A_c^1)), B_c^1, C_c^1, D_c^1)$ and $(A_c^2, B_c^2, C_c^2, D_c^2)$ are unitarily equivalent.

The class of continuous-time transfer functions that we are interested in are those that have reachable and observable continuous time realizations on some state space X, where X is a separable Hilbert space. This class will be denoted by $TLC^{U,Y}$, where U and Y are the input and output spaces, respectively. We characterize those transfer functions in terms of their Hankel operators in §4.

3. Connection between continuous- and discrete-time systems. What is essential in our development is to relate discrete-time systems to continuous-time systems using a generalization of the well-known bilinear transformation for finite-dimensional systems. Thereby it is possible to carry some of the results in [24] for discrete-time systems over to continuous-time systems. It should be noted, however, that not all results of discrete-time systems can be translated to the continuous-time case in this way. For example, under this bilinear map an exponentially stable continuous-time system does not necessarily correspond to a power stable discrete-time system.

3.1. Admissible discrete-time systems. We recall [24] that an *admissible discrete-time system* with input space U, output space Y, and state space X, with U, X, and Y being separable Hilbert spaces, is a quadruple of operators (A_d, B_d, C_d, D_d) that satisfy the following:

1. $A_d \in L(X)$ is a contraction and $-1 \notin \sigma_p(A_d)$;
2. $B_d \in L(U, X)$, $C_d \in L(X, Y)$ and $D_d \in L(U, Y)$;
3. the limit $\lim_{r>1, r\to 1} C_d(rI + A_d)^{-1}B_d$ exists in the norm topology.

The set of all such systems is denoted by $D_X^{U,Y}$. For $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$, the function

$$G_d(z) = C_d(zI - A_d)^{-1}B_d + D_d : \mathbb{D}_e \to L(U, Y)$$

is called the *transfer function* of (A_d, B_d, C_d, D_d) and (A_d, B_d, C_d, D_d) is called a *realization* of G_d. Evidently, the transfer function G_d is analytic on $\mathbb{D}_e$ and at infinity.

For $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$, its *observability operator* $\mathcal{O}_d : D(\mathcal{O}_d) \to H_Y^2$ is defined as

$$(\mathcal{O}_d x)(z) = \sum_{n\geq 0}(C_d A_d^n x)z^n, \qquad x \in D(\mathcal{O}_d) := \left\{ x \mid \sum_{n\geq 0}(C_d A_d^n x)z^n \in H_Y^2 \right\}.$$

If $D(\mathcal{O}_d) = X$, $\mathcal{O}_d$ is bounded and $\operatorname{Ker}(\mathcal{O}_d) = \{0\}$, then the system (A_d, B_d, C_d, D_d) is said to be *observable*. The system (A_d, B_d, C_d, D_d) is said to be *reachable* if its *reachability operator* $\mathcal{R}_d : D(\mathcal{R}_d) \to X$ defined by

$$\mathcal{R}_d\left(\sum_{n\geq 0} u_n z^n\right) = \sum_{n\geq 0} A_d^n B_d u_n \quad \left(\sum_{n\geq 0} u_n z^n \in D(\mathcal{R}_d)\right),$$

where $D(\mathcal{R}_d) = \{\sum_{n=0}^{N} u_n z^n \mid N = 0, 1, \ldots, u_n \in U\}$ can be extended to a bounded operator with range dense in X. The set of all reachable and observable discrete-time admissible systems with input space U, output space Y, and state space X is denoted by $LD_X^{U,Y}$. The set of all discrete-time transfer functions that have realizations $(A_d, B_d, C_d, D_d) \in LD_X^{U,Y}$ for some state space X is denoted by $TLD^{U,Y}$. A characterization will be given of this class of transfer functions in the next section.

For $(A_d, B_d, C_d, D_d) \in LD_X^{U,Y}$, we define its *reachability Gramian* $\mathcal{W}_d : X \to X$ as

$$\mathcal{W}_d x = \mathcal{R}_d \mathcal{R}_d^* x, \quad x \in X,$$

and its *observability Gramian* $\mathcal{M}_d : X \to X$ as

$$\mathcal{M}_d x = \mathcal{O}_d^* \mathcal{O}_d x, \quad x \in X.$$

If $\mathcal{W}_d = \mathcal{M}_d$ and (A_d, B_d, C_d, D_d) is reachable and observable, then (A_d, B_d, C_d, D_d) is said to be a *parbalanced* realization. If the Gramian of a parbalanced realization has a diagonal representation with respect to an orthonormal basis, the realization is said to be *balanced*. If $\mathcal{W}_d = I$, then the reachable and observable admissible system is called *input-normal*. If $\mathcal{M}_d = I$, then the reachable and observable admissible system is called *output-normal*.

3.2. Bilinear transform. In the following theorems (see [23]) we introduce the map $T : D_X^{U,Y} \to C_X^{U,Y}$, which transforms discrete-time systems to continuous-time systems. Throughout the rest of this paper T will denote this map.

THEOREM 3.1. *Let* $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$; *then* $T((A_d, B_d, C_d, D_d)) := (A_c, B_c, C_c, D_c) \in C_X^{U,Y}$, *where the operators* A_c, B_c, C_c, *and* D_c *are defined as follows:*

1. $A_c := (I + A_d)^{-1}(A_d - I) = (A_d - I)(I + A_d)^{-1}$, $D(A_c) := D((I + A_d)^{-1})$. *It generates a strongly continuous semigroup of contractions on* X *given by* $\varphi_t(A_d)$, $t \geq 0$, *with* $\varphi_t(z) = e^{t\frac{z-1}{z+1}}$.

2. *The operator B_c is given by*

$$\begin{aligned} B_c := \sqrt{2}(I + A_d)^{-1}B_d : U &\to D(A_c^*)^{(\prime)}, \\ u &\longmapsto \sqrt{2}(I + A_d)^{-1}B_d(u)[x] \\ &:= \sqrt{2} < B_d(u), (I + A_d^*)^{-1}(x) >_X . \end{aligned}$$

3. *The operator C_c is given by*

$$\begin{aligned} C_c : D(C_c) &\to Y, \\ x &\longmapsto \lim_{\substack{\lambda \to 1 \\ \lambda > 1}} \sqrt{2}C_d(\lambda I + A_d)^{-1}x, \end{aligned}$$

where $D(C_c) = D(A_c) + (I - A_c)^{-1}B_cU$. On $D(A_c)$ we have

$$C_{c|D(A_c)} = \sqrt{2}C_d(I + A_d)^{-1}.$$

4. $D_c := D_d - \lim_{\substack{\lambda \to 1 \\ \lambda > 1}} C_d(\lambda I + A_d)^{-1}B_d$.

Moreover, let the admissible discrete-time system (A_d, B_d, C_d, D_d) be a realization of the transfer function

$$G_d : \mathbb{D}_e \to L(U, Y) ,$$

i.e., $G_d(z) = C_d(zI - A_d)^{-1}B_d + D_d$ for $z \in \mathbb{D}_e$. Then

$$(A_c, B_c, C_c, D_c) = T((A_d, B_d, C_d, D_d))$$

is an admissible continuous-time realization of the transfer function

$$G_c(s) := G_d\left(\frac{1+s}{1-s}\right) : RHP \to L(U, Y). \qquad \square$$

The inverse map is considered in the next theorem [23].

THEOREM 3.2. *Let $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$; then $T^{-1}((A_c, B_c, C_c, D_c)) := (A_d, B_d, C_d, D_d) \in D_X^{U,Y}$, where the operators A_d, B_d, C_d, and D_d are defined as follows:*

1. $A_d := (I+A_c)(I-A_c)^{-1}$, *and for $x \in D(A_c)$ we have $A_dx = (I-A_c)^{-1}(I+A_c)x$.*
2. $B_d := \sqrt{2}(I - A_c)^{-1}B_c$.
3. $C_d := \sqrt{2}C_c(I - A_c)^{-1}$.
4. $D_d := C_c(I - A_c)^{-1}B_c + D_c$.

Moreover, let the admissible continuous time system (A_c, B_c, C_c, D_c) be a realization of the transfer function

$$G_c : RHP \to L(U, Y) ,$$

i.e., $G_c(s) = C_c(sI - A_c)^{-1}B_c + D_c$ for $s \in RHP$. Then

$$(A_d, B_d, C_d, D_d) = T^{-1}((A_c, B_c, C_c, D_c))$$

is an admissible discrete-time realization of the transfer function

$$G_d(z) := G_c\left(\frac{z-1}{z+1}\right) : \mathbb{D}_e \to L(U, Y). \qquad \square$$

We recall that two discrete-time systems $(A_{di}, B_{di}, C_{di}, D_{di}) \in D_{X_i}^{U,Y}$ $(i = 1, 2)$ are equivalent (unitarily equivalent) if there is a bounded operator (a unitary operator) V from X_1 onto X_2 such that

$$(A_{d1}, B_{d1}, C_{d1}, D_{d1}) = (VA_{d2}V^{-1}, VB_{d2}, C_{d2}V^{-1}, D_{d2}).$$

In [23] it was shown that T preserves (unitary) equivalence of systems and respects duality of systems.

Note that in the previous two theorems the state spaces for the continuous- and discrete-time realizations are the same. As will be seen in later sections for continuous-time systems it is more natural to work on a different yet unitarily equivalent state space that is a subspace of $H_Y^2(RHP)$. Here we point out the equivalence of the Hilbert spaces $H_Y^2(\mathbb{D})$ and $H_Y^2(RHP)$, where Y is a separable Hilbert space (see [25, Thm. 4.6]).

PROPOSITION 3.3. *The spaces $H_Y^2(\mathbb{D})$ and $H_Y^2(RHP)$ are unitarily equivalent by the map*

$$\begin{aligned} V_Y : \quad H_Y^2(\mathbb{D}) &\to H_Y^2(RHP), \\ f_d &\mapsto (V_Y f_d)(\bullet) := f_c(\bullet) := \frac{1}{\sqrt{\pi}(1+\bullet)} f_d\left(\frac{1-\bullet}{1+\bullet}\right). \end{aligned}$$

The inverse of V is given by

$$\begin{aligned} V_Y^{-1} : \quad H_Y^2(RHP) &\to H_Y^2(\mathbb{D}) \\ f_c &\mapsto (V_Y^{-1} f_c)(\bullet) := f_d(\bullet) := \frac{2\sqrt{\pi}}{(1+\bullet)} f_c\left(\frac{1-\bullet}{1+\bullet}\right). \end{aligned}$$

□

The next result shows that observability and reachability properties as well as the Gramians are preserved under T. This implies that the transformation preserves parbalancing of systems. This result is the translation of a result in [23] to the frequency domain.

THEOREM 3.4. *Let $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$ and $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$ be such that*

$$(A_c, B_c, C_c, D_c) = T((A_d, B_d, C_d, D_d)).$$

Then

1. *(A_c, B_c, C_c, D_c) is observable (reachable) if and only if (A_d, B_d, C_d, D_d) is observable (reachable). In fact, if $\mathcal{O}_c$ ($\mathcal{R}_c$) and $\mathcal{O}_d$ ($\mathcal{R}_d$) are the observability (reachability) operators of (A_c, B_c, C_c, D_c) and (A_d, B_d, C_d, D_d), respectively, and if either (A_c, B_c, C_c, D_c) or (A_d, B_d, C_d, D_d) has a bounded observability (reachability) operator, then the following relations hold:*

$$V_1 \mathcal{O}_d x = \mathcal{L}\mathcal{O}_c x, \quad x \in X \quad (\mathcal{R}_d V_2^{-1} u = \mathcal{R}_c \mathcal{L}^{-1} u, \quad u \in H_U^2(RHP))$$

where $V_1 : H_Y^2(\mathbb{D}) \to H_Y^2(RHP)$ and $V_2 : H_U^2(\mathbb{D}) \to H_U^2(RHP)$ are unitary transformations as defined in Proposition 3.3:

$$V_i f_i = \frac{1}{\sqrt{\pi}(1+s)} f_i\left(\frac{1-s}{1+s}\right), \quad f_1 \in H_Y^2(\mathbb{D}), \text{ and } f_2 \in H_U^2(\mathbb{D}), \quad i = 1, 2,$$

and $\mathcal{L}$ is the Laplace transform.

2. *If the reachability Gramians $\mathcal{W}_c$ and $\mathcal{W}_d$ (observability Gramians $\mathcal{M}_c$ and $\mathcal{M}_d$) of (A_d, B_d, C_d, D_d) and (A_c, B_c, C_c, D_c) are defined, then*

$$\mathcal{W}_c = \mathcal{W}_d \qquad (\mathcal{M}_c = \mathcal{M}_d).$$

Proof. In [23] a "time domain" version of this result was proven. The present result follows from the result in [23] by applying the z-transform (respectively, Laplace transform) and using the unitary transformation of Proposition 3.3. □

Therefore if $G_d(z) = G_c(\frac{z-1}{z+1})$, then $G_d \in TLD^{U,Y}$, i.e., G_d has a reachable and observable discrete-time admissible realization, if and only if $G_c \in TLC^{U,Y}$, i.e., if and only if G_c has a reachable and observable continuous-time admissible realization.

The combination of Theorems 3.1, 3.2, and 3.4 gives us an effective machinery to transform discrete-time results to the continuous-time case. Before doing this, we need to study Hankel operators which will be important in the analysis of parbalanced, input-normal, or output-normal realizations treated in the sequel.

4. Linear systems and Hankel operators. In the study of discrete-time systems Hankel operators on $H^2(\mathbb{D})$ play an important role [11]. Given a discrete-time transfer function, a Hankel operator can be associated with it in a natural way. The so-called *restricted shift realization* of the transfer function is constructed by using the range of the Hankel operator as its state space (see [11], [30], [24], and §5 below). When the Hankel operator is compact, a balanced realization can be obtained whose Gramians have diagonal representations with diagonal entries equal to the singular values of the Hankel operator [30]. In the continuous-time situation Hankel operators on $H^2(RHP)$ will be of equal importance. We therefore examine the relationship between discrete-time Hankel operators and their continuous-time counterparts.

4.1. Hankel operators and realizability. Let G_d be analytic on $\mathbb{D}_e$ and at infinity so that $G_d^\perp(z) = z^{-1}[G_d(z^{-1}) - G_d(\infty)]$ is analytic on $\mathbb{D}$. We define the operator $H_{G_d^\perp,\mathbb{D}} : D(H_{G_d^\perp,\mathbb{D}}) \to H_Y^2(\mathbb{D})$ by

$$(H_{G_d^\perp,\mathbb{D}} f)(z) = P_+ G_d^\perp J f \quad (f \in D(H_{G_d^\perp})),$$

where $D(H_{G_d^\perp,\mathbb{D}}) = \{f \in H_U^2(\mathbb{D}) : f \text{ polynomial, } G_d^\perp Jf \text{ has nontangential limit in } \mathbb{D} \text{ almost everywhere (a.e.) at } \partial\mathbb{D} \text{ with limit in } L_Y^2(\partial\mathbb{D})\}$ and $(Jf)(z) = f(1/z)$. The operator $H_{G_d^\perp}$ is called the Hankel operator with symbol $G_d^\perp$. If $D(H_{G_d^\perp,\mathbb{D}})$ is dense in $H_U^2(\mathbb{D})$ and $H_{G_d^\perp,\mathbb{D}}$ extends to a bounded operator on $H_U^2(\mathbb{D})$, this extension is also called the Hankel operator with symbol $G_d^\perp$ and is denoted by $H_{G_d^\perp,\mathbb{D}}$.

The following lemma [24] relates the existence of a reachable and observable realization of a discrete-time transfer function G_d to the boundedness of the Hankel operator $H_{G_d^\perp,\mathbb{D}}$.

LEMMA 4.1. *The transfer function G_d is in $TLD^{U,Y}$; i.e., G_d has an admissible reachable and observable realization if and only if* (i) *G_d is analytic on $\mathbb{D}_e$ and at infinity,* (ii) *the limit $\lim_{r<-1,\ r\to-1,\ r\in\mathbb{R}} G_d(r)$ exists in the norm topology, and* (iii) *the Hankel operator $H_{G_d^\perp,\mathbb{D}}$ is bounded.* □

We analogously define Hankel operators for continuous-time transfer functions.

DEFINITION 4.2. *If G_c is an $L(U,Y)$-valued function analytic on RHP, then the operator*

$$\begin{aligned} H_{G_c,RHP} : \quad D(H_{G_c,RHP}) &\to H_Y^2(RHP), \\ f &\mapsto P_+ M_{G_c} R f, \end{aligned}$$

where

$$Rf(s) = f(-s),$$

M_{G_c} *is the multiplication operator by* G_c,

P_+ *is the projection on* $H_Y^2(RHP)$,

with $D(H_{G_c,RHP}) = \{f \in H_U^2(RHP) : f$ *rational,* $G_c Rf$ *has a nontangential limit in* RHP *on a.e. on* $i\mathbb{R}$ *that is in* $L_Y^2(i\mathbb{R})\}$, *is called the* Hankel operator $H_{G_c,RHP}$ with symbol G_c. □

If $D(H_{G_c,RHP})$ is dense in $H_U^2(RHP)$ and $H_{G_c,RHP}$ extends to a bounded operator on $H_U^2(RHP)$, this extension is also called the Hankel operator with symbol G_c and is denoted by $H_{G_c,RHP}$.

If it is clear from the context that the Hankel operator is defined on RHP, we will drop the subscript RHP and write H_G instead of $H_{G,RHP}$.

It is important in our context that Hankel operators defined on the disk are unitarily equivalent to Hankel operators in the right half plane in the following way (see, e.g., [25, Thm. 4.6]).

PROPOSITION 4.3. *Let* V_U *and* V_Y *be the unitary operators defined in Proposition* 3.3.

1. *Let* $G_d \in TLD^{U,Y}$ *and* $G_c \in TLC^{U,Y}$. *If*

$$G_d(z) = G_c\left(\frac{z-1}{z+1}\right) \quad \text{for } z \in \mathbb{D}_e,$$

or equivalently

$$G_c(s) = G_d\left(\frac{1+s}{1-s}\right) \quad \text{for } s \in RHP,$$

then the Hankel operators $H_{G_d^\perp,\mathbb{D}}$ *and* $H_{G_c,RHP}$ *are unitarily equivalent, i.e.,*

$$H_{G_c,RHP} = V_Y H_{G_d^\perp,\mathbb{D}} V_U^{-1},$$

where $G_d^\perp(z) = z^{-1}[G_d(z^{-1}) - G_d(\infty)]$ $(z \in \mathbb{D})$.

2. *Let* $K_d \in H_{L(U,Y)}^\infty(\mathbb{D})$ *and* $K_c \in H_{L(U,Y)}^\infty(RHP)$ *be such that*

$$K_d(z) = K_c\left(\frac{1-z}{1+z}\right), \qquad z \in \mathbb{D},$$

or equivalently

$$K_c(s) = K_d\left(\frac{1-s}{1+s}\right), \qquad s \in RHP.$$

Then

$$V_Y(K_d H_U^2(\mathbb{D})) = K_c H_U^2(RHP), \quad V_Y\left((K_d H_U^2(\mathbb{D}))^\perp\right) = (K_c H_U^2(RHP))^\perp$$

and

$$V_Y^{-1}(K_c H_U^2(RHP)) = K_d H_U^2(\mathbb{D}), \quad V_Y^{-1}\left((K_c H_U^2(RHP))^\perp\right) = (K_d H_U^2(\mathbb{D}))^\perp.$$

Proof. The proposition follows from direct verification. □

Using this proposition we can give a characterization for a continuous-time transfer function G_c to be in $TLC^{U,Y}$, i.e., to have an observable, reachable, and admissible continuous-time realization.

COROLLARY 4.4. *The following two statements are equivalent.*

1. $G_c \in TLC^{U,Y}$, *that is,* G_c *has a reachable, observable, and admissible realization on some Hilbert space.*

2. $G_c(s)$ *is analytic on* RHP, *the limit* $\lim_{r\in\mathbb{R},r\to+\infty} G_c(r)$ *exists in the norm topology, and the Hankel operator* $H_{G_c} : H_U^2(RHP) \to H_Y^2(RHP)$ *is bounded.*

Proof. This follows from Theorem 3.4, Proposition 4.3, and Lemma 4.1. □

4.2. Range spaces of Hankel operators and factorizations of transfer functions. It is known that the orthogonal complement,

$$(\text{range}H_{G_c})^{\perp} = H_Y^2(RHP) \ominus \overline{\text{range}}H_{G_c},$$

of the range of the Hankel operator H_{G_c} is invariant under any multiplication operator with symbol in H_Y^∞. Hence by Beurling's theorem, the subspace $(\text{range}H_{G_c})^{\perp}$ is either $\{0\}$ or $QH_Y^2(RHP)$, where $Q \in H_Y^\infty(RHP)$ is a rigid function. A rigid function is a function $Q \neq 0$ such that $Q(iy)$ is for a.e. $y \in \mathbb{R}$ a partial isometry with a fixed initial space (see, e.g., [11, p. 186], and [15]). In particular, inner functions are rigid functions.

Using the above-defined notions, we introduce the concept of cyclicity of continuous-time transfer functions, which relates Hankel operators with their symbols. The discrete-time case was studied in, e.g., Fuhrmann [11]. A general study of strictly noncyclic transfer functions can also be found in Dewilde [7].

DEFINITION 4.5. *Let $G_c \in H^\infty_{L(U,Y)}(RHP)$. Then G_c is called*

1. cyclic *if* $(\text{range}H_{G_c,RHP})^{\perp} = \{0\}$;
2. noncyclic *if* $(\text{range}H_{G_c,RHP})^{\perp} = QH_Y^2(RHP)$, *where* $Q \in H_Y^\infty(RHP)$ *is a rigid function*;
3. strictly noncyclic *if* $(\text{range}H_{G_c,RHP})^{\perp} = QH_Y^2(RHP)$, *where* $Q \in H_Y^\infty(RHP)$ *is an inner function.* □

Evidently in the scalar case G_c is strictly noncyclic if and only if it is noncyclic.

In the sequel it will be seen that the cyclicity of the transfer functions has much to do with the stability and other properties of their realizations. Here we present more information on cyclicity of H^∞ transfer functions.

DEFINITION 4.6. *Let G be in $H^\infty_{L(U,Y)}(RHP)$. Then the $L(U, Y)$-valued function $\hat{G}$ defined on LHP is called a* meromorphic pseudocontinuation of bounded type *of G if*

1. *$\hat{G}$ is of bounded type, i.e.,*

$$\hat{G} = \frac{F}{h},$$

where F is a $L(U, Y)$-valued function and h is a scalar-valued function and both functions are bounded and analytic in LHP.

2. *G and $\hat{G}$ have the same strong radial limits on $i\mathbb{R}$, i.e., for a.e. $y \in \mathbb{R}$*

$$\lim_{x<0,x\to 0} \hat{G}(x+iy) = \lim_{x>0,x\to 0} G(x+iy). \qquad \square$$

The following proposition summarizes the connection between discrete- and continuous-time transfer functions in terms of cyclicity, meromorphic pseudocontinuation of bounded type, and factorizations. We refer to [11] for a discussion of these concepts for discrete-time transfer functions, which are analogous to those that have been defined here for continuous-time transfer functions.

PROPOSITION 4.7. *Let $G_c \in TLC^{U,Y}$, $G_d \in TLD^{U,Y}$, and set $G_d^{\perp}(z) = z^{-1}[G_d(z^{-1}) - G_d(\infty)]$. Assume that*

$$G_d(z) = G_c\left(\frac{z-1}{z+1}\right) \quad (z \in \mathbb{D}_e),$$

or equivalently

$$G_c(s) = G_d\left(\frac{1+s}{1-s}\right) \quad (s \in RHP).$$

Then

1. *$G_d^{\perp}$ is strictly noncyclic (cyclic, noncyclic) if and only if G_c is strictly noncyclic (cyclic, noncyclic).*

2. *Let $Q_{d,1} \in H^{\infty}_{L(Y)}(\mathbb{D})$, $Q_{d,2} \in H^{\infty}_{L(U)}(\mathbb{D})$, $Q_{c,1} \in H^{\infty}_{L(Y)}(RHP)$, and $Q_{c,2} \in H^{\infty}_{L(U)}(RHP)$ be inner functions. Let $F_{d,1} \in H^{\infty}_{L(Y,U)}(\mathbb{D})$, $F_{d,2} \in H^{\infty}_{L(U,Y)}(\mathbb{D})$, $F_{c,1} \in H^{\infty}_{L(Y,Y)}(RHP)$, and $F_{c,2} \in H^{\infty}_{L(U,Y)}(RHP)$. Assume*

$$F_{d,1}(z) = F_{c,1}\left(\frac{1-z}{1+z}\right) - G_c(1)^* Q_{c,1}\left(\frac{1-z}{1+z}\right), \quad z \in \mathbb{D},$$

$$F_{d,2}(z) = F_{c,2}\left(\frac{1-z}{1+z}\right) - Q_{c,2}\left(\frac{1-z}{1+z}\right) G_c(1)^*, \quad z \in \mathbb{D},$$

$$Q_{d,i}(z) = Q_{c,i}\left(\frac{1-z}{1+z}\right), \quad z \in \mathbb{D}, \quad i = 1, 2,$$

or equivalently

$$F_{c,1}(s) = F_{d,1}\left(\frac{1-s}{1+s}\right) + G_d(\infty)^* Q_{d,1}\left(\frac{1-s}{1+s}\right), \quad s \in RHP,$$

$$F_{c,2}(s) = F_{d,2}\left(\frac{1-s}{1+s}\right) + Q_{d,2}\left(\frac{1-s}{1+s}\right) G_d(\infty)^*, \quad s \in RHP,$$

$$Q_{c,i}(s) = Q_{d,i}\left(\frac{1-s}{1+s}\right), \quad s \in RHP, \quad i = 1, 2.$$

Then G_c can be factored on $i\mathbb{R}$ as

$$G_c = Q_{c,1} F_{c,1}^* = F_{c,2}^* Q_{c,2}$$

if and only if $G_d^{\perp}$ can be factored on $\partial\mathbb{D}$ as

$$G_d^{\perp}(z) = Q_{d,1}(z)(zF_{d,1}(z))^* = (zF_{d,2}(z))^* Q_{d,2}(z) \quad (z \in \partial\mathbb{D}).$$

3. *Assume that $G_c \in H^{\infty}_{L(U,Y)}(RHP)$ and $G_d^{\perp} \in H^{\infty}_{L(U,Y)}(\mathbb{D})$. Let F_d (F_c) be a $L(U, Y)$-valued analytic function in $\mathbb{D}_e$ (LHP) and h_d (h_c) be a scalar-valued analytic function in $\mathbb{D}_e$ (LHP), both bounded, such that*

$$F_d(z) = \frac{1}{z}\left[F_c\left(\frac{1-z}{1+z}\right) - G_c(1) h_c\left(\frac{1-z}{1+z}\right)\right], \quad z \in \mathbb{D}_e,$$

$$h_d(z) = h_c\left(\frac{1-z}{1+z}\right), \quad z \in \mathbb{D}_e,$$

or equivalently

$$F_c(s) = \frac{1-s}{1+s} F_d\left(\frac{1-s}{1+s}\right) + G_d(\infty) h_d\left(\frac{1-s}{1+s}\right), \quad s \in LHP,$$

$$h_c(s) = h_d\left(\frac{1-s}{1+s}\right), \quad s \in LHP.$$

Then $G_d^{\perp}$ has a meromorphic pseudocontinuation $\hat{G}_d^{\perp}$ of bounded type in $\mathbb{D}_e$, which is given by

$$\hat{G_d^{\perp}} = \frac{F_d}{h_d}$$

if and only if G_c has a meromorphic pseudocontinuation $\hat{G}_c$ of bounded type in LHP, which is given by

$$\hat{G}_c = \frac{F_c}{h_c}.$$

Proof. The results can be directly verified. □

The next theorem provides some convenient ways to determine whether a transfer function is cyclic, noncyclic, or strictly noncyclic. Note that $Q \in H^\infty_{L(Z,Y)}(RHP)$ and $F \in H^\infty_{L(U,Y)}(RHP)$ are said to be *weakly left coprime* if $QH^2_Z(RHP) \vee FH^2_U(RHP) = H^2_Y(RHP)$, where $\vee$ denotes the closed linear span. In this case we write $(Q, F)_L = I_Y$. If two functions $Q_1 \in H^\infty_{L(U,Y)}(RHP)$ and $F_1 \in H^\infty_{L(U,Z)}(RHP)$ are such that $\tilde{Q}_1$ and $\tilde{F}_1$ are weakly left coprime, where $\tilde{Q}_1(s) = (Q_1(\bar{s}))^*$ and $\tilde{F}_1(s) = (F_1(\bar{s}))^*$ ($s \in RHP$), they are said to be weakly right coprime, and we denote this by $(Q_1, F_1)_R = I_U$ (see Fuhrmann [11]).

THEOREM 4.8. *Let $G_c \in H^\infty_{L(U,Y)}(RHP)$ with finite-dimensional U and Y. Then the following statements are equivalent:*

1. *G_c is strictly noncyclic.*
2. *G_c has a meromorphic pseudocontinuation of bounded type on LHP.*
3. *On $i\mathbb{R}$ the function G_c can be factored as*

$$G_c = Q_1 F_1^* = F_2^* Q_2,$$

where Q_1 and Q_2 are inner functions in $H^\infty_{L(Y)}(RHP)$ and $H^\infty_{L(U)}(RHP)$, respectively. The functions F_1 and F_2 are in $H^\infty_{L(Y,U)}(RHP)$ and $H^\infty_{L(U,Y)}(RHP)$, respectively, and the coprimeness conditions

$$(Q_1, F_1)_R = I_Y, \qquad (Q_2, F_2)_L = I_U$$

hold. If part 3 holds, then $Q_1 H^2_U(RHP) = (\text{range} H_{G_c})^\perp$ and $\tilde{Q}_2 H^2_U(RHP) = (\text{range} H_{\tilde{G}_c})^\perp$, where $\tilde{Q}_2(s) = (Q_2(\bar{s}))^*$ and $\tilde{G}_c(s) = (G_c(\bar{s}))^*$.

Proof. Analogous results are shown in [11] for discrete-time transfer functions. Thus the theorem follows from Proposition 4.7. □

The factorization in the theorem is Fuhrmann's generalization of the Douglas, Shapiro, and Shields factorization [8] to matrix-valued functions. For a given function, part 2 of the theorem may be easy to check. For example, the function $e^{-\alpha s} R(s)$ is strictly noncyclic, where $\alpha > 0$ and $R(s)$ is any rational function in $H^\infty_{L(U,Y)}(RHP)$. This is because $e^{-\alpha s} R(s)$ has a meromorphic pseudocontinuation of bounded type on LHP of the form $F(s)/e^{\alpha s} h(s)$, where if $a_1, \ldots, a_n$ denote the poles of $R(s)$, then,

$$h(s) = \frac{(s - a_1) \cdots (s - a_n)}{(s + a_1) \cdots (s + a_n)},$$

and $F(s) = h(s)R(s)$. Part 2 of the theorem also gives the following corollary.

COROLLARY 4.9. *Under the assumption of the theorem, $G \in H^\infty_{L(U,Y)}(RHP)$ is strictly noncyclic if and only if $\tilde{G} \in H^\infty_{L(Y,U)}(RHP)$ is strictly noncyclic.* □

4.3. Hankel operators with closed range. Similarly to Theorem 4.8, the following theorem (see [11]) gives necessary and sufficient conditions for the Hankel operator to have closed range.

THEOREM 4.10. *Let $G_c \in H_{L(U,Y)^\infty}(RHP)$ with U and Y finite dimensional. Then the Hankel operator H_{G_c} has closed range if and only if on $i\mathbb{R}$ the function G_c has the factorization*

$$G_c(s) = Q(s) F(s)^*,$$

where $Q \in H^\infty_{L(Y)}(RHP)$, $F \in H^\infty_{L(Y,U)}(RHP)$, *and the equality*

$$WQ + VF = I_Y$$

holds for some $W \in H^\infty_{L(Y)}(RHP)$ *and* $V \in H^\infty_{L(U,Y)}(RHP)$; *that is,* Q *and* F *are* strongly right coprime. *In this case*

$$H_{G_c}\left(H^2_{L(U)}(RHP)\right) = H^2_{L(Y)}(RHP) \ominus QH^2_{L(Y)}(RHP). \qquad \square$$

This section essentially established that the unitary equivalence of the spaces $H^2(\mathbb{D})$ and $H^2(RHP)$ implies the unitary equivalence of the Hankel operators $H_{G_d^\perp}$ and H_{G_c}, where $G_d(z) = G_c(\frac{z-1}{z+1})$, $z \in \mathbb{D}$. Therefore the spaces $\overline{\text{range}}H_{G_d^\perp}$ and $\overline{\text{range}}H_{G_c}$ are unitarily equivalent. As a consequence, the discrete-time transfer function G_d and the continuous-time transfer function G_c have the same cyclicity properties.

These results will be repeatedly used in the next section when we obtain the restricted shift realization of a continuous-time system by applying the bilinear map in §3 to the corresponding discrete-time system.

5. Continuous-time shift realizations via a bilinear transformation. As a direct application of the bilinear transformation T given in §3 the continuous-time restricted and *-restricted shift realizations can be obtained from the corresponding discrete realizations. These realizations can be further analyzed via the connection between continuous and discrete-time transfer functions shown in §§3 and 4.

Restricted and *-restricted shift realizations are central to the development here since they serve as prototypes of output-normal (respectively, input-normal) realizations. It will be shown in Proposition 6.2 that each output-(input-)normal realization of an admissible transfer function G is unitarily equivalent to the restricted (*-restricted) shift realization. The concrete representations of the continuous-time shift realizations obtained in this section will allow us to analyze input- and output-normal realizations in some detail in later sections.

Another important result of this section is Proposition 5.11, in which the state spaces of the restricted shift realizations for strictly noncyclic transfer functions are characterized through the inner factors in the Douglas–Shapiro–Shields factorizations of the transfer function.

5.1. Discrete-time shift realizations. We first recall the discrete-time restricted and *-restricted shift realizations of a discrete-time transfer function (see [11], [30], and [24]).

THEOREM 5.1. *Let* $G_d \in TLD^{U,Y}$. *Then* G_d *has two state-space realizations:* (A_d, B_d, C_d, D_d) *with state space* X_d *and* $(A_{d,*}, B_{d,*}, C_{d,*}, D_{d,*})$ *with state space* $X_{d,*}$, *i.e., for* $z \in \mathbb{D}_e$

$$G_d(z) = C_d(zI - A_d)^{-1}B_d + D_d = C_{d,*}(zI - A_{d,*})^{-1}B_{d,*} + D_{d,*}.$$

They are given in the following way:

1. *The state space* X_d *is given by* $X_d = \overline{\text{range}}H_{G^\perp} \subseteq H^2_Y(\mathbb{D})$, *where*

$$G_d^\perp(z) = \frac{1}{z}\left[G_d\left(\frac{1}{z}\right) - G_d(\infty)\right]$$

and $H_{G_d^\perp}$ *is the Hankel operator with symbol* $G_d^\perp$. *The operators* A_d, B_d, C_d, *and* D_d *are given as follows:*

$$(A_d f)(z) := (S^* f)(z) = \frac{f(z) - f(0)}{z}, \quad f \in X, \quad z \in \mathbb{D},$$

$$(B_d u)(z) := G_d^{\perp}(z)u, \quad u \in U, \quad z \in \mathbb{D},$$

$$C_d f := f(0), \quad f \in X,$$

$$D_d u := G_d(+\infty)u, \quad u \in U,$$

where S is the (forward) shift operator $(Sf)(z) = zf(z)$, $f \in H_Y^2(\mathbb{D})$, $z \in \mathbb{D}$.

The realization (A_d, B_d, C_d, D_d) is called the restricted shift realization *of the transfer function G_d. It is admissible, observable, and reachable, and the observability and reachability operators $\mathcal{R}_d$ and $\mathcal{O}_d$ are, respectively,*

$$\mathcal{O}_d = I_{X_d}, \quad \mathcal{R} = H_{G_d^{\perp}}.$$

2. *The realization $(A_{d,*}, B_{d,*}, C_{d,*}, D_{d,*})$ is given as follows: The state space $X_{d,*}$ is given by $X_{d,*} = \overline{\text{range}} H_{\tilde{G}_d^{\perp}}$ with*

$$\tilde{G}_d(z) = (G_d(\bar{z}))^* \text{ and } \tilde{G}_d^{\perp}(z) = \frac{1}{z}\left(\tilde{G}_d\left(\frac{1}{z}\right) - \tilde{G}_d(\infty)\right), \qquad z \in \mathbb{D}.$$

The operators $A_{d,*}$, $B_{d,*}$, $C_{d,*}$, and $C_{d,*}$ are defined as

$$\begin{aligned} A_{d,*} &= P_{X_{d,*}} S|_{X_{d,*}}, \\ B_{d,*} &: U \to X_{d,*}; u \mapsto P_{X_{d,*}} u, \\ C_{d,*} &: X_{d,*} \to Y; x \mapsto \tfrac{1}{2\pi i} \int_{\partial \mathbb{D}} (z\tilde{G}_d^{\perp}(z))^* x(z) dz = P_Y H_{G_d^{\perp}} x = (H_{G_d^{\perp}} x)(0), \\ D_{d,*} &= G_d(+\infty), \end{aligned}$$

where Y is considered a subspace embedded in $H_Y^2(\mathbb{D})$: $Y = \{y_0 + 0z + 0z^2 + \cdots \mid y_0 \in Y\} \subseteq H_Y^2(\mathbb{D})$, and $P_{X_{d,*}}$ and P_Y are orthogonal projections from $H_Y^2(\mathbb{D})$ onto $X_{d,*}$ and Y, respectively.

The realization $(A_{d,*}, B_{d,*}, C_{d,*}, D_{d,*})$ is called the *-restricted shift realization *of the transfer function G_d. It is admissible, observable, and reachable, and the observability and reachability operators $\mathcal{O}_{d,*}$ and $\mathcal{R}_{d,*}$ are, respectively,*

$$\mathcal{R}_{d,*} = P_{X_{d,*}} : H_U^2(\mathbb{D}) \to X_* \text{ and } \mathcal{O}_{d,*} = H_{\tilde{G}_d^{\perp}}^*|_{X_{d,*}} = H_{G_d^{\perp}}|_{X_{d,*}}. \qquad \square$$

5.2. Continuous-time restricted shift realization. Now we apply T to the realizations given in the theorem to get the continuous-time realizations. We need some simple lemmas.

LEMMA 5.2. *For any $x \in H_Y^2(\mathbb{D})$, $\lim_{r \in \mathbb{R},\, r>-1,\, r \to -1}(1+r)x(r) = 0$ in the norm of Y. For any $f \in H_Y^2(RHP)$, $\lim_{r \in \mathbb{R},\, r \to +\infty} f(r) = 0$ in the norm of Y.*

Proof. For $x \in H_Y^2(\mathbb{D})$ and $z \in \mathbb{D}$ we have $x(z) = \sum_{n \geq 0} z^n \hat{x}_n$, where $\hat{x}_n \in Y$ and $\sum_{n \geq 0} \|\hat{x}_n\|^2 = \|x\|_{H_Y^2(\mathbb{D})}^2$. Thus

$$\begin{aligned} \|x(z)\|_Y &\leq \sum_{n \geq 0} |z^n| \, \|\hat{x}_n\| \leq \left(\sum_{n \geq 0} |z^n|^2\right)^{1/2} \left(\sum_{n \geq 0} \|\hat{x}_n\|^2\right)^{1/2} \\ &= (1 - |z|)^{-1/2}(1 + |z|)^{-1/2}\|x\|. \end{aligned}$$

Hence $\lim_{r>-1,\, r \to -1}(1+r)x(r) = 0$.

Now for $f \in H_Y^2(RHP)$ and any $s \in RHP$, it is shown in [22, p. 254] that

$$\|f(s)\| \leq \delta(Re(s))^{-1/2},$$

where δ is a constant depending on f. Thus the lemma is proven. □

LEMMA 5.3. *With the notation of Theorem* 5.1 *we have*

1. $$\text{range}(I + A_d) = \left\{ x \mid x(z) = \frac{(1+z)h(z) - h(0)}{z},\ (z \in \mathbb{D}),\ h \in X_d \right\},$$

and for $x \in \text{range}(I + A_d)$ *the limit* $\lim_{r \in \mathbb{R},\ r>-1,\ r \to -1} x(r)$ *exists*;

2. $$[(\lambda I + A_d)^{-1}x](z) = \frac{z}{1+\lambda z}x(z) + \frac{1}{\lambda(1+\lambda z)}x\left(-\frac{1}{\lambda}\right),$$

where $x \in X_d$, $\lambda \in \mathbb{D}_e$, $z \in \mathbb{D}$;

3. $$C_d(\lambda I + A_d)^{-1}x = \frac{1}{\lambda}x\left(-\frac{1}{\lambda}\right),\ \lambda \in \mathbb{D}_e,\ x \in X_d;$$

4. $$[(I + A_d)^{-1}x](z) = \frac{z}{1+z}\left[x(z) + \frac{1}{z}x(-1)\right],\ x \in \text{range}(I + A_d),$$

where $x(-1) = \lim_{r \in \mathbb{R},\ r>-1, r \to -1} x(r)$.

Proof. 1. Since $\text{range}(I + A_d) = \{x + A_d x \mid x \in X_d\} = \{\frac{x(z)-x(0)}{z} + x(z) \mid x \in X_d\}$, we have the equality in 1. If $x \in \text{range}(I + A_d)$, then $x(z) = \frac{(1+z)h(z)-h(0)}{z}$ for some $h \in X_d$. By Lemma 5.2,

$$\lim_{r \in \mathbb{R},\ r>-1,\ r \to -1} x(r) = \lim_{r \in \mathbb{R},\ r>-1,\ r \to -1} \frac{(1+r)h(r) - h(0)}{r} = h(0).$$

2. First we show that for $x \in X_d$ and $\lambda \in \mathbb{D}_e$, the element

$$\frac{z}{1+\lambda z}x(z) + \frac{1}{\lambda(1+\lambda z)}x\left(-\frac{1}{\lambda}\right) = P_+\left(\frac{z}{1+\lambda z}x(z)\right)$$

is in X_d. Take any y in the invariant space $H_Y^2(\mathbb{D}) \ominus X_d$. Since $\frac{1}{z+\bar{\lambda}} \in H^\infty$ for $\lambda \in \mathbb{D}_e$, we have $\frac{1}{z+\bar{\lambda}}y \in H_Y^2(\mathbb{D}) \ominus X_d$. Therefore,

$$\left\langle P_+\left(\frac{z}{1+\lambda z}x(z)\right),\ y\right\rangle_{H_Y^2(\mathbb{D})} = \left\langle x,\ \frac{1}{z+\bar{\lambda}}y\right\rangle = 0.$$

This shows that $P_+\left(\frac{z}{1+\lambda z}x(z)\right) \in X_d$. Since A_d is a contraction, $(\lambda I + A_d)^{-1}$ is a bounded operator on X_d for $\lambda \in \mathbb{D}_e$. Then the equality in 2. follows from the equality

$$(\lambda I + A_d)\left[\frac{z}{1+\lambda z}x(z) + \frac{1}{\lambda(1+\lambda z)}x\left(-\frac{1}{\lambda}\right)\right] = x(z).$$

3. Using 2. and the definition of C_d we get 3.

4. If $x \in \text{range}(I + A_d)$, then by 1. and its proof there exists $h \in X_d$ such that $x(z) = \frac{(1+z)h(z)-h(0)}{z}$ and $\lim_{r \in \mathbb{R},\ r>-1,\ r \to -1} x(r) = h(0)$. Set

$$x(-1) = \lim_{r \in \mathbb{R},\ r>-1,\ r \to -1} x(r) = h(0).$$

We have

$$\frac{z}{1+z}\left[x(z) + \frac{1}{z}x(-1)\right] = h(z),$$

which is in X_d. Note that $-1 \notin \sigma_p(A_d)$. Thus $(I + A_d)^{-1}$ is defined on range$(I + A_d)$. The equality in 4. then follows from

$$(I + A_d)\left(\frac{z}{1+z}\left[x(z) + \frac{x(-1)}{z}\right]\right) = [(I + A_d)h](z) = \frac{(1+z)h(z) - h(0)}{z} = x(z). \qquad \square$$

LEMMA 5.4. *Let $f \in L^2_Y(i\mathbb{R})$. Then in $L^2_Y(i\mathbb{R})$ norm,*

$$\lim_{n\to\infty} \frac{s}{n+s} f = \lim_{n\to\infty} \frac{s}{n-s} f = 0.$$

Proof. Since $\|\frac{s}{n+s} f(s)\|^2_Y \le \|f(s)\|^2_Y$ for any $s \in i\mathbb{R}$ and $n > 0$ and

$$\lim_{n\to\infty} \left\| \frac{s}{n+s} f(s) \right\|^2_Y = 0$$

for a.e. $s \in i\mathbb{R}$, the lemma follows from the Lebesgue dominated convergence theorem. $\square$

LEMMA 5.5. *Let $G_c \in TLC^{U,Y}$. Set $X = \overline{\text{range}} H_{G_c}$ and $\mathcal{D} = P_X\{\frac{u}{1+s} : u \in U\}$. Then the map*

$$\begin{aligned} M_1 : \mathcal{D} &\to Y, \\ P_X \tfrac{u}{1+s} &\mapsto [\tilde{G}_c(1) - \tilde{G}(\infty)]u \end{aligned}$$

is well defined and the map

$$\begin{aligned} M_2 : X &\to X, \\ f &\mapsto P_X \tfrac{f}{1+s} \end{aligned}$$

is injective.

Proof. Assume $P_X \frac{u_1}{1+s} = P_X \frac{u_2}{1+s}$. Then $P_X \frac{u_1-u_2}{1+s} = 0$. This shows that $\frac{u_1-u_2}{1+s} \in H^2_Y(RHP) \ominus X$. Therefore, for any $f \in H^2_Y(RHP)$,

$$\begin{aligned} 0 &= \left\langle \frac{u_1 - u_2}{1+s},\ H_{G_c} f \right\rangle_{H^2_U(RHP)} = \left\langle \frac{u_1 - u_2}{1+s},\ P_+ G_c f(-s) \right\rangle_{H^2_U(RHP)} \\ &= \left\langle \frac{u_1 - u_2}{1+s},\ P_+[G_c - G_c(+\infty)] f(-s) \right\rangle \\ &= \left\langle \frac{u_1 - u_2}{1+s},\ [G_c(s) - G_c(+\infty)] f(-s) \right\rangle \\ &= \left\langle [G_c(s) - G_c(+\infty)]^* \frac{u_1 - u_2}{1+s},\ f(-s) \right\rangle \\ &= \left\langle [\tilde{G}_c(s) - \tilde{G}_c(\infty)] \frac{u_1 - u_2}{1-s},\ f(s) \right\rangle. \end{aligned}$$

Hence $[\tilde{G}(s) - \tilde{G}(1)]\frac{u_1-u_2}{1-s} = 0$. So we have $[\tilde{G}(s) - \tilde{G}(1)](u_1 - u_2) = 0$. Taking the limit on the real line, we get

$$[\tilde{G}_c(1) - \tilde{G}_c(+\infty)](u_1 - u_2) = 0.$$

This shows that indeed M_1 is well defined.

To show that M_2 is injective, assume $P_X \frac{h_1(s)}{1+s} = P_X \frac{h_2(s)}{1+s}$, $h_1, h_2 \in X$. Then $P_X \frac{h_1(s)-h_2(s)}{1+s} = 0$. Hence

$$\frac{h_1(s) - h_2(s)}{1+s} \in H_Y^2 \ominus X.$$

By Lemma 5.4, we have

$$\lim_{n\to\infty} \left\| \frac{1+s}{1+s/n} \frac{h_1(s) - h_2(s)}{1+s} - (h_1 - h_2) \right\|_{H_Y^2(RHP)} = \lim_{n\to\infty} \left\| \frac{-s}{n+s}(h_1 - h_2) \right\| = 0.$$

Hence $\lim_{n\to+\infty} \frac{1+s}{1+s/n} \frac{h_1(s)-h_2(s)}{1+s} = h_1 - h_2$ in H_Y^2. Note that $H_Y^2 \ominus X$ is an invariant space and $\frac{1+s}{1+s/n} \in H^\infty$ for $n > 0$. So $\frac{1+s}{1+s/n} \frac{h_1(s)-h_2(s)}{1+s} \in H_Y^2 \ominus X$, and hence $h_1 - h_2 \in H_Y^2 \ominus X$. Since $h_1 - h_2 \in X$, we therefore have $h_1 - h_2 = 0$. This shows that M_2 is injective. □

We will need the following result on the reproducing kernel in $H^2(RHP)$ (see, e.g., [10]).

LEMMA 5.6. *For $f \in H_U^2(RHP)$, $u \in U$, and $\alpha \in RHP$ the following hold:*

$$\left\langle f, \frac{u}{s+\bar{\alpha}} \right\rangle_{H_U^2(RHP)} = 2\pi \langle f(\alpha), u \rangle_U,$$

$$\left\langle \frac{u}{s+\bar{\alpha}}, f \right\rangle_{H_U^2(RHP)} = 2\pi \langle u, f(\alpha) \rangle_U.$$

We are ready to present the continuous-time restricted shift realization using the bilinear transform T. For a continuous-time transfer function G_c we first realize the discrete-time transfer function G_d defined by

$$G_d(z) = G_c\left(\frac{z-1}{z+1}\right)$$

in terms of the restricted shift realization. Applying T to this discrete-time realization we obtain a realization of G_c with the same state space. Then we use a unitary transformation to get the continuous-time restricted shift realization with state space $\overline{\text{range}} H_{G_c}$.

THEOREM 5.7. *Let $G_c \in TLC^{U,Y}$. Then G_c has a state-space realization $(A_c, B_c, D_c, C_c) \in C_X^{U,Y}$, which is given in the following way:*

1. *The state space is given by*

$$X = \overline{\text{range}} H_{G_c,RHP} \subseteq H_Y^2(RHP).$$

2. *The semigroup $(e^{tA_c})_{t\geq 0}$ corresponding to the realization is given by*

$$\begin{aligned} e^{tA_c}: \quad X &\to X, \\ f &\mapsto (e^{tA_c} f)(s) = P_+ e^{ts} f(s). \end{aligned}$$

The infinitesimal generator $(A_c, D(A_c))$ of the semigroup $(e^{tA_c})_{t\geq 0}$ is given by

$$\begin{aligned} A_c: \quad D(A_c) &\to X, \\ f &\mapsto (A_c f)(s) = sf(s) - \lim_{\substack{r\in\mathbb{R} \\ r\to\infty}} rf(r). \end{aligned}$$

The domain $D(A_c)$ is dense in X, and we have

$$D(A_c) = \left\{ f \mid f(s) = \frac{1}{1-s}[h(s) - h(1)] : (s \in RHP),\ h \in X \right\}.$$

The domain of the adjoint A_c^* of the operator A_c is

$$D(A_c^*) = \left\{ f \mid f(s) = P_X \frac{h(s)}{1+s} : (s \in RHP),\ h \in X \right\}$$

and

$$A_c^* f = f - h \text{ for } f(s) = P_X \frac{h(s)}{1+s} \in D(A_c^*).$$

On $\mathcal{L}^{-1}(X) \subseteq L^2_U([0,+\infty))$ the semigroup is given by

$$\begin{aligned} e^{tA_c} :\ \mathcal{L}^{-1}(X) &\to \mathcal{L}^{-1}(X), \\ f &\mapsto (e^{tA_c} f)(\tau) = P_{L^2_U([0,+\infty))}(f(\tau+t))_{\tau \geq 0}. \end{aligned}$$

3. *The input operator is given by*

$$\begin{aligned} B_c :\ U &\to D(A_c^*)^{(\prime)}, \\ u &\mapsto B_c(u), \end{aligned}$$

where for $u \in U$ and $x(s) = P_X \frac{h(s)}{1+s} \in D(A_c^*)$,

$$\begin{aligned} [B_c(u)](x) &= \frac{1}{\sqrt{2\pi}} \left\langle \frac{1}{1-s}[G_c(s) - G_c(1)]u,\ (1 - A_c^*)x \right\rangle \\ &= \frac{1}{\sqrt{2\pi}} \left\langle H_{G_c} \frac{u}{1+s},\ (1 - A_c^*)x \right\rangle \\ &= \frac{1}{\sqrt{2\pi}} \left\langle H_{G_c} \frac{u}{1+s},\ h \right\rangle \\ &= \sqrt{2\pi} \left\langle u, (H_{\tilde{G}_c} h)(1) \right\rangle_U. \end{aligned}$$

4. *The output operator is given by*

$$\begin{aligned} C_c : D(C) = D(A_c) + (I - A_c)^{-1} B_c U &\to Y, \\ x &\mapsto \sqrt{2\pi} \lim_{\substack{r \in \mathbb{R} \\ r \to \infty}} rx(r). \end{aligned}$$

5. *The feedthrough operator is given by*

$$\begin{aligned} D_c :\ U &\to Y, \\ u &\mapsto G_c(+\infty)u := \lim_{\substack{r \in \mathbb{R} \\ r \to +\infty}} G_c(r)u. \end{aligned}$$

The realization (A_c, B_c, C_c, D_c) of G_c is called the restricted shift realization.

Proof. These results are obtained by applying the map T of Theorem 3.1 to the restricted shift realization (A_d, B_d, C_d, D_d) of $G_d(z) = G_c(\frac{z-1}{z+1})$, $(z \in \mathbb{D}_e)$, with the state space then transformed by the unitary operator $V = V_Y$ defined in Proposition 3.3.

1. Let $(A_{c1}, B_{c1}, C_{c1}, D_{c1}) = T((A_d, B_d, C_d, D_d))$ and

$$(A_c, B_c, C_c, D_c) = (VA_{c1}V^{-1}, VB_{c1}, CV^{-1}, D_{c1}).$$

We use the following notation: $G_d^{\perp}(z) = \frac{1}{z}[G_d(\frac{1}{z}) - G_d(\infty)] = \frac{1}{z}[G_c(\frac{1-z}{1+z}) - G_c(1)]$ $(z \in \mathbb{D})$, $X_d = \overline{\text{range}} H_{G_d^{\perp}}$, and $\phi_t(z) = e^{t\frac{z-1}{z+1}}$, $t \geq 0$. Then by Proposition 4.3 $X = VX_d$, and $\phi_t(A_d)$ is

the semigroup of contractions on X_d with infinitesimal generator $(A_d - I)(A_d + I)^{-1} = A_{c1}$ (see [28, p. 141]). Specifically, for $x \in X_d$ we have

$$\phi_t(A_d)x = P_+\phi_t\left(\frac{1}{z}\right)x = P_+e^{t\frac{\frac{1}{z}-1}{\frac{1}{z}+1}}x = P_+e^{t\frac{1-z}{1+z}}x.$$

Then it is easy to see that $A_c = VA_{c1}V^{-1}$ generates the semigroup of contractions $V\phi_t(A_d)V^{-1}$ on X. If we extend the unitary transformation $V : H^2(\mathbb{D}) \to H^2(RHP)$ naturally to

$$\begin{aligned} V : L^2(\partial\mathbb{D}) &\to L^2(i\mathbb{R}), \\ x_d &\mapsto (Vx_d)(\bullet) = \frac{1}{\sqrt{\pi}(1+\bullet)}x_d\left(\frac{1-\bullet}{1+\bullet}\right), \end{aligned}$$

we still have a unitary transformation. Moreover, by considering $z^n y$ for $n \in \mathbb{Z}$ and $y \in Y$ we can show that

$$VP_+^{\mathbb{D}} = P_+^{RHP}V,$$

where $P_+^{\mathbb{D}} : L^2(\partial\mathbb{D}) \to H^2(\mathbb{D})$ and $P_+^{RHP} : L^2(i\mathbb{R}) \to H^2(RHP)$ are the orthogonal projections. From this it follows that for $f \in X$,

$$e^{tA_c}f = V\phi_t(A_d)V^{-1}f = VP_+^{\mathbb{D}}e^{t\frac{1-z}{1+z}}V^{-1}f = P_+^{RHP}Ve^{t\frac{1-z}{1+z}}V^{-1}f = P_+e^{ts}f.$$

Clearly, $D(A_{c1}) = \text{range}(A_d+I)$, and by Lemma 5.3 $\text{range}(A_d+I) = \{\frac{(1+z)x(z)-x(0)}{z} \mid x \in X_d\}$. Since $D(A_c) = VD(A_{c1})$ and $x(0) = 2\sqrt{\pi}(Vx)(1)$ for $x \in X_d$, we have

$$D(A_c) = V\text{range}(A_d + I) = V\left\{\frac{(1+z)x(z) - x(0)}{z} \mid x \in X_d\right\}$$

$$= \left\{V\left(\frac{(1+z)x(z) - x(0)}{z}\right) \mid x \in X_d\right\} = \left\{\frac{(1+\frac{1-s}{1+s})f(s) - \frac{2f(1)}{1+s}}{\frac{1-s}{1+s}} \mid f \in X\right\}$$

$$= \left\{\frac{f(s) - f(1)}{1-s} \mid f \in X\right\}.$$

For $x \in D(A_{c1}) = \text{range}(A_d + I)$ the limit $\lim_{r\in\mathbb{R},\, r>-1,\, r\to-1} x(r)$ exists by Lemma 5.3. Denoting it by $x(-1)$ and using Lemma 5.3, part 4, we have

$$(A_{c1}x)(z) = [(A_d - I)(A_d + I)^{-1}x](z)$$

$$= (A_d - I)\left(\frac{z}{1+z}\left[x(z) + \frac{1}{z}x(-1)\right]\right) = \frac{(1-z)x(z) - 2x(-1)}{1+z}.$$

From this we obtain, for $f \in D(A_c) = VD(A_{c1})$,

$$(A_cf)(s) = (VA_{c1}V^{-1}f)(s) = V\left(\frac{(1-z)(V^{-1}f)(z) - 2(V^{-1}f)(-1)}{1+z}\right)$$
$$= sf(s) - \lim_{r\in\mathbb{R},\, r\to+\infty} rf(r),$$

where we have used the fact that for $f \in D(A_c)$

$$(V^{-1}f)(-1) = \sqrt{\pi}\lim_{\substack{r\in\mathbb{R} \\ r\to+\infty}}(1+r)f(r) = \sqrt{\pi}\lim_{\substack{r\in\mathbb{R} \\ r\to+\infty}} rf(r).$$

Now we show the form of A_c^*. Recall that A_c^* is the generator of the strongly continuous semigroup $(e^{tA_c})^*$. Let

$$D(\hat{A}) = \left\{ f \mid f(s) = P_X \frac{h(s)}{1+s} \text{ for some } h \in X \right\}$$

and

$$\begin{aligned} \hat{A} : D(\hat{A}) &\to X, \\ \hat{A}f &= f - h \quad \left(f = P_X \frac{h(s)}{1+s} \in D(\hat{A}) \right). \end{aligned}$$

By Lemma 5.5, the operator $\hat{A}$ is well defined.

For $f \in D(\hat{A})$ and $g \in D(A_c)$ there are v and w in X such that $f = P_X \frac{v}{1+s}$ and $g = \frac{w-w(1)}{1-s}$. By the definition of A_c and $\hat{A}$, we have $A_c g = w$ and $\hat{A}f = v$. It then follows that

$$\langle A_c g,\ f \rangle = \left\langle w,\ P_X \frac{v}{1+s} \right\rangle = \left\langle \frac{w}{1-s},\ v \right\rangle = \left\langle \frac{w-w(1)}{1-s}, v \right\rangle = \langle g,\ \hat{A}f \rangle.$$

This shows that $D(\hat{A}) \subseteq D(A_c^*)$ and $\hat{A} = A_c^*|_{D(\hat{A})}$. On the other hand, we clearly have $(I - \hat{A})D(\hat{A}) = X$ and hence

$$(I - A_c^*)D(\hat{A}) = X.$$

Let $x \in D(A_c^*)$. Then there exists $x_1 \in D(\hat{A})$ such that

$$(I - A_c^*)x_1 = (I - A_c^*)x.$$

Since A_c^* is the infinitesimal generator of a semigroup of contractions, the number 1 is not in the spectrum of A_c^*. Thus we must have $x_1 = x$. This shows that

$$D(A_c^*) \subseteq D(\hat{A}).$$

Therefore $D(A_c^*) = D(\hat{A})$ and hence $A_c^* = \hat{A}$.

2. For the operator B_c we first compute B_{c1}, following the definition of T:

$$\begin{aligned} B_{c1} := \sqrt{2}(I + A_d)^{-1} B_d : U &\to D(A_c^*)^{(\prime)}, \\ u &\longmapsto \sqrt{2}(I + A_d)^{-1} B_d(u)[\cdot], \\ &:= \sqrt{2} < B_d(u), (I + A_d^*)^{-1}(\cdot) >_{X_d}. \end{aligned}$$

Note that $V^* = V^{-1}$, $(I + A_d^*)^{-1} = \frac{1}{2}(I - A_{c1}^*)$, and

$$(VB_d u)(s) = \frac{1}{\sqrt{\pi}} \frac{G_c(s) - G_c(1)}{1-s} u \quad (s \in RHP).$$

Thus for $x = P_X \frac{h(s)}{1+s} \in D(A_c^*) \subseteq X$, we have $(I - A_c^*)x = h$ and

$$\begin{aligned} (B_c u)(x) &= (VB_{c1})(x) = (B_{c1}u)(V^*x) = \sqrt{2}\langle B_d u,\ (I + A_d^*)^{-1} V^{-1} x \rangle_{X_d} \\ &= \sqrt{2}\langle VB_d u,\ V(I + A_d^*)^{-1} V^{-1} x \rangle_X = \sqrt{2} \left\langle VB_d u,\ \frac{1}{2} V(I - A_{c1}^*) V^{-1} x \right\rangle_X \\ &= \frac{1}{\sqrt{2}} \langle VB_d u,\ (I - A_c^*)x \rangle_X \end{aligned}$$

$$= \frac{1}{\sqrt{2\pi}} \left\langle \frac{G_c(s) - G_c(1)}{1-s} u,\ (I - A_c^*)x \right\rangle_X$$

$$= \frac{1}{\sqrt{2\pi}} \left\langle \frac{G_c(s) - G_c(1)}{1-s} u,\ h \right\rangle_X$$

$$= \frac{1}{\sqrt{2\pi}} \left\langle H_{G_c} \frac{u}{1+s},\ h \right\rangle_X .$$

Since $(H_{G_c})^* = H_{\tilde{G}_c}$, we have

$$(B_c u)(x) = \frac{1}{\sqrt{2\pi}} \left\langle \frac{u}{1+s},\ H_{\tilde{G}_c} h \right\rangle .$$

By Lemma 5.6 the right-hand side is $\sqrt{2\pi}\langle u,\ (H_{\tilde{G}_c} h)(1)\rangle_U$.

3. To compute C_c we use Lemma 5.3, part 3, to get

$$C_d(\lambda I + A_d)^{-1} x = \frac{1}{\lambda} x\left(-\frac{1}{\lambda}\right), \quad \lambda \in \mathbb{D}_e .$$

So for $x \in D(C_{c1}) = D(A_{c1}) + (I - A_{c1})^{-1} B_{c1} U$ we have

$$C_{c1} x = \sqrt{2} \lim_{\substack{\lambda > 1 \\ \lambda \to 1}} C_d(\lambda I + A_d)^{-1} x$$

$$= \sqrt{2} \lim_{\substack{\lambda > 1 \\ \lambda \to 1}} \frac{1}{\lambda} x\left(-\frac{1}{\lambda}\right) = \sqrt{2} \lim_{\substack{\lambda > -1 \\ \lambda \to -1}} x(\lambda).$$

The existence of the limit for $x \in D(A_{c1})$ follows from Lemma 5.3, part 1, because $D(A_{c1}) = \text{range}(A_d + I)$. For $x \in (I - A_{c1})^{-1} B_{c1} U$ we have that the limit

$$\sqrt{2} \lim_{\substack{\lambda > 1 \\ \lambda \to 1}} C_d(\lambda I + A_d)^{-1} x$$

also exists by the admissibility of G_d, since $(I - A_{c1}) B_{c1} u = \frac{1}{\sqrt{2}} B_d u = \frac{1}{\sqrt{2}} G_d^{\perp} u$ (see [23]). Now it can be verified that $V D(C_{c1}) = D(A_c) + (I - A_c)^{-1} B_c U$, i.e., $V D(C_{c1}) = D(C_c)$. Hence we get, for $f \in D(C_c)$,

$$C_c f = C_{c1} V^{-1} f = \sqrt{2} \lim_{\substack{\lambda > -1 \\ \lambda \to -1}} \frac{2\sqrt{\pi}}{1+\lambda} f\left(\frac{1-\lambda}{1+\lambda}\right)$$

$$= \sqrt{2\pi} \lim_{\substack{r \in \mathbb{R} \\ r \to +\infty}} (1+r) f(r) = \sqrt{2\pi} \lim_{\substack{r \in \mathbb{R} \\ r \to +\infty}} r f(r).$$

Finally, the obvious expression $D_c = G_c(\infty)$ can also be verified as follows:

$$D_c u = D_{c1} u = D_d u - \lim_{\substack{\lambda > 1 \\ \lambda \to 1}} C_d(\lambda I + A_d)^{-1} B_d u$$

$$= G_d(\infty) u - \lim_{\substack{\lambda > 1 \\ \lambda \to 1}} C_d(\lambda I + A_d)^{-1} G_d^{\perp} u = G_c(1) u - \lim_{\substack{\lambda > 1 \\ \lambda \to 1}} \frac{1}{\lambda} G_d^{\perp}\left(-\frac{1}{\lambda}\right) u$$

$$= G_c(1) u - \lim_{\substack{\lambda > 1 \\ \lambda \to 1}} (-1)\left[G_c\left(\frac{\lambda+1}{\lambda-1}\right) - G_c(1) \right] u = \lim_{\substack{\lambda > 1 \\ \lambda \to 1}} G_c\left(\frac{\lambda+1}{\lambda-1}\right)$$

$$= G_c(+\infty). \qquad \square$$

Regarding the expressions for the operator B_c in the theorem we have the following corollary.

COROLLARY 5.8. *$B_c u \in X$, $(u \in U)$ if and only if $[G_c - G_c(+\infty)]u \in X$ $(u \in U)$. In this case*

$$(B_c u)(s) = \frac{1}{\sqrt{2\pi}}[G_c(s) - G_c(+\infty)]u \quad (u \in U).$$

In particular, if G_c satisfies

$$\sup_{x>0} \int_{-\infty}^{+\infty} \|G_c(x+iy) - G_c(+\infty)\|^2 dy < \infty,$$

where for $s \in RHP$ the expression $\|G_c(s) - G_c(+\infty)\|$ denotes the operator norm of the operator $G_c(s) - G_c(+\infty) \in L(U, Y)$, then $B_c u \in X$ and

$$(B_c u)(s) = \frac{1}{\sqrt{2\pi}}[G_c(s) - G_c(+\infty)]u$$

for any $u \in U$.

Proof. First we assume that $[G_c - G_c(+\infty)]u \in X$ $(u \in U)$. Define $F(s) = G_c(s) - G_c(+\infty)$. Then $Fu \in X$. It follows from the formula for $D(A_c)$ that

$$\frac{1}{1-s}[G_c(s) - G_c(1)]u = \frac{1}{1-s}[F(s)u - F(1)u] \in D(A_c)$$

and hence for $x \in D(A_c^*)$,

$$[B_c(u)](x) = \frac{1}{\sqrt{2\pi}}\left\langle \frac{1}{1-s}[G_c(s) - G_c(1)]u,\ (I - A_c^*)x \right\rangle_{H_Y^2(RHP)}$$

$$= \frac{1}{\sqrt{2\pi}}\left\langle (I - A_c)\frac{1}{1-s}[G_c(s) - G_c(1)]u,\ x \right\rangle$$

$$= \frac{1}{\sqrt{2\pi}}\left\langle \left(\frac{1}{1-s} - \frac{s}{1-s}\right)[G_c(s) - G_c(1)]u + \lim_{\substack{r\in\mathbb{R} \\ r\to+\infty}} \frac{r}{1-r}[G_c(r) - G_c(1)]u,\ x \right\rangle$$

$$= \frac{1}{\sqrt{2\pi}}\langle [G_c(s) - G_c(+\infty)]u,\ x\rangle = \frac{1}{\sqrt{2\pi}}\langle Fu,\ x\rangle.$$

Here we have used the definition of A_c and the fact that the limit $\lim_{\substack{r\in\mathbb{R} \\ r\to+\infty}} G_c(r)u$ exists, which follows from the admissibility of G_c. Thus we have shown that $B_c(u) \in X$ and $B_c(u) = \frac{1}{\sqrt{2\pi}}Fu$ for any $u \in U$.

On the other hand, if $B_c(u) \in X$, $(u \in U)$, then there is $f_u \in X$ such that $[B_c(u)](x) = \langle f_u,\ x\rangle$ for any $x \in X$. Therefore

$$\frac{1}{\sqrt{2\pi}}\left\langle \frac{1}{1-s}[G_c(s) - G_c(1)]u,\ (I - A_c^*)x \right\rangle = \langle f_u,\ x\rangle \ (x \in X).$$

This shows that $\frac{1}{1-s}[G_c(s) - G_c(1)]u \in D((I - A_c^*)^*) = D(I - A_c) = D(A_c)$. So there is $h \in X$ such that

$$\frac{1}{1-s}[G_c(s) - G_c(1)]u = \frac{h(s) - h(1)}{1-s}.$$

Hence $G_c(s) - G_c(1) = h(s) - h(1)$. Since $\lim_{\substack{s\in\mathbb{R}\\ s\to\infty}} h(s) = 0$ (see Lemma 5.2), we have $G_c - G_c(+\infty) = h \in X$.

To complete the proof of the corollary, it suffices to show that the condition that G_c is analytic for $Re(s) > 0$ and satisfies

$$\sup_{x>0} \int_{-\infty}^{+\infty} \|G_c(x+iy) - G_c(+\infty)\|^2 dy < \infty$$

implies that $[G_c - G_c(+\infty)]u \in X$ for any $u \in U$.

Again let $F(s) = G_c(s) - G_c(+\infty)$. We have the equality of Hankel operators:

$$H_{G_c} = H_F.$$

The assumption on G_c implies that $Fu \in L^2_Y(i\mathbb{R})$ for any $u \in U$. Now we show that in $L^2_Y(i\mathbb{R})$ norm

$$Fu = \lim_{n\to\infty} H_F \frac{n}{n+s} u$$

and hence $Fu \in X = \overline{\text{range}} H_F$. The proof will then be complete.

Consider

$$\left\| Fu - H_F \frac{n}{n+s} u \right\|_{L^2_Y(i\mathbb{R})} = \left\| P_+ F \frac{-s}{n-s} u \right\| = \left\| P_+ \frac{-s}{n-s} Fu \right\|.$$

By Lemma 5.4, we have $\lim_{n\to\infty} \| \frac{-s}{n-s} Fu \|_{L^2_Y(i\mathbb{R})} = 0$. Therefore

$$\lim_{n\to\infty} \| P_+ \frac{-s}{n-s} Fu \| = 0.$$

So we indeed have $Fu = \lim_{n\to\infty} H_F \frac{n}{n+s} u$, converging in $L^2_Y(i\mathbb{R})$ norm. □

5.3. Continuous-time $*$-restricted shift realization. If we apply the map T in Theorem 3.1 to the *-restricted shift realization of $G_d(z) = G_c(\frac{z-1}{z+1})$ and then transform the state space by the unitary operator of Proposition 3.3, we obtain the *-restricted shift realization of G_c. Alternatively, we can find the restricted shift realization of the transfer function $\tilde{G}_c \in TLC^{Y,U}$ first, and then the dual system of this restricted shift realization will be the *-restricted shift realization of G_c.

THEOREM 5.9. *Let $G_c \in TLC^{U,Y}$. Then G_c has a state-space realization $(A_{c,*}, B_{c,*}, C_{c,*}, D_{c,*}) \in C_X^{U,Y}$, which is given in the following way:*

1. *The state space is given by*

$$X_* = \overline{\text{range}} H_{\tilde{G}_c, RHP} \subseteq H^2_U(RHP),$$

where $\tilde{G}_c(s) = (G(\bar{s}))^*$ for $s \in RHP$.

2. *The semigroup $(e^{tA_{c,*}})_{t\geq 0}$ corresponding to the realization is given by*

$$\begin{aligned} e^{tA_{c,*}} : \quad X_* &\to X_*, \\ f &\mapsto (e^{tA_{c,*}} f)(s) = P_{X_*} e^{-ts} f(s), \end{aligned}$$

where the operator $A_{c,*}$ has domain

$$D(A_{c,*}) = P_{X_*} \left\{ \frac{1}{1+s} h(s) : h \in X_* \right\}$$

and for $f(s) = P_{X_*}\frac{1}{1+s}h(s) \in D(A_{c,*})$,

$$A_{c,*}f = f - h.$$

On $\mathcal{L}^{-1}(X_*) \subseteq L^2_U([0,\infty))$ *the semigroup is given by*

$$\begin{aligned} e^{tA_{c,*}}: \quad \mathcal{L}^{-1}(X_*) &\to \quad \mathcal{L}^{-1}(X_*), \\ f &\mapsto \quad (e^{tA_{c,*}}f)(s) = P_{\mathcal{L}^{-1}(X)_*}f(s-t). \end{aligned}$$

3. *The input operator* $B_{c,*}: U \to D(A^*_{c,*})^{(\prime)}$ *is given by*

$$u \mapsto B_c(u)$$

with

$$\begin{aligned} \left[B_{c,*}(u)\right](x) &= \frac{1}{\sqrt{2\pi}}\left\langle \frac{1}{1+s}u,\ (1-A^*_{c,*})x\right\rangle \\ &= \frac{1}{\sqrt{2\pi}}\left\langle \frac{1}{1+s}u,\ h\right\rangle \\ &= \sqrt{2\pi}\langle u, h(1)\rangle_U, \quad x = \frac{h(s)-h(1)}{1-s} \in D(A^*_{c,*}),\ h \in X^*. \end{aligned}$$

4. *The output operator has the following form:*

$$\begin{aligned} D(C_{c,*}) &= D(A_{*,c}) + (I - A_{*,c})^{-1}B_{c,*}U \\ &= P_{X_*}\left\{\frac{h(s)}{1+s} \mid h \in X_*\right\} + P_{X_*}\left\{\frac{u}{1+s} \mid u \in U\right\}. \end{aligned}$$

If $x = P_{X_*}\frac{h(s)}{1+s}$, *then*

$$C_{c,*}x = \sqrt{2\pi}(H_{G_c}h)(1),$$

and if $x = P_{X_*}\frac{u}{1+s}$, *then*

$$C_{c,*}x = \sqrt{2\pi}[G_c(1) - G(+\infty)]u.$$

5. *The feedthrough operator is given by*

$$\begin{aligned} D_{c,*}: \quad U &\to \quad Y, \\ u &\mapsto \quad G_c(+\infty)u := \lim_{\substack{r\in\mathbb{R}\\ r\to\infty}} G_c(r)u. \end{aligned}$$

The realization $(A_{c,*}, B_{c,*}, C_{c,*}, D_{c,*})$ *of* G_c *is called the* *- restricted shift realization.

Proof. Let (A, B, C, D) be the restricted shift realization of the transfer function $\tilde{G}_c(s) = (G(\bar{s}))^*$. Take $(A_{c,*}, B_{c,*}, C_{c,*}, D_{c,*})$ to be the dual system $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ of (A, B, C, D). Then $(A_{c,*}, B_{c,*}, C_{c,*}, D_{c,*})$ is a realization of G (see Definition 2.2). We show that $(A_{c,*}, B_{c,*}, C_{c,*}, D_{c,*})$ obtained this way has the expressions as given in the theorem. Notice that $\tilde{A} = A^*$, i.e., $A_{c,*} = A^*$.

1. By Theorem 5.1 the state space of the realization (A, B, C, D) is $\overline{\text{range}}H_{\tilde{G}_c}$. Thus the dual system $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ has the same state space. That is,

$$X_* = \overline{\text{range}}H_{\tilde{G}_c}.$$

2. The semigroup generated by A is defined as

$$e^{tA}f = P_+ e^{ts} f \quad (f \in X_*).$$

It is easy to verify

$$e^{tA^*}f = (e^{tA})^* f = P_{X_*} e^{-ts} f \quad (f \in X_*).$$

That is,

$$e^{tA_{c,*}} f = P_{X_*} e^{-ts} f \quad (f \in X_*).$$

By Theorem 5.1,

$$D(A_{c,*}) = D(A^*) = \left\{ P_{X_*} \frac{1}{1+s} h(s) \mid h \in X_* \right\},$$

$$A_{c,*} f = f - h \text{ for } f(s) = P_{X_*} \frac{1}{1+s} h(s) \in D(A_{c,*}),$$

and $A_{c,*}$ is well defined.

3. By the definition of the dual system (Definition 2.2), we have

$$\tilde{B} : U \to D(A)^{(\prime)}; \qquad u \longmapsto \tilde{B}(u)[\cdot] := \langle u,\ C(\cdot)\rangle.$$

For $x(s) = \frac{1}{1-s}[h(s) - h(1)] \in D(A)$ $(h \in X_*)$, we have

$$\tilde{B}(u)[x] = \langle u,\ Cx\rangle = \left\langle u, \sqrt{2\pi} \lim_{\substack{r\in\mathbb{R}\\ r\to\infty}} rx(r) \right\rangle = \sqrt{2\pi}\langle u, h(1)\rangle.$$

By Lemma 5.6, $\sqrt{2\pi}\langle u, h(1)\rangle_U = \frac{1}{\sqrt{2\pi}}\langle \frac{1}{1+s} u,\ h\rangle_{H_U^2(RHP)}$.

4. Now we compute $C_{c,*} = \tilde{C}$. Again use Definition 2.2:

$$D(\tilde{C}) = D(\tilde{A}) + (I - \tilde{A})^{-1}\tilde{B}U = D(A_{c,*}) + (I - A_{c,*})^{-1} B_{c,*} U,$$

and $\tilde{C}x_0$ is defined by

$$\begin{cases} \langle y, \tilde{C}x_0\rangle = B(y)[x_0], & x_0 \in D(A_{c,*}) \\ \langle \tilde{C}x_0, y\rangle = \langle u_0, C(I - A)^{-1} By\rangle, & x_0 = (I - A_{c,*})^{-1} B_{c,*} u_0,\ u_0 \in U,\ y \in Y. \end{cases}$$

Since by Theorem 5.7

$$B(y)[x] = \sqrt{2\pi}\langle y,\ (H_{G_c} h)(1)\rangle_Y \quad \left(x = P_X \frac{h(s)}{1+s} \in D(A_{c,*}) \right),$$

we have

$$\tilde{C}x = \sqrt{2\pi}(H_{G_c} h)(1) \text{ for } x = P_X \frac{h(s)}{1+s} \in D(A_{c,*}).$$

From Lemma 5.5 it follows that $\tilde{C}$ is well defined for $x \in D(A_{c,*})$.

Note that $C(I - A)^{-1}By = [\tilde{G}_c(1) - \tilde{G}_c(+\infty)]y$. Thus

$$\tilde{C}x_0 = [G_c(1) - G_c(+\infty)]u_0 \text{ for } x_0 = (I - A_{c,*})^{-1} B_{c,*} u_0.$$

Now we show that $(I - A_{c,*})^{-1}B_{c,*}u_0 = \frac{1}{\sqrt{2\pi}}P_{X_*}\frac{u_0}{1+s}$. Let $x \in D(A_{c,*}^*)$. Since by Theorem 5.7 $(I - A_{c,*}^*)^{-1}x = \frac{x-x(1)}{1-s}$, we have

$$\begin{aligned}
&\langle (I - A_{c,*})^{-1}B_{c,*}u_0,\ x\rangle = [(I - A_{c,*})^{-1}B_{c,*}u_0](x)\\
&= [B_{c,*}u_0]((I - A_{c,*}^*)^{-1}x)\\
&= [B_{c,*}u_0]\left(\frac{x - x(1)}{1-s}\right)\\
&= \frac{1}{\sqrt{2\pi}}\left\langle \frac{u_0}{1+s},\ x\right\rangle = \frac{1}{\sqrt{2\pi}}\left\langle P_{X_*}\frac{u_0}{1+s},\ x\right\rangle.
\end{aligned}$$

This shows that $(I - A_{c,*})^{-1}B_{c,*}u_0 = \frac{1}{\sqrt{2\pi}}P_{X_*}\frac{u_0}{1+s}$. Hence, to sum up, the operator $C_{c,*} = \tilde{C}$ is defined in the following way:

$$D(C_{c,*}) = P_{X_*}\left\{\frac{h(s)}{1+s} \mid h \in X_*\right\} + P_{X_*}\left\{\frac{u}{1+s} \mid u \in U\right\}.$$

If $x = P_{X_*}\frac{h(s)}{1+s}$, then

$$C_{c,*}x = \sqrt{2\pi}(H_{G_c}h)(1),$$

and if $x = P_{X_*}\frac{u}{1+s}$, then

$$C_{c,*}x = \sqrt{2\pi}[G_c(1) - G_c(+\infty)]u.$$

Note that by Lemma 5.5 $C_{c,*}x$ is also well defined for $x \in P_{X_*}\{\frac{u}{1+s} \mid u \in U\}$.

4. It is straightforward to get

$$D_{c,*} = \tilde{D}^* = ((G_c(+\infty))^*)^* = G_c(+\infty). \qquad \square$$

Note that the restricted and *-restricted shift realizations of admissible transfer functions in H^∞ are well posed in the sense of Curtain and Weiss [5] and Salamon [27]. Indeed we have the following corollary concerning the reachability and observability of the restricted and *-restricted shift realizations.

COROLLARY 5.10. 1. *The reachability operator of the restricted shift realization is given by*

$$\mathcal{R}_c : L_U^2[0, +\infty) \to X,\quad f \mapsto H_{G_c,RHP}\mathcal{L}f.$$

The observability operator of the restricted shift realization is given by

$$\mathcal{O}_c|_X : X \to L_Y^2[0, +\infty),\quad x \mapsto \mathcal{L}^{-1}x.$$

2. *The reachability operator of the *-restricted shift realization is given by*

$$\mathcal{R}_{c,*} : L_U^2[0, +\infty) \to X_*,\quad f \mapsto P_{X_*}\mathcal{L}f.$$

*The observability operator of the *-restricted shift realization is given by*

$$\mathcal{O}_{c,*} : X_* \to L_Y^2[0, +\infty),\quad x \mapsto \mathcal{L}^{-1}H_{G_c,RHP}x.$$

Here $\mathcal{L}$ denotes the Laplace transform.

Proof. This follows from Theorem 3.4 and Theorem 5.7 □

We categorize the state spaces of the restricted and *-restricted shift realizations here for later use.

PROPOSITION 5.11. *Let X and X_* be, respectively, the state spaces of restricted and *-restricted shift realizations of $G_c \in TLC^{U,Y}$. Then*

1. *if G_c is cyclic, then $X = H^2_Y(RHP)$ and $X_* = H^2_U(RHP)$;*
2. *if G_c is noncyclic, then $X = H^2_Y(RHP) \ominus Q_1 H^2_Y(RHP)$ and $X_* = H^2_U(RHP) \ominus Q_2 H^2_U(RHP)$, where $Q_1 \in H^\infty_{L(Y)}(RHP)$ and $Q_2 \in H^\infty_{L(U)}(RHP)$ are rigid functions;*
3. *if G_c is in $H^\infty_{L(U,Y)}(RHP)$, is strictly noncyclic, and has factorization $G_c = Q_1 F_1^* = \tilde{F}_2^* \tilde{Q}_2$, where $Q_1 \in H^\infty_{L(Y)}(RHP)$ and $Q_2 \in H^\infty_{L(U)}(RHP)$ are inner, Q_1 and F_1 are left coprime, and Q_2 and F_2 are also left coprime, then $X = H^2_Y(RHP) \ominus Q_1 H^2_Y(RHP)$ and $X_* = H^2_U(RHP) \ominus Q_2 H^2_U(RHP)$.*

Proof. This follows from Definition 4.5 and Theorems 4.8, 5.7, and 5.9.

6. Continuous-time input-normal, output-normal, parbalanced realizations and their asymptotic stability. Recall that a reachable and observable admissible system (A_c, B_c, C_c, D_c) is said to be input-normal if $\mathcal{W}_c = I$. It is output-normal if $\mathcal{M}_c = I$. The reachable and observable admissible systems are said to be parbalanced if

$$\mathcal{W}_c = \mathcal{M}_c.$$

Here $\mathcal{W}_c$ and $\mathcal{M}_c$ are, respectively, the reachability and observability Gramians of the system. Given a transfer function $G_c \in TLC^{U,Y}$, by Corollary 5.10 the restricted and *-restricted shift realizations are examples of, respectively, output-normal and input-normal realizations of G_c. Proposition 6.2 shows that up to unitary equivalence all observable input-normal and reachable output-normal realizations of an admissible transfer function G are up to unitary equivalence *-restricted and restricted shift realizations, respectively.

In this section we establish the existence of a parbalanced realization for any $G_c \in TLC^{U,Y}$ and study the stability properties of input-normal, output-normal, and parbalanced realizations.

A parbalanced realization of a continuous-time transfer function $G_c \in TLC^{U,Y}$ can be obtained from the map T in Theorem 3.1 applied to a discrete-time parbalanced realization of the corresponding discrete-time transfer function G_d. The existence of parbalanced realizations was shown by Young [30]. In [23] Young's results are cast into the continuous-time situation and the following theorem is proven.

THEOREM 6.1. 1. *For $G_c \in TLC^{U,Y}$, there exists a parbalanced realization $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$ of G_c. The state space of this realization is given by the closure of the range of the Hankel operator with symbol G_c, i.e., $X = \overline{\text{range}} H_{G_c}$. If $(\bar{A}_c, \bar{B}_c, \bar{C}_c, \bar{D}_c)$ is another parbalanced realization of G_c with state space $\bar{X}$, then (A_c, B_c, C_c, D_c) and $(\bar{A}_c, \bar{B}_c, \bar{C}_c, \bar{D}_c)$ are unitarily equivalent.*

2. *If in addition $G_c(s)$ is continuous on the extended $i\mathbb{R}$ (i.e. on $i\mathbb{R} \cup \{i\infty\}$) and is a compact operator for each $s \in i\mathbb{R}$, then there is a basis of $X = \overline{\text{range}} H_{G_c}$ on which the Gramians of the above realization have a diagonal matrix representation with its diagonal consisting of the Hankel singular values of G_c. We will call this realization a balanced realization of G_c.* □

6.1. Characterization of the realizations. Concerning the equivalence of different realizations, we have the following proposition.

PROPOSITION 6.2. 1. *Any two input-normal (output-normal) realizations of $G_c \in TLC^{U,Y}$ are unitarily equivalent. Hence every input-normal (output-normal) realization of G_c is unitarily equivalent to the *-restricted (restricted) shift realization of G_c.*

2. *An input-normal realization and an output-normal realization of* $G_c \in TLC^{U,Y}$ *are equivalent if and only if the Hankel operator* H_{G_c} *has closed range.*

3. *All reachable and observable admissible realizations of* G_c *are equivalent if and only if the Hankel operator* H_{G_c} *has closed range.*

Proof. Analogous results in the discrete-time case are shown in [24] (see Theorem 3.1, Corollary 3.1, and Proposition 4.1 therein). Applying Theorem 3.1, Theorem 3.4, and Proposition 4.3 to these results we have the proposition. □

A consequence of the proposition is that the study of input-normal (output-normal) realizations reduces to the study of the *-restricted (restricted) shift realizations. This point will be used repeatedly.

Part 2 of the proposition shows when the state-space isomorphism theorem holds. Note that the Hankel operator H_{G_c} to have closed range is a very strong condition. This condition can be stated in terms of the Douglas–Shapiro–Shields factorization of the transfer function G_c (see Theorem 4.10 and [11]).

6.2. Asymptotic stability. Now we turn to the study of stability properties of continuous-time systems and use the classes $C_{i,j}$ to describe different asymptotic stability properties of systems [28].

DEFINITION 6.3. *Let* $(e^{tA_c})_{t\geq 0}$ *be a semigroup of contractions on the Hilbert space* H. *Then*

1. $(e^{tA_c})_{t\geq 0} \in C_{0\cdot}$ *if* $\lim_{t\to\infty} e^{tA_c}h = 0$ *for all* $h \in H$,
2. $(e^{tA_c})_{t\geq 0} \in C_{\cdot 0}$ *if* $\lim_{t\to\infty} e^{tA_c^*}h = 0$ *for all* $h \in H$,
3. $(e^{tA_c})_{t\geq 0} \in C_{1\cdot}$ *if* $\lim_{t\to\infty} e^{tA_c}h \neq 0$ *for all* $h \in H$,
4. $(e^{tA_c})_{t\geq 0} \in C_{\cdot 1}$ *if* $\lim_{t\to\infty} e^{tA_c^*}h \neq 0$ *for all* $h \in H$.

We further set

$$C_{ij} = C_{i\cdot} \cap C_{\cdot j}, \qquad i, j = 0, 1. \qquad \square$$

The notions of stability that we consider are the following.

DEFINITION 6.4. *A continuous-time system* $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$ *is*

1. asymptotically stable *if for all* $x \in X$,

$$e^{tA_c}x \to 0$$

as $t \to \infty$, *i.e.,* $(e^{tA_c})_{t\geq 0} \in C_0$;

2. exponentially stable *if*

$$\omega := \inf\{\alpha \in \mathbb{R} \mid \text{there exists } M_\alpha \geq 0 \text{ such that } \|e^{tA_c}\| \leq M_\alpha e^{\alpha t} \ (t \geq 0)\} < 0.$$

The number ω *is called the* growth bound *of the semigroup.* □

We comment that the asymptotic and exponential stability of a system is preserved by system equivalence. Moreover, if two systems are unitarily equivalent, they will have the same growth bound.

An important result in [28, Prop. 9.1, p. 148] implies that a continuous-time system is asymptotically stable if and only if the corresponding discrete-time system is asymptotically stable.

PROPOSITION 6.5. *Let* $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$ *and* $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$ *such that*

$$(A_c, B_c, C_c, D_c) = T((A_d, B_d, C_d, D_d)).$$

Then for all $x \in X$,

$$\lim_{n\to\infty} \|A_d^n x\| = \lim_{t\to\infty} \|e^{tA_c}x\|$$

and

$$\lim_{n\to\infty} \|(A_d^*)^n x\| = \lim_{t\to\infty} \|e^{tA_c^*} x\|. \qquad \square$$

Therefore, the study of asymptotic stability of a continuous-time system reduces to the study of the asymptotic stability of the corresponding discrete-time system.

Now we state the main result of this section, which asserts that any admissible parbalanced realization of an admissible continuous-time transfer function is asymptotically stable.

THEOREM 6.6. *Let* $G_c \in TC^{U,Y}$. *Let* (A_b, B_b, C_b, D_b), (A_i, B_i, C_i, D_i), *and* (A_o, B_o, C_o, D_o) *be, respectively, a parbalanced, an input-normal, and an output-normal observable and reachable realization of* G_c. *Then*

1. (a) $(e^{tA_i})_{t\geq 0} \in C_{\cdot 0}$,
   (b) $(e^{tA_i})_{t\geq 0} \in C_{00}$ *if* $G_c^{\perp}$ *is strictly noncyclic,*
   (c) $(e^{tA_i})_{t\geq 0} \in C_{10}$ *if* $G_c^{\perp}$ *is cyclic,*
2. (a) $(e^{tA_o})_{t\geq 0} \in C_{0\cdot}$, *i.e., asymptotically stable,*
   (b) $(e^{tA_o})_{t\geq 0} \in C_{00}$ *if* G_c *is strictly noncyclic,*
   (c) $(e^{tA_o})_{t\geq 0} \in C_{01}$ *if* G_c *is cyclic,*
3. $(e^{tA_b})_{t\geq 0} \in C_{00}$.

Proof. The corresponding asymptotic stability results for discrete-time systems were obtained in Theorem 3.2 and Theorem 4.2 of [24]. Hence, combining those theorems with Proposition 6.5 and part 1 of Proposition 4.7, we have the theorem. $\square$

Since by Proposition 6.2 all reachable and observable realizations of G_c are equivalent when the Hankel operator H_{G_c} has closed range, and equivalent realizations have the same asymptotic stability properties, the theorem has the following corollary.

COROLLARY 6.7. *If the Hankel operator* H_{G_c} *has closed range, then all reachable, observable, and admissible realizations of* G_c *are asymptotically stable.* $\square$

7. Spectral minimality and exponential stability of input-normal, output-normal, and parbalanced realizations. This section aims to examine the exponential stability of continuous-time input-normal, output-normal, and parbalanced realizations of certain classes of transfer functions. The results are mainly based upon a detailed spectral analysis of input-normal and output-normal realizations. While the asymptotic stability properties of continuous-time systems can be obtained directly from the discrete-time case as we did in the previous section, exponential stability properties of continuous-time systems do not follow in the same way. However, we can relate the spectrum of the discrete-time system to that of the continuous-time system and thus establish the exponential stability results.

Recall that a continuous-time system (A_c, B_c, C_c, D_c) is exponentially stable if

$$\inf\{\alpha \in \mathbb{R} | \text{ there exists } M_\alpha \geq 0 \text{ such that } \|e^{tA_c}\| \leq M_\alpha e^{\alpha t} \text{ for } t \geq 0\} < 0.$$

The following proposition gives an interpretation of the growth bound of a semigroup in terms of the spectral radius of the semigroup (see, e.g., [21, p. 60]).

PROPOSITION 7.1. *Let* ω *be the growth bound of the semigroup* $(e^{tA_c})_{t\geq 0}$ *and* $r(e^{tA_c})$ *the spectral radius of* e^{tA_c}; *then*

$$r(e^{tA_c}) = e^{\omega t}$$

for $t \geq 0$. $\square$

Note that it follows from this proposition that equivalent systems have the same growth bound.

7.1. Spectral analysis. Thus we have to investigate the spectral properties of a continuous-time linear system (A_c, B_c, C_c, D_c) in order to study its exponential stability.

The way we do this is to relate the spectral properties of (A_c, B_c, C_c, D_c) to those of the corresponding discrete-time system (A_d, B_d, C_d, D_d). First we have the following relation between $\sigma(A_d)$ and $\sigma(A_c)$.

PROPOSITION 7.2. *Let A_c be the infinitesimal generator of a semigroup of contractions and A_d the co-generator such that $A_c = (A_d - I)(A_d + I)^{-1}$. Then*

$$\sigma_p(A_c) = \left\{ \frac{z-1}{z+1} : \ z \in \sigma_p(A_d) \right\} \text{ and } \sigma_p(A_d) = \left\{ \frac{1+s}{1-s} : \ s \in \sigma_p(A_c) \right\},$$

$$\sigma(A_c) = \left\{ \frac{z-1}{z+1} : \ z \in \sigma(A_d),\ z \neq -1 \right\} \text{ and } \sigma(A_d) \setminus \{-1\} = \left\{ \frac{1+s}{1-s} : \ s \in \sigma(A_c) \right\}.$$

Proof. First note that $1 \notin \sigma(A_c)$ since e^{tA_c} is a semigroup of contractions and that by Theorem 3.1,

$$A_c x = (A_d - I)(A_d + I)^{-1}x = (A_d + I)^{-1}(A_d - I)x \text{ for } x \in D(A_c),$$

where $D(A_c) = \text{range}(A_d + I)$. Hence the following relations hold:

$$\begin{aligned} (7.1) \quad (sI - A_c)(A_d + I)x &= [sI - (A_d - I)(A_d + I)^{-1}](A_d + I)x \\ &= [s(A_d + I) - (A_d - I)]x \\ &= (1-s)\left(\frac{1+s}{1-s}I - A_d\right)x, \qquad x \in X_d,\ s \neq 1; \end{aligned}$$

$$\begin{aligned} (7.2) \quad (A_d + I)(sI - A_c)x &= (A_d + I)[sI - (A_d - I)(A_d + I)^{-1}]x \\ &= (1-s)\left(\frac{1+s}{1-s}I - A_d\right)x, \qquad x \in D(A_c),\ s \neq 1. \end{aligned}$$

The equations (7.1) and (7.2) show that

$$\sigma_p(A_d) = \left\{ \frac{1+s}{1-s} : \ s \in \sigma_p(A_c) \right\}.$$

Now if $\frac{1+s}{1-s} \notin \sigma(A_d)$, i.e., if $(\frac{1+s}{1-s}I - A_d)^{-1}$ exists and is bounded, then

$$(A_d + I)\left(\frac{1+s}{1-s}I - A_d\right)^{-1} = \left(\frac{1+s}{1-s}I - A_d\right)^{-1}(A_d + I).$$

Thus by (7.1) and (7.2)

$$\begin{aligned} (7.3) \quad (sI - A_c)^{-1}x &= (1-s)^{-1}(A_d + I)\left(\frac{1+s}{1-s}I - A_d\right)^{-1}x \\ &= (1-s)^{-1}\left(\frac{1+s}{1-s}I - A_d\right)^{-1}(A_d + I)x, \ x \in D(A_c). \end{aligned}$$

So $(sI - A_c)^{-1}$ is bounded and densely defined, i.e., $s \notin \sigma(A_c)$. Hence

$$(7.4) \qquad \left\{ \frac{1+s}{1-s} : \ s \in \sigma(A_c) \right\} \subseteq \sigma(A_d).$$

On the other hand, if $s \neq 1$ and $s \notin \sigma(A_c)$, then $(sI - A_c)^{-1}$ is a bounded operator. It is easy to verify in this case that

$$\left(\frac{1+s}{1-s}I - A_d\right)^{-1} x = \frac{(1-s)^2}{2}(sI - A_c)^{-1} + \frac{1-s}{2}x, \qquad x \in D(A_c).$$

In fact from (7.3) we have

$$\left(\frac{1+s}{1-s}I - A_d\right)\left[\frac{(1-s)^2}{2}(sI - A_c)^{-1} + \frac{1-s}{2}I\right]x$$

$$= \frac{(1-s)^2}{2}\left(\frac{1+s}{1-s}I - A_d\right)(sI - A_c)^{-1}x + \frac{1-s}{2}\left(\frac{1+s}{1-s}I - A_d\right)x$$

$$= \frac{(1-s)^2}{2}(1-s)^{-1}(A_d + I)x + \frac{1-s}{2}\left(\frac{1+s}{1-s}I - A_d\right)x$$

$$= x, \quad x \in D(A_c).$$

Similarly,

$$\left[\frac{(1-s)^2}{2}(sI - A_c)^{-1} + \frac{1-s}{2}x\right]\left(\frac{1+s}{1-s}I - A_d\right)x = x, \qquad x \in D(A_c).$$

Thus $\frac{1+s}{1-s} \notin \sigma(A_d)$. So we have

$$\left\{s : \frac{1+s}{1-s} \in \sigma(A_d)\right\} \subseteq \sigma(A_c). \tag{7.5}$$

Combining (7.4) and (7.5) we have that

$$\sigma(A_c) = \left\{s : \frac{1+s}{1-s} \in \sigma(A_d)\right\} = \left\{\frac{z-1}{z+1} : z \in \sigma(A_d), z \neq -1\right\},$$

which implies

$$\sigma(A_d) \setminus \{-1\} = \left\{\frac{1+s}{1-s} : s \in \sigma(A_c)\right\}. \qquad \square$$

In our application of the proposition, A_c is the state propagation operator of a continuous-time system $(A_c, B_c, C_c, D_c) \in C_X^{U,Y}$ and A_d is the state propagation operator of the corresponding discrete-time system $(A_d, B_d, C_d, D_d) \in D_X^{U,Y}$, which is related to (A_c, B_c, C_c, D_c) by

$$(A_c, B_c, C_c, D_c) = T((A_d, B_d, C_d, D_d)),$$

where T is the bilinear mapping in Theorem 3.1.

A powerful tool in spectral analysis is the spectral mapping theorem for C_0 operators (see, e.g., [22, p. 74]). A contraction $W \in L(M)$, where M is a separable Hilbert space, is called a C_0 operator, denoted $W \in C_0$, if there exists no subspace $V \in M$ such that $W|_V : V \to V$ is unitary and if there exists an inner function $m \in H^\infty(\mathbb{D})$ such that $m(W) = 0$. The least common divisor of all such inner functions is called the minimal function of W, denoted

m_W, which is still an inner function such that $m_W(W) = 0$. Note that if W is a C_0 operator, W is unitarily equivalent to a left shift S^* restricted to a left invariant space of the form $H_Y^2(\mathbb{D}) \ominus QH_Y^2(\mathbb{D})$, where Q is inner (see, [22, p. 72]). It can be seen that minimal functions are the generalizations of minimal polynomials of matrices. As in the matrix case, the spectrum of a C_0 operator is given by the "zeros" of its minimal function in the following sense (see [22, p. 72]).

LEMMA 7.3. *If $W \in C_0$, then $\sigma(W) = \sigma(m_W)$ and $\sigma_p(W) = \sigma(m_W) \cap \mathbb{D}$, where for an inner function $Q \in H^\infty_{L(Y)}(\mathbb{D})$, Y is a Hilbert space, and the spectrum of Q is defined as*

$$\sigma(Q) = \left\{ \lambda \in \overline{\mathbb{D}} \mid \lim_{\delta \to 0} \inf_{\substack{|\xi - \lambda| < \delta \\ \xi \in \mathbb{D}}} \inf_{\substack{\|y\| = 1 \\ y \in Y}} \|Q(\xi) y\| = 0 \right\}. \qquad \square$$

Given a C_0 operator W and a function $\phi \in H^\infty(\mathbb{D})$, the operator

$$\phi(W) := \lim_{r<1,\ r \to 1} \phi(rW)$$

is well-defined. The following theorem relates the spectra of these two operators (see [22, p. 74]).

THEOREM 7.4 (the spectral mapping theorem). *Let $\phi \in H^\infty(\mathbb{D})$ and $W \in C_0$. Then*

$$\sigma(\phi(W)) \subseteq \left\{ \xi \in \mathbb{C} \mid \inf_{z \in \mathbb{D}} (|\phi(z) - \xi| + |m_W(z)|) = 0 \right\},$$

where m_W is the minimal function of W. $\square$

7.2. Spectral minimality. We are going to use these results to transpose the spectral properties of the discrete-time input- and output-normal realizations to those of the continuous-time case. First we recall the discrete-time results. Assume that the input and output spaces are of finite dimension. If the transfer function G_d is such that $G_d^\perp$ is strictly noncyclic, then G_d has a pseudomorphic continuation of bounded type to the unit disk $\mathbb{D}$ (see [11]). Take this continuation as the definition of G_d on $\mathbb{D}$, wherever defined. Consider the analytic continuation of the extended G_d. Let $\sigma_s(G_d)$ be the set of points at which G_d has no analytic continuation. We are interested in the relationship between $\sigma_s(G_d)$ and $\sigma(A_d)$. The following theorem shows $\sigma_s(G_d) = \sigma(A_d)$ for input-normal or output-normal realizations. If G_d is not strictly noncyclic, the spectrum of A_d can also be characterized (see [24] and [11]).

THEOREM 7.5. *Let $G_d \in TLD^{U,Y}$ with U and Y finite dimensional and let $(A_{d,o}, B_{d,o}, C_{d,o}, D_{d,o})$, $(A_{d,i}, B_{d,i}, C_{d,i}, D_{d,i})$, and $(A_{d,b}, B_{d,b}, C_{d,b}, D_{d,b})$ be, respectively, an output-normal, an input-normal, and a parbalanced realization of G_d.*

1. *If $G_d^\perp$ is in $H^\infty_{L(Y)}(\mathbb{D})$ and strictly noncyclic, then $(A_{d,o}, B_{d,o}, C_{d,o}, D_{d,o})$, $(A_{d,i}, B_{d,i}, C_{d,i}, D_{d,i})$, and $(A_{d,b}, B_{d,b}, C_{d,b}, D_{d,b})$ are* spectrally minimal, *i.e.,*

$$\sigma(A_{d,o}) = \sigma(A_{d,i}) = \sigma(A_{d,b}) = \sigma_s(G_d).$$

In this case $A_{d,o}$, $A_{d,i}$, and $A_{d,b}$ are all C_0 operators and have the same minimal function—say, m. Moreover, if $G_c(e^{it}) = Q_1(e^{it})(e^{it}F_1(e^{it}))^*$ is the Douglas–Shapiro–Shields factorization of G_d (see Theorem 4.8) and $\tilde{G}_d(e^{it}) = Q_2(e^{it})(e^{it}F_2(e^{it}))^*$ is the factorization of $\tilde{G}_d$, then the following equalities hold:

$$\sigma(m) = \sigma_s(G_d) = (\sigma(Q_1))^* = \sigma(Q_2),$$

$$\sigma_p(A_{d,o}) = \{\overline{\lambda} \in \mathbb{D} \mid \operatorname{Ker} Q_1(\lambda)^* \neq \{0\}\},$$

and

$$\sigma_p(A_{d,i}) = \{\lambda \in \mathbb{D} | \operatorname{Ker} Q_2(\lambda) \neq \{0\}\},$$

where $(\sigma(Q_1))^* = \{\bar{\lambda} | \lambda \in \sigma(Q_1)\}$ *and*

$$\sigma(Q_i) = \{\lambda \in \overline{\mathbb{D}} \mid \liminf_{\substack{\xi \to \lambda \\ \xi \in \mathbb{D}}} \inf_{\substack{\|y\|=1 \\ y \in Y}} \|Q_i(\xi) y\| = 0\} \quad (i = 1, 2).$$

2. *If* $G_d^{\perp}$ *is noncyclic but not strictly noncyclic, then*

$$\sigma(A_o) = \sigma(A_i) = \overline{\mathbb{D}}.$$

3. *If* $G_d^{\perp}$ *is cyclic, then*

$$\sigma_p(A_o) = \mathbb{D}, \quad \sigma_p(A_i) = \emptyset, \quad \sigma(A_o) = \sigma(A_i) = \overline{\mathbb{D}} \ . \qquad \square$$

Corresponding to this theorem we have the following continuous-time analogue. For a strictly noncyclic continuous-time transfer function G_c, we define $\sigma_s(G_c)$ similarly as in the discrete-time situation. Specifically, G_c has a pseudomorphic continuation of bounded type to LHP (see Theorem 4.8), which is taken to be the definition of G_c on LHP. We consider the analytic continuation of the redefined G_c and denote by $\sigma_s(G_c)$ the set of points in the complex plane at which G_c has no analytic continuation.

We note that results in part 1 of the following theorem can be found in the thesis by Gearheart [12] and a paper by Moeller [19].

THEOREM 7.6. *Let* $G_c \in TLC^{U,Y}$ *and let* $(A_{c,o}, B_{c,o}, C_{c,o}, D_{c,o}) \in C_X^{U,Y}$ *be an output-normal realization with* U *and* Y *finite dimensional. Then*

1. *if* G_c *is in* $H^{\infty}_{L(U,Y)}(RHP)$ *and is strictly noncyclic with factorization* $G_c = Q_1 F_1^*$, *where* $Q_1 \in H^{\infty}_{L(Y)}(RHP)$ *is inner and* Q_1 *and* $F_1 \in H^{\infty}_{L(Y,U)}(RHP)$ *are weakly coprime, then* $\lambda \notin \sigma(e^{A_{c,o}})$, $|\lambda| < 1$, *if and only if for*

$$w_n = -\log \bar{\lambda} + 2\pi n i, \qquad n \in \mathbb{Z},$$

$Q_1(w_n)$ *is invertible for all* $n \in \mathbb{Z}$ *and*

$$\sup_{-\infty < n < \infty} \|Q_1(w_n)^{-1}\| < \infty.$$

For $|\lambda| = 1$, $\lambda \notin \sigma(e^{A_{c,o}})$ *if and only if there exists a* $\delta > 0$ *and* $M > 0$ *such that* $Q_1(w_n)^{-1}$ *exists for all* $n \in \mathbb{Z}$ *and* $Q_1(s)^{-1}$ *is bounded by* M *in the* δ *neighborhood of each point* w_n.

For the point spectrum, we have

$$\sigma_p(e^{A_{c,o}}) \setminus \{0\} = \{e^{-\bar{s}} : s \in RHP \text{ and } \operatorname{Ker} Q_1(s)^* \neq \{0\}\}.$$

2. *under the same assumptions on* G_c *as in* 1., *we have*

$$\sigma(A_{c,o}) = \{-\bar{s} : s \in \sigma(Q_1)\} = \sigma_s(G_c)$$
$$\sigma_p(A_{c,o}) = \{-\bar{s} : s \in RHP \text{ and } \operatorname{Ker} Q_1(s)^* \neq \{0\}\}.$$

3. *if* G_c *is noncyclic but not strictly noncyclic, then*

$$\sigma(A_{c,o}) = \text{the closed left half plane.}$$

4. *if G_c is cyclic, then*

$$\sigma(e^{A_{c,o}}) = \bar{\mathbb{D}}, \qquad \sigma_p(e^{A_{c,o}}) \setminus \{0\} = \mathbb{D} \setminus \{0\},$$
$$\sigma(A_{c,o}) = \textit{the closed left half plane}, \quad \sigma_p(A_{c,o}) = LHP.$$

Proof. Without loss of generality we may assume that $(A_{c,o}, B_{c,o}, C_{c,o}, D_{c,o})$ is the restricted shift realization. We write (A_c, B_c, C_c, D_c) for $(A_{c,o}, B_{c,o}, C_{c,o}, D_{c,o})$. Let $G_d(z) = G_c(\frac{z-1}{z+1})$ for $z \in \mathbb{D}_e$ and let

$$G_d^{\perp}(z) = z^{-1}[G_d(z^{-1}) - G_d(\infty)] = z^{-1}\left[G_c\left(\frac{1-z}{1+z}\right) - G_c(1)\right], \quad z \in \mathbb{D}.$$

Suppose (A_d, B_d, C_d, D_d) is the restricted shift realization of the discrete-time transfer function G_d. We use the mapping T defined in Theorem 3.1.

1. The formula for $\sigma(e^{A_c})$ can be found in [19] in the case $0 < |\lambda| < 1$. If $|\lambda| = 1$ or $\lambda = 0$, see [12].

For the formula of $\sigma_p(e^{A_c})$ see the proof of 2. below.

2. Note that by Proposition 4.7 $G_d^{\perp}$ is also strictly noncyclic and has a factorization $G_d = Q_{d,1}F_{d,1}^*$, where $Q_{d,1}(z) = Q_1(\frac{1-z}{1+z})$ and $F_{d,1}(z) = F_1(\frac{1-z}{1+z})$. The spectra of $Q_{d,1}$ and Q_1 are related as

$$\sigma(Q_{d,1}) = \left\{\frac{1-s}{1+s} \mid s \in \sigma(Q_1)\right\},$$

and the sets $\sigma_s(G_d)$ and $\sigma_s(G_c)$ are related as

$$\sigma_s(G_c) = \left\{s : \frac{1+s}{1-s} \in \sigma_s(G_d)\right\}.$$

Then the equalities about $\sigma(A_c)$ and $\sigma_s(G_c)$ follow from Proposition 7.2 and Theorem 7.5. Similarly the expression for $\sigma_p(A_c)$ also follows from Proposition 7.2 and Theorem 7.5.

The point spectrum $\sigma_p(e^{A_c})$ can be obtained by the general formula (see [26, Thm. 2.4, p. 46])

$$\sigma_p(e^{tA_c}) \setminus \{0\} = e^{\sigma_p(tA_c)}.$$

3. This also follows from Proposition 7.2 and Theorem 7.5.

4. We offer a direct proof here, although the result again follows from Proposition 7.2 and Theorem 7.5.

If G_c is cyclic, then the state space is $X_c = H_Y^2(RHP)$. It is easy to see that for any $\mu \in LHP, t \geq 0$, and $y \in Y$ we have $\frac{y}{s-\mu} \in X_c = H_Y^2(RHP)$ and

$$\frac{e^{ts} - e^{t\mu}}{s-\mu} y \in H_Y^2(LHP) = (H_Y^2(RHP))^{\perp},$$

where the orthogonal complement is taken in $L_Y^2(i\mathbb{R})$. Therefore,

$$e^{tA_c}\frac{1}{s-\mu}y = P_+\frac{e^{ts}}{s-\mu}y = P_+\left[\frac{e^{t\mu}}{s-\mu}y + \frac{e^{ts}-e^{t\mu}}{s-\mu}y\right] = \frac{e^{t\mu}}{s-\mu}y.$$

Hence $e^{t\mu} \in \sigma_p(e^{tA_c})$. This shows that $\sigma_p(e^{tA_c}) \setminus \{0\} = \mathbb{D} \setminus \{0\}$ and hence $\sigma(e^{tA_c}) = \bar{\mathbb{D}}$.

Also for any $\mu \in LHP$ and $y \in Y$ we have $h = \frac{1-\mu}{s-\mu} y \in H_Y^2(RHP)$ and

$$\frac{y}{s-\mu} = \frac{h(s)-h(1)}{1-s}.$$

Hence $\frac{y}{s-\mu} \in D(A_c)$. Using the definition of A_c we have

$$A_c \frac{y}{s-\mu} = \frac{sy}{s-\mu} - \lim_{\substack{r \in \mathbb{R} \\ r \to \infty}} \frac{ry}{r-\mu} = \mu \frac{y}{s-\mu}.$$

Therefore $\mu \in \sigma_p(A_c)$. This shows that $\sigma_p(A_c) = LHP$ and hence $\sigma(A_c) = \overline{LHP}$. □

For input-normal realizations we have results analogous to the results above. The proof is similar to the proof of the previous results.

THEOREM 7.7. *Let $G_c \in TLC^{U,Y}$ and let $(A_{c,i}, B_{c,i}, C_{c,i}, D_{c,i}) \in C_X^{U,Y}$ be an observable input-normal realization with U and Y finite dimensional. Then*

1. *if G_c is in $H^\infty_{L(U,Y)}(RHP)$ and is strictly noncyclic with $\tilde{G}_c = Q_2 F_2^*$, where $Q_2 \in H^\infty_{L(U)}(RHP)$ is inner and Q_2 and $F_2 \in H^\infty_{L(U,Y)}(RHP)$ are weakly coprime, then $\lambda \notin \sigma(e^{A_{c,i}})$, $|\lambda| < 1$, if and only if for*

$$w_n = -\log \lambda + 2\pi n i, \qquad n \in \mathbb{Z},$$

$Q_2(w_n)$ is invertible for all $n \in \mathbb{Z}$ and

$$\sup_{-\infty < n < \infty} \|Q_2(w_n)^{-1}\| < \infty.$$

For $|\lambda| = 1$, $\lambda \notin \sigma(e^{A_{c,i}})$ if and only if there exists a $\delta > 0$ and $M > 0$ such that $Q_2(w_n)^{-1}$ exists for all $n \in \mathbb{Z}$ and $Q_2(s)^{-1}$ is bounded by M in a δ neighborhood of each point w_n. As to the point spectrum, we have

$$\sigma_p(e^{A_{c,i}}) \setminus \{0\} = \{e^{-s} : s \in RHP, \ \mathrm{Ker} Q_2(s) \neq \{0\}\}.$$

2. *Under the same assumption as in 1., for the generator $A_{c,i}$ we have*

$$\sigma(A_{c,i}) = \{-\lambda : \lambda \in \sigma(Q_2)\} = \sigma_s(G_c),$$
$$\sigma_p(A_{c,i}) = \{-s : s \in RHP, \ \mathrm{Ker} Q_2(s) \neq \{0\}\}.$$

3. *If $\tilde{G}_c$ is noncyclic and $\overline{\mathrm{range}}(H_{\tilde{G}_c}) = (Q_2 H_U^2(RHP))^\perp$, where Q_2 is a non-inner rigid function, then*

$$\sigma(A_{c,i}) = \text{the closed left half plane.}$$

4. *If G_c is cyclic, then*

$$\sigma(e^{A_{c,i}}) = \bar{\mathbb{D}}, \qquad \sigma_p(e^{A_{c,i}}) \setminus \{0\} = \emptyset,$$
$$\sigma(A_{c,i}) = \text{the closed left half plane.}$$ □

The following proposition gives the spectral properties of parbalanced realizations in the case of strictly noncyclic transfer functions.

PROPOSITION 7.8. *If $G_c \in H^\infty_{L(U,Y)}(RHP)$ is strictly noncyclic with finite dimensional U and Y, then*

$$\sigma(A_{c,o}) = \sigma(A_{c,i}) = \sigma(A_{c,b}) = \sigma_s(G_c),$$

where $(A_{c,b}, B_{c,b}, C_{c,b}, D_{c,b})$ is a parbalanced realization of G_c.

Proof. The analogous results in the discrete-time case are proven in [24, Cor. 4.3]. Since $\sigma_s(G_c) = \{\frac{z-1}{z+1} | \ z \in \sigma_s(G_d), \ z \neq -1\}$, where $G_d(z) = G_c(\frac{z-1}{z+1})$, $(z \in \mathbb{D}_e)$, the statement follows from Propositions 4.7 and 7.2 and Theorem 7.5. □

7.3. Exponential stability. Before we can give a criterion for the exponential stability of input- and output-normal realizations, we need some results concerning the relation between the spectrum of a semigroup and the spectrum of its generator. The following lemma can be deduced from [9, p. 622] (see also [21, p. 84]).

LEMMA 7.9. *Let e^{tA} be a strongly continuous semigroup of operators on a Hilbert space X with infinitesimal generator A. If $\sigma(e^{tA}) \subseteq \{\lambda : |\lambda| \leq e^{\alpha t}\}$ $(t > 0)$, then $\sigma(A) \subseteq \{s : Re(s) \leq \alpha\}$.* □

Note that in particular if $\|e^{tA}\| \leq Me^{\alpha t}$ for some $M > 0$, then $r(e^{tA}) \leq e^{\alpha t}$ and hence $\sigma(A) \subseteq \{s : Re(s) \leq \alpha\}$, where $r(e^{tA})$ is the spectral radius.

It is well known that in general the converse of the lemma is not true (see [21, Chap. A-III]). However, the converse can be proven in some particular cases.

PROPOSITION 7.10. *Let e^{tA} be a strongly continuous semigroup of contractions on a Hilbert space X with infinitesimal generator A. Let A_d be its co-generator; that is, A_d is a contraction with $-1 \notin \sigma(A_d)$ and*

$$Ax = (A_d - I)(I + A_d)^{-1}x \quad (x \in D(A) = \text{range}(I + A_d)).$$

Assume that A_d is a C_0 operator with minimal function m. Then $\sigma(e^{tA}) \subseteq \{\lambda : |\lambda| \leq e^{\alpha t}\}$ $(t > 0)$, if and only if $\sigma(A) \subseteq \{s : Re(s) \leq \alpha\}$. Here α is a real number.

Proof. The necessity part follows from Lemma 7.9.

Now assume $\sigma(A) \subseteq \{s : Re(s) \leq \alpha\}$. Since $\sigma(e^{tA}) \subseteq \{\lambda : |\lambda| \leq 1\}$, we may assume $\alpha < 0$.

By Lemma 7.3, we have $\sigma(A_d) = \sigma(m)$. On the other hand Proposition 7.2 shows that

$$\sigma(A_d) \setminus \{-1\} = \left\{\frac{1+s}{1-s} : \ s \in \sigma(A)\right\}.$$

Since $\sigma(A) \subseteq \{s : Re(s) \leq \alpha\}$, we have

$$\sigma(A_d) \setminus \{-1\} \subseteq \left\{\frac{1+s}{1-s} : \ Re(s) \leq \alpha\right\}.$$

Thus

$$\sigma(m) \setminus \{-1\} \subseteq \left\{\frac{1+s}{1-s} : \ Re(s) \leq \alpha\right\}.$$

Let $\xi = \frac{1+s}{1-s}$. Then $Re(s) \leq \alpha$ if and only if $|\xi - \frac{\alpha}{2-\alpha}| \leq 1 + \frac{\alpha}{2-\alpha}$. This shows that

$$\sigma(m) \subseteq \left\{\xi : \ |\xi - \frac{\alpha}{2-\alpha}| \leq 1 + \frac{\alpha}{2-\alpha}\right\}.$$

Therefore if $\xi \in \bar{\mathbb{D}}$ and $|\xi - \frac{\alpha}{2-\alpha}| > 1 + \frac{\alpha}{2-\alpha}$, then $\xi \notin \sigma(m)$. Hence there exist $\delta_1 > 0$ and $\delta_2 > 0$ such that

$$|m(z)| \geq \delta_1 \text{ for any } z \in \mathbb{D} \text{ satisfying } |z - \xi| \leq \delta_2.$$

Now for each $t > 0$, let $u(z) = e^{t\frac{z-1}{z+1}}$. Then $u \in H^\infty(\mathbb{D})$ and $e^{tA} = u(A_d)$. Using the spectral mapping theorem (Theorem 7.4) we have (note again that if $\xi = \frac{1+s}{1-s}$, then $Re(s) \leq \alpha$ if and only if $|\xi - \frac{\alpha}{2-\alpha}| \leq 1 + \frac{\alpha}{2-\alpha}$)

$$\sigma(e^{tA}) = \sigma(u(A_d))$$

$$\subseteq \left\{\lambda : \inf_{\xi\in\mathbb{D}}\{|u(\xi) - \lambda| + |m(\xi)|\} = 0\right\}$$

$$= \left\{\lambda : \inf_{|\xi - \frac{\alpha}{2-\alpha}|\le 1+\frac{\alpha}{2-\alpha}}\{|u(\xi) - \lambda| + |m(\xi)|\} = 0\right\}$$

$$\subseteq \left\{\lambda : \inf_{|\xi - \frac{\alpha}{2-\alpha}|\le 1+\frac{\alpha}{2-\alpha}}\{|u(\xi) - \lambda|\} = 0\right\}$$

$$= \left\{\lambda : \inf_{|\xi - \frac{\alpha}{2-\alpha}|\le 1+\frac{\alpha}{2-\alpha}}\{|e^{t\frac{\xi-1}{\xi+1}} - \lambda|\} = 0\right\}$$

$$= \text{closure}\left\{e^{t\frac{\xi-1}{\xi+1}} : \ |\xi - \frac{\alpha}{2-\alpha}| \le 1 + \frac{\alpha}{2-\alpha}\right\}$$

$$= \text{closure}\{e^{st} : \ Re(s) \le \alpha\}$$

$$= \{\lambda : \ |\lambda| \le e^{\alpha t}\}, \quad (t > 0).$$

This completes the proof. □

We are ready to show when an input- or output-normal realization is exponentially stable. For exponentially stable realizations we also characterize the growth bound in terms of the analyticity of the transfer function. The results remarkably resemble the related results for finite dimensional systems.

THEOREM 7.11. *Let G_c be in $H^\infty_{L(U,Y)}(RHP)$ with finite dimensional U and Y. Then an input-normal or output-normal realization of G_c is exponentially stable if and only if G_c is strictly noncyclic and there is $\alpha < 0$ such that G_c has analytic continuation on $Re(s) > \alpha$.*

In this case the growth bound is given by

$$\omega = \inf\{\alpha : \ G_c \text{ has analytic continuation on } Re(s) > \alpha\}.$$

Proof. We prove the theorem for output-normal realizations. The proof in the input-normal case is exactly the same. For output-normal realizations, it suffices to prove the result for the restricted shift realization.

Thus we assume that the restricted shift realization (A, B, C, D) of G_c is exponentially stable. Hence there are $\alpha < 0$ and $M > 0$ such that

$$\|e^{tA}\| \le Me^{\alpha t} \text{ for } t \ge 0.$$

Then by the remark after Lemma 7.9 $\sigma(A) \subseteq \{s|\ Re(s) \le \alpha\}$. As $\alpha < 0$, from Theorem 7.6 it follows that G_c has to be strictly noncyclic since otherwise $\sigma(A) = \overline{LHP}$. Now applying Proposition 7.8 we have

$$\sigma_s(G_c) = \sigma(A) \subseteq \{s : \ Re(s) \le \alpha\}.$$

Hence G_c has analytic continuation on $Re(s) > \alpha$. This also shows that

$$\inf\{\alpha' : \ G_c \text{ has analytic continuation on } Re(s) > \alpha'\}$$

is not greater than the growth bound of (A, B, C, D).

Conversely, assume that G_c is strictly noncyclic and there is $\alpha < 0$ such that G_c has an analytic continuation on $Re(s) > \alpha$. Let (A, B, C, D) be the restricted shift realization of G_c and (A_d, B_d, C_d, D_d) be the discrete-time restricted shift realization of $G_d(z) = G_c(\frac{z-1}{z+1})$. Note that $(A, B, C, D) = T((A_d, B_d, C_d, D_d))$.

Again by Proposition 7.8 we have

$$\sigma(A) = \sigma_s(G_c).$$

Therefore $\sigma(A) \subseteq \{s : Re(s) \leq \alpha\}$. Note that G_d is also strictly noncyclic. It follows from Theorem 7.5 that A_d is a C_0 operator. Now we can apply Proposition 7.10 to get

$$\sigma(e^{tA}) \subseteq \{\lambda : |\lambda| \leq e^{\alpha t}\}.$$

This shows that $r(e^{tA}) \leq e^{\alpha t}$. Thus by Proposition 7.1, (A, B, C, D) is exponentially stable. This also implies that the growth bound of (A, B, C, D) is not greater than

$$\inf\{\alpha' : G_c \text{ has analytic continuation on } Re(s) > \alpha'\}.$$

The proof is now complete. □

The following proposition shows that for strictly noncyclic transfer functions parbalanced realizations have the same exponential stability properties as input- and output-normal realizations.

PROPOSITION 7.12. *Let $G_c \in H^\infty_{L(U,Y)}(RHP)$ be strictly noncyclic with U and Y finite dimensional and let $(A_{c,o}, B_{c,o}, C_{c,o}, D_{c,o})$, $(A_{c,i}, B_{c,i}, C_{c,i}, D_{c,i})$, and $(A_{c,b}, B_{c,b}, C_{c,b}, D_{c,b})$ be, respectively, an output-normal, an input-normal, and a parbalanced realization of G_c. Then the following are equivalent:*

1. *$(A_{c,o}, B_{c,o}, C_{c,o}, D_{c,o})$ is exponentially stable with growth bound α;*
2. *$(A_{c,i}, B_{c,i}, C_{c,i}, D_{c,i})$ is exponentially stable with growth bound α;*
3. *$(A_{c,b}, B_{c,b}, C_{c,b}, D_{c,b})$ is exponentially stable with growth bound α.*

Proof. By Theorem 7.11, 1. and 2. are equivalent. Hence we need only to prove the equivalence of 1. and 3. Assume that 1. is true. Then there exist $M > 0$ and $\alpha < 0$ such that

$$\|e^{tA_{c,o}}\| \leq Me^{t\alpha} \quad (t > 0).$$

From the remark after Lemma 7.9 it follows that $\sigma(A_{c,o}) \subseteq \{s | \ Re(s) \leq \alpha\}$. Since now by Proposition 7.8 $\sigma(A_{c,b}) = \sigma(A_{c,o})$ we have

$$\sigma(A_{c,b}) \subseteq \{s \mid \ Re(s) \leq \alpha\}.$$

Let $A_{d,b} = (I + A_{c,b})(I - A_{c,b})^{-1}$ be the propagation operator of the corresponding discrete-time parbalanced realization of $G_d(z) = G_c(\frac{z-1}{z+1})$ $(z \in \mathbb{D}_e)$. That is, $A_{d,b}$ is the co-generator of the semigroup $e^{tA_{c,b}}$. Note that $G_d^\perp$ is strictly noncyclic. By Theorem 7.5 $A_{d,b}$ is a C_0 operator. Therefore it follows from Proposition 7.10 that

$$\sigma(e^{tA_{c,b}}) \subseteq \{\lambda : |\lambda| \leq e^{\alpha t}\}.$$

This, by Proposition 7.1, shows that $(A_{c,b}, B_{c,b}, C_{c,b}, D_{c,b})$ is exponentially stable with growth bound no greater than α and hence no greater than the growth bound of $(A_{c,o}, B_{c,o}, C_{c,o}, D_{c,o})$.

If we assume 3., a similar argument will lead to 1. □

If the Hankel operator H_{G_c} has closed range, then by Proposition 6.2 all reachable, observable, and admissible realizations of G_c are equivalent. Hence we have the following corollary.

COROLLARY 7.13. *Assume that the spaces U and Y are finite dimensional and the Hankel operator H_{G_c} has closed range. Then a reachable, observable, and admissible realization of G_c is exponentially stable if and only if there is a number $\alpha < 0$ such that G_c is strictly noncyclic and can be analytically continued to the half plane $\{s|\ Re(s) > \alpha\}$. The growth bound of these systems is* $\inf \alpha$. □

8. Boundedness of the system operators. We have seen that for an admissible continuous-time transfer function $G_c(s)$ there are always output-normal, input-normal, and parbalanced realizations with well-defined bounded observability and reachability operators. In this sense those realization are well posed. As expected for all infinite-dimensional continuous-time realizations, the propagation, input, and output operators of those realizations are in general unbounded. The input operators are defined in such a way that the range may not be contained in the state space. In this section we are going to investigate when those operators are bounded. We will use the specific form of the restricted and *-restricted shift realizations obtained in §5.

8.1. Boundedness of A_c. First we have the following lemma which shows that the input and output operators are bounded when the propagation operators are.

LEMMA 8.1. *Let (A_c, B_c, C_c, D_c) be an admissible system in $C_X^{U,Y}$. If $A_c : X \to X$ is bounded, then $C_c \in L(X, Y)$ and B_c can be considered as an operator in $L(U, X)$.*

Proof. By definition $C_c|_{D(A_c)} : \big(D(A_c), \|\cdot\|_{A_c}\big) \to Y$ is bounded. Now that A_c is bounded, $D(A_c) = X$. Hence for any $x \in X = D(A_c)$,

$$\begin{aligned}\|C_c x\| &\le \|C_c|_{D(A_c)}\|\,(\|x\|^2 + \|A_c x\|^2)^{1/2}\\ &\le \|C_c|_{D(A_c)}\|\,(\|x\|^2 + \|A_c\|^2\,\|x\|^2)^{1/2}\\ &= \|C_c|_{D(A_c)}\|(1 + \|A_c\|^2)^{1/2}\|x\|.\end{aligned}$$

So $C_c \in L(X, Y)$.

For B we know that $B_c u \in D(A_c^*)^{(')}$ and by definition $\|B_c u\|^{(')} \le b\|u\|$ for any $u \in U$ and some fixed number $b > 0$. By the Riesz representation theorem, there exists $x_u \in D(A_c^*) = X$ such that

$$\|B_c u\|^{(')} = \|x_u\|_{A_c^*}$$

and for $x \in D(A_c^*) = X$,

$$(B_c u)(x) = \langle x_u, x\rangle_{A_c^*} = \langle x_u, x\rangle + \langle A_c^* x_u, A_c^* x\rangle = \langle (1 + A_c A_c^*)x_u, x\rangle.$$

Therefore $B_c u = (1 + A_c A_c^*)x_u \in X$ and

$$\begin{aligned}\|B_c u\| &= \|(1 + A_c A_c^*)x_u\|\\ &\le \|1 + A_c A_c^*\|\,\|x_u\| \le \|1 + A_c A_c^*\|\,\|x_u\|_{A_c^*}\\ &= \|1 + A_c A_c^*\|\,\|B_c u\|^{(')} \le \|1 + A_c A_c^*\|\,b\,\|u\|.\end{aligned}$$

Hence $B_c \in L(U, X)$. □

Now we give a necessary and sufficient condition for the propagation operators in the input-normal and output-normal realizations to be bounded.

THEOREM 8.2. *Let G_c be in $H^\infty_{L(U,Y)}(RHP)$ with U and Y finite dimensional and let (A_c, B_c, C_c, D_c) be an input-normal (or output-normal) realization of G_c. Then A_c is bounded*

if and only if $G_c(s)$ is strictly noncyclic and analytic at infinity. Here analyticity at infinity means that $G_c(\frac{1}{s})$ is analytic at the origin.

Proof. Since all output-normal (input-normal) realizations are unitarily equivalent to the restricted (*-restricted) shift realizations, we prove the theorem for the restricted shift realization and *-restricted shift realization. Let $G_d(z) = G_c(\frac{z-1}{z+1})$ $(z \in \mathbb{D}_e)$ and (A_d, B_d, C_d, D_d) be the restricted realization of G_d on $X_d = \overline{\text{range}}(H_{G_d^\perp})$. Then $A_c = V(A_d - I)(A_d + I)^{-1}V^{-1}$, where V is the unitary operator defined in Proposition 3.3.

If $G_c(s)$ is strictly noncyclic and analytic at infinity, then $G_d(z)$ is strictly noncyclic and analytic at -1. Hence by the spectral minimality of the discrete-time restricted shift realization (see [11]) $-1 \notin \sigma(A_d)$; i.e., $(A_d + I)^{-1}$ is bounded and so $A_c = V(A_d - I)(A_d + I)^{-1}V^{-1}$ is bounded.

Conversely, if A_c is bounded, then $(A_d+I)^{-1} = \frac{1}{2}(I - V^{-1}A_cV)$ is also bounded and thus $-1 \notin \sigma(A_d)$. By Theorem 7.5 G_d has to be strictly noncyclic since otherwise $\sigma(A_d) = \bar{\mathbb{D}}$. Also $G_d(z) = C_d(zI - A_d)^{-1}B_d + D_d$ is analytic at -1. Therefore $G_c(s)$ is strictly noncyclic and analytic at infinity.

Exactly the same argument can also be applied to the *-restricted case. □

Regarding the boundedness of a parbalanced realization, we have the following.

COROLLARY 8.3. *Let $G_c \in H^\infty_{L(U,Y)}(RHP)$ be strictly noncyclic with finite dimensional U and Y and let $(A_{c,o}, B_{c,o}, C_{c,o}, D_{c,o})$, $(A_{c,i}, B_{c,i}, C_{c,i}, D_{c,i})$, and $(A_{c,b}, B_{c,b}, C_{c,b}, D_{c,b})$ be, respectively, an output-normal, an input-normal, and a parbalanced realization of G_c. Then the boundedness of one of $A_{c,o}$, $A_{c,i}$, and $A_{c,b}$ implies the boundedness of the other two.*

Proof. By Theorem 8.2, it suffices to prove that the boundedness of $A_{c,o}$ implies and is implied by that of $A_{c,b}$. We do this by connecting the continuous-time and discrete-time systems as in Theorem 3.1.

Assume that $A_{c,o}$ is bounded. Then, as in the proof of Theorem 8.2, $-1 \notin \sigma(A_{d,o})$. Since G_c and hence $G_d^\perp$ are strictly noncyclic, $\sigma(A_{d,o}) = \sigma(A_{d,b})$. Thus $-1 \notin \sigma(A_{d,b})$ and hence $A_{c,b} = (A_{d,b} - I)(A_{d,b} + I)^{-1}$ is bounded. The same argument can also go the other direction, and the result is proven. □

8.2. Boundedness of B_c in output-normal realizations. We now consider the boundedness of the input and output operators. First we recall that for the input operator B_c of the restricted shift realization with state space X, we have that $B_cu \in X$ $(u \in U)$ if and only if

$$[G_c(s) - G_c(+\infty)]u \in X \quad (u \in U),$$

and in this case

$$(B_cu)(s) = \frac{1}{\sqrt{2\pi}}[G_c(s) - G_c(+\infty)]u$$

(see Theorem 5.7 and Corollary 5.8).

PROPOSITION 8.4. *Let $G_c \in TLC^{U,Y}$.*

1. *The input operator of an output-normal realization of G_c is bounded if and only if there is $M > 0$ such that*

$$\sup_{x>0} \int_{-\infty}^{+\infty} \|[G_c(x + iy) - G_c(+\infty)]u\|^2 dy \le (M\|u\|)^2 \text{ for any } u \in U,$$

where $M > 0$ is a constant.

2. *The output operator of an input-normal realization of G_c is bounded if and only if there is $M > 0$ such that*

$$\sup_{x>0} \int_{-\infty}^{+\infty} \|[\tilde{G}_c(x + iy) - \tilde{G}_c(+\infty)]v\|^2 dy \le (M\|v\|)^2 \text{ for any } v \in Y.$$

3. *If*

$$\sup_{x>0}\int_{-\infty}^{+\infty}\|[G_c(x+iy)-G_c(+\infty)]\|^2dy<\infty,$$

then the input operator of any output-normal realization and the output operator of any input-normal realization are bounded. If in addition the Hankel operator H_{G_c} has closed range, then both the input and the output operators of any output-normal, input-normal, and parbalanced realizations are bounded.

Proof. 1. It suffices to prove the result for the restricted shift realization of G_c. Let B_c be the input operator of the restricted shift realization (see Theorem 5.7).

Assume

$$\sup_{x>0}\int_{-\infty}^{+\infty}\|[G_c(x+iy)-G_c(+\infty)]u\|^2dy\le(M\|u\|)^2 \text{ for any } u\in U.$$

This condition implies that $[G_c-G_c(+\infty)]u\in X$ for any $u\in U$ because as in the proof of Corollary 5.8 we have in $L^2(i\mathbb{R})$ norm

$$[G_c-G_c(+\infty)]u=\lim_{n\to\infty}\frac{G_c(s)-G_c(n)}{1-s/n}u=\lim_{n\to\infty}H_{G_c}\frac{n}{n+s}u.$$

Hence $(B_cu)(s)=\frac{1}{\sqrt{2\pi}}[G_c(s)-G_c(+\infty)]u$ $(u\in U)$ and $\|B_cu\|\le\frac{1}{\sqrt{2\pi}}M\|u\|$.

Conversely, if B_c is bounded, then there is $M>0$ such that $\|B_cu\|\le M\|u\|$. Also $B_cu\in X$ for any $u\in U$. By Corollary 5.8

$$(B_cu)(s)=\frac{1}{\sqrt{2\pi}}[G_c(s)-G_c(+\infty)]u\ \ (u\in U).$$

Thus we have

$$\sup_{x>0}\int_{-\infty}^{+\infty}\|[G_c(x+iy)-G_c(+\infty)]u\|^2dy=\frac{1}{2\pi}\|B_cu\|^2\le\frac{1}{2\pi}M^2\|u\|^2.$$

2. Similarly we only need to prove the result for the *-restricted shift realization. Since the *-restricted shift realization of G_c is the dual of the restricted shift realization of $\tilde{G}_c$, the result follows from 1.

3. First note that

$$\sup_{x>0}\int_{-\infty}^{+\infty}\|[G_c(x+iy)-G_c(+\infty)]\|^2dy<\infty$$

if and only if

$$\sup_{x>0}\int_{-\infty}^{+\infty}\|\tilde{G}_c(x+iy)-\tilde{G}_c(+\infty)\|^2dy<\infty.$$

Clearly these conditions imply

$$\sup_{x>0}\int_{-\infty}^{+\infty}\|[G_c(x+iy)-G_c(+\infty)]u\|^2dy\le(M\|u\|)^2 \text{ for any } u\in U$$

and

$$\sup_{x>0}\int_{-\infty}^{+\infty}\|[\tilde{G}_c(x+iy)-\tilde{G}c(+\infty)]u\|^2dy\le(\tilde{M}\|u\|)^2 \text{ for any } u\in U$$

for some constants $M < \infty$ and $\tilde{M} < \infty$. Hence by 1. and 2. the input operator of any output-normal realization and the output operator of any input-normal realization are bounded. If in addition H_{G_c} has closed range, then by Proposition 6.2 all reachable and observable admissible realizations of G_c are equivalent. Thus all have bounded input and output operators. □

The following corollary gives a simple condition for the input operator of an output-normal realization and the output operator of an input-normal realization to be bounded.

COROLLARY 8.5. *Let $G_c \in H^\infty_{L(U,Y)}(RHP)$ be analytic at ∞. Then the input operator of the output-normal realizations and the output operator of the input-normal realizations are bounded.*

Proof. Let $F_d(z) = G_c(\frac{1-z}{1+z}) - G_c(\infty)$. The analyticity of G_c at ∞ means that F_d and $\frac{F_d(z)}{1+z}$ are both analytic at -1. Hence $\frac{F_d(z)}{1+z} \in H^\infty_{L(U,Y)}(\mathbb{D})$ and for any $u \in U$,

$$\left\| \frac{F_d(z)}{1+z} u \right\|_{H^2_Y(\mathbb{D})} \leq M \|u\|_U,$$

where $M = \sup_{z\in\mathbb{D}} \| \frac{F_d(z)}{1+z} \|_{L(U,Y)}$. Applying the unitary transformation V in Proposition 3.3, we have

$$\|(G_c - G_c(\infty))u\|_{H^2_Y(RHP)} = 2\sqrt{\pi} \| V \frac{F_d(z)}{1+z} u \|_{H^2_Y(\mathbb{D})} \leq 2\sqrt{\pi} M \|u\|_U \quad (u \in U).$$

Since the analyticity of G_c at ∞ implies the analyticity of $\tilde{G}_c$ at ∞, we have similarly

$$\|(\tilde{G}_c - \tilde{G}_c(\infty))y\|_{H^2_U(RHP)} \leq 2\sqrt{\pi} M \|y\|_Y \quad (y \in Y).$$

By Proposition 8.4 it follows that the input operator of the restricted shift realization and the output operator of the *-restricted shift realization are bounded. This proves the corollary. □

8.3. Boundedness of C_c in output-normal realization. Now we consider the boundedness of the input operator of the *-restricted shift realization and the output operator of the restricted shift realization. We present here results for noncyclic scalar transfer functions.

It is well known that a scalar inner function $q_d \in H^\infty(\mathbb{D})$ admits a factorization of the form $q_d(z) = \lambda \mathcal{B}_d(z) \mathcal{S}_d(z)$, where λ is a complex number, $|\lambda| = 1$;

$$\mathcal{B}_d(z) = \prod_{n=1}^{\infty} \frac{\bar{\alpha}_n}{|\alpha_n|} \frac{\alpha_n - z}{1 - \bar{\alpha}_n z}$$

is a Blaschke product, and

$$\mathcal{S}_d(z) = \exp\left[- \int_0^{2\pi} \frac{e^{i\theta} + z}{e^{i\theta} - z} d\mu_d(\theta) \right]$$

is a singular inner function with μ_d a finite singular positive measure on the unit circle $\partial\mathbb{D}$ (see [17]). Here we take $\frac{\bar{\alpha}_n}{|\alpha_n|}$ to be 1 when $\alpha_n = 0$. Ahern and Clark [1] have proved the following theorem.

THEOREM 8.6. *Set $X = H^2(\mathbb{D}) \ominus q_d H^2(\mathbb{D})$ and denote the compressed shift operator on X by $S(q_d) := P_{X_d} z|_{X_d}$. Then the following statements are equivalent.*

1. *For every $x \in X$ the nontangential limit of $x(z)$ exists at -1.*
2. $P_X 1 \in \text{range}(I + S(q_d))$.
3. *For the function q_d*

$$\sum_{n=1}^{\infty} \frac{1 - |\alpha_n|^2}{|1 + \alpha_n|^2} < \infty \quad \text{and} \quad \int_0^{2\pi} \frac{d\mu_d(\theta)}{|1 + e^{i\theta}|} < \infty.$$

Furthermore, if one of these conditions hold, then there exists a function $k \in X$ such that the nontangential limit of any $x \in X$ at -1 is

$$x(-1) := \lim_{\substack{z \to -1,\ z \in \mathbb{D} \\ \text{nontangential}}} x(z) = \langle x, k \rangle. \qquad \square$$

This theorem can be cast into left invariant spaces on the right half plane. Let q_c be an inner function in $H^\infty(RHP)$. Then q_c has the form $q_c(s) = \lambda \mathcal{B}_c(s)\mathcal{S}_c(s)$, where $\mathcal{B}_c$ is a Blaschke product on the right half plane,

$$\mathcal{B}_c(s) = \prod_{n=1}^{\infty} \frac{|1-\beta_n^2|}{1-\beta_n^2} \frac{s-\beta_n}{s+\bar{\beta}_n},$$

and

$$\mathcal{S}_c(s) = e^{-as} \exp\left[-\int_{-\infty}^{\infty} \frac{ys+i}{y+is} d\mu_c(y) \right]$$

is a singular inner function with μ_c a finite singular positive measure on $i\mathbb{R}$ and $a \geq 0$ (see [17]). Here $\frac{|1-\beta_n^2|}{1-\beta_n^2}$ is taken to be 1 if $\beta_n = 1$. Let V be the transformation defined in Proposition 3.3. Applying V to X_d in Theorem 8.6, we obtain the following theorem.

THEOREM 8.7. *Set $X = H^2(RHP) \ominus q_c H^2(RHP)$ and $\mathcal{D} = \{P_X \frac{h}{1+s} : h \in X\}$. Then the following statements are equivalent.*

1. *For every $f \in X$, the limit*

$$\lim_{\substack{s \to \infty,\ Re(s) > 0 \\ Re(s) > \epsilon |s|}} sf(s)$$

exists for any $\epsilon > 0$.

2. $P_X \frac{1}{1+s} \in \mathcal{D}$.

3. *For the inner function q_c,*

$$a = 0, \quad \sum_{n=1}^{\infty} Re\beta_n < \infty, \quad \text{and} \quad \int_{-\infty}^{\infty} \sqrt{1+y^2} d\mu_c(y) < \infty.$$

Moreover, if one of the statements holds, then there exists $k \in X$ such that

$$\lim_{\substack{s \to \infty, Re(s) > 0 \\ Re(s) > \epsilon |s|}} sf(s) = \langle f, k \rangle \quad (f \in X).$$

Proof. Let $q_d(z) = q_c(\frac{1-z}{1+z})$ $(z \in \mathbb{D})$. Then q_d admits factorization as in Theorem 8.6: $q_d(z) = \lambda \mathcal{B}_d(z)\mathcal{S}_d(z)$. It can be easily seen that the Blaschke products $\mathcal{B}_d(z)$ and $\mathcal{B}_c(z)$ can be related by $\beta_n = \frac{1-\alpha_n}{1+\alpha_n}$; that the functions $\mathcal{S}_c(s)$ and $\mathcal{S}_d(z)$ are related by

$$\mathcal{S}_c(s) = \mathcal{S}_d\left(\frac{1-s}{1+s}\right) \quad (s \in RHP),$$

with $a = \mu_d(\{-1\})$; and that the measure μ_c is the measure μ_d transformed by the bilinear transformation

$$s = \frac{1-z}{1+z} : \partial\mathbb{D} \backslash \{-1\} \to i\mathbb{R}.$$

Hence the condition $\sum_{n=1}^{\infty} \frac{1-|\alpha_n|^2}{|1+\alpha_n|^2} < \infty$ is equivalent to

$$\sum_{n=1}^{\infty} Re\beta_n < \infty,$$

and the condition $\int_0^{2\pi} \frac{d\mu_d(\theta)}{|1+e^{i\theta}|} < \infty$ is equivalent to

$$a = 0, \quad \int_{-\infty}^{\infty} \sqrt{1+y^2}\, d\mu_c(y) < \infty.$$

This shows that condition 3 of Theorem 8.6 and condition 3 of Theorem 8.7 are equivalent.

Let V be the unitary transformation defined in Proposition 3.3. Then $V(H^2(\mathbb{D}) \ominus q_d H^2(\mathbb{D})) = H^2(RHP) \ominus q_c H^2(RHP)$, $V(\text{range}(I + S(q_d))) = \{P_X \frac{h}{1+s} : h \in H^2(RHP) \ominus q_c H^2(RHP)\} = \mathcal{D}$, and

$$P_{H^2(RHP)\ominus q_c H^2(RHP)} \frac{1}{1+s} = \sqrt{\pi} V P_{H^2(\mathbb{D})\ominus q_d H^2(\mathbb{D})} 1.$$

Therefore condition 2 of Theorem 8.6 and condition 2 of Theorem 8.7 are equivalent.

Finally, for $x \in H^2(\mathbb{D}) \ominus q_d H^2(\mathbb{D})$, we have $f := Vx \in H^2(RHP) \ominus q_c H^2(RHP)$ and

$$\lim_{\substack{z\in\mathbb{D}, z\to -1 \\ \text{nontangential}}} x(z) := \lim_{\substack{z\in\mathbb{D}, z\to -1 \\ (1-|z|)>\epsilon|z+1|}} x(z) = \sqrt{\pi} \lim_{\substack{s\to\infty,\ Re(s)>0 \\ Re(s)>\epsilon|s|}} sf(s)$$

for any $\epsilon > 0$. This completes the proof. □

If in the theorem we replace q_c by $\tilde{q}_c(s) = \overline{q_c(\bar{s})}$ and $H^2(RHP) \ominus q_c H^2(RHP)$ by $H^2(RHP) \ominus \tilde{q}_c H^2(RHP)$, then we have the results to be true for the space $H^2(RHP) \ominus \tilde{q}_c H^2(RHP)$ while condition 3 of Theorem 8.7 remains unchanged in terms of a, β_n's and the singular measure μ_c.

These results can be immediately used to show the boundedness of the output operators of output-normal realizations and the input operators of input-normal realizations.

COROLLARY 8.8. *Let $G_c \in H^\infty(RHP)$ be a scalar noncyclic transfer function admitting the factorization $G_c = q_c f^*$, where $q_c \in H^\infty(RHP)$ is inner and q_c and $f \in H^\infty(RHP)$ are weakly coprime. Assume q_c has decomposition as in Theorem* 8.7, *and set $X = H^2(RHP) \ominus q_c H^2(RHP)$. Then the following statements are equivalent:*

1. *The output operator C_c of the restricted shift realization of G_c is bounded.*
2. *The input operator $B_{c,*}$ of the *-restricted shift realization of G_c is bounded.*
3. *One of the statements in Theorem* 8.7 *is true.*

Hence the output operator of every output-normal realization and the input operator of every input-normal realization are bounded if and only if one of the statements in Theorem 8.7 *is true.*

If in addition the Hankel operator H_{G_c} has closed range, then both the input operator and the output operator of every reachable and observable admissible realization of G_c are bounded.

Proof. First assume one and hence all of the statements in Theorem 8.7 to be true. We prove 1. and 2.

By Theorem 5.7, the output operator of the shift realization of G_c is given by

$$\begin{aligned} C_c : D(C) = D(A_c) + (I - A)^{-1} B_c U \subseteq X_c &\to Y, \\ x &\mapsto \sqrt{2\pi} \lim_{\substack{r\in\mathbb{R} \\ r\to\infty}} rx(r), \end{aligned}$$

where $X_c = H^2(RHP) \ominus q_c H^2(RHP)$. Now by Theorem 8.7 there exists $k \in X_c$ such that

$$\lim_{\substack{r\in\mathbb{R} \\ r\to\infty}} rx(r) = \langle x, k\rangle \quad (x \in X_c).$$

Hence

$$C_c x = \sqrt{2\pi} \lim_{\substack{r\in\mathbb{R} \\ r\to\infty}} rx(r) = \sqrt{2\pi}\langle x, k\rangle$$

for any $x \in D(C_c)$, and it follows that C_c is bounded: $\|C_c\| \le \sqrt{2\pi}\|k\|$.

To show the boundedness of the input operator $B_{c,*}$ of the *-restricted shift realization we use the expression of $B_{c,*}$ as given in Theorem 5.9: The state space is $X_{c,*} = H^2(RHP) \ominus \tilde{q}_c H^2(RHP)$ and

$$\begin{aligned} B_{c,*}: \qquad U &\to D(A_{c,*}^*)', \\ u &\mapsto B_c(u), \\ [B_{c,*}(u)](x) &= \frac{1}{\sqrt{2\pi}} \left\langle \frac{1}{1+s} u,\ (1 - A_{c,*}^*)x \right\rangle,\ x \in D(A_{c,*}^*), \end{aligned}$$

where the operator $A_{c,*}$ has domain $D(A_{c,*}) = \{P_{X_{c,*}} \frac{h}{1+s} : h \in X_{c,*}\}$. Here $U = \mathbb{C}$. Since by Theorem 8.7 we have $P_{X_{c,*}} \frac{1}{1+s} \in D(A_{c,*})$, it follows that

$$\begin{aligned} [B_{c,*}(u)](x) &= \frac{1}{\sqrt{2\pi}} \left\langle \frac{1}{1+s} u,\ (1 - A_{c,*}^*)x \right\rangle \\ &= \frac{1}{\sqrt{2\pi}} \left\langle P_{X_{c,*}} \frac{1}{1+s} u,\ (1 - A_{c,*}^*)x \right\rangle \\ &= \frac{1}{\sqrt{2\pi}} \left\langle (1 - A_{c,*}) P_{X_{c,*}} \frac{1}{1+s} u,\ x \right\rangle. \end{aligned}$$

This shows that $B_{c,*}(u) \in X_c$ and $B_{c,*}(u) = \frac{1}{\sqrt{2\pi}}(I - A_{c,*}) P_{X_{c,*}} \frac{1}{1+s} u$ for $u \in \mathbb{C}$. Hence B_c is bounded: $\|B_c\| \le \frac{1}{\sqrt{2\pi}} \|(I - A_{c,*}) P_{X_{c,*}} \frac{1}{1+s}\|_{H^2(RHP)}$.

Now we assume 2. and prove that this implies 3. As in the above, $X_{c,*} = H^2(RHP) \ominus \tilde{q}_c H^2(RHP)$ and

$$[B_{c,*}(u)](x) = \frac{1}{\sqrt{2\pi}} \left\langle P_{X_{c,*}} \frac{1}{1+s} u,\ (1 - A_{c,*}^*)x \right\rangle \quad (x \in D(A_{c,*}^*))$$

where $D(A_{c,*}) = \{P_{X_{c,*}} \frac{h}{1+s} : h \in X_{c,*}\}$. As $B_{c,*}$ is assumed to be bounded, for any $u \in \mathbb{C}$ there exists $k(u) \in X_{c,*}$ such that

$$[B_{c,*}(u)](x) = \langle k(u),\ x \rangle \ (x \in D(A_{c,*}^*)).$$

Hence $\langle k(u),\ x \rangle = \frac{1}{\sqrt{2\pi}} \langle P_{X_{c,*}} \frac{1}{1+s} u,\ (1 - A_{c,*}^*)x \rangle$ for any $x \in D(A_{c,*}^*)$. This shows that

$$P_{X_{c,*}} \frac{u}{1+s} \in D((1 - A_{c,*}^*)^*) = D(1 - A_{c,*}) = D(A_{c,*}) = \left\{ P_{X_{c,*}} \frac{h}{1+s} : h \in X_{c,*} \right\}.$$

Thus statement 2 in Theorem 8.7 is true if the space X is replaced by $X_{c,*}$ and the inner function q_c is replaced by $\tilde{q}_c$. By the remark following Theorem 8.7, we know that the statements in Theorem 8.7 are true for X and q_c.

Finally, we show that 1. implies 3. Let (A_c, B_c, C_c, D_c) be the restricted shift realization of $G_c = q_c f^*$ and assume that C_c is bounded. Denote by $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ the dual system of (A_c, B_c, C_c, D_c). Then $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ is the *-restricted shift realization of $\tilde{G}_c = \tilde{q}_c \tilde{f}^*$ with state space $X = H^2(RHP) \ominus q_c H^2(RHP)$, and

$$\tilde{B}_c = C_c^*.$$

Hence $\tilde{B}_c$ is bounded. By the preceding proof, statement 2 in Theorem 8.7 is true for the space X and the inner function q_c. This completes the proof. □

8.4. Boundedness of B_c, C_c for parbalanced realizations. To conclude this section we show some results on the boundedness of the input and output operators of parbalanced realizations.

PROPOSITION 8.9. *If the output operator of an output-normal realization of $G_c \in TLC^{U,Y}$ is bounded, then the output operator of a parbalanced realization of G_c is bounded.*

Proof. Consider the discrete-time transfer function $G_d(z) = G_c(\frac{z-1}{z+1})$. Let $(A_{do}, B_{do}, C_{do}, D_{do})$ be the discrete-time restricted shift realization of G_d with state space X and $(A_{co}, B_{co}, C_{co}, D_{co}) = T((A_{do}, B_{do}, C_{do}, D_{do}))$ be the continuous-time restricted shift realization of G_c with the same state space X. Denote their observability operators by $\mathcal{O}_{do}$ and $\mathcal{O}_{co}$, respectively. By Theorem 3.4 we have

$$V\mathcal{O}_{do}x = \mathcal{L}\mathcal{O}_{co}x, \quad x \in X,$$

where $V : H_Y^2(\mathbb{D}) \to H_Y^2(RHP)$ is the unitary transformation as defined in Proposition 3.3, and $\mathcal{L}$ is the Laplace transform.

In [30] (see also [24]) it has been shown that there is a parbalanced realization $(A_{db}, B_{db}, C_{db}, D_{db})$ of G_d with state space X that satisfies the following:

$$\mathcal{W}^{1/4}A_{db} = A_{do}\mathcal{W}^{1/4}$$

and

$$\mathcal{O}_{db} = \mathcal{W}^{1/4},$$

where $\mathcal{W} = H_{G^\perp}H_{G^\perp}^*|_X$, and $\mathcal{O}_{db}$ is the observability operator of $(A_{db}, B_{db}, C_{db}, D_{db})$.

Let $(A_{cb}, B_{cb}, C_{cb}, D_{cb}) = T((A_{db}, B_{db}, C_{db}, D_{db}))$, where T is the transformation defined in Theorem 3.1. Since $D(A_{cb}) = \text{range}(A_{db} + I)$ and $D(A_{co}) = \text{range}(A_{do} + I)$, we have, by the equality $\mathcal{W}^{1/4}A_{db} = A_{do}\mathcal{W}^{1/4}$, that $\mathcal{W}^{1/4}D(A_{cb}) \subseteq D(A_{co})$. By Theorem 3.4, $(A_{cb}, B_{cb}, C_{cb}, D_{cb})$ is a parbalanced realization of G_c and for the observability operator $\mathcal{O}_{cb}$ of $(A_{cb}, B_{cb}, C_{cb}, D_{cb})$ we have

$$\mathcal{L}\mathcal{O}_{cb}x = V\mathcal{O}_{db}x = V\mathcal{W}^{1/4}x, \quad x \in X.$$

Notice that in fact by Theorem 5.1 we have $\mathcal{O}_{do} = I_X$. Thus

$$\mathcal{L}\mathcal{O}_{cb}x = V\mathcal{O}_{do}\mathcal{W}^{1/4}x = \mathcal{L}\mathcal{O}_{co}\mathcal{W}^{1/4}x, \quad x \in X.$$

Since $\mathcal{L}$ is unitary, this shows that $\mathcal{O}_{cb}x = \mathcal{O}_{co}\mathcal{W}^{1/4}x$ for $x \in X$. By the definition of $\mathcal{O}_{co}$ and $\mathcal{O}_{cb}$ we have

$$C_{cb}e^{tA_{cb}}x = C_{co}e^{tA_{co}}\mathcal{W}^{1/4}x, \quad x \in D(A_{cb}).$$

Note that C_{cb} is a bounded operator from $(D(A_{cb}), \|\cdot\|_{A_{cb}})$ to Y (see Definition 2.1) and C_{co} has the analogous property. For $x \in D(A_{cb})$ the function $e^{tA_{cb}}x$ is continuous in t in the graph norm $\|\cdot\|_{A_{cb}}$. Similarly, since $\mathcal{W}^{1/4}x \in D(A_{co})$ for $x \in D(A_{cb})$, $e^{tA_{co}}\mathcal{W}^{1/4}x$ is continuous in t in the graph norm $\|\cdot\|_{A_{co}}$. Therefore both $C_{cb}e^{tA_{cb}}x$ and $C_{co}e^{tA_{co}}\mathcal{W}^{1/4}x$ are continuous in t in the norm of Y. Taking $t = 0$, we have

$$C_{cb}x = C_{co}\mathcal{W}^{1/4}x, \quad x \in D(A_{cb}).$$

Since by assumption C_{co} and hence $C_{co}\mathcal{W}^{1/4}$ are bounded, the operator

$$C_{cb}|_{A_{cb}} : D(A_{cb}) \to Y$$

is bounded, where $D(A_{cb})$ is equipped with the norm of X. As $D(A_{cb})$ is dense in X, C_{cb} can be boundedly extended to X.

To complete the proof we just note that the restricted shift realization is unitarily equivalent to any output-normal realization of G_c and that all parbalanced realizations of G_c are equivalent. □

COROLLARY 8.10. *If the input operator of an input-normal realization of* $G_c \in TLC^{U,Y}$ *is bounded, then the input operator of a parbalanced realization of* G_c *is bounded.*

Proof. Let (A_*, B_*, C_*, D_*) be the *-restricted shift realization of G_c and let (A_o, B_o, C_o, D_o) be its dual realization. Then (A_o, B_o, C_o, D_o) is the restricted shift realization of $\tilde{G}_c$. By the assumption, the operator B_* is bounded. Hence so is the operator C_o. By Proposition 8.9 the output operator of a parbalanced realization of $\tilde{G}_c$ is bounded. Consider the dual system (A, B, C, D) of this parbalanced realization of $\tilde{G}_c$. We have B to be bounded. Notice that the dual system of a parbalanced realization of $\tilde{G}_c$ is a parbalanced realization of G_c. Therefore the input operator of any parbalanced realization of G_c is bounded. □

COROLLARY 8.11. *Let* G_c *be in* $TLC^{U,Y}$. *Assume that the Hankel operator* H_{G_c} *has closed range. Then the input (output) operator of a parbalanced realization of* G_c *is bounded if and only if there is a constant* $M > 0$ *such that*

$$\sup_{x>0} \int_{-\infty}^{+\infty} \|[G_c(x+iy) - G_c(+\infty)]u\|^2 dy \le (M\|u\|)^2 \text{ for any } u \in U,$$

$$\left(\sup_{x>0} \int_{-\infty}^{+\infty} \|[\tilde{G}_c(x+iy) - \tilde{G}c(+\infty)]v\|^2 dy \le (M\|v\|)^2 \text{ for any } v \in Y\right).$$

Proof. Since the Hankel operator H_{G_c} has closed range, by Proposition 6.2 all input-normal, output-normal, and parbalanced realizations of G_c are equivalent. The corollary then follows from Proposition 8.4. □

9. Examples.

Example 1: Rational transfer function. Let $g(s)$ be a scalar-valued rational proper transfer function in $H^\infty(RHP)$, i.e., $g(s)$ has all its poles in the open left half plane.

Note that $g(s)$ has, up to a unitary scalar, a unique factorization as

$$g(s) = q(s)f(-s),$$

where $q(s)$ is an inner function, i.e., a Blaschke product with poles in LHP, and $f(s)$ is a rational function in $H^\infty(RHP)$, i.e., a proper rational function with poles in LHP. The functions $q(s)$ and $f(s)$ are strongly coprime, which is for rational functions equivalent to both functions not having common zeros in the extended RHP, i.e., $\{s \in \mathbb{C} \mid Re(s) \ge 0\} \cup \{\infty\}$.

The Blaschke product q is determined by the poles of g. For example if

$$g(s) = \frac{(s-1)(s+2)}{(s+3)(s+4)(s+5)},$$

then the Blaschke product is given by

$$q(s) = \frac{(s-3)(s-4)(s-5)}{(s+3)(s+4)(s+5)}$$

and

$$f(s) = \frac{(s+1)(s-2)}{(s+3)(s+4)(s+5)}.$$

It follows from the results in §6 that the state space of the restricted and *-restricted shift realization of the transfer function g is given by

$$X = (qH^2(RHP))^\perp.$$

Note that by Kronecker's theorem (see, e.g., [22]) X is a finite-dimensional space with dimension equal to the number of zeros or poles (counted with multiplicities) of the Blaschke product. From the construction it is clear that the Blaschke product is completely determined by the poles of the transfer function. Hence we have recovered the well-known result that the dimension of a minimal state-space realization equals the number of poles of the transfer function.

Example 2: Delay system with strictly proper rational part. In this example we consider single-input single-output delay systems. We continue with the notation in the above example and let the transfer function have the form $g_1(s) = e^{-\alpha s}g(s)$ with $\alpha > 0$. Let $p(s) = e^{-\alpha s}q(s)$. Clearly p is in $H^\infty(RHP)$ and inner. Later we will show that in fact p and f are weakly coprime. For now assume that this is true. Thus by Theorem 4.8 g_1 is strictly noncyclic, and by Proposition 5.11 the state space X of the restricted shift realization (A_c, B_c, C_c, D_c) has the form

$$X = H^2(RHP) \ominus pH^2(RHP).$$

The domain of A_c is $D(A_c) = \{\frac{x(s)-x(1)}{1-s} \mid x \in X\}$. Hence for $h \in D(A_c)$ we will have $h(s) = \frac{x(s)-x(1)}{1-s}$ for some $x \in X$, $\lim_{r\in\mathbb{R}, r\to+\infty} rh(r) = x(1)$ and

$$(A_c h)(s) = sh(s) - \lim_{r\in\mathbb{R}, r\to+\infty} rh(r) = sh(s) - x(1).$$

Note that g_1 satisfies the condition in Proposition 8.4. So the operator B_c is defined as

$$(B_c u)(s) = \frac{1}{\sqrt{2\pi}}[g_1(s) - g_1(+\infty)]u = \frac{1}{\sqrt{2\pi}}g_1(s)u, \qquad u \in \mathbb{C},$$

and B_c is bounded. Hence $(I - A_c)^{-1}B\mathbb{C} \subseteq D(A_c)$ and

$$D(C_c) = D(A_c) + (I - A_c)^{-1}BU = D(A_c).$$

We have, for $h \in D(A_c)$,

$$C_c h = \sqrt{2\pi} \lim_{r\in\mathbb{R}, r\to+\infty} rh(r).$$

Note that because $\alpha \neq 0$, by Corollary 8.8 C_c is unbounded.

The operator D_c is $D_c = g_1(+\infty) = 0$.

We can directly verify that this is a realization of g_1. Let $\xi \in RHP$. An easy calculation will show that for $h \in D(A_c)$

$$((\xi I - A_c)^{-1}h)(s) = \frac{h(s) - h(\xi)}{\xi - s}.$$

(We remark here that this formula is true in general, not just for this particular example.) Then

$$((\xi I - A_c)^{-1}B_c u)(s) = \frac{1}{\sqrt{2\pi}}(\xi I - A_c)^{-1}g_1(s)u = \frac{1}{\sqrt{2\pi}}\frac{g_1(s) - g_1(\xi)}{\xi - s}.$$

Hence

$$C_c(\xi I - A_c)^{-1}B_c u = \lim_{r\in\mathbb{R},\ r\to+\infty} r\frac{g_1(r) - g_1(\xi)}{\xi - r} = g_1(\xi).$$

This realization is exponentially stable by Theorem 7.11 since g_1 is clearly analytic on $Re(s) > -3$. It also follows from Theorem 7.11 that the degree of stability is $-3 = \max\{s : s \text{ is a pole of } g\}$. Consequently the parbalanced realization will also be exponentially stable

with the same degree of stability. Notice that g_1 is continuous in the extended $i\mathbb{R}$. Hence the Hankel operator H_{g_1} is compact. Therefore by Theorem 6.1 there exists a balanced realization.

To show that p and f are weakly coprime, consider the closed linear span $S := pH^2(RHP) \vee fH^2(RHP)$. We need to show that $S = H^2(RHP)$. The space S is obviously a (right) invariant subspace of $H^2(RHP)$. Hence by Beurling's theorem [22] there is an inner function $\Theta \in H^\infty(RHP)$ such that

$$S = \Theta H^2(RHP).$$

Hence $pH^2(RHP) \subseteq \Theta H^2(RHP)$ and $\overline{fH^2(RHP)} \subseteq \Theta H^2(RHP)$. Let $q_1(s) = \frac{s-2}{s+2}$ (which is the inner part of the inner-outer factorization of f; see [22, p. 11]). Then

$$q_1 H^2(RHP) = \overline{fH^2(RHP)}.$$

So by [22, Cor. 5, p. 13] we must have that p/Θ and q_1/Θ are both inner functions. Note that $\Theta(2) \neq 0$ since otherwise $h(2) = 0$ for any $h \in pH^2(RHP) \subseteq \Theta H^2(RHP)$, and this is certainly not true. Thus the inner function $q_1(s)/\Theta(s)$ has a zero at 2. Hence the function $\frac{s+2}{s-2}\frac{q_1(s)}{\Theta(s)}$ will still be in $H^\infty(RHP)$. That is, $1/\Theta \in H^\infty(RHP)$. Hence $H^2(RHP) = \Theta(1/\Theta)H^2(RHP) \subseteq \Theta H^2(RHP) = S$.

Note that exactly the same argument in this example will apply for any transfer function $g_1 = e^{-\alpha s}g(s)$, where g is a stable and strictly proper rational function and $\alpha > 0$. Also, in a similar manner we can obtain the *-restricted shift realization which will have bounded output operator and has the same stability properties as the restricted shift realization.

We summarize these as follows.

PROPOSITION 9.1. *If a scalar transfer function G has the form $G(s) = e^{-\alpha s}g(s)$, $\alpha > 0$, where g is a stable and strictly proper rational function, then*

1. *G has a balanced realization;*

2. *all reachable output-normal realizations of G have bounded input operator and unbounded output operator, whereas all observable input-normal realizations have bounded output operator and unbounded input operator;*

3. *all reachable and observable input- and output-normal realizations and all parbalanced realizations are exponentially stable with growth bound equal to* $\max\{Re(s) : s$ *is a pole of* $G\}$. □

Example 3: Delay system with not strictly proper rational part. When the rational transfer function g in the previous example is not strictly proper, the resulting realizations will be different: the input operator of the restricted shift realization is not necessarily bounded, and it is not clear whether there is a balanced realization of g_1 because the Hankel operator H_{g_1} is not compact. A parbalanced realization, however, exists by Theorem 6.1. We first consider the simplest case with $g(s) = 1$. This is a simple delay $g_1(s) = e^{-\alpha s}$ ($\alpha > 0$). The state space of the restricted shift realization is $X = H^2 \ominus e^{-\alpha s}H^2$, which is the image of the Laplace transform $\mathcal{L}$ on $L^2([0, \alpha])$. Let (A_c, B_c, C_c, D_c) be the restricted shift realization and let

$$(A, B, C, D) = (\mathcal{L}^{-1}A_c\mathcal{L}, \mathcal{L}^{-1}B_c, C_c\mathcal{L}, D_c).$$

We know that (see Theorem 5.7)

$$(e^{tA_c}f)(x) = f(x + t)|_{[0,\alpha]}, \qquad f \in L^2([0, \alpha]),\ x \in [0, \alpha], t \geq 0,$$

where $f(x + t)|_{[0,\alpha]} = f(t + x)$ if $t + x \in [0, \alpha]$ and 0 otherwise. Thus

$$Af = f', \qquad f \in D(A),$$

with $D(A) = \{x \in L^2([0,\alpha]) : x \text{ is absolutely continuous}, x' \in L^2([0,\alpha]),\ x(\alpha) = 0\}$. By Theorem 5.7, for $x \in D(A_c^*)$ and $u \in \mathbb{C}$,

$$\begin{aligned}
[B_c(u)](x) &= \frac{1}{\sqrt{2\pi}}\left\langle \frac{1}{1-s}[G_c(s) - G_c(1)]u,\ (1 - A_c^*)x \right\rangle \\
&= \frac{1}{\sqrt{2\pi}}\left\langle \frac{1}{1-s}(e^{-\alpha s} - e^{-\alpha})u, (1 - A_c^*)x \right\rangle \\
&= \frac{1}{\sqrt{2\pi}}\left\langle \mathcal{L}^{-1}\frac{e^{-\alpha s} - e^{-\alpha}}{1-s}u, \mathcal{L}^{-1}(1 - A_c^*)\mathcal{L}\mathcal{L}^{-1}x \right\rangle \\
&= \langle e^{t-\alpha}u|_{[0,\alpha]}, \mathcal{L}^{-1}(1 - A_c^*)\mathcal{L}\mathcal{L}^{-1}x\rangle_{L^2([0,\alpha])} \\
&= \langle e^{t-\alpha}u, (1 - A^*)\mathcal{L}^{-1}x\rangle_{L^2([0,\alpha])},
\end{aligned}$$

where $e^{t-\alpha}u|_{[0,\alpha]} = \left(\frac{1}{\sqrt{2\pi}}\mathcal{L}^{-1}\frac{e^{-\alpha s} - e^{-\alpha}}{1-s}u\right)(t)$ is $e^{t-\alpha}u$ for $t \in [0,\alpha]$ and 0 otherwise. This shows that for $x \in D(A^*) \subseteq L^2([0,\alpha])$,

$$[B(u)](x) = [\mathcal{L}^{-1}B_c u](x) = [\mathcal{L}^* B_c u](x) = [B_c u](\mathcal{L}x) = \langle e^{t-\alpha}u, (1 - A^*)x\rangle_{L^2([0,\alpha])}.$$

It can be shown that

$$D(A^*) = \{x \in L^2([0,\alpha]) : x \text{ is absolutely continuous}, x' \in L^2([0,\alpha]),\ x(0) = 0\},$$

and $A^*x = -x'$ for $x \in D(A^*)$. Hence

$$[B(u)](x) = \langle e^{t-\alpha}u, (1 - A^*)x\rangle_{L^2([0,\alpha])} = \langle e^{t-\alpha}u, x + x'\rangle_{L^2([0,\alpha])} = ux(\alpha).$$

Since for $x \in D(C_c)$,

$$C_c x = \sqrt{2\pi}\lim_{\substack{r\in\mathbb{R}\\ r\to\infty}} rx(r),$$

we have for $x \in D(C) \subseteq L^2([0,\alpha])$,

$$Cx = C_c\mathcal{L}x = \sqrt{2\pi}\lim_{\substack{r\in\mathbb{R}\\ r\to\infty}} r(\mathcal{L}x)(r) = \lim_{\substack{\lambda\to 0\\ \lambda>0}} x(\lambda) = x(0).$$

Finally, $D_c = g(+\infty) = 0$.

This realization is, by Theorem 7.11, exponentially stable. In fact, the spectrum of e^{tA} is $\{0\}$ $(t > 0)$. The operators B and C are both unbounded.

Now consider the factorization $e^{-\alpha s} = qf^*$, where $q(s) = e^{-\alpha s}$ and $f(s) = 1$. Clearly this is a strongly coprime factorization. Therefore by Proposition 6.2 all reachable and observable realizations of $e^{-\alpha s}$ are equivalent. This shows that all reachable and observable realizations are exponentially stable and have unbounded input, output, and state propagation operators.

As in the previous example, we can generalize this result.

PROPOSITION 9.2. *If a scalar transfer function G has the form $G(s) = e^{-\alpha s}g(s)$, where g is a stable proper rational function and $g(\infty) \neq 0$, $\alpha > 0$, then*

1. *all reachable and observable admissible realizations of G are equivalent;*

2. *if (A, B, C, D) is a reachable and observable admissible realization of G, then the operators A, B, and C are all unbounded;*

3. *every reachable and observable admissible realization of G is exponentially stable with growth bound equal to* $\max\{Re(s) : s \text{ is a pole of } G\}$.

Proof. Since g is a stable proper rational function, g has a factorization $g = qf^*$ such that q and f are stable proper rational and strongly coprime (see Theorem 4.10 for the definition of strong coprimeness). Hence

$$\inf_{s\in RHP}[|q(s)| + |f(s)|] > 0.$$

Since $g(\infty) \neq 0$, we must have that $f(\infty) \neq 0$. Therefore

$$\inf_{s\in RHP}[|q(s)e^{-\alpha s}| + |f(s)|] > 0.$$

This, by the Corona theorem (see [22, p. 66]), shows that $qe^{-\alpha s}$ and f are strongly coprime. So by Theorem 4.10 the Hankel operator H_G has closed range and by Proposition 6.2 all reachable and observable realizations of G are equivalent. Thus 1. is proven.

Since G is not analytic at infinity, by Theorem 8.2 the state propagation operator of any reachable output-normal realization is unbounded. Note that in the factorization $G = (qe^{-\alpha s})f^*$ the inner function does not satisfy condition 3 in Theorem 8.6 because now $\alpha \neq 0$. Therefore by Corollary 8.8 the output operator of the restricted shift realization and the input operator of the *-restricted shift realization are unbounded. Thus 2. follows from 1.

Since G is strictly noncyclic and

$$\begin{aligned}&\inf\{\alpha : G(s) \text{ has analytic continuation on } Re(s) > \alpha\}\\ &\quad = \max\{Re(s) : s \text{ is a pole of } g\}\\ &\quad < 0,\end{aligned}$$

by Theorem 7.11 all reachable output-normal realizations of G are exponentially stable with growth bound $\max\{Re(s) : s \text{ is a pole of } g\}$. As equivalent systems have the same exponential stability property and growth bound, 3. also follows from 1. □

Example 4: Systems with infinite Blaschke product. In this example we consider transfer functions of the form $g(s) = R(s)B(s)$, where $R(s)$ is a proper rational function in $H^\infty(RHP)$ and $B(s)$ is an infinite Blaschke product also in $H^\infty(RHP)$. We assume that there is no pole-zero cancellation. That is, the zeros of $R(s)$ ($B(s)$) do not coincide with any of the poles of $B(s)$ (respectively, $R(s)$). We point out that B has the form

$$B(s) = \prod_{n=1}^{\infty} \frac{|1-\beta_n^2|}{1-\beta_n^2}\,\frac{s-\beta_n}{s+\bar{\beta}_n},$$

where $\frac{|1-\beta_n^2|}{1-\beta_n^2}$ is assumed to be 1 if $\beta_n = 1$. The zeros $\beta_n (n = 1, 2, \ldots)$ of B satisfy the condition (see [17])

$$\sum_{n=1}^{\infty} \frac{Re(\beta_n)}{1+|\beta_n|^2} < \infty.$$

Note that either infinity is an accumulation point of the zeros (and the poles) of B, or else, the zeros (and the poles) of B are bounded and have accumulation points which are on the imaginary line.

First we consider the case that $R(s)$ is not strictly proper and the zeros of $R(s)$ do not coincide with any of the accumulation points of the poles of $B(s)$.

Write $R(s) = n(s)/d(s)$, where $n(s)$ and $d(s)$ are coprime polynomials. Then we have

$$g(s) = \frac{d^*(s)}{d(s)}B(s)\frac{n(s)}{d^*(s)},$$

where $d^*(s) = \overline{d(-\bar{s})}$. Set $q(s) = \frac{d^*(s)}{d(s)} B(s)$ and $f(s) = \frac{n^*(s)}{d(s)}$. We have $g = qf^*$. The inner function $q(s)$ is again a Blaschke product and $f(s)$ is in $H^\infty(RHP)$ and rational. Furthermore, from the assumption on $R(s)$ and $B(s)$ it follows that the zeros of $f(s)$ do not coincide with any of the zeros or accumulation points of the zeros of $q(s)$. Thus we must have

$$\inf_{s \in RHP} |f(s)| + |q(s)| > 0.$$

This shows that g has a strongly coprime Douglas–Shapiro–Shields factorization. Hence the Hankel operator H_g has closed range. Thus by Proposition 6.2 all reachable and observable admissible realizations of g are equivalent. Therefore all these realizations are asymptotically stable. They are exponentially stable if and only if there exists $\alpha > 0$ such that g is analytic on $Re(s) > -\alpha$. Since $R(s)$ is rational and in $H^\infty(RHP)$, we know that g is analytic on $Re(s) > -\alpha$ for some $\alpha > 0$ if and only if there is $\lambda > 0$ such that $B(s)$ is analytic on $Re(s) > -\lambda$. Note that the last condition on $B(s)$ is equivalent to that there is $\lambda > 0$ such that $Re(\beta_n) > \lambda$, $n = 1, 2, \ldots$.

By Corollary 8.8 we know that the input and output operators of any reachable and observable admissible realization of g are bounded if and only if $\sum Re(\beta_n) < \infty$.

The second case is that $R(s)$ is strictly proper, no zero of $R(s)$ coincides with any accumulation point of the poles of $B(s)$, and infinity is not an accumulation point of the poles of $B(s)$. In this case B is analytic at infinity and the poles of B have accumulation points on the imaginary line. As in the first case, g has a strongly coprime factorization and hence H_g has closed range. Thus all reachable and observable admissible realizations of g are equivalent and asymptotically stable. However, no reachable and observable realization of g is exponentially stable, since the poles of B have accumulation points on the imaginary line and hence g is not analytic on $Re(s) > -\alpha$ for any $\alpha > 0$.

Since in this case we have $g \in H^2(RHP)$ by Proposition 8.4, the input and output operators of any reachable and observable realization of g are bounded.

The third case is that $R(s)$ is strictly proper, no zero of $R(s)$ coincides with any accumulation point of the poles of $B(s)$, and infinity is an accumulation point of the poles of $B(s)$. In this case we can show as was done in Example 2 that the factorization of g in the first case is a weakly coprime factorization. Hence g is strictly noncyclic. Thus all input-normal, output-normal, and parbalanced realizations of g are asymptotically stable. As in the first case, an input-normal, an output-normal, or a parbalanced realization of g is exponentially stable if and only if there exists $\lambda > 0$ such that $Re(\beta_n) > \lambda$, $(n = 1, 2, \ldots)$.

From Corollary 8.8 it follows that the input operator of an input-normal realization or the output operator of an output-normal realization is bounded if and only if $\sum Re(\beta_n) < \infty$. Thus by Proposition 8.9 and Corollary 8.10 the input operator and output operator of any parbalanced realization of g are bounded if $\sum Re(\beta_n) < \infty$.

Since clearly $g \in H^2(RHP)$, by Proposition 8.4 the input operator of an output-normal realization and the output operator of an input-normal realization of g are bounded. If in addition no accumulation point of the poles of $B(s)$ is on the imaginary line, then g is continuous in the extended imaginary line and therefore g has a balanced realization.

We observe that in this case an output-normal realization cannot have a bounded output operator and still be exponentially stable. An analogous fact holds for an input-normal realization and its input operator.

The fourth and final case is that at least one of the zeros of $R(s)$ coincides with an accumulation point of the poles of $B(s)$. Note that this accumulation point must be on the imaginary line.

As in the previous case, the factorization of g in the first case is a weakly coprime factorization. Hence g is strictly noncyclic. Thus all input-normal, output-normal, and parbalanced realizations of g are asymptotically stable. They are not exponentially stable because g is not analytic on $Re(s) > -\alpha$ for any $\alpha > 0$.

Again by Corollary 8.8 the input operator of an input-normal realization or the output operator of an output-normal realization is bounded if and only if $\sum Re(\beta_n) < \infty$. Thus by Proposition 8.9 and Corollary 8.10 the input operator and output operator of any parbalanced realization of g are bounded if $\sum Re(\beta_n) < \infty$.

If every accumulation point of the poles of B is a zero of R, then g is continuous on the extended imaginary line. Hence g has a balanced realization.

We now summarize the results as follows.

PROPOSITION 9.3. *Consider $g(s) = R(s)B(s)$, where $R(s)$ is a proper rational function and $B(s)$ is an infinite Blaschke product, both in $H^\infty(RHP)$, and B and R have no pole-zero cancellation.*

1. *If $R(s)$ is not strictly proper and no zero of $R(s)$ coincides with any accumulation point of the poles of $B(s)$, then*
   (a) *all reachable and observable admissible realizations of g are equivalent;*
   (b) *all reachable and observable admissible realizations of g are asymptotically stable;*
   (c) *all reachable and observable admissible realizations of g are exponentially stable if and only if there exists $\alpha > 0$ such that $Re(\beta_n) > \alpha$, $n = 1, 2, \ldots$, where β_n, $n = 1, 2, \ldots$, are the zeros of $B(s)$;*
   (d) *all reachable and observable admissible realizations of g have bounded input and output operators if and only if $\sum Re(\beta_n) < \infty$.*

2. *If $R(s)$ is strictly proper, no zero of $R(s)$ coincides with any accumulation point of the poles of $B(s)$, and infinity is not an accumulation point of the poles of $B(s)$, then*
   (a) *all reachable and observable admissible realizations of g are equivalent;*
   (b) *all reachable and observable admissible realizations of g are asymptotically stable;*
   (c) *no reachable and observable admissible realization of g is exponentially stable;*
   (d) *all reachable and observable admissible realizations of g have bounded input and output operators.*

3. *If $R(s)$ is strictly proper, no zero of $R(s)$ coincides with any accumulation point of the poles of $B(s)$, and infinity is an accumulation point of the poles of $B(s)$, then*
   (a) *all input-normal, output-normal, and parbalanced realizations of g are asymptotically stable;*
   (b) *all input-normal, output-normal, and parbalanced realizations of g are exponentially stable if and only if there exists $\alpha > 0$ such that $Re(\beta_n) > \alpha$ $(n = 1, 2, \ldots)$;*
   (c) *the input operator of an input-normal realization or the output operator of an output-normal realization of g is bounded if and only if $\sum Re(\beta_n) < \infty$. The input operator and output operator of any parbalanced realization of g are bounded if $\sum Re(\beta_n) < \infty$;*
   (d) *the input operator of an output-normal realization and the output operator an input-normal realization of g are bounded.*

If, in addition, no accumulation point of the poles of B is on the imaginary line, then g has a balanced realization.

4. *If at least one of the zeros of R coincides with an accumulation point of the poles of B, then*
   (a) *all input-normal, output-normal, and parbalanced realizations of g are asymptotically stable;*
   (b) *no input-normal, output-normal, or parbalanced realization of g is exponentially stable;*
   (c) *the input operator of an input-normal realization or the output operator of an output-normal realization of g is bounded if and only if $\sum Re(\beta_n) < \infty$. The input operator and output operator of any parbalanced realization of g are bounded if $\sum Re(\beta_n) < \infty$.*

If every accumulation point of the poles of B is a zero of R, then g has a balanced realization.

REFERENCES

[1] P. R. AHERN AND D. N. CLARK, *Radial limits and invariant subspaces*, Amer. J. Math., 92 (1970), pp. 332-342.

[2] J. BARAS AND R. BROCKETT, *H^2-functions and infinite-dimensional realization theory*, SIAM J. Control, 13 (1973), pp. 221-241.

[3] J. BARAS, R. BROCKETT, AND P. FUHRMANN, *State-space models for infinite-dimensional systems*, IEEE Trans. Automat. Control, 19 (1974), pp. 693-700.

[4] R. CURTAIN, *Equivalence of input-output stability and exponential stability for infinite dimensional systems*, Math. Systems Theory, 21 (1988), pp. 19-48.

[5] R. CURTAIN AND G. WEISS, *Well posedness of triples of operators (in the sense of linear systems theory)*, in Control and Estimation of Distributed Parameter Systems, Birkhäuser, Basel, 1989, pp. 41-59.

[6] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Inform. Sci. 8, Springer-Verlag, New York, 1978.

[7] P. DEWILDE, *Input-output description of roomy systems*, SIAM J. Control Optim., 14 (1976), pp. 712-736.

[8] R. G. DOUGLAS, H. S. SHAPIRO, AND A. L. SHIELDS, *Cyclic vectors and invariant subspaces for the backward shift operator*, Ann. Inst. Fourier (Grenoble), 20 (1970), pp. 37-76.

[9] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators Part I: General Theory*, Interscience, New York, 1959.

[10] H. DYM, *J Contractive Matrix Functions, Reproducing Kernal Hilbert Spaces and Interpolation*, CBMS Lecture Notes, Washington, DC, 1988.

[11] P. A. FUHRMANN, *Linear Systems and Operators in Hilbert Space*, McGraw-Hill, New York, 1981.

[12] L. GEARHEART, *On the Spectral Theory of the Translation Semigroup and Its Commutant*, Ph.D. thesis, University of Illinois, 1975.

[13] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds*, Internat. J. Control, 39 (1984), pp. 1115-1193.

[14] K. GLOVER, R. F. CURTAIN, AND J. R. PARTINGTON, *Realisation and approximation of linear infinite dimensional systems with error bounds*, SIAM J. Control Optim., 26 (1988), pp. 863-898.

[15] H. HELSON, *Lectures on Invariant Subspaces*, Academic Press, New York, 1964.

[16] J. W. HELTON, *Systems with infinite dimensional state space*, Proc. IEEE, 64 (1976), pp. 145-160.

[17] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[18] H. LOGEMANN, *On the transfer matrix of a neutral system: Characterization of exponential stability in input-output terms*, Systems Control Lett., 9 (1987), pp. 393-400.

[19] J. MOELLER, *Translation invariant spaces with zero-free spectra*, Duke Math. J., 31 (1964), pp. 98-108.

[20] B. C. MOORE, *Principal component analysis in linear systems: Controllability, observability and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17-32.

[21] R. NAGEL, ED., *One-Parameter Semigroups of Positive Operators*, Lecture Notes in Math., Springer-Verlag, New York, 1986.

[22] N. K. NIKOL'SKIĬ, *Treatise on the Shift Operator: Spectral Function Theory*, Springer-Verlag, New York, 1986.

[23] R. OBER AND S. MONTGOMERY-SMITH, *Bilinear transformation of infinite dimensional state space systems and balanced realizations of nonrational transfer functions*, SIAM J. Control Optim., 28 (1990), pp. 439-465.

[24] R. OBER AND Y. WU, *Asymptotic stability of infinite dimensional discrete-time balanced realizations*, SIAM J. Control Optim., 31 (1993), pp. 1321-1339.

[25] J. R. PARTINGTON, *An Introduction to Hankel Operators*, London Math. Soc. Stud. Texts, Cambridge University Press, Cambridge, UK, 1988.

[26] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[27] D. SALAMON, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147-164.

[28] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North Holland, Amsterdam, 1970.

[29] Y. YAMAMOTO, *Equivalence of internal and external stability for a class of distributed systems*, Math. Control Signals Systems, 4 (1991), pp. 391-409.

[30] N. YOUNG, *Balanced Realizations in Infinite Dimensions*, Oper. Theory Adv. Appl. 19, Birkhäuser Verlag, Basel, 1986, pp. 449-470.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 813-832, May 1996

002

DYNAMIC POLE ASSIGNMENT AND SCHUBERT CALCULUS*

M. S. RAVI†, JOACHIM ROSENTHAL‡, AND XIAOCHANG WANG§

Abstract. The output feedback pole assignment problem is a classical problem in linear systems theory. In this paper we calculate the number of complex dynamic compensators of order q assigning a given set of poles for a q-nondegenerate m-input, p-output system of McMillan degree $n = q(m + p - 1) + mp$. As a corollary it follows that when this number is odd, the generic system can be arbitrarily pole assigned by output feedback with a real dynamic compensator of order at most q if and only if $q(m + p - 1) + mp \geq n$.

Key words. output feedback pole assignment, dynamic compensator, holomorphic curves in Grassmannian, degree of variety

AMS subject classifications. 93B55, 93B27, 14M15

1. Introduction.

The output feedback pole assignment of linear systems with static or dynamic compensators is a classical problem in control theory and many theoretical and numerical research papers have been devoted to this problem.

Although the systems involved are linear, the problem is in fact not linear. It was Brockett and Byrnes [4] who first explained the pole assignment problem with static compensators as an intersection problem in a compactified set of static compensators, the Grassmann manifold $\text{Grass}(m, m + p)$. In making the connection to the classical Schubert calculus they were able to show that there are

$$d(m, p) = \deg \text{Grass}(m, m + p) = \frac{1!2! \cdots (p-1)!(mp)!}{m!(m+1)! \cdots (m+p-1)!} \tag{1.1}$$

complex static output feedback laws which assign a set of poles for a nondegenerate m-input, p-output linear system of McMillan degree $n = mp$. In particular if the number $d(m, p)$ is odd, pole assignment by real static feedback is possible, because the set of complex solutions has to be invariant under complex conjugation. Moreover even if $d(m, p)$ is even, Wang, using algebrogeometric techniques, showed in [25] that a real solution exists for the generic system as soon as $mp > n$.

People have been looking for similar results for the dynamic pole assignment problem for a long time. A first attempt was made by Byrnes in [5]. Recently Rosenthal interpreted in [16, 17] the pole assignment problem with dynamic compensators, again as an intersection problem in a compactified space of dynamic compensators which we denote by $K^q_{m,p}$. It was also proven in [17] that if a plant has McMillan degree $n = q(m + p - 1) + mp$ and is q-nondegenerate, then there exist

$$d(m, p, q) = \deg K^q_{m,p} \tag{1.2}$$

complex dynamic feedback compensators of order q which assign a set of $n + q$ closed-loop poles. At this point we want to mention that all major results derived in [4, 17, 25] are based on a careful study of the associated pole assignment map. (See §2 for more details.) Indeed the number $d(m, p, q)$ can also be viewed as the mapping degree of the associated pole assignment map and this map has geometrically the format of a central projection.

*Received by the editors July 26, 1993; accepted for publication (in revised form) December 22, 1994.

†Department of Mathematics, East Carolina University, Greenville, NC 27858 (maravi@ecuvax.cis.ecu.edu).

‡Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556-5683 (Joachim.Rosenthal@nd.edu). The research of this author was supported in part by National Science Foundation grant DMS-9400965.

§Department of Mathematics, Texas Tech University, Lubbock, TX 79409-2013 (mdxia@ttacs1.ttu.edu). The research of this author was supported in part by National Science Foundation grant DMS-9224541.

One goal of our paper is to derive a formula for $d(m, p, q)$. Historically the formula (1.1) for $d(m, p) = d(m, p, 0)$ was first discovered in 1886 by Hannibal Schubert [18], a German high school teacher, using a symbolic formalism known as Schubert calculus. Using modern language the number $d(m, p)$ is equal to σ_1^{mp}, where σ_1 denotes the first Chern class of the classifying bundle over the Grassmann manifold Grass$(m, m+p)$. By applying Pieri's formula (see §3 for more details)

$$(i_1, i_2, \ldots, i_m) \cdot \sigma_1 = \sum_{i_l - 1 > i_{l-1}} (i_1, \ldots, i_l - 1, \ldots, i_m) \tag{1.3}$$

repeatedly to $(p, p+2, p+3, \ldots, p+m) = \sigma_1$, we can compute the number $d(m, p) = \deg \text{Grass}(m, m+p)$.

In [15] we defined a set of subvarieties of $K_{m,p}^q$ similar to the Schubert varieties of Grass$(m, m+p)$ and proved a geometric formula similar to Pieri's formula (1.3). This enables us to express $d(m, p, q) = \deg K_{m,p}^q$ as the solution of a partial difference equation with boundary condition. In this paper (§3) we will solve this difference equation and derive a closed formula for $d(m, p, q)$ which is valid for all positive integers m, p, and q. From this formula we finally will derive several new results which predict real and complex solutions assigning a specific set of closed-loop poles. One of the main results of this paper is Theorem 1.1.

THEOREM 1.1. *The poles of an m-input, p-output, q-nondegenerate, linear system of McMillan degree*

$$n = q(m + p - 1) + mp \tag{1.4}$$

can be assigned arbitrarily by using output feedback with complex dynamic compensators of order at most q, and there are

$$d(m, p, q) = (-1)^{q(m+1)}(mp + q(m+p))! \sum_{n_1 + \cdots + n_m = q} \frac{\prod_{k<j} (j - k + (n_j - n_k)(m+p))}{\prod_{j=1}^{m} (p + j + n_j(m+p) - 1)!} \tag{1.5}$$

complex solutions for each set of poles. In particular, if $d(m, p, q)$ is odd, a real solution always exists. Moreover when $d(m, p, q)$ is odd, the generic system can be arbitrarily pole assigned by output feedback with real dynamic compensators of order at most q if and only if

$$n \leq q(m + p - 1) + mp. \tag{1.6}$$

The variety $K_{m,p}^q$ which parameterizes the set of m-input, p-output compensators of McMillan degree q can also be viewed as a parameterization of the space of rational curves of degree q on the Grassmann variety Grass$(m, m+p)$. This geometric link originates from the well-known Hermann-Martin identification [12]. (Compare also with [6, 17].)

We were surprised to learn that there has recently been a tremendous interest in the intersection theory of parameterized curves (of arbitrary genus) on Grassmann varieties and other homogeneous spaces [2, 8, 24, 31]. Researchers working in conformal quantum field theory conjectured several new intersection numbers and an interesting formula for all numbers $d(m, p, q)$, different from (1.5), was part of this conjecture. Readers interested in the physics behind this conjecture are referred to Vafa [24]. The conjecture itself is formulated by Intriligator in [8] as well as in [2]. In [15] we were able to verify this conjecture for all numbers $d(m, p, q)$. More recently Siebert and Tian [22] presented a proof covering the conjecture for all spaces of parameterized curves on a Grassmann variety. For readers interested in these connections we will give some more details at the end of §3.

The paper is organized as follows. In the next section we review the notion of an autoregressive system. This class of systems generalizes the class of transfer functions and it allows us to define the pole placement map by using the behavioral approach to systems modeling as proposed by Willems [28, 29]. In this framework the points of the variety $K^q_{m,p}$ naturally parameterize all autoregressive compensators of a fixed number of inputs, outputs, and a bounded McMillan degree. We also restate the main results derived in [17], which were in part the motivation of this paper. We conclude this section with two new theorems (Theorems 2.14 and 2.15) which sharpen the main results derived in [17].

The main theorem (Theorem 1.1) is proven in §3. The proof involves the review of the generalized Pieri formula which was derived in [15]. To derive the new formula (1.5) describing the degree of the pole placement map in the critical dimension, we solve the partial recurrence relation mentioned earlier. This leads not only to a closed formula for the degree of the pole placement map in the critical dimension but also to a formula of the degree of some generalized Schubert varieties (Theorem 3.5). The section is concluded with several simplified formulas covering particular situations.

In §4 we concentrate on the question of for which triples m, p, q the degree $d(m, p, q)$ is odd, respectively, even. In Theorem 4.2 and Corollary 4.4 we present a relatively simple combinatorial procedure which computes the mod 2 degree of the variety $K^q_{m,p}$ for arbitrary m, p, q. Using this procedure we prove the existence of odd degrees even if $\min(m, p) \geq 3$, covering in this way many multi-input, multi-output feedback situations. (If $\min(m, p) \geq 3$ the degree of all Grassmann varieties is even. In part because of this there do not exist any positive pole placement results over $\mathbb{R}$ in the critical dimension, i.e., when $n = mp$.) We conclude the section with a complete description of all odd numbers $d(m, p, q)$ for $q = 0, 1, 2$.

Finally in the last section we merge the derived results and provide a collection of corollaries and consequences. In this section we also cover situations when the plant is represented by a "traditional" strictly proper transfer function or when the compensator is supposed to be a proper transfer function only.

2. The set of autoregressive systems $A^q_{m,p}$, the projective variety $K^q_{m,p}$, and the pole placement map. In this section we collect some mathematical preliminaries and simultaneously establish our notation. We develop the theory by using the behavioral approach of Willems [28] because we believe that the problem formulation in this setting is very natural. For the relation of this formulation to the traditional transfer function formulation we refer to [17, 28, 29].

First we review the notion of signal space, behavior, and autoregressive system. For this let $\mathbb{K}$ denote either the set of real numbers or the set of complex numbers, i.e., $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. Let $\mathbb{K}^{\mathbb{R}}$ denote the set of all functions $\psi : \mathbb{R} \to \mathbb{K}$. With respect to the usual addition and scalar multiplication of functions, $\mathbb{K}^{\mathbb{R}}$ is a real vector space. A linear subspace $\mathcal{H} \subset \mathbb{K}^{\mathbb{R}}$ which consists of functions that are arbitrarily many times differentiable will be called a signal space (see [3, 27]). In other words, $\mathcal{H}$ is a linear subspace which is invariant under the linear transformation $\frac{d}{dt}$. Usually we will assume that $\mathcal{H} = C^\infty(\mathbb{R}, \mathbb{K})$, though other function spaces are well possible. (Compare with [3, p. 76] and [28].)

Let $p(s)$ be a polynomial with coefficients in $\mathbb{K}$, i.e., $p(s) \in \mathbb{K}[s]$. Such a polynomial induces a linear transformation $\hat{p} : \mathcal{H} \to \mathcal{H}$, $w(t) \mapsto p(\frac{d}{dt})w(t)$. More generally consider a $p \times k$ polynomial matrix $P(s)$ with entries in $\mathbb{K}[s]$. $P(s)$ induces a linear transformation

$$\begin{aligned} (2.1) \qquad \hat{P} : \mathcal{H}^k &\longrightarrow \mathcal{H}^p, \\ *w(t) &\longmapsto P\left(\frac{d}{dt}\right) w(t). \end{aligned}$$

Using the language of Willems [28], we call the kernel of the linear transformation $\hat{P}$ the behavior and will denote this subset of the signal set $\mathcal{H}^k$ by $\mathcal{B}$.

In general the behavior $\mathcal{B} = \ker(P(\frac{d}{dt}))$ is an infinite-dimensional $\mathbb{R}$-vector space of the signal space $\mathcal{H}^k$. In the case where $P(s)$ is square and invertible it is, however, well known that the behavior $\mathcal{B}$ has real dimension $n = \deg \det P(s)$. Moreover the dynamics of this autonomous system are described by the roots of the characteristic polynomial $\det P(s) = 0$.

Recall that two $p \times k$ polynomial matrices $P(s)$ and $\tilde{P}(s)$ are called (row) equivalent if there is a unimodular matrix $U(s)$ with $\tilde{P}(s) = U(s)P(s)$. Clearly row equivalent matrices define the same behavior. On the other hand if the signal space is sufficiently rich, e.g., if $C^\infty(\mathbb{R}, \mathbb{R}) \subset \mathcal{H}$, we have the following result. (Compare with [3, §6.2] and [10, Thm. 3.9].)

LEMMA 2.1 (cf. [19, Cor. 2.5]). *If $C^\infty(\mathbb{R}, \mathbb{R}) \subset \mathcal{H}$, then $P(s)$ and $\tilde{P}(s)$ define the same behavior if and only if they are row equivalent.*

Based on this result we have the following definition.

DEFINITION 2.2. *An equivalence class of full rank $p \times k$ polynomial matrices is called an* autoregressive system.

The class of autoregressive systems generalizes the class of transfer functions in the following way. Consider a proper or improper $p \times m$ transfer function $G(s)$. Assume $G(s)$ has a left (polynomial) coprime factorization $D^{-1}(s)N(s) = G(s)$. If $\tilde{D}^{-1}(s)\tilde{N}(s) = G(s)$ is a second left coprime factorization, then it is well known that the $p \times (m + p)$ polynomial matrices $(N(s)\ D(s))$ and $(\tilde{N}(s)\ \tilde{D}(s))$ are row equivalent. In other words $(N(s)\ D(s))$ defines an autoregressive system.

The following definition extends the notion of McMillan degree to the class of autoregressive systems.

DEFINITION 2.3 (see [17, 26, 28]). *The degree of an autoregressive system $P(s)$ is given by the maximal degree of the full-size minors of $P(s)$.*

Next we would like to introduce feedback. For this consider a $p \times (m + p)$ autoregressive system $P(s)$ (the plant) and an $m \times (m + p)$ autoregressive system $C(s)$ (the compensator). The closed-loop system is then the dynamical system described through the system of autoregressive equations

$$\begin{pmatrix} P(\frac{d}{dt}) \\ C(\frac{d}{dt}) \end{pmatrix} \cdot w(t) = 0. \tag{2.2}$$

Note that the square polynomial matrix $\binom{P(s)}{C(s)}$ is in general not of full rank, i.e., (2.2) does not describe an autoregressive system as defined in Definition 2.2. To single out the compensators which give rise to a closed-loop autoregressive system we need the following definition (compare with [20]).

DEFINITION 2.4. *A compensator $C(s)$ is called* admissible *if the closed-loop characteristic polynomial*

$$\phi(s) := \det \begin{pmatrix} P(s) \\ C(s) \end{pmatrix} \neq 0. \tag{2.3}$$

We are now in a position to define the pole placement map. Let $P(s)$ be a $p \times (m + p)$ autoregressive system of McMillan degree n and denote by $A^q_{m,p}$ the set of all $m \times (m + p)$ autoregressive systems of McMillan degree at most q. Let $B_p \subset A^q_{m,p}$ be the set of autoregressive systems which are not admissible compensators. Finally identify the set of nonzero polynomials of degree at most d with the projective space $\mathbb{P}^d$. Then define the pole placement map as follows.

DEFINITION 2.5. *The pole placement map for a plant $P(s)$ is defined as the rational map given by*

$$\rho_P : A^q_{m,p} - B_P \longrightarrow \mathbb{P}^{n+q}, \tag{2.4}$$
$$C(s) \longmapsto \phi(s) = \det \begin{pmatrix} P(s) \\ C(s) \end{pmatrix}.$$

We want to note at this point that the roots of $\phi(s)$ do not depend on the particular representation of the plant $P(s)$ or the compensator $C(s)$. Indeed if $\tilde{P}(s) = U_1(s)P(s)$ and $\tilde{C}(s) = U_2(s)C(s)$, then $\tilde{\phi}(s) = \det U_1(s) \cdot \det U_2(s) \cdot \phi(s)$. Finally the roots of $\phi(s)$ correspond to the poles of the closed-loop system in the transfer function formulation. (See [17] for details.)

For a given plant $P(s)$ we usually say that $P(s)$ is pole assignable (almost pole assignable) in the class of feedback compensators of degree at most q if the map ρ_P is onto (almost onto). Though many results are known when a system is pole assignable in the class of feedback compensators of order at most q over the complex numbers $\mathbb{C}$ [17], the question is still far from being solved over the reals and in the ungeneric situation. (Compare with [28].) Clearly the following property is a necessary condition for pole assignability.

DEFINITION 2.6 (see [26, 28]). *An autoregressive system $P(s)$ is called* controllable *or* irreducible *if the matrix $P(s)$ is of full row rank for all $s \in \mathbb{C}$.*

Indeed if the system $P(s)$ is not controllable, the full-size minors of $P(s)$ have a common factor which is necessarily a factor of the closed-loop characteristic polynomial $\phi(s)$. Clearly, even if $P(s)$ is controllable we cannot expect that $P(s)$ is pole assignable in the set of feedback compensators of degree at most q. The following definition singles out an interesting class of systems which has the pole assignability property in the critical dimension (i.e., when $\dim A^q_{m,p} = \dim \mathbb{P}^{n+q}$) over the complex numbers.

DEFINITION 2.7 (see [17]). *A plant $P(s)$ is called* q-nondegenerate *if all compensators $C(s)$ of order at most q are admissible. To put it in other words, $P(s)$ is q-nondegenerate if the set B_p introduced in* (2.4) *is empty.*

In the last part of this section we establish the connection to our earlier work in [17, 26]. First we would like to point out the following observation. The pole placement map ρ_P as introduced in (2.4) actually depends only on the full-size minors of $P(s)$ and $C(s)$. In other words if $C(s)$ and $\tilde{C}(s)$ have the same full-size minors, then the resulting closed-loop characteristic polynomial $\rho_P(C(s))$ and $\rho_P(\tilde{C}(s))$ have the same roots. Based on this fact we assign to each autoregressive system $C(s) \in A^q_{m,p}$ its full-size minors, i.e., we consider the following Plücker map:

$$\pi : A^q_{m,p} \longrightarrow \mathbb{P}(\mathbb{K}^{q+1} \otimes \wedge^m \mathbb{K}^{m+p}) \tag{2.5}$$
$$C(s) \longmapsto c_1(s) \wedge \cdots \wedge c_m(s).$$

Here $c_l(s)$ denotes the lth row vector of the $m \times (m+p)$ matrix $C(s)$. Of course when describing the map π with respect to the standard basis

$$\{e_{i_1} \wedge \cdots \wedge e_{i_m} \mid 1 \le i_1 < \cdots < i_m \le m+p\}, \tag{2.6}$$

it is well known that the coordinates are exactly the full-size minors of the matrix $C(s)$. In particular the map π is well defined.

In the following, whenever we work with coordinates, we will assume the standard basis (2.6). More specifically, if

$$\pi(C(s)) = \sum_{i \in I(m)} f_i(s) \cdot e_{i_1} \wedge \cdots \wedge e_{i_m}, \tag{2.7}$$

we will use the coordinates

$$f_i(s) = z_{(i;q)}s^q + z_{(i;q-1)}s^{q-1} + \cdots + z_{(i;0)}. \tag{2.8}$$

The map π is in general not an embedding as it is for the classical Plücker embedding (the case $q = 0$). Indeed as shown in [17], $\pi(C(s)) = \pi(\tilde{C}(s))$ if and only if the matrices $C(s)$ and $\tilde{C}(s)$ are H-equivalent. (See [17] for details.) On the other hand if $C(s)$ and $\tilde{C}(s)$ are both controllable (see Definition 2.6), then $C(s)$ and $\tilde{C}(s)$ are H-equivalent if and only if they are row equivalent. The following lemma summarizes these statements.

LEMMA 2.8. *π restricted to the set of controllable autoregressive systems is an embedding, in particular π is generically one-to-one.*

From the earlier remarks it is clear that the pole placement map ρ_p factors over the image of π. We introduce therefore the following notation.

DEFINITION 2.9. *$K^q_{m,p}$ denotes the image of $A^q_{m,p}$ under the map π.*

By definition the set $K^q_{m,p}$ is a subset of the projective space

$$\mathbb{P}^N := \mathbb{P}(\mathbb{K}^{q+1} \otimes \wedge^m \mathbb{K}^{m+p}).$$

Note that the Plücker coordinates $\{f_i(s)\}$ introduced in (2.8) satisfy a set of quadratic relations coming from the description of Grass$(m, m+p)$ in $\mathbb{P}^{\binom{m+p}{m}-1}$ [7, p. 65]. Those relations must hold for all $s \in \mathbb{K}$. Equating coefficients we get a necessary set of quadratic relations for the coordinates $z_{(i;d)}$ as well. The following theorem states that those relations define $K^q_{m,p}$.

THEOREM 2.10 (see [17]). *$K^q_{m,p}$ is a projective (sub)variety of $\mathbb{P}^N$. The defining relations are given by a set of homogeneous quadratic polynomials obtained from equating the coefficients in the Plücker relations. The variety $K^q_{m,p}$ is in general singular and has dimension $q(m+p)+mp$.*

The following example explains the situation.

Example 2.11 (see [16]). The only Plücker relation of Grass(2, 4) in $\mathbb{P}^5$ is given by

$$x_{12}x_{34} - x_{13}x_{24} + x_{14}x_{23} = 0. \tag{2.9}$$

Let $f_{ij}(s) = z_{(ij;1)}s + z_{(ij;0)}$ and

$$f_{12}(s)f_{34}(s) - f_{13}(s)f_{24}(s) + f_{14}(s)f_{23}(s) = 0; \tag{2.10}$$

we then have three quadratic equations

$$z_{(12;1)}z_{(34;1)} - z_{(13;1)}z_{(24;1)} + z_{(14;1)}z_{(23;1)} = 0,$$

$$z_{(12;1)}z_{(34;0)} - z_{(13;1)}z_{(24;0)} + z_{(14;1)}z_{(23;0)} + z_{(12;0)}z_{(34;1)} - z_{(13;0)}z_{(24;1)} + z_{(14;0)}z_{(23;1)} = 0,$$

$$z_{(12;0)}z_{(34;0)} - z_{(13;0)}z_{(24;0)} + z_{(14;0)}z_{(23;0)} = 0,$$

which define the projective variety $K^1_{2,2}$ in $\mathbb{P}^{11}$. Because $\dim K^1_{2,2} = 8$, it follows that $K^1_{2,2}$ is a complete intersection and by Bézout's theorem [7], the degree is equal to $2^3 = 8$.

As we can describe the compensator $C(s)$ through the vector

$$c(s) = c_1(s) \wedge \cdots \wedge c_m(s), \tag{2.11}$$

we can describe the plant $P(s)$ through the vector

$$p(s) = p_1(s) \wedge \cdots \wedge p_p(s). \tag{2.12}$$

Finally the closed-loop characteristic polynomial is given through the linear pairing

$$\langle p(s), c(s)\rangle := c_1(s) \wedge \cdots \wedge c_m(s) \wedge p_1(s) \wedge \cdots \wedge p_p(s) = \phi(s). \tag{2.13}$$

Note that the linear pairing $\langle , \rangle$ originally defined on $K^n_{p,m} \times K^q_{m,p}$ extends linearly to the product space $\mathbb{P}(\mathbb{K}^{n+1} \otimes \wedge^p \mathbb{K}^{m+p}) \times \mathbb{P}(\mathbb{K}^{q+1} \otimes \wedge^m \mathbb{K}^{m+p})$.

Next we show that the pole placement map ρ_p induces a central projection in the projective space $\mathbb{P}^N = \mathbb{P}(\mathbb{K}^{q+1} \otimes \wedge^m \mathbb{K}^{m+p})$. For this consider a fixed plant $P(s)$ represented through the vector $p(s) = p_1(s) \wedge \cdots \wedge p_p(s)$. Consider the subspace

$$E_P := \{c(s) | \langle p(s), c(s)\rangle \equiv 0\} \subset \mathbb{P}^N. \tag{2.14}$$

Then we have a central projection (compare with [17, 25]):

$$\begin{aligned} L_P \; : \; \mathbb{P}^N - E_P &\longrightarrow \mathbb{P}^{n+q}, \\ f(s) &\longmapsto \langle\, g(s)\,,\, f(s)\,\rangle. \end{aligned} \tag{2.15}$$

Let χ_P be the restriction map $L_P \mid_{(K^q_{m,p} - E_P)}$, i.e.,

$$\chi_P \; : \; K^q_{m,p} - E_P \longrightarrow \mathbb{P}^{n+q}. \tag{2.16}$$

The next lemma explains the relation between the maps χ_P, L_P and the pole placement map ρ_P.

LEMMA 2.12. *The pole placement map ρ_P introduced in* (2.4) *factors over the variety $K^q_{m,p}$ through*

$$\rho_P = L_P \circ \pi. \tag{2.17}$$

The map ρ_P is onto (almost onto) if and only if χ_P is. Finally a plant $P(s)$ is q-nondegenerate if and only if $K^q_{m,p} \cap E_P = \emptyset$.

Proof. From the definition of the linear pairing $\langle , \rangle$ it is clear that $\rho_P = L_P \circ \pi$. Moreover because

$$\pi \; : \; A^q_{m,p} \to K^q_{m,p} \tag{2.18}$$

is onto, the second statement follows. Finally if $P(s)$ is q-degenerate there is a compensator $C(s) \in A^q_{m,p}$ which is not admissible. But this is equivalent to the statement

$$c_1(s) \wedge \cdots \wedge c_m(s) \in K^q_{m,p} \cap E_P. \tag{2.19}$$

□

This lemma will allow us to study the pole assignment problem completely in the projective space $\mathbb{P}^N$. In the geometric picture the set $K^q_{m,p} \cap E_P$ will be of crucial importance. Note that

$$\pi(B_P) = K^q_{m,p} \cap E_P.$$

By abuse of notation we will denote $K^q_{m,p} \cap E_P$ by B_P as well; Lemma 2.12 justifies this choice. The set B_P is sometimes called the base locus of the central projection χ_P and by Lemma 2.12 this set is empty if and only if the plant $P(s)$ is q-nondegenerate. The following theorem gives the result which mainly motivated this paper.

THEOREM 2.13 (see [17]). *For a q-nondegenerate system of McMillan degree $n = q(m + p - 1) + mp$, the pole assignment map χ_P is onto over $\mathbb{C}$ and there are* $\deg K^q_{m,p}$ *(counted with multiplicity) complex dynamic compensators assigning each set of poles. In particular, a real solution always exists if* $\deg K^q_{m,p}$ *is odd.*

Proof. Since $B_P = \emptyset$, the pole placement map $\chi_P : K^q_{m,p} \to \mathbb{P}^{n+q}$ is a finite morphism [21, Chap. I, §5, Thm. 7]. Therefore χ_P is onto over $\mathbb{C}$ [21, Chap. I, §5, Thm. 4] and $\deg \chi_P = \deg K^q_{m,p}$ [13, Cor. (5.6)]. □

Actually we can strengthen this result with the following theorem.

THEOREM 2.14. *Let P be a system of degree $n < q(m + p - 1) + mp$. If*

$$\dim B_P = \dim E_P \cap K^q_{m,p} = q(m + p) + mp - n - q - 1, \tag{2.20}$$

then χ_P is onto over $\mathbb{C}$ (and over $\mathbb{R}$ if $\deg K^q_{m,p}$ is also odd).

Proof. Let H be the $q(m + p) + mp - n - q$ codimensional projective subspace in $\mathbb{P}^N$ such that

$$B_P \cap H = \emptyset \tag{2.21}$$

(such H exists by [13, Cor. (2.29)]), $\pi_1 : K^q_{m,p} \to \mathbb{P}^{q(m+p)+mp}$ is the central projection with center $E_P \cap H$, and $\pi_2 : \mathbb{P}^{q(m+p)+mp} - \pi_1(B_P) \to \mathbb{P}^{n+q}$ is the central projection with center $\pi_1(E_P)$. Then π_1 is onto over $\mathbb{C}$ and is onto over $\mathbb{R}$ if $\deg K^q_{m,p}$ is also odd, and

$$\chi_P = \pi_2 \circ \pi_1. \qquad \square$$

THEOREM 2.15. *The pole assignment map χ_P is onto over $\mathbb{C}$ for the generic system if and only if*

$$n \leq q(m + p - 1) + mp. \tag{2.22}$$

This condition is also sufficient over $\mathbb{R}$ if $\deg K^q_{m,p}$ is odd.

Proof. The necessity was proven by Willems and Hesselink in [30]. On the other hand if $n = q(m + p - 1) + mp$, the generic system is q-nondegenerate by [17, Cor. 5.6] and the sufficiency follows from Theorem 2.13. If $n < q(m + p - 1) + mp$, then it follows for the generic system from [17, Thm. 5.5] that

$$\dim B_P = q(m + p) + mp - n - q - 1. \tag{2.23}$$

By Theorem 2.14 the sufficiency follows. □

3. The subvarieties Z_α of the variety $K^q_{m,p}$ and a closed formula of their degrees. In this section we derive a closed formula for the degree of a set of generalized Schubert subvarieties of the variety $K^q_{m,p}$. As a corollary we will obtain a formula for the mapping degree of the pole placement map in the critical dimension. For the convenience of the reader we quickly review some geometric aspects of the classical Pieri formula (1.3). For this consider the index set

$$I = \{i = (i_1, \ldots, i_m) | 1 \leq i_1 < \cdots < i_m\}$$

equipped with the partial order

$$(i_1, \ldots, i_m) \leq (j_1, \ldots, j_m) \quad \Leftrightarrow \quad i_l \leq j_l \forall l. \tag{3.1}$$

If an m-dimensional plane $P \in \mathrm{Grass}(m, m + p) \subset \mathbb{P}(\wedge^m \mathbb{K}^{m+p})$ is expanded in terms of the standard basis (2.6), i.e., if P is represented by the vector

$$x := \sum_{i \in I} x_i \cdot e_{i_1} \wedge \cdots \wedge e_{i_m}, \tag{3.2}$$

we will call the coordinates x_i the Plücker coordinates (see [7, p. 64]) of the plane P. The set

$$S_i := \{x \in \mathrm{Grass}(m, m+p) | x_j = 0 \text{ for all } j \not\leq i\} \tag{3.3}$$

is called a Schubert variety. Let H_i be the hyperplane defined by setting $x_i = 0$ and let $|i| := \sum_{l=1}^{m}(i_l - l)$. Then the geometric version of Pieri's formula states that

$$S_i \cap H_i = \bigcup_{\substack{j \in I \\ j<i,\ |j|=|i|-1}} S_j \tag{3.4}$$

and that the intersection multiplicity along each S_j is 1. In terms of the intersection ring, S_i represents a Schubert cycle $(i_1, i_2, \ldots, i_m)$, H_i represents the Schubert cycle $\sigma_1 := (p, p+2, p+3, \ldots, p+m)$, and the geometric intersection is expressed through a formal multiplication as given in (1.3). Readers who want to learn more about Schubert calculus are referred to the excellent survey article of Kleiman and Laksov [9].

In [15] we proved a similar formula as given in (3.4) for subvarieties of $K^q_{m,p}$. To explain this generalized Pieri formula we first re-index the coordinates $z_{(i;d)}$ of $K^q_{m,p}$.

DEFINITION 3.1. *For each* $(i; d)$, $i = (i_1, \ldots, i_m)$, $1 \leq i_1 < \cdots < i_m \leq m+p$, *let* $\alpha := (\alpha_1, \ldots, \alpha_m)$ *be defined as*

$$\alpha_l = \begin{cases} [d/m](m+p) + i_{l+d-m[d/m]} & \text{for } l = 1, 2, \ldots, m[d/m] + m - d, \\ ([d/m]+1)(m+p) + i_{l+d-m[d/m]-m} & \text{for } l = m[d/m] + m - d + 1, \ldots, m. \end{cases}$$

Using this re-indexing we can associate to every coordinate $z_{(i;d)}$ of $K^q_{m,p}$ a new coordinate z_α. The following example shows the relation between the indices $(i; d)$ and α:

$$\begin{aligned} z_{(i;0)} &= z_i, \\ z_{(i;1)} &= z_{(i_2,\ldots,i_m,i_1+m+p)}, \\ z_{(i;2)} &= z_{(i_3,\ldots,i_m,i_1+m+p,i_2+m+p)}, \\ &\vdots \\ z_{(i;m)} &= z_{(i_1+m+p,\ldots,i_m+m+p)}, \\ z_{(i;m+1)} &= z_{(i_2+m+p,\ldots,i_m+m+p,i_1+2(m+p))}, \\ &\vdots \end{aligned}$$

Note that the indices α belong to the index set

$$\tilde{I} := \{\alpha \in I | \alpha_m - \alpha_1 < m + p\}, \tag{3.5}$$

which is by definition a subset of the index set I. In particular $\tilde{I}$ is also equipped with a partial order. Using this partial order we can now define an interesting set of subvarieties of $K^q_{m,p}$.

DEFINITION 3.2.

$$Z_\alpha := \{z \in K^q_{m,p} | z_\beta = 0 \textit{ for all } \beta \not\leq \alpha\}. \tag{3.6}$$

The main results of [15] are summarized in the following proposition and corollary.

PROPOSITION 3.3 (see [15]). *For each index* α, Z_α *is a subvariety of dimension* $|\alpha|$. *If* H_α *is the hyperplane of* $\mathbb{P}^N$ *defined by* $z_\alpha = 0$, *then*

$$Z_\alpha \cap H_\alpha = \bigcup_{\substack{\beta \in \tilde{I} \\ \beta<\alpha,\ |\beta|=|\alpha|-1}} Z_\beta \tag{3.7}$$

and the intersection multiplicity along each Z_β *is* 1.

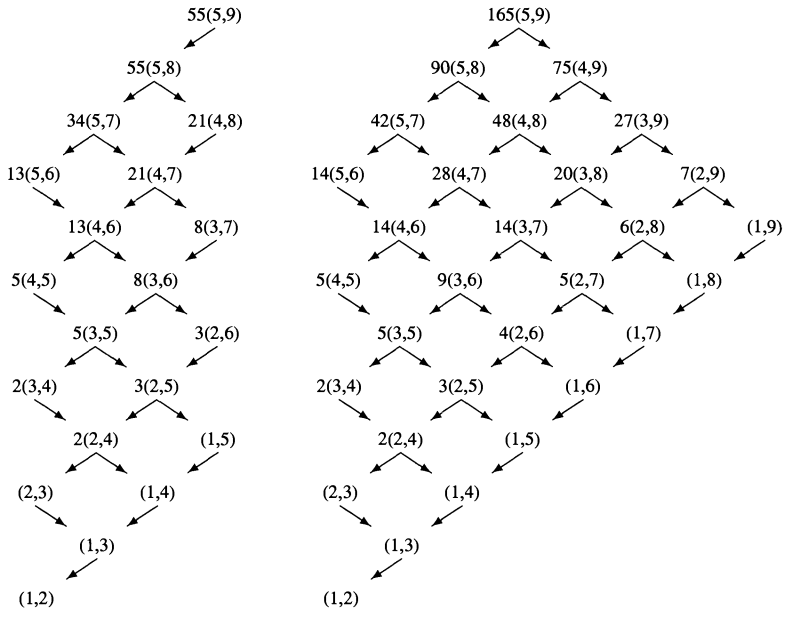


FIG. 1. (a) *Hasse Diagram of* $Z_{(5,9)}$. (b) *Hasse Diagram of* $S_{(5,9)}$.

Using Bézout's theorem [7, Thm. 18.3] the expression (3.7) translates into a partial recurrence relation, which the degrees of the varieties Z_α have to satisfy.

COROLLARY 3.4 (see [15]).

$$\deg Z_\alpha = \sum_{\substack{\beta \in \tilde{I} \\ \beta < \alpha,\ |\beta| = |\alpha| - 1}} \deg Z_\beta. \tag{3.8}$$

The partial recurrence relation (3.8) has to be satisfied for the whole index set $\tilde{I}$. It is possible to depict this relation with the help of a Hasse diagram. A Hasse diagram corresponding to the variety Z_α is a directed graph, whose vertices are all $\beta \in \tilde{I}$, $\beta \leq \alpha$. The directed edges $\beta \to \gamma$ are precisely those ordered pairs such that β covers γ (i.e., $\beta > \gamma$ and $|\beta| = |\gamma| + 1$). Then according to Corollary 3.4, the degree of Z_α can be computed graphically in the following way: If we label the vertices in such a way that the number on $(1, 2, \ldots, m)$ is 1 and the number on β is the sum of the numbers on the vertices covered by β, then the number on α is $\deg Z_\alpha$. Figure 1 provides an example of $Z_{(5,9)} = K^1_{2,3}$. Note that Rosenthal obtained $\deg K^1_{m,p} = 55$ by computing the coefficients of the Hilbert polynomial using the computer program CoCoA in [16]. For comparison we also include the Hasse diagram of the Schubert variety $S_{(5,9)}$, whose underlying diagram corresponds to all indices $i \in I, i \leq (5, 9)$.

From Fig. 1 we can see that the Hasse diagram of Z_α an be obtained by "cutting off" all the vertices of I that are not in $\tilde{I}$ in the Hasse diagram of S_α. If we use $d(\alpha_1, \ldots, \alpha_m)$ for the degree, then both $\deg Z_\alpha$ and $\deg S_\alpha$ satisfy the partial difference equation

$$d(\alpha_1, \ldots, \alpha_m) = \sum_{l=1}^{m} d(\alpha_1, \ldots, \alpha_l - 1, \ldots, \alpha_m) \tag{3.9}$$

subject to the initial condition

$$d(1, 2, \ldots, m) = 1 \tag{3.10}$$

and subject to the boundary conditions

$$d(0, \ldots, \alpha_m) = 0, \tag{3.11}$$

$$d(\ldots, k, k, \ldots) = 0. \tag{3.12}$$

$\deg Z_\alpha$ is subject to one more boundary condition, namely,

$$d(k, \ldots, k + m + p) = 0. \tag{3.13}$$

The computation of the degrees of the varieties Z_α is therefore reduced to the solution of a partial difference equation with boundary conditions. The next theorem provides a closed formula for this problem.

THEOREM 3.5.

$$\deg Z_\alpha = |\alpha|! \sum_{n_1+\cdots+n_m=0} \frac{\prod_{k<j}(\alpha_j - \alpha_k + (n_j - n_k)(m+p))}{\prod_j (\alpha_j + n_j(m+p) - 1)!} \tag{3.14}$$

with the convention that $1/k! = 0$ *if* $k < 0$.

Proof. Let $g(\alpha) = \deg S_\alpha$. Then (see [11, p. 103] and [23])

$$g(\alpha) = |\alpha|! \frac{\prod_{k<j}(\alpha_j - \alpha_k)}{\prod_{j=1}^{m}(\alpha_j - 1)!} = |\alpha|! \det \begin{bmatrix} \frac{1}{(\alpha_1-1)!} & \frac{1}{(\alpha_1-2)!} & \cdots & \frac{1}{(\alpha_1-m)!} \\ \frac{1}{(\alpha_2-1)!} & \frac{1}{(\alpha_2-2)!} & \cdots & \frac{1}{(\alpha_2-m)!} \\ \vdots & \vdots & & \vdots \\ \frac{1}{(\alpha_m-1)!} & \frac{1}{(\alpha_m-2)!} & \cdots & \frac{1}{(\alpha_m-m)!} \end{bmatrix} \tag{3.15}$$

and (3.14) becomes

$$\deg Z_\alpha = \sum_{n_1+\cdots+n_m=0} g(\alpha_1 + n_1(m+p), \ldots, \alpha_m + n_m(m+p)). \tag{3.16}$$

Let

$$d(\alpha) = \sum_{n_1+\cdots+n_m=0} g(\alpha_1 + n_1(m+p), \ldots, \alpha_m + n_m(m+p)).$$

Then $d(\alpha)$ satisfies the equation (3.9) because $g(\alpha)$ does. Moreover since

$$d(\alpha_1, \ldots, \alpha_m) = g(\alpha_1, \ldots, \alpha_m)$$

for $\alpha_1 < \cdots < \alpha_m < m + p$, $d(\alpha)$ satisfies the conditions (3.10) and (3.11) for $\alpha_m < m + p$. We only need to verify (3.12) and (3.13).

Notice that by (3.15)

$$g(\ldots, \alpha_j, \ldots, \alpha_k, \ldots) = -g(\ldots, \alpha_k, \ldots, \alpha_j \ldots). \tag{3.17}$$

If $\alpha_j = \alpha_{j+1}$, then

$$\begin{aligned} &g(\ldots, \alpha_j + k_j(m+p), \alpha_{j+1} + k_{j+1}(m+p), \ldots) \\ &= -g(\ldots, \alpha_j + k_{j+1}(m+p), \alpha_{j+1} + k_j(m+p), \ldots). \end{aligned}$$

On the other hand if $\alpha_1 + m + p = \alpha_m$, then

$$\begin{aligned} &g(\alpha_1 + k_1(m+p), \ldots, \alpha_m + k_m(m+p)) \\ &= g(\alpha_m + (k_1 - 1)(m+p), \ldots, \alpha_1 + (k_m + 1)(m+p)) \\ &= -g(\alpha_1 + (k_m + 1)(m+p), \ldots, \alpha_m + (k_1 - 1)(m+p)). \end{aligned}$$

In either case $d(\alpha) = -d(\alpha)$; i.e., $d(\alpha) = 0$. □

Applying the formula (3.14) to $K^q_{m,p}$ we then have a formula for $\deg K^q_{m,p}$.

THEOREM 3.6.

$$\deg K^q_{m,p} = (-1)^{q(m+1)}(mp+q(m+p))! \sum_{n_1+\cdots+n_m=q} \frac{\prod_{k<j}(j-k+(n_j-n_k)(m+p))}{\prod_{j=1}^{m}(p+j+n_j(m+p)-1)!}. \tag{3.18}$$

Proof. Let $k = [q/m]$ and $r = q - km$. Then

$$K^q_{m,p} = Z_{(p+r+1+k(m+p),\ldots,p+m+k(m+p),p+1+(k+1)(m+p),\ldots,p+r+(k+1)(m+p))}.$$

So

$$\deg K^q_{m,p} = (\operatorname{sgn}\sigma) \sum_{n_1+\cdots+n_m=q} g(p+1+n_1(m+p),\ldots,p+m+n_m(m+p)),$$

where σ is the permutation

$$(r+1, r+2, \ldots, m, 1, 2, \ldots, r) \to (1, 2, \ldots, m),$$

and the sign of this permutation is given by

$$\begin{aligned}\operatorname{sgn}\sigma &= (-1)^{r(m-r)} = (-1)^{(q-km)((k+1)m-q)} = (-1)^{2qkm+mq-q^2-k(k+1)m^2} \\ &= (-1)^{mq}(-1)^{q^2} = (-1)^{mq}(-1)^q.\end{aligned}$$

Therefore

$$\deg K^q_{m,p} = (-1)^{q(m+1)} \sum_{n_1+\cdots+n_m=q} g(p+1+n_1(m+p),\ldots,p+m+n_m(m+p)),$$

which is the formula (3.18). □

Combining Theorems 2.13, 2.15, and 3.6, we then have Theorem 1.1.

We conclude this section with several simplified formulas. First recall the definition of the Fibonacci numbers given by the recurrence relation $f_1 = 1$, $f_2 = 1$, and $f_{n+1} = f_n + f_{n-1}$ for $n > 1$. From Corollary 3.4 it follows immediately that

$$\deg K^q_{2,3} = f_{5q+5}. \tag{3.19}$$

Using a well-known expression for the Fibonacci sequence we therefore get

$$\deg K^q_{2,3} = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^{5(q+1)} - \left(\frac{1-\sqrt{5}}{2}\right)^{5(q+1)}\right). \tag{3.20}$$

Note that formula (3.20) has also been given by Intriligator [8, p. 3554] as an illustration of the conjectured intersection numbers arising from some computation in conformal quantum field theory. For $q = 1$, we again get $\deg K^1_{2,3} = 55$ (compare with the Hasse diagram of $Z_{(5,9)}$) and for $q = 2$, we get $\deg K^2_{2,3} = 610$.

In general, for $m = 2$, formula (3.18) can be simplified to

$$\deg K^q_{2,p} = (-1)^q(q(p+2)+2p)! \sum_{j=0}^{q} \frac{(q-2j)(p+2)+1}{(p+j(p+2))!(p+1+(q-j)(p+2))!}.$$

To illustrate "the nonlinear character" of the pole placement map we derive a table that shows all degrees of the variety $K^q_{2,p}$ for $p = 1, \ldots, 9$ and $q = 0, \ldots, 5$; see Table 1.

TABLE 1.

$p\backslash q$	0	1	2	3	4	5
1	1	1	1	1	1	1
2	2	8	32	128	512	2048
3	5	55	610	6765	75025	832040
4	14	364	9842	265720	7174454	193710244
5	42	2380	147798	9112264	562110290	34673583028
6	132	15504	2145600	290926848	39541748736	5372862566400
7	429	100947	30664890	8916942687	2610763825782	763562937059280
8	1430	657800	435668420	266668876540	165745451110910	102703589621825280
9	4862	4292145	6186432967	7853149169635	10262482704258873	13319075453502743045

4. Odd or even degrees. In this section we introduce some methods that can be used to determine whether the deg $K^q_{m,p}$ is odd or even without computing the degree itself.

For Grass$(m, m+p) = K^0_{m,p}$, it is a well-known fact that deg Grass$(m, p+m)$ is even whenever $\min(m, p) \geq 3$ [1]. This is not the case for general $K^q_{m,p}$, i.e., in a certain sense there are many more odd numbers for a fixed $q > 0$ than there are for $q = 0$.

The main result of this section is Theorem 4.2, which provides a short combinatorial description of all triples m, p, q which result in an odd degree. Using this theorem we derive several corollaries classifying the odd- and even-degree varieties.

To prepare for the main theorem we first rewrite formula (3.18):

$$\deg K^q_{m,p} = (-1)^{q(m+1)}(mp+q(m+p))! \sum_{n_1+\cdots+n_m=q} (-1)^{\frac{m(m-1)}{2}}$$

$$\cdot \det \begin{bmatrix} \frac{1}{(p-m+1+n_1(m+p))!} & \frac{1}{(p-m+2+n_1(m+p))!} & \cdots & \frac{1}{(p+n_1(m+p))!} \\ \frac{1}{(p-m+2+n_2(m+p))!} & \frac{1}{(p-m+3+n_2(m+p))!} & \cdots & \frac{1}{(p+1+n_2(m+p))!} \\ \vdots & \vdots & & \vdots \\ \frac{1}{(p+n_m(m+p))!} & \frac{1}{(p+1+n_m(m+p))!} & \cdots & \frac{1}{(p+m-1+n_m(m+p))!} \end{bmatrix} \tag{4.1}$$

$$= (-1)^{q(m+1)}(-1)^{\frac{m(m-1)}{2}} \sum_{n_1+\cdots+n_m=q} \sum_{\sigma} \operatorname{sgn}\sigma$$

$$\cdot \frac{(mp+q(m+p))!}{(p-m+\sigma(1)+n_1(m+p))!\cdots(p-1+\sigma(m)+n_m(m+p))!}. \tag{4.2}$$

Note that $\sum_{i=1}^m((p-m-1)+i+\sigma(i)+n_i(m+p)) = mp+q(m+p)$. It therefore follows that every summand in the expression (4.2) is a multinomial coefficient

$$\binom{k}{k_1,\ldots,k_m} := \frac{(k_1+k_2+\cdots+k_m)!}{k_1!k_2!\cdots k_m!}. \tag{4.3}$$

For multinomial coefficients there is a well-known criterion frequently used by topologists which guarantees that such a coefficient is odd. We formulate this criterion as a lemma.

LEMMA 4.1. *The multinomial coefficient* $\binom{k}{k_1,\ldots,k_m}$ *is odd if and only if there are no "carry overs" in the summation* $k_1+\cdots+k_m$ *when calculated using binary representation.*

Proof. Let

$$k = 2^{n_1}+\cdots+2^{n_l},\ 0 \leq n_1 < \cdots < n_l.$$

Then

$$(x_1+\cdots+x_m)^k = \prod_{i=1}^{l}(x_1^{2^{n_i}}+\cdots+x_m^{2^{n_i}}) \bmod 2. \qquad \square$$

In particular it follows from this lemma that $\binom{k}{k_1,\ldots,k_m}$ is even as soon as two numbers among $\{k_1, \ldots, k_m\}$ are equal or two numbers are odd.

We will call a set $\{k_1, k_2, \ldots, k_m\}$ of positive integers *a disjoint binary partition of* k if the multinomial coefficient $\binom{k}{k_1,\ldots,k_m}$ is odd. To put it in other words, $\{k_1, k_2, \ldots, k_m\}$ is a disjoint binary partition of k if $k_1 + k_2 + \cdots + k_m = k$ and if their binary representations

$$k_i = 2^{n_{i1}} + 2^{n_{i2}} + \cdots + 2^{n_{ir_i}},\ 0 \le n_{i1} < n_{i2} < \cdots < n_{ir_i},\quad i = 1, 2, \ldots, m,$$

have disjoint exponents; i.e., $n_{ij} \neq n_{rs}$ for all i, j, r, s.

THEOREM 4.2. *Let* $a = \min(m, p)$. *Then* $\deg K^q_{m,p}$ *is odd if and only if the number of disjoint binary partitions* $\{k_1, \ldots, k_a\}$ *of* $q(m + p) + mp$ *having the property that*

$$\{k_1, \ldots, k_a\} = \{m + p - 1, m + p - 3, \ldots, m + p - 2a + 1\} \bmod m + p$$

is odd.

Before we give the proof we will illustrate Theorem 4.2 with several examples.

Example 4.3. a. $m = 2, p = 9, q = 4, q(m + p) + mp = 62 = 2 + 2^2 + 2^3 + 2^4 + 2^5$. The disjoint binary partitions equal to $\{10, 8\}$ mod 11 are

$$\{2^5, 2 + 2^2 + 2^3 + 2^4\} = \{32, 30\},$$
$$\{2 + 2^2 + 2^4 + 2^5, 2^3\} = \{54, 8\},$$
$$\{2 + 2^3, 2^2 + 2^4 + 2^5\} = \{10, 52\}.$$

So $\deg K^4_{2,9} = \deg K^4_{9,2}$ is odd.

b. $m = 3, p = 4, q = 5, q(m + p) + mp = 47 = 1 + 2 + 2^2 + 2^3 + 2^5$. The disjoint binary partitions equal to $\{6, 4, 2\}$ mod 7 are

$$\{2 + 2^2, 2^5, 1 + 2^3\} = \{6, 32, 9\},$$
$$\{2 + 2^5, 2^2, 1 + 2^3\} = \{34, 4, 9\},$$
$$\{1 + 2^2 + 2^3, 2^5, 2\} = \{13, 32, 2\},$$
$$\{1 + 2^3 + 2^5, 2^2, 2\} = \{41, 4, 2\}.$$

So $\deg K^5_{3,4} = \deg K^5_{4,3}$ is even.

c. $m = 3, p = 6, q = 3, q(m + p) + mp = 45 = 1 + 2^2 + 2^3 + 2^5$. There is only one disjoint binary partition equal to $\{8, 6, 4\}$ mod 9:

$$\{2^3, 1 + 2^5, 2^2\} = \{8, 33, 4\}.$$

So $\deg K^3_{3,6} = \deg K^3_{6,3}$ is odd.

d. $m = 5, p = 6, q = 3, q(m + p) + mp = 63 = 1 + 2 + 2^2 + 2^3 + 2^4 + 2^5$. There is only one disjoint binary partition equal to $\{10, 8, 6, 4, 2\}$ mod 11:

$$\{2^5, 2^3, 1 + 2^4, 2^2, 2\} = \{32, 8, 17, 4, 2\}.$$

So $\deg K^3_{5,6} = \deg K^3_{6,5}$ is odd.

Proof. Without loss of generality, assume $m \le p$. Consider again the description of the degree of the variety $K^q_{m,p}$ as it was provided in formula (4.2). It is our goal to show that in the summation mod 2 the only relevant permutation is $\sigma = id$. In other words, we will show by clever "book keeping" that all other multinomial coefficients are either 0 or cancel each other.

First assume σ is not an idempotent, i.e., $\sigma^2 \neq id$ or $\sigma \neq \sigma^{-1}$. In this case we immediately verify that the sets

$$\{((p-m-1)+i+\sigma(i)+n_i(m+p)) \mid i=1,\ldots,m\}$$

and

$$\{((p-m-1)+i+\sigma^{-1}(i)+n_{\sigma^{-1}(i)}(m+p)) \mid i=1,\ldots,m\}$$

are equal as unordered sets. But this just means that the corresponding multinomial coefficients in the summation (4.2) cancel each other mod 2. So it follows that we only have to sum over idempotent permutations.

Assume therefore that $\sigma^2 = id$. If $\sigma \neq id$ then σ contains a pure transposition, i.e., there are two distinct integers a, b having the property that $\sigma(a) = b$ and $\sigma(b) = a$. Consider the ordered set

$$(((p-m-1)+i+\sigma(i)+n_i(m+p)) \mid i=1,\ldots,m).$$

If $n_a = n_b$, then the corresponding multinomial coefficient is zero since two numbers in the binary partition (namely the numbers at positions a and b) are equal. On the other hand if $n_a \neq n_b$, then the corresponding multinomial coefficient cancels with the multinomial coefficient obtained by interchanging n_a and n_b.

It therefore follows that the only relevant summand in (4.2) is $\sigma = id$. The mod 2 degree of $K^q_{m,p}$ reduces, therefore, to the evaluation of the mod 2 sum of

(4.4)

$$\sum_{n_1+\cdots+n_m=q} \binom{(mp+q(m+p))}{p-m+1+n_1(m+p),\ p-m+3+n_2(m+p),\ \ldots,\ p+m-1+n_m(m+p)}.$$

Since a summand

$$\binom{(mp+q(m+p))}{p-m+1+n_1(m+p),\ p-m+3+n_2(m+p),\ \ldots,\ p+m-1+n_m(m+p)}$$

is odd if and only if $\{p-m+1+n_1(m+p), \ldots, p+m-1+n_m(m+p)\}$ is a disjoint binary partition of $mp+q(m+p)$, the deg $K^q_{m,p}$ is odd if and only if the number of disjoint binary partitions equal to $\{p-m+1, p-m+3, \ldots, p+m-1\}$ mod $m+p$ of $q(m+p)+mp$ is odd. □

For the Grassmann variety it is possible to identify the first Chern class c_1 (respectively, the first Stiefel-Whitney class w_1) of the classifying bundle with the first elementary symmetric function

$$x_1+\cdots+x_m \in \mathcal{Z}[x_1,\ldots,x_m].$$

The degree (respectively, the mod 2 degree) of the Grassmann variety is then represented through the coefficient of a certain monom (see [23] for details) in the expansion of

$$(x_1+\cdots+x_m)^{\dim \mathrm{Grass}(m,m+p)}.$$

For the mod 2 degree of the variety $K^q_{m,p}$, Theorem 4.2 gives a way to do a similar computation. For this consider the polynomial ring $\mathcal{Z}_2[x_1,\ldots,x_m]$, the ideal

$$I := \left\langle x_1^{m+p}-1, \ldots, x_m^{m+p}-1 \right\rangle,$$

and the factor ring $R := \mathcal{Z}_2[x_1,\ldots,x_m]/I$. Then we have the following corollary.

COROLLARY 4.4. *If $m \leq p$ the* mod 2 *degree of the variety $K^q_{m,p}$ is equal to the coefficient of the monom $x_1^{m+p-1}x_2^{m+p-3}\cdots x_m^{p-m+1}$ in the expansion of*

$$(x_1 + \cdots + x_m)^{\dim K^q_{m,p}} \in R.$$

Proof. From the proof of Theorem 4.2 it follows that the mod 2 degree of $K^q_{m,p}$ is equal to the sum of certain multinomial coefficients of the form

$$\binom{\dim K^q_{m,p}}{k_1, \ldots, k_m} = \binom{q(m+p)+mp}{k_1, \ldots, k_m}.$$

Since the mod $(m+p)$ identification of disjoint binary partitions in Theorem 4.2 corresponds to the ideal theoretic identification of monoms in the factor ring R, the total mod 2 number of identified monoms is exactly the mod 2 degree of $K^q_{m,p}$. □

In practice we can often use the "freshman's dream"

$$(x_1 + \cdots + x_m)^{2^k} = x_1^{2^k} + \cdots + x_m^{2^k} \mod 2.$$

The following examples illustrate the corollary.

Example 4.5. a. $m = 2$, $p = 3$, $q = 3$. In this case $\dim K^3_{2,3} = 21$ and we know from the table at the end of §3 that $\deg K^3_{2,3} = 6765$. Using the corollary we compute

$$(x+y+z)^{21} = (x^{16}+y^{16}+z^{16})(x+y+z)^5 = (x^2+y^2+z^2)^3.$$

Since the coefficient in front of the monom x^4y^2 is indeed 1, we conclude once more that $K^3_{2,3}$ is of odd degree.

b. $m = 3, p = 4, q = 69, q(m+p)+mp = 495 = 1+2+2^2+2^3+2^5+2^6+2^7+2^8$. Since the dimension is quite large we reduce mod 7 in the first step:

$$(1, 2, 2^2, 2^3, 2^5, 2^6, 2^7, 2^8) = (1, 2, 4, 1, 4, 1, 2, 4) \mod 7.$$

Using this reduction we have

$$\begin{aligned}(x+y+z)^{495} &= (x+y+z)^3(x^2+y^2+z^2)^2(x^4+y^4+z^4)^3\\ &= (x+y+z)(x^4+y^4+z^4).\end{aligned}$$

Since there is no monom $x^6y^4z^2$ in this expansion, we conclude that $K^{69}_{3,4}$ and $K^{69}_{4,3}$ both have an even degree.

COROLLARY 4.6. *Let* $\min(m, p) > 1$. *Then any of the following conditions implies that* $\deg K^q_{m,p}$ *is even:*

a) $m + p$ *is even.*
b) $mp + 3 > (m+p)(q+2)$.
c) $\min(m, p) \geq q + 2 + \sqrt{q^2+4q+1}$.
d) *The binary number of* $q(m+p)+mp$ *has less than* m 1*'s other than the digit on the* 2^0 *position.*
e) $2^{\min(m,p)+1} > q(m+p)+mp+2$.
f) $\sum_{i=1}^{l} r_i < mp$, *where* $r_i \in [0, m+p)$ *is the number, equals the* $2^{n_i} \bmod m+p$ *in the binary representation* $q(m+p)+mp = 2^{n_1}+\cdots+2^{n_l}$.
g) $m + p = 2^k - 1$.

Proof. Without loss of generality assume that $m \leq p$.

a) When $m+p$ is even, all the integers $p-m+1+n_1(m+p), \ldots, p+m-1+n_m(m+p)$ are odd. By the remark after Lemma 4.1, all multinomial coefficients appearing in (4.4) are even.

b) Let $2^r \le q(m+p)+mp < 2^{r+1}$. Then a necessary condition for $\{p-m+1+n_1(m+p), \dots, p+m-1+n_m(m+p)\}$ to be a disjoint binary partition is

$$p-m+2i-1+n_i(m+p) \ge 2^r$$

for some i. In particular

$$p+m-1+q(m+p) \ge 2^r \ge (1/2)(q(m+p)+mp+1),$$

which implies

$$(m+p)(q+2) \ge mp+3.$$

c) Consider $-(m^2-2(q+2)m+3)$. It has two roots: $q+2 \pm \sqrt{q^2+4q+1}$. So when $m \ge q+2+\sqrt{q^2+4q+1}$,

$$-(m^2-2(q+2)m+3) \le 0.$$

The degree is even if $m=p$ by a). If $m<p$ (note that $q+2-m<0$),

$$\begin{aligned}(m+p)(q+2)-mp-3 &= -(m^2-2(q+2)m+3)+(q+2-m)(p-m)\\ &< -(m^2-2(q+2)m+3)\\ &\le 0.\end{aligned}$$

So c) implies b).

d) Under the condition, $q(m+p)+mp$ cannot have a disjoint binary partition $\{k_1, \dots, k_m\}$ such that none of the k_i is 1.

e) The smallest number such that d) is not satisfied is $2^{m+1}-2$. So e) implies d).

f) A necessary condition for

$$(k_1, \dots, k_m) = (p-m+1, \dots, p+m-1) \mod m+p, \quad k_i > 0,$$

is $\sum_{i=1}^m k_i \ge mp$.

g) Notice that $m+p$ is odd. So $2 \le m < p$. For any $n \ge k$, let $n = ak+r$, $0 \le r < k$. Then

$$2^n = 2^r(1+2^k+\cdots+2^{k(a-1)})(2^k-1)+2^r.$$

So $2^n = 2^r \mod m+p$. Let the binary representation of $q(m+p)+mp$ be $2^{n_1}+2^{n_2}+\cdots+2^{n_l}$ and consider

$$\prod_{i=1}^{l}(x_1^{2^{n_i}}+\cdots+x_m^{2^{n_i}}).$$

By replacing 2^{n_i} with 2^{r_i} for $r_i = n_i \mod k$, $r_i \in [0,k)$, and using the property

$$(x_1^{2^r}+\cdots+x_m^{2^r})^2 = (x_1^{2^{r+1}}+\cdots+x_m^{2^{r+1}}) \mod 2$$

repeatedly, we get

$$\prod_{i=1}^{j}(x_1^{2^{r_i}}+\cdots+x_m^{2^{r_i}}), \tag{4.5}$$

with $\{r_1, \dots, r_j\} \subset [0,k)$ distinct. The polynomial (4.5) has degree at most $1+2+\cdots+2^{k-1} = m+p$, which is always less than mp under the condition $2 \le m < p$. By the same argument as in the proof of (f), the degree is even. □

An immediate corollary of Theorem 4.2 is the result of [1]: deg Grass$(m, m+p)$ is odd if and only if

1. $\min(m, p) = 1$ or
2. $\min(m, p) = 2$, $\max(m, p) = 2^k - 1$.

We say this corollary is immediate because when $m = 2 \leq p$, $\{p+1, p-1\}$ is a disjoint binary partition if and only if $p = 2^k - 1$, and when $\min(m, p) \geq 3$, all the degrees are even by Corollary 4.6 (c).

COROLLARY 4.7. deg $K^1_{m,p}$ *is odd if and only if either*

1. $\min(m, p) = 1$ *or*
2. $\min(m, p) = 2$, $\max(m, p) = 2^{n_1} + 2^{n_2} + \cdots + 2^{n_l} - 1$ *with* $n_{i+1} > n_i + 1$, $i = 1, \ldots, l-1$.

Proof. By letting $m + p = 1 + 2^{n_1} + \cdots + 2^{n_l}$ we can easily show that neither of the sets

$$\{m+p-1, m+p-3, m+p-5\},\ \{2(m+p)-1, m+p-3, m+p-5\},$$

$$\{m+p-1, 2(m+p)-3, m+p-5\},\ \{m+p-1, m+p-3, 2(m+p)-5\}$$

can have disjoint exponents in the binary representations of the elements. So the degree is even if $\min(m, p) \geq 3$. Now let $m = 2$, $p > m$, be odd and

$$p + 1 = 2^{n_1} + \cdots + 2^{n_l}.$$

Then 2^{n_1} appears in both $p + 1 = m + p - 1$ and $2p + 1 = 2(m + p) - 3$. So deg $K^1_{2,p}$ is odd if and only if

$$\{p-1, 2p+3\} = \{2 + 2^2 + \cdots + 2^{n_1 - 1} + 2^{n_2} + \cdots + 2^{n_l},\ 1 + 2^{n_1+1} + \cdots + 2^{n_l+1}\}$$

is a disjoint binary partition, i.e., if and only if $n_{i+1} > n_i + 1$ for $i = 1, \ldots, l-1$. □

Similar results can also be proven for $q > 1$. The combinatorics however becomes very involved. We provide without proof the result for $q = 2$.

COROLLARY 4.8. deg $K^2_{m,p}$ *is odd if and only if either*

1. $\min(m, p) = 1$,
2. $\min(m, p) = 2$, $\max(m, p) = 8(\frac{4^k - 1}{3}) + 1$, *or*
3. $\min(m, p) = 3$, $\max(m, p) = 8$.

5. Corollaries and additional new positive pole placement results. In this section we establish the connection to the classical state space and transfer function formulation of the pole placement problem. We also derive several results which combine the results derived in §3 with some results derived in [17].

Consider a controllable observable linear system

$$\dot{x} = Ax + Bu, \quad y = Cx, \tag{5.1}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^p$, respectively. If a controllable observable dynamic compensator of order q,

$$\dot{u} = Fu + Ey, \quad u = Hu + Ky, \tag{5.2}$$

is applied to the system, the closed-loop system becomes

$$\begin{pmatrix} \dot{x} \\ \dot{u} \end{pmatrix} = \begin{pmatrix} A + BKC & BH \\ EC & F \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}, \quad y = Cx. \tag{5.3}$$

So the closed-loop characteristic polynomial is

$$\phi(s) = \det \begin{pmatrix} sI - A - BKC & -BH \\ -EC & sI - F \end{pmatrix}. \tag{5.4}$$

If $G(s) = C(sI - A)^{-1}B$ and $T(s) = K + H(sI - F)^{-1}E$ are the transfer functions of the system (5.1) and compensator (5.2), respectively, and if $G(s) = D(s)^{-1}N(s)$ and $T(s) = T_d^{-1}(s)T_n(s)$ are left coprime fractions such that $\det(sI - A) = \det D(s)$ and $\det(sI - F) = \det T_d(s)$, then $\phi(s)$ can also be written as

$$\phi(s) = \det(sI - A)\det(I - G(s)T(s))\det(sI - F) = \det \begin{pmatrix} D(s) & N(s) \\ T_n(s) & T_d(s) \end{pmatrix}. \tag{5.5}$$

Let $P(s) = (D(s)\ N(s))$ and $C(s) = (T_n(s)\ T_d(s))$. Then $P(s)$ and $C(s)$ can be viewed as autoregressive systems describing the behavior of the plant and the compensator, respectively. The combined dynamics are then described by

$$\begin{pmatrix} D(\frac{d}{dt}) & N(\frac{d}{dt}) \\ T_n(\frac{d}{dt}) & T_d(\frac{d}{dt}) \end{pmatrix} \cdot \begin{pmatrix} y \\ -u \end{pmatrix}(t) = 0. \tag{5.6}$$

The following result combines Theorem 2.15 with [17, Cor. 5.8].

THEOREM 5.1. *Consider a generic set of matrices* $(A, B, C) \in \mathbb{R}^{n(n+m+p)}$ *describing a plant as in* (5.1) *and consider an arbitrary monic polynomial* $\phi(s) \in \mathbb{R}[s]$ *of degree* $n + q$. *If*

$$n \leq q(m + p - 1) + mp, \tag{5.7}$$

then there exists a complex dynamic compensator of the form (5.2) *resulting in the closed-loop characteristic polynomial* $\phi(s)$. *If in addition the number* $d(m, p, q)$ *introduced in* (1.5) *is odd, then there even exists a real compensator assigning the closed-loop characteristic polynomial* $\phi(s)$.

Proof. We only outline the main steps. Using the same argument as in Theorem 2.15, we verify that $\dim B_p = q(m + p) + mp - n - q - 1$ for the generic and strictly proper plant. Theorem 2.14 therefore still applies and the pole placement map is onto if we allow all autoregressive systems. Since the plant is strictly proper a closed-loop characteristic polynomial of degree $n + q$ can only be achieved if the compensator is proper. □

The following example illustrates how this theorem can be applied.

Example 5.2. Assume the matrices (A, B, C) describe the plant parameters of a generic real 2-input, 9-output plant of McMillan degree n. From Table 1 it follows immediately that there exists a real compensator of degree 1 as long as $n \leq 28$. If, e.g., $n \leq 58$, then it follows that there is a real compensator of degree 4 assigning an arbitrary set of self-conjugate closed-loop poles.

Combining Theorem 2.15 with [17, Cor. 5.9] one can finally prove the following result.

THEOREM 5.3. *Let* $G(s)$ *be a generic* m*-input,* p*-output proper transfer function of McMillan degree* n *and let* $\phi(s) \in \mathbb{K}[s]$ *be a generic polynomial of degree* $n + q$. *If*

$$n \leq q(m + p - 1) + mp, \tag{5.8}$$

then there exists a proper complex compensator $T(s)$ *of McMillan degree* q *such that the closed-loop transfer function*

$$G_T(s) := (I - G(s)T(s))^{-1}G(s)$$

has characteristic polynomial $\phi(s)$. *If in addition the number* $d(m, p, q)$ *introduced in* (1.5) *is odd, then there even exists a real transfer function* $T(s)$.

REFERENCES

[1] I. BERSTEIN, *On the Lusternik-Šnirel'mann category of real Grassmannians*, Proc. Cambridge Philos. Soc., 79 (1976), pp. 129-239.

[2] A. BERTRAM, G. DASKALOPOULOS, AND R. WENTWORTH, *Gromov Invariants for Holomorphic Maps from Riemann Surfaces to Grassmanians*, preprint, 1993.

[3] H. BLOMBERG AND R. YLINEN, *Algebraic Theory for Multivariable Linear Systems*, Academic Press, London, 1983.

[4] R. W. BROCKETT AND C. I. BYRNES, *Multivariable Nyquist criteria, root loci and pole placement: A geometric viewpoint*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 271-284.

[5] C. I. BYRNES, *On compactifications of spaces of systems and dynamic compensation*, in IEEE Conf. on Decision and Control, vol. 4, San Antonio, TX, 1983, pp. 889-894.

[6] H. GLÜSING-LÜERSSEN, *On Various Topologies for Finite-Dimensional Linear Systems*, Tech. Report 273, University of Bremen, Germany, 1992.

[7] J. HARRIS, *Algebraic Geometry, A First Course*, Graduate Text in Mathematics, Springer-Verlag, New York, Berlin, 1992.

[8] K. INTRILIGATOR, *Fusion residues*, Mod. Phys. Lett. A, 6 (1991), pp. 3543-3556.

[9] S. L. KLEIMAN AND D. LAKSOV, *Schubert calculus*, Amer. Math. Monthly, 79 (1972), pp. 1061-1082.

[10] M. KUIJPER, *First-Order Representation of Linear Systems*, Ph.D. thesis, Centrum voor Wiskunde en Informatica, Amsterdam, the Netherlands, 1992.

[11] M. P. MACMAHON, *Combinatory Analysis*, vol. I, Cambridge University Press, Cambridge, 1915.

[12] C. F. MARTIN AND R. HERMANN, *Applications of algebraic geometry to system theory: The McMillan degree and Kronecker indices as topological and holomorphic invariants*, SIAM J. Control Optim., 16 (1978), pp. 743-755.

[13] D. MUMFORD, *Algebraic Geometry* I: *Complex Projective Varieties*, Springer-Verlag, Berlin, New York, 1976.

[14] M. S. RAVI AND J. ROSENTHAL, *A smooth compactification of the space of transfer functions with fixed McMillan degree*, Acta Appl. Math, 34 (1994), pp. 329-352.

[15] M. S. RAVI, J. ROSENTHAL, AND X. WANG, *Degree of the Generalized Plücker Embedding of a Quot Scheme and Quantum Cohomology*, preprint alg-geom/9402011, 1994.

[16] J. ROSENTHAL, *On minimal order dynamical compensators of low order systems*, in Proc. of European Control Conference, Hermes, France, 1991, pp. 374-378.

[17] ———, *On dynamic feedback compensation and compactification of systems*, SIAM J. Control Optim., 32 (1994), pp. 279-296.

[18] H. SCHUBERT, *Kalkühl der abzählenden Geometrie*, Teubner, Leipzig, 1879.

[19] J. M. SCHUMACHER, *Transformations of linear systems under external equivalence*, Linear Algebra Appl., 102 (1988), pp. 1-33.

[20] ———, *A pointwise criterion for controller robustness*, Systems Control Lett., 18 (1992), pp. 1-8.

[21] I. R. SHAFAREVICH, *Basic Algebraic Geometry*, Springer-Verlag, Berlin, New York, 1974.

[22] B. SIEBERT AND G. TIAN, *On Quantum Cohomology Rings of Fano Manifolds and a Formula of Vafa and Intriligator*, preprint alg-geom/9403010, 1994.

[23] R. STANLEY, *Some combinatorial aspects of the Schubert calculus*, in Lecture Notes in Mathematics vol. 579, Springer-Verlag, Berlin, New York, 1977, pp. 217-251.

[24] C. VAFA, *Topological mirrors and quantum rings*, in Essays on Mirror Manifolds, S. T. Yau, ed., International Press, Hong Kong, 1992.

[25] X. WANG, *Pole placement by static output feedback*, J. Math. Systems Estim. Control, 2 (1992), pp. 205-218.

[26] X. WANG AND J. ROSENTHAL, *A cell structure for the set of autoregressive systems*, Linear Algebra Appl., 205/206 (1994), pp. 1203-1226.

[27] J. C. WILLEMS, *Input–output and state–space representations of finite-dimensional linear time-invariant systems*, Linear Algebra Appl., 50 (1983), pp. 581-608.

[28] ———, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 259-294.

[29] ———, *Feedback in a behavioral setting*, in Systems, Models and Feedback: Theory and Applications, Birkhäuser-Verlag, Basel, Switzerland, 1992, pp. 179-191.

[30] J. C. WILLEMS AND W. H. HESSELINK, *Generic properties of the pole placement problem*, in Proc. of the IFAC, 1978, pp. 1725-1729.

[31] E. WITTEN, *The Verlinde Algebra and the Cohomology of the Grassmannian*, preprint IASSNS-HEP-93/41, hep-th/9312104, 1993.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 833-854, May 1996

003

ON THE GENERICITY OF STABILIZABILITY FOR TIME-DELAY SYSTEMS*

LUC C. G. J. M. HABETS†

Abstract. Conditions for the stabilizability of time-delay systems with incommensurable point delays by dynamic state feedback are known in the literature. In this paper it is shown that these conditions are satisfied generically.

Although an algebraic approach is used to describe the class of all time-delay systems with point delays, the concept of genericity is formulated in a topological framework. In the metric space consisting of all parametrizations of time-delay systems, the subset of all stabilizable systems is an open and dense subset.

The proof is given for the commensurable delay case first. It is shown that the incommensurable delay case is not significantly more difficult and that the same arguments prove also that systems with incommensurable time-delays are generically stabilizable.

Key words. time-delay systems with point delays, stabilizability, genericity

AMS subject classifications. 93B25, 93D15, 15A54

1. Introduction.

Time-delay systems with point delays can be seen as rather straightforward generalizations of ordinary linear time-invariant systems. In the delay case, $\dot{x}(t)$, the derivative of the evolution variable x at time t, and $y(t)$, the output y at time t, do not depend only on the evolution variable x and the input u at time t but also on the evolution variable and input from specific time instants in the past. Let $\sigma_1, \ldots, \sigma_k$ denote k delay operators with incommensurable time-delays $\tau_1, \ldots, \tau_k$, acting on the trajectories of the evolution variable and the input:

$$\sigma_i x(t) = x(t - \tau_i), \qquad \sigma_i u(t) = u(t - \tau_i), \qquad (i = 1, \ldots, k). \tag{1}$$

Then a system with k *incommensurable* time-delays $\tau_1, \ldots, \tau_k$ can be written as

$$\begin{cases} \dot{x}(t) = A(\sigma_1, \ldots, \sigma_k)x(t) + B(\sigma_1, \ldots, \sigma_k)u(t), \\ y(t) = C(\sigma_1, \ldots, \sigma_k)x(t) + D(\sigma_1, \ldots, \sigma_k)u(t), \end{cases} \tag{2}$$

where $A(\sigma_1, \ldots, \sigma_k)$, $B(\sigma_1, \ldots, \sigma_k)$, $C(\sigma_1, \ldots, \sigma_k)$, and $D(\sigma_1, \ldots, \sigma_k)$ are polynomial matrices in the delay operators $\sigma_1, \ldots, \sigma_k$ of appropriate dimensions. Note that the state of this system at time t is not the evolution variable $x(t)$ but the time-trajectory $\{x(\xi) \mid \xi \in [t-T, t]\}$ of this evolution variable. Here T denotes the length of the largest time-delay occurring in (2).

After substitution of indeterminates $s_1, \ldots, s_k$ for the delay operators $\sigma_1, \ldots, \sigma_k$ in (2), the system $\Sigma = (A(s_1, \ldots, s_k), B(s_1, \ldots, s_k), C(s_1, \ldots, s_k), D(s_1, \ldots, s_k))$ over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$ is obtained. Together with the delays $\tau_1, \ldots, \tau_k$, this quadruple of matrices is a complete description of the delay system (2); since the time-delays $\tau_1, \ldots, \tau_k$ are incommensurable, there is a 1-1 correspondence between time-delay systems of the form (2) and systems $\Sigma = (A(s_1, \ldots, s_k), B(s_1, \ldots, s_k), C(s_1, \ldots, s_k), D(s_1, \ldots, s_k))$ over the ring $\mathbb{R}[s_1, \ldots, s_k]$.

To study the concept of internal stability for time-delay systems, consider the differential-difference equation for the evolution variable x, and assume that no input is applied. Then the system is called *internally stable* if, independent of the given initial conditions, the evolution

*Received by the editors April 23, 1993; accepted for publication (in revised form) December 23, 1994. This research was supported by the Netherlands Organization for Scientific Research (NWO) and carried out while the author was with the Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven, the Netherlands.

†Institut für Dynamische Systeme, Department of Mathematics, University of Bremen, P.O. Box 330 440, D-28334 Bremen, Germany (luc@mathematik.uni-Bremen.de).

variable $x(t)$ tends to zero for $t \to \infty$. According to [7, Cor. 4.1, p. 182], this notion of stability is equivalent to the following condition on the matrix $A(s_1, \ldots, s_k)$:

$$\forall \lambda \in \overline{\mathbb{C}^+} : \det(\lambda I - A(e^{-\tau_1 \lambda}, \ldots, e^{-\tau_k \lambda})) \neq 0.$$

Here $\tau_1, \ldots, \tau_k$ are the time-delays of the delay operators $\sigma_1, \ldots, \sigma_k$ corresponding to the indeterminates $s_1, \ldots, s_k$, and $\mathbb{C}^+$ denotes the open complex right half plane.

If a system is not internally stable, this property may be achieved by a proper choice of a static or dynamic feedback compensator. Completely analogous to the case of systems without delays, this so-called stabilizability problem can be split into two dual parts: the problem of stabilization by (static or dynamic) *state* feedback and the detectability problem. In the rest of this paper we confine ourselves to the problem of stabilizability by state feedback and therefore assume that $C = I$ and $D = 0$.

In the literature, the problem of stabilizability of time-delay systems has been solved in at least two different ways. Surprisingly, both the infinite-dimensional systems approach and the systems over rings approach yield the same conditions for the solvability of this problem. However, there are also important differences between these results. In the infinite-dimensional systems approach, a static state feedback (possibly containing distributed time-delays) suffices to achieve internal stability, whereas in the algebraic approach a dynamic feedback compensator (containing only point delays) is required for this.

THEOREM 1.1 (see [12], [3], [14], [4]). *Consider a time-delay system* Σ*:*

$$\dot{x}(t) = A(\sigma_1, \ldots, \sigma_k)x(t) + B(\sigma_1, \ldots, \sigma_k)u(t), \tag{3}$$

where σ_i $(i = 1, \ldots, k)$ *denotes the delay operator with time-delay* τ_i *and where* $A(\sigma_1, \ldots, \sigma_k)$ *and* $B(\sigma_1, \ldots, \sigma_k)$ *are matrices of polynomials in the delay operators* $\sigma_1, \ldots, \sigma_k$*, of size* $n \times n$ *and* $n \times m$*, respectively. Substitute indeterminates* $s_1, \ldots, s_k$ *for* $\sigma_1, \ldots, \sigma_k$ *and regard* $\Sigma = (A(s_1, \ldots, s_k), B(s_1, \ldots, s_k))$ *as a linear system over the polynomial ring* $\mathcal{R} = \mathbb{R}[s_1, \ldots, s_k]$*. Then the following three conditions are equivalent:*

(i) Σ *is internally stabilizable by a dynamic state feedback compensator only containing point delays,*

(ii) Σ *is internally stabilizable by a static state feedback, possibly containing distributed time-delays,*

(iii) $\forall z \in \overline{\mathbb{C}^+} : \operatorname{rank}(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) \mid B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})) = n.$ (4)

Rank condition (iii), which can be seen as a generalization of the well-known Hautus test (see [8]) to the case of time-delay systems with point delays, is the starting point of this paper. We shall prove that this condition is generically satisfied on the parameter-space describing all time-delay systems with point delays of the form (3). This means that condition (iii) is very weak; it is satisfied for most time-delay systems.

The condition of reachability for systems over polynomial rings (see, e.g., [9], [16]) can be stated as a rank condition in almost the same way as the stabilizability condition (for a short proof see, for example, [5]). In [11], Lee and Olbrot prove that this condition is generically satisfied if and only if the number of inputs to the system is larger than the number of indeterminates of the polynomial ring (i.e., the number of incommensurable time-delays). Their approach is completely algebraic; they compare the number of polynomial equations that have to be satisfied with the number of unknowns and apply some results from algebraic geometry to prove their result (except on some hypersurfaces in the parameter-space of all time-delay systems, the reachability condition is always satisfied).

At first sight, this approach also looks very promising for solving the genericity problem of stabilizability for time-delay systems. In this case, however, each indeterminate s_i in the polynomial ring corresponds to a delay operator σ_i of length τ_i, and in the Laplace domain

σ_i and τ_i are interrelated via an exponential function. In this way some extra (exponential) equations are obtained that are probably enough to remove the condition on the number of inputs. Unfortunately, this method fails because we are now dealing with both polynomial and exponential equations, which do not fit into the algebraic-geometric framework any more.

In this paper we choose a completely different approach; we describe the concept of genericity in a topological way. In this setting, a certain property is called *generic* if it holds on an open and dense subset of the parameter-space describing all time-delay systems. However, before we can speak of open or dense subsets, we first have to introduce a topology on this space. This topology formalizes our intuitive ideas on the following question: when are the parametrizations of two time-delay systems said to be close to each other? In §2 this topological framework is treated in more detail.

Then all tools are available to prove that the set of stabilizable time-delay systems is open, which is described in our §3. The proof of denseness is more involved. In §4 we start with some preliminary results on matrices over the ring of analytic functions. These are used in §5 to show that the set of stabilizable time-delay systems is indeed a dense subset of the parameter-space describing all time-delay systems.

Remark 1.2. In the rest of this paper it is always tacitly assumed that we are dealing with time-delay systems with commensurable delays. This implies that there is only one delay operator σ required to describe the system equations (2). In general, this situation is much simpler than the incommensurable delay case. Fortunately this distinction does not make any difference for the approach we take to the problem. All results are easily generalized to the incommensurable delay case because the assumption of the presence of only one time-delay operator is never used explicitly. This assumption is only made to simplify notation to highlight the really important ideas more clearly. In §6 we return to this subject briefly and explain why the methods developed in this paper are also applicable in the incommensurable delay case.

2. A topological framework for time-delay systems. This section is devoted to the introduction of a topology on the parameter-space describing all time-delay systems with commensurable time-delays. This topology reflects our intuitive notion of the concept of genericity. Also the space of all 2-dimensional polynomials is equipped with a suitable norm. These polynomials, and especially characteristic polynomials, play a vital role in the characterization of stability. Some of the topological aspects of this relationship are discussed in more detail.

Consider a triple $\Sigma = (A(s), B(s), \tau)$, with $A(s) \in \mathbb{R}[s]^{n\times n}$, $B(s) \in \mathbb{R}[s]^{n\times m}$, and $\tau \in \mathbb{R}^+$. After substitution of the delay operator σ with time-delay τ for the indeterminate s, such a triple is a complete description of the time-delay system:

$$\begin{cases} \dot{x}(t) = A(\sigma)x(t) + B(\sigma)u(t), \\ \sigma x(t) = x(t-\tau), \qquad \sigma u(t) = u(t-\tau). \end{cases} \tag{5}$$

On the other hand, the triple $\Sigma = (A(s), B(s), \tau)$ can be seen as a point in the parameter-space

$$\mathcal{V} = \{(A(s), B(s), \tau) \mid A(s) \in \mathbb{R}[s]^{n\times n}, B(s) \in \mathbb{R}[s]^{n\times m}, \tau \in \mathbb{R}^+\}. \tag{6}$$

Clearly, to each element of $\mathcal{V}$ there corresponds a time-delay system as defined in (5). By imposing a metric on each of the three components of $\mathcal{V}$, the parameter-space $\mathcal{V}$ is turned into a metric space, and thereby its topology is fixed. We start with the introduction of a norm on polynomial matrices in $\mathbb{R}[s]^{p\times q}$.

Let $P(s)$ be a $p \times q$ polynomial matrix over $\mathbb{R}[s]$. Then there exists an $\ell \in \mathbb{N} \cup \{0\}$ and real matrices $P_0, P_1, \ldots, P_\ell$, with $P_\ell \neq 0$, such that

$$P(s) = \sum_{i=0}^{\ell} P_i s^i.$$

This ℓ is called the *degree* of the polynomial matrix $P(s)$ and is denoted by $\ell = \deg(P(s))$. Defining $P_i := 0$ for $i > \ell$, we can map the polynomial matrix $P(s)$ to the sequence $(P_i)_{i=0}^{\infty}$ of real matrices. In this way we obtain an explicit description of $P(s)$ in terms of its parameters. In fact, there is a 1-1 correspondence between polynomial matrices and the space $\ell_0(\mathbb{R}^{p\times q})$ consisting of all real matrix sequences with only a finite number of nonzero elements (i.e., matrices with at least one nonzero entry), via the bijection

$$\psi : \ell_0(\mathbb{R}^{p\times q}) \to \mathbb{R}[s]^{p\times q} : \psi((P_i)_{i=0}^{\infty}) = \sum_{i=0}^{\infty} P_i s^i.$$

The space $\ell_0(\mathbb{R}^{p\times q})$ is easily turned into a normed space by defining the norm of $(P_i)_{i=0}^{\infty}$ by

$$\|(P_i)_{i=0}^{\infty}\| = \sum_{i=0}^{\infty} \|P_i\|,$$

where $\|P_i\|$ is the operator induced norm of the real matrix P_i. It is evident that the same norm can also be used for polynomial matrices.

DEFINITION 2.1. *Let $P(s)$ be a $p \times q$ matrix over $\mathbb{R}[s]$. Let $(P_i)_{i=0}^{\infty} \in \ell_0(\mathbb{R}^{p\times q})$ be such that*

$$P(s) = \sum_{i=0}^{\infty} P_i s^i. \tag{7}$$

Then the norm *of $P(s)$ is defined as*

$$\|P(s)\|_{pm} := \sum_{i=0}^{\infty} \|P_i\|, \tag{8}$$

where $\|P_i\|$ is the operator induced matrix norm of P_i for all $i \in \mathbb{N} \cup \{0\}$.

The norm $\|\cdot\|_{pm}$ for polynomial matrices has a very important property. In the Introduction we have seen that for the investigation of the stability properties of a time-delay system, the exponential function $e^{-\tau z}$ has to be substituted for the indeterminate s in a polynomial matrix $A(s)$. Since for all $z \in \overline{\mathbb{C}^+}$, the norm $|e^{-\tau z}|$ is bounded above by 1, the norm $\|P(s)\|_{pm}$ of the polynomial matrix $P(s)$ is a uniform upper bound for the norm of $P(e^{-\tau z})$ in the closed right half plane.

LEMMA 2.2. *Let $P(s) \in \mathbb{R}[s]^{p\times q}$. Then for all $\tau > 0$ and for all $z \in \overline{\mathbb{C}^+}$, we have*

$$\|P(e^{-\tau z})\| \leq \|P(s)\|_{pm}.$$

With condition (iii) of Theorem 1.1 in mind, we see that Lemma 2.2 has a very interesting consequence for square polynomial matrices.

COROLLARY 2.3. *Let $A(s) \in \mathbb{R}[s]^{n\times n}$. Then for all $\tau > 0$ and $z \in \overline{\mathbb{C}^+}$, and for all $w \in \mathbb{C}$ satisfying $|w| > \|A(s)\|_{pm}$, we have*

$$\operatorname{rank}(wI - A(e^{-\tau z})) = n.$$

Using Definition 2.1, the parameter-space $\mathcal{V}$ may be equipped with a suitable metric.

DEFINITION 2.4. *Let* $\Sigma_1 = (A_1(s), B_1(s), \tau_1)$ *and* $\Sigma_2 = (A_2(s), B_2(s), \tau_2)$ *be two elements of the parameter-space* $\mathcal{V}$. *Then the* distance *between* Σ_1 *and* Σ_2 *is defined as*

$$d_{\mathcal{V}}(\Sigma_1, \Sigma_2) := \|A_1(s) - A_2(s)\|_{pm} + \|B_1(s) - B_2(s)\|_{pm} + |\tau_1 - \tau_2|. \tag{9}$$

With this distance function $d_{\mathcal{V}}(\cdot, \cdot)$, *the parameter-space* $\mathcal{V}$ *becomes a* metric space.

Once the topology on the parameter-space $\mathcal{V}$ has been fixed, the concept of genericity is easily defined. For each triple $\Sigma = (A(s), B(s), \tau)$ in $\mathcal{V}$, it is possible to check the stabilizability of the corresponding time-delay system using Theorem 1.1. Let

$$S := \{(A(s), B(s), \tau) \in \mathcal{V} \mid \forall z \in \overline{\mathbb{C}^+} : \operatorname{rank}(zI - A(e^{-\tau z}) | B(e^{-\tau z})) = n\}$$

be the set of all stabilizable delay systems. Then the property of stabilizability is called *generic* if the set S is an open and dense subset of the parameter-space $\mathcal{V}$. In the topology on $\mathcal{V}$ generated by the metric $d_{\mathcal{V}}(\cdot, \cdot)$, this implies that the set S covers almost the whole space $\mathcal{V}$:

(i) *S is open.* A stabilizable time-delay system remains stabilizable after a small perturbation of the parameters describing the system (i.e., the property of stabilizability is a robust property).

(ii) *S is a dense subset of $\mathcal{V}$.* Every element $\Sigma \in \mathcal{V}$ can be approximated arbitrarily close by a sequence of stabilizable systems (i.e., a sequence in S).

We see that the topology generated by the metric of Definition 2.4 leads to a formal definition of genericity that looks very natural and that is completely in accordance with our intuitive notion of this concept.

In almost the same way as for polynomial matrices, it is possible to regard the polynomial ring $\mathbb{R}[s, z]$ as a linear space and to define a norm on this space.

DEFINITION 2.5. *Let* $p(s, z) \in \mathbb{R}[s, z]$, *and write* $p(s, z)$ *as*

$$p(s, z) = \sum_{i=0}^{\ell} \sum_{j=0}^{k} p_{ij} s^j z^i. \tag{10}$$

Then the norm of $p(s, z)$ *is defined as*

$$\|p(s, z)\|_p := \sum_{i=0}^{\ell} \sum_{j=0}^{k} |p_{ij}|. \tag{11}$$

With this norm, $\mathbb{R}[s, z]$ *becomes a normed ring.*

Analogously to the polynomial matrix case, there exists a 1-1 correspondence between polynomials $p(s, z)$ in two variables and exponential polynomials of the form $p(e^{-\tau z}, z)$. Characteristic polynomials of this form determine the stabilizability of a time-delay system. From this point of view, the norm of Definition 2.5 has several interesting properties. For example, the norm $\|p(s, z)\|_p$ is a good measure for the magnitude of $|p(e^{-\tau z}, z)|$ in a bounded part of the closed right half plane.

LEMMA 2.6. *Let* $p(s, z) \in \mathbb{R}[s, z]$, *and assume that the degree of* p *in* z *is* n, *i.e.,*

$$p(s, z) = \sum_{i=0}^{n} \sum_{j=0}^{k} p_{ij} s^j z^i,$$

and there exists a $j \in \{0, \ldots, k\}$ *such that* $p_{nj} \neq 0$. *Let* $M > 1$ *and* $\varepsilon > 0$. *If*

$$\|p(s, z)\|_p < \varepsilon \cdot \frac{M - 1}{M^{n+1} - 1}, \tag{12}$$

then

$$\forall \tau > 0 \, \forall z \in \overline{\mathbb{C}^+} \text{ s.t. } |z| \leq M : |p(e^{-\tau z}, z)| < \varepsilon. \tag{13}$$

Finally there is a clear relationship between polynomial matrices in one indeterminate on the one hand and 2-dimensional polynomials on the other. In this relationship the characteristic polynomial plays the leading role. In the rest of this paper we need only the following result.

PROPOSITION 2.7. *Let* $A(s) \in \mathbb{R}[s]^{n \times n}$. *Then*

$$\forall \varepsilon > 0 \, \exists \delta > 0 \, \forall B(s) \in \mathbb{R}[s]^{n \times n} :$$
$$\|A(s) - B(s)\|_{pm} < \delta \Longrightarrow \|\det(zI - A(s)) - \det(zI - B(s))\|_p < \varepsilon.$$

According to Proposition 2.7, the map χ,

$$\chi : \mathbb{R}[s]^{n \times n} \longrightarrow \mathbb{R}[s, z] : \chi(A(s)) = \det(zI - A(s)), \tag{14}$$

is continuous with respect to the norms on $\mathbb{R}[s]^{n \times n}$ and $\mathbb{R}[s, z]$ as defined in (8) and (11), respectively. The validity of this result follows from the fact that the determinant of a matrix is a sum of products of its entries. Since both addition and multiplication are continuous operations, a proof of Proposition 2.7 follows straightforwardly.

Remark 2.8. With the norms and metrics defined in this section, none of the spaces $\mathcal{V}$, $\mathbb{R}[s]^{p \times q}$, or $\mathbb{R}[s, z]$ becomes a complete metric space. All these spaces basically consist of sequences (of scalars or matrices) with a finite number of nonzero elements. However, we imposed a sort of ℓ_1-norm on these spaces that does not distinguish between sequences with a finite and an infinite number of nonzero elements. Therefore it is easy to construct a Cauchy sequence that does not converge.

Fortunately, this somewhat unsatisfactory situation is not troublesome because completeness is never used in the proofs of our genericity result. Moreover, this problem may be solved by introducing so-called *inductive limit topologies* (see, e.g., [1, Chap. IV, §5]). For the study of genericity of more general concepts of stabilizability, this topology is indispensable (see [6, §3.3]), but in our case, inductive limit topologies would make things unnecessarily complicated.

3. On the robustness of the property of stabilizability. In this section the first part of our genericity result is proved. Based on the topological framework introduced in the previous section, it is shown that the subset S of $\mathcal{V}$, consisting of all parametrizations of stabilizable time-delay systems, is an open subset of $\mathcal{V}$. In practice this means that stabilizability of a time-delay system is a robust property: it is preserved after small perturbations of the parameters. In this section an upper bound is derived for the distance between a nominal stabilizable system and all the perturbed systems that are allowed. If the distance between a perturbed system and the nominal system is smaller than this upper bound, the perturbed system is still stabilizable. Since this upper bound is always larger than zero, this immediately implies that S is open.

From Theorem 1.1 we know that the stabilizability condition for time-delay systems is a full rank condition on a matrix in the variable z, which has to be satisfied for all $z \in \overline{\mathbb{C}^+}$. Now the proof of the main theorem of this section is based on the fact that in $\mathbb{C}^{n \times (n+m)}$ the set of all matrices of full row rank is open, i.e., a full row rank matrix in $\mathbb{C}^{n \times (n+m)}$ remains of full row rank after small perturbations of its entries.

THEOREM 3.1. *Let* $\Sigma_0 = (A_0(s), B_0(s), \tau_0)$ *be a point in* $\mathcal{V}$, *and assume that the time-delay system* (5) *corresponding to* Σ_0 *is stabilizable, i.e.,*

$$\forall z \in \overline{\mathbb{C}^+} : \operatorname{rank}(zI - A_0(e^{-\tau_0 z}) | B_0(e^{-\tau_0 z})) = n.$$

Then there exists a $\rho > 0$ such that all systems Σ in the ball around Σ_0 with radius ρ,

$$\mathcal{B}(\Sigma_0, \rho) := \{\Sigma \in \mathcal{V} \mid d_{\mathcal{V}}(\Sigma, \Sigma_0) < \rho\},$$

are stabilizable.

Proof. First of all there exists an $\ell \in \mathbb{N}$ such that $A_0(s)$ and $B_0(s)$ can be written as

$$A_0(s) = \sum_{i=0}^{\ell} A_i s^i, \qquad B_0(s) = \sum_{i=0}^{\ell} B_i s^i.$$

Next define G as

$$G := \{z \in \mathbb{C} \mid \operatorname{Re} z \geq 0 \text{ and } |z| \leq \|A_0(s)\|_{pm} + 1\}. \tag{15}$$

Since the delay system corresponding to $\Sigma_0 = (A_0(s), B_0(s), \tau_0)$ is stabilizable, it follows from [3] or [14] that the matrix $(zI - A_0(e^{-\tau_0 z})|B_0(e^{-\tau_0 z}))$ has a right-inverse $T(z)$ that is analytic on $\overline{\mathbb{C}^+}$. Now G is a compact subset of $\overline{\mathbb{C}^+}$, so $T(z)$ is bounded on G, and thus

$$K := \max\{\|T(z)\| \mid z \in G\} \tag{16}$$

is well defined.

Choose

$$\rho := \min\left(1, \frac{1}{4K}, \frac{1}{\|A_0(s)\|_{pm} + 1} \cdot \frac{1}{4K\ell} \cdot \min\left(\frac{1}{\|A_0(s)\|_{pm}}, \frac{1}{\|B_0(s)\|_{pm}}\right)\right), \tag{17}$$

then clearly $\rho > 0$. We show that all systems in $\mathcal{B}(\Sigma_0, \rho)$ are stabilizable.

Let $\Sigma = (A(s), B(s), \tau) \in \mathcal{V}$ be such that $d_{\mathcal{V}}(\Sigma, \Sigma_0) < \rho$. The proof that Σ is stabilizable, i.e., that

$$\forall z \in \overline{\mathbb{C}^+} : \operatorname{rank}(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) = n,$$

is divided into two parts: the case $|z| > \|A_0(s)\|_{pm} + 1$ and the case $|z| \leq \|A_0(s)\|_{pm} + 1$.

Let $z \in \overline{\mathbb{C}^+}$, and assume that $|z| > \|A_0(s)\|_{pm} + 1$. Because $d_{\mathcal{V}}(\Sigma, \Sigma_0) < \rho$, we have

$$\|A(s)\|_{pm} \leq \|A_0(s)\|_{pm} + \|A(s) - A_0(s)\|_{pm} < \|A_0(s)\|_{pm} + \rho.$$

Using (17) it follows that $|z| > \|A_0(s)\|_{pm} + 1 \geq \|A_0(s)\|_{pm} + \rho > \|A(s)\|_{pm}$ and, according to Corollary 2.3 (with $w = z$), this implies that

$$\operatorname{rank}(zI - A(e^{-\tau z})) = n.$$

But then certainly $\operatorname{rank}(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) = n$.

The second part of the proof is more complicated. Let $z \in \overline{\mathbb{C}^+}$, $|z| \leq \|A_0(s)\|_{pm} + 1$. We start by proving that

$$\|(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) - (zI - A_0(e^{-\tau_0 z}) \mid B_0(e^{-\tau_0 z}))\| < \frac{1}{K}. \tag{18}$$

First note that

$$\|(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) - (zI - A_0(e^{-\tau_0 z}) \mid B_0(e^{-\tau_0 z}))\|$$

$$\leq \|A_0(e^{-\tau_0 z}) - A(e^{-\tau z})\| + \|B(e^{-\tau z}) - B_0(e^{-\tau_0 z})\|. \tag{19}$$

Now clearly

$$\|A_0(e^{-\tau_0 z}) - A(e^{-\tau z})\| \leq \|A_0(e^{-\tau_0 z}) - A_0(e^{-\tau z})\| + \|A_0(e^{-\tau z}) - A(e^{-\tau z})\|. \tag{20}$$

With Lemma 2.2 it is easy to see that the second term in (20) is bounded from above. Since $\|A(s) - A_0(s)\|_{pm} \leq d_{\mathcal{V}}(\Sigma, \Sigma_0) < \rho$ and $\rho \leq \frac{1}{4K}$, we obtain

$$\|A_0(e^{-\tau z}) - A(e^{-\tau z})\| \leq \|A_0(s) - A(s)\|_{pm} < \rho \leq \frac{1}{4K}. \tag{21}$$

To estimate the other term, we apply the mean value theorem:

$$\|A_0(e^{-\tau_0 z}) - A_0(e^{-\tau z})\| = \left\| \sum_{i=0}^{\ell} A_i \cdot (e^{-i\tau_0 z} - e^{-i\tau z}) \right\| \leq \sum_{i=0}^{\ell} \|A_i\| \cdot i|z| \int_{\tau_0}^{\tau} |e^{-i\xi z}| d\xi$$

$$\leq \|A_0(s)\|_{pm} \cdot \ell \cdot (\|A_0(s)\|_{pm} + 1) \cdot \rho < \frac{1}{4K}, \tag{22}$$

where in the last inequality (17) was used.

Completely analogously we can prove that

$$\|B(e^{-\tau z}) - B_0(e^{-\tau_0 z})\| \leq \|B(e^{-\tau z}) - B_0(e^{-\tau z})\| + \|B_0(e^{-\tau z}) - B_0(e^{-\tau_0 z})\| < \frac{1}{2K}. \tag{23}$$

Combining the previous inequality with (19)–(22), we get (18).

Now recall that $(zI - A_0(e^{-\tau_0 z}) \mid B_0(e^{-\tau_0 z}))$ is right-invertible, with right-inverse $T(z)$. Moreover $\|T(z)\| \leq K$. So

$$\|(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) - (zI - A_0(e^{-\tau_0 z}) \mid B_0(e^{-\tau_0 z}))\| < \frac{1}{K} \leq \frac{1}{\|T(z)\|}. \tag{24}$$

Finally, according to [10, p. 399], (24) implies that the matrix $(zI - A(e^{-\tau z}) \mid B(e^{-\tau z}))$ is right-invertible. This completes the proof. □

Remark 3.2. Although Theorem 3.1 seems to have much in common with the result of Pandolfi in [13, §5], there are several differences. First of all, Pandolfi's result is obtained within the more general framework of distributed parameter systems, of which the class of time-delay systems considered in this paper is only a small subclass. Moreover, in the setting of Pandolfi, perturbations of systems are perturbations of the linear operators describing the system, and robustness of stabilizability is studied in this context. In our approach, perturbations are described within the metric space $\mathcal{V}$ of all parametrizations. Although Pandolfi's result holds in a much more general setting, our result is more specialized to capture the concept of genericity for the class of time-delay systems with point delays.

Remark 3.3. For robustness of stabilizability it is crucial that the rank condition for stabilizability, $\operatorname{rank}(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) = n$, has to be satisfied only on the half plane $\overline{\mathbb{C}^+}$. In [13, §4] it is shown that modal controllability, i.e., the property that the same rank condition is satisfied on the whole complex plane, is not robust. In that situation, the division of the proof into two parts (the cases $|z| > \|A_0(s)\|_{pm} + 1$ and $|z| \leq \|A_0(s)\|_{pm} + 1$) is of no use because the norm of $A(e^{-\tau z})$ may become arbitrarily large when $\operatorname{Re} z \to -\infty$.

4. Some results on matrices over the ring of analytic functions. In the second part of the proof of our genericity result, matrices of analytic functions play an important role. The relationship between the rank of these matrices and their determinants is of special interest. This section can be seen as an intermezzo in which this relationship between rank and determinant is studied using projection matrices.

The first lemma describes how projections can be helpful for the computation of the determinant of a matrix.

LEMMA 4.1. *Let A_1 and A_2 be two arbitrary square matrices, and let E be a projection. Define $\rho(E) := \operatorname{rank}(E)$. Let α be an indeterminate. Then*

$$\det(\alpha E A_1 + (I - E)A_2) = \alpha^{\rho(E)} \cdot \det(E A_1 + (I - E)A_2). \tag{25}$$

Proof. Choose a basis $\{x_1, \ldots, x_n\}$ such that $\operatorname{range}(E) = \langle x_1, \ldots, x_{\rho(E)}\rangle$ and $\operatorname{range}(I - E) = \langle x_{\rho(E)+1}, \ldots, x_n\rangle$. Let $B = \left(\frac{B_1}{B_2}\right)$ denote the matrix of $EA_1 + (I - E)A_2$ with respect to this new basis, where B_1 consists of the first $\rho(E)$ rows of B and B_2 consists of the last $n - \rho(E)$ rows. Then $\left(\frac{\alpha B_1}{B_2}\right)$ is the matrix of $\alpha E A_1 + (I - E)A_2$ with respect to this basis. Hence

$$\begin{aligned}\det(\alpha E A_1 + (I - E)A_2) &= \det\left(\frac{\alpha B_1}{B_2}\right) = \alpha^{\rho(E)} \cdot \det\left(\frac{B_1}{B_2}\right) \\ &= \alpha^{\rho(E)} \cdot \det(E A_1 + (I - E)A_2). \qquad \square\end{aligned}$$

Let $Q(z)$ be an $n \times n$ matrix over the ring of analytic functions on $\mathbb{C}$, i.e., all entries of $Q(z)$ are analytic functions in z. Define

$$p(z) := \det(Q(z)). \tag{26}$$

It is clear that in a point $\lambda \in \mathbb{C}$, the matrix $Q(\lambda)$ is of full rank if and only if $p(\lambda) \neq 0$. Also, when $p(\lambda) = 0$, it is possible to obtain more precise information on the rank of $Q(\lambda)$ from the determinant function $p(z)$, by using a suitable projection E.

PROPOSITION 4.2. *Let $Q(z)$ be an $n \times n$ matrix of analytic functions and $p(z) = \det(Q(z))$. Assume that for a certain $\lambda \in \mathbb{C}$, $p(\lambda) = 0$. Define the matrix of analytic functions $Q_1(z)$ as*

$$Q_1(z) := \frac{Q(z) - Q(\lambda)}{z - \lambda} = \sum_{j=1}^{\infty} \frac{1}{j!} Q^{(j)}(\lambda)(z - \lambda)^{j-1}.$$

Let E be a projection such that $EQ(\lambda) = 0$. Then

$$p(z) = (z - \lambda)^{\rho(E)} \cdot \det(E Q_1(z) + (I - E)Q(z)). \tag{27}$$

Moreover, if $\rho(E) = k$, then

$$\begin{cases} p^{(j)}(\lambda) = 0 \quad \text{for } j = 1, \ldots, k - 1, \\ p^{(k)}(\lambda) = k! \cdot \det(E Q'(\lambda) + (I - E)Q(\lambda)). \end{cases} \tag{28}$$

Proof. $Q(z)$ can be written as $Q(z) = Q(\lambda) + (z - \lambda)Q_1(z)$. Therefore

$$\begin{aligned}p(z) &= \det(EQ(z) + (I - E)Q(z)) = \det((z - \lambda)EQ_1(z) + (I - E)Q(z)) \\ &= (z - \lambda)^{\rho(E)} \cdot \det(EQ_1(z) + (I - E)Q(z)),\end{aligned}$$

where in the last step Lemma 4.1 is used. The result on the derivatives of $p(z)$ in λ when $\rho(E) = k$ is an easy consequence of (27) and the definition of $Q_1(z)$. $\square$

COROLLARY 4.3. *Let $Q(z)$ be an $n \times n$ matrix of analytic functions and $p(z) = \det(Q(z))$. Then*

$$\forall \lambda \in \mathbb{C} : \left[\begin{matrix} p(\lambda) = 0 \\ p'(\lambda) \neq 0 \end{matrix} \right\} \Longrightarrow \operatorname{rank}(Q(\lambda)) = n - 1 \Bigg]. \tag{29}$$

Proof. Let $\lambda \in \mathbb{C}$ be such that $p(\lambda) = 0$ and $p'(\lambda) \neq 0$. Choose a projection E with $\text{range}(Q(\lambda)) = \ker(E)$. Since $Q(\lambda)$ is singular, $\rho(E) = \text{rank}(E) \geq 1$. According to Proposition 4.2 we have

$$p(z) = (z - \lambda)^{\rho(E)} \cdot \det(EQ_1(z) + (I - E)Q(z)).$$

Suppose that $\rho(E) > 1$. Then $p'(\lambda) = 0$. This contradicts our assumption, and therefore, $\rho(E) = 1$. This implies that $\dim(\text{range}(Q(\lambda))) = n - 1$. □

Remark 4.4. From Proposition 4.2 it is clear that ($p(\lambda) = 0$ and $p'(\lambda) \neq 0$) is a sufficient condition for $Q(\lambda)$ to have rank $n - 1$, but it is not a necessary one. It is also possible that $\text{rank}(Q(\lambda)) = n - 1$ while $p'(\lambda) = 0$. In that case the matrix $EQ'(\lambda) + (I - E)Q(\lambda)$ is singular.

In §5 we are especially interested in matrices $Q(z)$ of analytic functions for which the determinant $p(z)$ has only simple zeros. According to Corollary 4.3 this implies that

$$\text{if } p(\lambda) = 0, \text{ then } \text{rank}(Q(\lambda)) = n - 1.$$

Let $Q(z)$ be given, and assume that $\lambda \in \mathbb{C}$ is such that $p(\lambda) = \det(Q(\lambda)) = 0$ and also $p'(\lambda) = 0$. Then it is possible to perturb $Q(z)$ in such a way that λ becomes a simple zero of $p(z)$. However, to prove this result, we first need a lemma that describes how a constant matrix can be perturbed to increase its rank.

LEMMA 4.5. *Let A be an $n \times n$ matrix over $\mathbb{C}$, and assume that* $\text{rank}(A) = \ell$. *For each $j \in \{1, \ldots, n - \ell\}$, there exists a matrix $B \in \mathbb{R}^{n \times n}$ satisfying the following properties:*

(i) $\|B\| = 1$ *and* $\text{rank}(B) = j$,

(ii) $\forall \alpha, \beta \neq 0 :$ $\text{range}(\alpha A + \beta B) = \text{range}(A) \oplus \text{range}(B)$.

Proof. Let $e_1, \ldots, e_n$ denote the standard basis in $\mathbb{C}^n$. Then there exists a permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ such that

$$\text{range}(A) = \langle Ae_{\pi(1)}, \ldots, Ae_{\pi(\ell)} \rangle. \tag{30}$$

Choose vectors $e_{i_1}, \ldots, e_{i_{n-\ell}}$ from the standard basis satisfying

$$\langle e_{i_1}, \ldots, e_{i_{n-\ell}} \rangle \oplus \text{range}(A) = \mathbb{C}^n. \tag{31}$$

Let $j \in \{1, \ldots, n - \ell\}$, and define B as

$$\begin{cases} Be_{\pi(k)} = 0 & \text{for } k = 1, \ldots, \ell, \ell + j + 1, \ldots, n, \\ Be_{\pi(k)} = e_{i_{k-\ell}} & \text{for } k = \ell + 1, \ldots, \ell + j. \end{cases}$$

With this choice of B, it is obvious that (i) is satisfied.

From the construction of B it is immediately clear that $\text{range}(A) \cap \text{range}(B) = \{0\}$. Moreover, the inclusion $\text{range}(\alpha A + \beta B) \subset \text{range}(A) + \text{range}(B)$ is trivial. So, to prove (ii), we only have to show the correctness of the other inclusion.

Let $x_1 \in \text{range}(A)$. Then there exists a $y_1 \in \langle e_{\pi(1)}, \ldots, e_{\pi(\ell)} \rangle$ such that $x_1 = Ay_1$. But clearly $By_1 = 0$. Hence

$$(\alpha A + \beta B)\left(\frac{1}{\alpha} y_1\right) = Ay_1 + \frac{\beta}{\alpha} By_1 = x_1,$$

and $x_1 \in \text{range}(\alpha A + \beta B)$.

Let $x_2 \in \text{range}(B)$. Then there exists a $y_2 \in \langle e_{\pi(\ell+1)}, \ldots, e_{\pi(\ell+j)} \rangle$ such that $By_2 = x_2$. Since $Ay_2 \in \text{range}(A)$, there exists a $y_3 \in \langle e_{\pi(1)}, \ldots, e_{\pi(\ell)} \rangle$ such that $Ay_2 = Ay_3$. Now

$$(\alpha A + \beta B) \cdot \frac{1}{\beta}(y_2 - y_3) = \frac{\alpha}{\beta}(Ay_2 - Ay_3) + By_2 - By_3 = By_2 = x_2,$$

and $x_2 \in \text{range}(\alpha A + \beta B)$. This completes the proof of (ii). □

At this stage all ingredients to prove the main result of this section are available. This result describes how a matrix of analytic functions may be perturbed to reduce the multiplicity of one of the zeros of its determinant to 1.

PROPOSITION 4.6. *Let $Q(z)$ be an $n \times n$ matrix of analytic functions, and define $p(z) = \det(Q(z))$. Assume that $\lambda \in \mathbb{C}$ satisfies $p(\lambda) = 0$. Let $g(z)$ be an analytic function such that $g'(\lambda) \neq 0$.*

Then for each $\varepsilon > 0$ there exists an $n \times n$ polynomial matrix $\Delta(s)$ over $\mathbb{R}[s]$ that satisfies the following properties (where $\tilde{Q}(z) := Q(z) + \Delta(g(z))$ and $\tilde{p}(z) := \det(\tilde{Q}(z))$):

(i) $\|\Delta(s)\|_{pm} < \varepsilon$,

(ii) $\deg(\Delta(s)) \leq 1$ *if $g(\lambda)$ is real, and* $\deg(\Delta(s)) \leq 2$ *if $g(\lambda)$ is complex,*

(iii) $\tilde{p}(\lambda) = 0$ *and* $\tilde{p}'(\lambda) \neq 0$.

Proof. If $p'(\lambda) \neq 0$, the proof is trivial: take $\Delta(s) = 0$.

Assume $p'(\lambda) = 0$. Let $\varepsilon > 0$. If $\text{rank}(Q(\lambda)) = n-1$, define $B_1 := 0$. Otherwise, choose a matrix B_1 according to Lemma 4.5, with $\|B_1\| = 1$ and $\text{rank}(B_1) = n - 1 - \text{rank}(Q(\lambda))$ in such a way that

$$\forall \alpha \neq 0 : \ \text{range}(Q(\lambda) + \alpha B_1) = \text{range}(Q(\lambda)) \oplus \text{range}(B_1).$$

This implies that for all $\alpha \neq 0$, $\text{rank}(Q(\lambda) + \alpha B_1) = n - 1$.

Fix $\alpha := \frac{1}{3}\varepsilon$ and apply Lemma 4.5 again, but now to the matrix $Q(\lambda) + \alpha B_1$. In this way we find a matrix B_2 (possibly depending on α), satisfying $\|B_2\| = 1$, $\text{rank}(B_2) = 1$, and

$$\forall \beta \neq 0 : \ \text{range}(Q(\lambda) + \alpha B_1 + \beta B_2) = \text{range}(Q(\lambda)) \oplus \text{range}(B_1) \oplus \text{range}(B_2).$$

So for every $\beta \neq 0$, the matrix $(Q(\lambda) + \alpha B_1 + \beta B_2)$ has rank n.

Let E denote the projection on $\text{range}(B_2)$ along $\text{range}(Q(\lambda) + \alpha B_1)$, so that $E(Q(\lambda) + \alpha B_1) = 0$ and $EB_2 = B_2$. Then $\rho(E) = \text{rank}(E) = 1$. Define $Q_\alpha(z) := Q(z) + \alpha B_1$ and $p_\alpha(z) := \det(Q_\alpha(z))$. So $p_\alpha(\lambda) = \det(Q(\lambda) + \alpha B_1) = 0$ and using formula (28) from Proposition 4.2 we obtain

$$p'_\alpha(\lambda) = \det(EQ'_\alpha(\lambda) + (I - E)Q_\alpha(\lambda)) = \det(EQ'(\lambda) + (I - E)Q_\alpha(\lambda)).$$

First note that $\ker(EQ'(\lambda) + (I - E)Q_\alpha(\lambda)) \subset \ker(Q_\alpha(\lambda))$. Moreover, we know that $\dim(\ker(Q_\alpha(\lambda))) = 1$. Therefore the problem can be divided into two different cases.

Case 1. $\ker(EQ'(\lambda) + (I - E)Q_\alpha(\lambda)) = \{0\}$. Then $p'_\alpha(\lambda) = \det(EQ'(\lambda) + (I - E)Q_\alpha(\lambda)) \neq 0$, and $\Delta(s) := \alpha B_1$ satisfies (ii), (iii), and also (i) because $\|\Delta(s)\|_{pm} = \|\alpha B_1\| \leq \frac{1}{3}\varepsilon$.

Case 2. $\ker(EQ'(\lambda) + (I - E)Q_\alpha(\lambda)) = \ker(Q_\alpha(\lambda))$. If $g(\lambda)$ is real, define for all $\beta \in \mathbb{R}\backslash\{0\}$

$$\Delta_\beta(s) := \alpha B_1 + \beta(s - g(\lambda))B_2;$$

if $g(\lambda)$ is complex, define for all $\beta \in \mathbb{R}\backslash\{0\}$

$$\Delta_\beta(s) := \alpha B_1 + \beta(s - g(\lambda))(s - \overline{g(\lambda)})B_2.$$

Then in each case $\Delta_\beta(s) \in \mathbb{R}[s]^{n \times n}$, and moreover (ii) is satisfied.

Let $\beta \in \mathbb{R}\backslash\{0\}$ and define $\tilde{Q}(z) := Q(z) + \Delta_\beta(g(z))$. Then $\tilde{Q}(\lambda) = Q(\lambda) + \alpha B_1$, and in both the real and the complex cases there exists a $\gamma \neq 0$ such that $\tilde{Q}'(\lambda) = Q'(\lambda) + \gamma B_2$. Since $\tilde{Q}(\lambda) = Q(\lambda) + \alpha B_1$ is singular, we still have that $\tilde{p}(\lambda) = 0$, and according to Proposition 4.2,

$$\tilde{p}'(\lambda) = \det(E\tilde{Q}'(\lambda) + (I - E)\tilde{Q}(\lambda)) = \det(E(Q'(\lambda) + \gamma B_2) + (I - E)Q_\alpha(\lambda)).$$

Assume that $x \in \ker(E(Q'(\lambda) + \gamma B_2) + (I - E)Q_\alpha(\lambda))$. Then $x \in \ker(Q_\alpha(\lambda))$. So by assumption, $x \in \ker(EQ'(\lambda) + (I - E)Q_\alpha(\lambda))$. Moreover we have that $EB_2 = B_2$, and thus we obtain

$$\gamma B_2 x = (E(Q'(\lambda) + \gamma B_2) + (I - E)Q_\alpha(\lambda))x - (EQ'(\lambda) + (I - E)Q_\alpha(\lambda))x = 0.$$

So $(Q_\alpha(\lambda) + \gamma B_2)x = 0$. By construction $Q_\alpha(\lambda) + \gamma B_2 = Q(\lambda) + \alpha B_1 + \gamma B_2$ has full rank, and thus $x = 0$. This implies that $\operatorname{rank}(E(Q'(\lambda) + \gamma B_2) + (I - E)Q_\alpha(\lambda)) = n$. Therefore $\tilde{p}'(\lambda) \neq 0$, and $\tilde{Q}(z)$ satisfies condition (iii) for all $\beta \neq 0$.

To satisfy (i), we choose

$$\beta := \frac{1}{4}\varepsilon \cdot \min\left(\frac{1}{|g(\lambda)|}, 1\right),$$

when $g(\lambda)$ is real, and

$$\beta := \frac{1}{8}\varepsilon \cdot \min\left(\frac{1}{|g(\lambda) + \overline{g(\lambda)}|}, \frac{1}{g(\lambda) \cdot \overline{g(\lambda)}}, 1\right),$$

when $g(\lambda)$ is complex. Then it is easily verified that $\|\Delta_\beta(s)\|_{pm} < \varepsilon$. This completes the proof. □

Remark 4.7. When the matrix $Q(z)$ of analytic functions has the property that $\overline{Q(\overline{z})} = Q(z)$, its determinant $p(z)$ also has that property. This implies that λ is a zero of $p(z)$ of multiplicity k if and only if $\overline{\lambda}$ is a zero of $p(z)$ of the same multiplicity. Note also that if $\overline{g(\overline{z})} = g(z)$ and $g(\lambda)$ is complex, the reduction process described in the proof of Proposition 4.6 reduces the multiplicity of both the zeros λ and $\overline{\lambda}$ to 1 in only one step. Although in general a perturbation matrix $\Delta(s)$ of degree 2 is needed to fix this problem, this matrix handles both zeros λ and $\overline{\lambda}$ at the same time.

Remark 4.8. Corollary 4.3 and Proposition 4.6 are formulated in a very general context of matrices over analytic functions, but in the next section they are only used for a very specific case. It is clear that for the time-delay system corresponding to the point $\Sigma = (A(s), B(s), \tau) \in \mathcal{V}$, the matrix $(zI - A(e^{-\tau z}))$ is very important for its stabilizability properties. Therefore the results of this section are applied to the case $Q(z) = (zI - A(e^{-\tau z}))$ and $g(z) = e^{-\tau z}$. Then clearly $g'(\lambda) = -\tau e^{-\tau\lambda} \neq 0$ for all $\lambda \in \mathbb{C}$. In this perspective, Proposition 4.6 describes how the matrix $A(s)$ has to be perturbed within the normed ring $\mathbb{R}[s]^{n\times n}$ in such a way that $(zI - (A(e^{-\tau z}) + \Delta(e^{-\tau z})))$ satisfies the condition of Corollary 4.3.

5. Approximation by stabilizable time-delay systems. In this section, the second and final part of our genericity result is proven. We show that the subset S of $\mathcal{V}$, consisting of all parametrizations of stabilizable time-delay systems, is a dense subset of $\mathcal{V}$. This means that in any arbitrary small neighborhood of a point $\Sigma \in \mathcal{V}$, corresponding to a nonstabilizable time-delay system, there is a point $\tilde{\Sigma} \in S \subset \mathcal{V}$ that describes a stabilizable time-delay system. In this section such an approximation by stabilizable time-delay systems is constructed explicitly.

The main idea of the proof is as follows. Let a point $\Sigma = (A(s), B(s), \tau) \in \mathcal{V}$ be given such that the corresponding time-delay system is not stabilizable. First of all it may be shown (use, e.g., Corollary 2.3) that for all matrices $\tilde{A}(s) \in \mathbb{R}[s]^{n\times n}$, the analytic function $\tilde{p}(z) = \det(zI - \tilde{A}(e^{-\tau z}))$ has only a finite number of zeros in $\overline{\mathbb{C}^+}$. Using Rouché's theorem,

Corollary 4.3, and Proposition 4.6, it is possible to prove that for every $\varepsilon > 0$, there exists a matrix $A_\varepsilon(s) \in \mathbb{R}[s]^{n\times n}$ such that $\|A(s) - A_\varepsilon(s)\|_{pm} < \frac{1}{2}\varepsilon$ and

$$(32) \qquad \forall z \in \overline{\mathbb{C}^+} : \left[\operatorname{rank}(zI - A_\varepsilon(e^{-\tau z})) < n \Longrightarrow \operatorname{rank}(zI - A_\varepsilon(e^{-\tau z})) = n - 1 \right].$$

So, in all points $z \in \overline{\mathbb{C}^+}$ where the matrix $(zI - A_\varepsilon(e^{-\tau z}))$ loses rank, it loses only rank 1. This loss of rank has to be compensated by the matrix $B(s)$. Therefore this matrix has to be perturbed in such a way that the perturbed version $B_\varepsilon(s)$ satisfies the inequality $\|B_\varepsilon(s) - B(s)\|_{pm} < \frac{1}{2}\varepsilon$ and is such that

$$(33) \quad \forall z \in \overline{\mathbb{C}^+} : \left[\operatorname{rank}(zI - A_\varepsilon(e^{-\tau z})) < n \Longrightarrow \operatorname{rank}(zI - A_\varepsilon(e^{-\tau z}) \mid B_\varepsilon(e^{-\tau z})) = n \right].$$

Since the analytic function $p_\varepsilon(z) = \det(zI - A_\varepsilon(e^{-\tau z}))$ has only a finite number of zeros in the closed right half plane, it is possible to satisfy this condition. In this way we find a stabilizable time-delay system $\Sigma_\varepsilon = (A_\varepsilon(s), B_\varepsilon(s), \tau)$ such that $d_\nu(\Sigma, \Sigma_\varepsilon) < \varepsilon$, and the proof is complete.

The rest of this section is devoted to a detailed elaboration of the scheme of the proof given above. The first lemma (which can be seen as a direct consequence of Corollary 2.3) describes the location of the zeros of the analytic function $p(z) = \det(zI - A(e^{-\tau z}))$ corresponding to the square polynomial matrix $A(s)$.

LEMMA 5.1 (see [7, p. 18]). *Let $A(s) \in \mathbb{R}[s]^{n\times n}$ and $\tau > 0$ be given. Then the analytic function $p(z) = \det(zI - A(e^{-\tau z}))$ has only a finite number of zeros in the closed right half plane $\overline{\mathbb{C}^+}$. Moreover, all the zeros of $p(z)$ in $\overline{\mathbb{C}^+}$ are located within the semi-disc*

$$(34) \qquad D := \{z \in \mathbb{C} \mid \operatorname{Re} z \geq 0 \text{ and } |z| \leq \|A(s)\|_{pm}\}.$$

In Lemma 5.1, the role of the right half plane $\overline{\mathbb{C}^+}$ is not crucial. By shifting to the left and to the right it is possible to show that $p(z) = \det(zI - A(e^{-\tau z}))$ has a finite number of zeros in any arbitrary right half plane.

In the proof of the main results of this section, we often assume that the function $p(z) = \det(zI - A(e^{-\tau z}))$ has no zeros on the boundary of $\mathbb{C}^+$, i.e., $p(z)$ has no zeros on the imaginary axis. Fortunately, this is not really a restriction. By an arbitrarily small perturbation of the matrix $A(s)$ corresponding to $p(z) = \det(zI - A(e^{-\tau z}))$, it is possible to shift the zeros of $p(z)$ in the horizontal direction. In this way we can prove the following result.

PROPOSITION 5.2. *Let $A(s) \in \mathbb{R}[s]^{n\times n}$ and $\tau > 0$ be given. Let $\varepsilon > 0$. Then there exists a polynomial matrix $A_1(s) \in \mathbb{R}[s]^{n\times n}$ satisfying the following properties:*

(i) $\|A(s) - A_1(s)\|_{pm} < \varepsilon$,

(ii) $\deg(A_1(s)) = \deg(A(s))$,

(iii) *the characteristic function $p_1(z) := \det(zI - A_1(e^{-\tau z}))$ has no zeros on the imaginary axis.*

The next theorem is a restatement of a well-known result from complex analysis. It plays a crucial role in the rest of this section because it describes how small perturbations of an analytic function influence the location of its zeros.

THEOREM 5.3 (Rouché's theorem (see, e.g., [15, Thm. 10.43])). *Let f and g be two functions that are analytic inside and on a Jordan curve $\mathcal{J}$. Suppose that f and g have no zeros on $\mathcal{J}$. Denote by N_f and N_g the total number of zeros of f and g inside $\mathcal{J}$, also counting multiplicities. Then*

$$(35) \qquad [\, \forall z \in \mathcal{J} : |f(z) - g(z)| < |f(z)| \,] \Longrightarrow N_g = N_f.$$

Let $\mathcal{J}$ be a Jordan curve, and let f and g be two functions satisfying the conditions of Theorem 5.3. Define $\delta := \min\{|f(z)| \mid z \in \mathcal{J}\}$. Then the condition $|f(z) - g(z)| < \delta$ implies

that f and g have the same number of zeros inside $\mathcal{J}$. This observation is exploited in the next lemma.

LEMMA 5.4. *Let $A(s) \in \mathbb{R}[s]^{n\times n}$ and $\tau > 0$ be given. Let $\mathcal{J}$ be a Jordan curve in $\overline{\mathbb{C}^+}$ such that $p(z) = \det(zI - A(e^{-\tau z}))$ has no zeros on $\mathcal{J}$. Then there exists an $\bar{\varepsilon} > 0$ such that for all polynomial matrices $\tilde{A}(s) \in \mathbb{R}[s]^{n\times n}$ satisfying $\|A(s) - \tilde{A}(s)\|_{pm} < \bar{\varepsilon}$, the characteristic function $\tilde{p}(z) := \det(zI - \tilde{A}(e^{-\tau z}))$ corresponding to $\tilde{A}(s)$ has the same number of zeros within $\mathcal{J}$ as $p(z)$ (counting multiplicities) and no zeros on $\mathcal{J}$.*

Proof. Define $p_c(s, z) := \det(zI - A(s))$. Then $p_c(s, z) \in \mathbb{R}[s, z]$, and the degree of $p_c(s, z)$ in z is n. Define

$$\delta := \min\{|p(z)| \mid z \in \mathcal{J}\}, \tag{36}$$

and $M := 1 + \max\{|z| \mid z \in \mathcal{J}\}$. Now apply Proposition 2.7. Choose an $\bar{\varepsilon} > 0$ in such a way that for all matrices $\tilde{A}(s) \in \mathbb{R}[s]^{n\times n}$ satisfying $\|A(s) - \tilde{A}(s)\|_{pm} < \bar{\varepsilon}$, the following inequality holds:

$$\|p_c(s, z) - \tilde{p}_c(s, z)\|_p < \delta \frac{M - 1}{M^{n+1} - 1}. \tag{37}$$

Here $\tilde{p}_c(s, z)$ denotes the characteristic polynomial $\det(zI - \tilde{A}(s))$ of $\tilde{A}(s)$, which is also of degree n in z. We show that for $\bar{\varepsilon}$ the claim of Lemma 5.4 holds.

Let $\tilde{A}(s) \in \mathbb{R}[s]^{n\times n}$ be such that $\|A(s) - \tilde{A}(s)\|_{pm} < \bar{\varepsilon}$. First apply Lemma 2.6 to $r(s, z) := p_c(s, z) - \tilde{p}_c(s, z)$ and use inequality (37). In this way we obtain

$$\forall z \in \overline{\mathbb{C}^+}, |z| \le M : \ |p(z) - \tilde{p}(z)| < \delta. \tag{38}$$

So in particular $|p(z) - \tilde{p}(z)| < \delta$ for all $z \in \mathcal{J}$. Apparently $\tilde{p}(z)$ has no zeros on $\mathcal{J}$. (Otherwise there would be a $\lambda \in \mathcal{J}$ such that $|p(\lambda)| < \delta$, which contradicts definition (36).) Finally, since both $p(z)$ and $\tilde{p}(z)$ are analytic functions without zeros on $\mathcal{J}$, Rouché's theorem and formulae (36) and (38) imply that $p(z)$ and $\tilde{p}(z)$ have the same number of zeros inside the Jordan curve $\mathcal{J}$ (counting multiplicities). □

Lemma 5.4 indicates that small perturbations of the matrix $A(s)$ affect the zeros of $p(z)$ only slightly: they cannot cross the Jordan curve $\mathcal{J}$. The idea is now to perturb $A(s)$ in such a way that the multiple zeros of $p(z)$ inside $\mathcal{J}$ become simple without changing the total number of zeros inside $\mathcal{J}$. In this approach, Rouché's theorem (in the disguised form of Lemma 5.4) again plays an important role.

PROPOSITION 5.5. *Let $A(s) \in \mathbb{R}[s]^{n\times n}$ and $\tau > 0$ be given. Let $\mathcal{J}$ be a Jordan curve in $\overline{\mathbb{C}^+}$, and assume that $p(z) = \det(zI - A(e^{-\tau z}))$ has no zeros on $\mathcal{J}$. Choose $\bar{\varepsilon} > 0$ such that Lemma 5.4 is satisfied. Let N_p denote the total number of zeros of $p(z)$ within $\mathcal{J}$, counting multiplicities. Then for all $\varepsilon \in (0, \bar{\varepsilon})$ there exists a matrix $\tilde{A}(s) \in \mathbb{R}[s]^{n\times n}$ such that*

(i) $\|A(s) - \tilde{A}(s)\|_{pm} < \varepsilon$,

(ii) $\deg(\tilde{A}(s)) \le \max(\deg(A(s)), 2)$,

(iii) *the analytic function $\tilde{p}(z) := \det(zI - \tilde{A}(e^{-\tau z}))$ has N_p zeros within $\mathcal{J}$ and all these zeros are simple.*

Proof. Let $\varepsilon \in (0, \bar{\varepsilon})$. Then it follows from Lemma 5.4 that for all $\tilde{A}(s) \in \mathbb{R}[s]^{n\times n}$ satisfying $\|A(s) - \tilde{A}(s)\|_{pm} < \varepsilon < \bar{\varepsilon}$, the number of zeros of $\tilde{p}(z) = \det(zI - \tilde{A}(e^{-\tau z}))$ inside $\mathcal{J}$ is equal to N_p. Let L_p denote the number of simple zeros of $p(z)$ within $\mathcal{J}$. The proposition is proved with the following induction argument:

$$\forall i \in \{0, 1, \ldots, N_p - L_p\} \ \exists A_i(s) \in \mathbb{R}[s]^{n\times n} \text{ such that}$$

(1) $\|A(s) - A_i(s)\|_{pm} \le \frac{2^i - 1}{2^i} \cdot \varepsilon$,

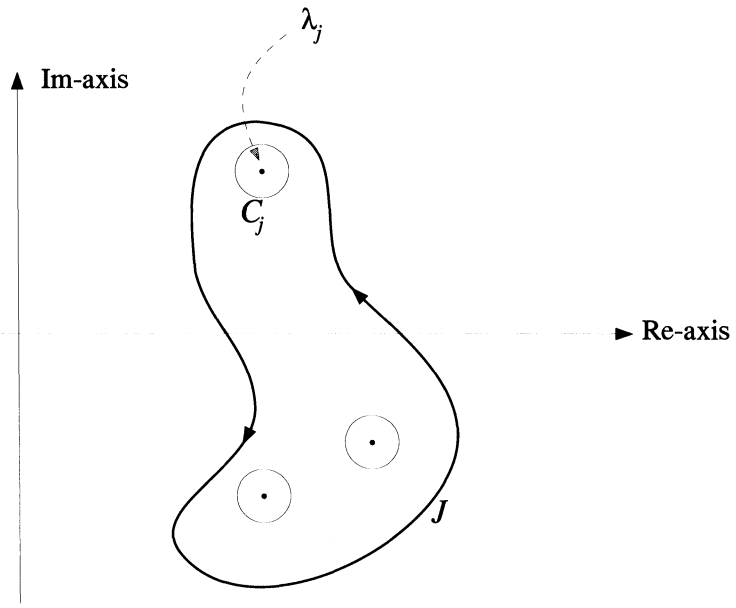


FIG. 5.1. *Location of the zeros inside a Jordan curve* $\mathcal{J}$.

(2) $\deg(A_i(s)) \leq \max(\deg(A(s)), 2)$,

(3) the analytic function $p_i(z) = \det(zI - A_i(e^{-\tau z}))$ has at least $L_p + i$ simple zeros within $\mathcal{J}$, i.e., $L_{p_i} \geq L_p + i$, where L_{p_i} denotes the number of simple zeros of $p_i(z)$ enclosed by $\mathcal{J}$.

When $i = 0$, this is trivial. Choose $A_0(s) = A(s)$.

Induction step. Suppose that for certain $i \in \{0, 1, \ldots, N_p - L_p - 1\}$ we have found a matrix $A_i(s)$ satisfying (1)-(3). If $L_{p_i} \geq L_p + i + 1$, choose $A_{i+1}(s) = A_i(s)$, and we are ready.

Next assume that $L_{p_i} = L_p + i$. Since $i < N_p - L_p$, we know that at least one of the N_p zeros of $p_i(z)$ inside $\mathcal{J}$ is a multiple zero. Let λ_j, $j \in \{1, \ldots, \ell\}$, denote all distinct zeros of $p_i(z)$ in $\mathcal{J}$. Then there exists a $\rho > 0$ such that the circles C_j defined by

$$C_j = \{z \in \mathbb{C} \mid |z - \lambda_j| = \rho\} \tag{39}$$

neither intersect one another nor the Jordan curve $\mathcal{J}$ (see Figure 5.1). Apply Lemma 5.4 to each of these circles C_j. Then for all $j = 1, \ldots, \ell$, we find an $\bar{\varepsilon}_j > 0$ such that for all $\hat{A}(s) \in \mathbb{R}[s]^{n \times n}$, the inequality $\|A(s) - \hat{A}(s)\|_{pm} < \bar{\varepsilon}_j$ implies that $p_i(z)$ and $\hat{p}(z) = \det(zI - \hat{A}(e^{-\tau z}))$ have the same number of zeros within C_j and no zeros on C_j. Define $\hat{\varepsilon} := \min\{\bar{\varepsilon}_j \mid j = 1, \ldots, \ell\}$.

Assume, without loss of generality, that λ_1 is a multiple zero of $p_i(z)$. Apply Proposition 4.6 to $Q(z) = (zI - A_i(e^{-\tau z}))$ with $g(z) = e^{-\tau z}$ and $\lambda = \lambda_1$. Clearly $g'(\lambda_1) = -\tau e^{-\tau \lambda_1} \neq 0$, so there exists a polynomial matrix $\Delta(s) \in \mathbb{R}[s]^{n \times n}$, with $\deg(\Delta(s)) \leq 2$, in norm bounded by

$$\|\Delta(s)\|_{pm} < \min\left(\hat{\varepsilon}, \frac{1}{2^{i+1}} \cdot \varepsilon\right)$$

and such that $\tilde{p}(z) = \det(Q(z) + \Delta(e^{-\tau z}))$ has only a simple zero in $z = \lambda_1$. We show that $A_{i+1}(s) := A_i(s) + \Delta(s)$ meets the requirements (1)-(3), with i replaced by $i + 1$.

(1) and (2) are very straightforward:

$$\begin{aligned}\|A(s) - A_{i+1}(s)\|_{pm} &\leq \|A(s) - A_i(s)\|_{pm} + \|A_i(s) - A_{i+1}(s)\|_{pm} \\ &\leq \frac{2^i - 1}{2^i} \cdot \varepsilon + \frac{1}{2^{i+1}} \cdot \varepsilon = \frac{2^{i+1} - 1}{2^{i+1}} \cdot \varepsilon,\end{aligned}$$

and $\deg(A_{i+1}(s)) \leq \max(\deg(A_i(s)), 2) \leq \max(\deg(A(s)), 2)$.

(3) Since $\|A_{i+1}(s) - A_i(s)\|_{pm} < \hat{\varepsilon}$, we can apply Lemma 5.4 to each of the circles $\mathcal{C}_j$ defined in (39) separately. In this way we determine that for all $j \in \{1, \ldots, \ell\}$, the number of zeros of $p_{i+1}(z)$ within $\mathcal{C}_j$ is equal to the number of zeros of $p_i(z)$ within $\mathcal{C}_j$ (counting multiplicities). This implies that the L_{p_i} circles containing a simple zero of $p_i(z)$ also contain exactly one (simple) zero of $p_{i+1}(z)$. Moreover, the multiple zero λ_1 has become simple by construction, and thus

$$L_{p_{i+1}} \geq L_{p_i} + 1 = L_p + i + 1.$$

This completes the proof of the induction argument. The correctness of Proposition 5.5 follows immediately by taking $\tilde{A}(s) = A_{N_p - L_p}(s)$. □

Proposition 5.5 shows that the matrix perturbations introduced in Proposition 4.6 can be used successively to reduce the multiplicity of zeros to 1. Rouché's theorem guarantees not only that the total number of zeros within the Jordan curve $\mathcal{J}$ remains constant but also that simple zeros remain simple. Combining Propositions 5.2 and 5.5, we can finish the first part of the proof as indicated in the introduction of this section by an appropriate choice of the Jordan curve $\mathcal{J}$.

THEOREM 5.6. *Let $A(s) \in \mathbb{R}[s]^{n \times n}$ and $\tau > 0$ be given. Then for all $\varepsilon > 0$ there exists a matrix $A_\varepsilon(s) \in \mathbb{R}[s]^{n \times n}$ such that*

(i) $\|A(s) - A_\varepsilon(s)\|_{pm} < \varepsilon$,

(ii) $\deg(A_\varepsilon(s)) \leq \max(\deg(A(s)), 2)$,

(iii) $\forall \lambda \in \overline{\mathbb{C}^+} : \operatorname{rank}(\lambda I - A_\varepsilon(e^{-\tau\lambda})) \geq n - 1$.

Proof. Let $\varepsilon > 0$. Choose according to Proposition 5.2 a matrix $A_1(s) \in \mathbb{R}[s]^{n \times n}$, of the same degree as $A(s)$, satisfying $\|A(s) - A_1(s)\|_{pm} < \frac{1}{2}\varepsilon$ and such that $p_1(z) = \det(zI - A_1(e^{-\tau z}))$ has no zeros on the imaginary axis.

Define $R := \|A_1(s)\|_{pm} + 1$ and the Jordan curve $\mathcal{J}$, as depicted in Figure 5.2, by

$$\mathcal{J} := \{z \in \mathbb{C} \mid (\operatorname{Re} z = 0 \text{ and } |z| < R) \text{ or } (\operatorname{Re} z \geq 0 \text{ and } |z| = R)\}. \tag{40}$$

So, according to Lemma 5.1, all zeros of $p_1(z) = \det(zI - A_1(e^{-\tau z}))$ in $\overline{\mathbb{C}^+}$ are located inside the Jordan curve $\mathcal{J}$. Let N_{p_1} denote this number of zeros of $p_1(z)$ within $\mathcal{J}$ (counting multiplicities). We choose $\bar{\varepsilon} > 0$ such that Lemma 5.4 is satisfied and apply Proposition 5.5 with $\tilde{\varepsilon} := \frac{1}{2} \cdot \min(1, \varepsilon, \bar{\varepsilon})$. Then we find a matrix $A_\varepsilon(s) \in \mathbb{R}[s]^{n \times n}$ such that

(1) $\|A_1(s) - A_\varepsilon(s)\|_{pm} \leq \frac{1}{2} \cdot \min(1, \varepsilon, \bar{\varepsilon}) \leq \frac{1}{2}\varepsilon$,

(2) $\deg(A_\varepsilon(s)) \leq \max(\deg(A_1(s)), 2)$,

(3) $p_\varepsilon(z) = \det(zI - A_\varepsilon(e^{-\tau z}))$ has N_{p_1} zeros within $\mathcal{J}$ that are all simple.

Clearly, the matrix $A_\varepsilon(s)$ satisfies both (i) and (ii), so we have to prove only (iii). Since $\|A_\varepsilon(s)\|_{pm} < \|A_1(s)\|_{pm} + \frac{1}{2}$, Lemma 5.1 implies that $p_\varepsilon(z)$ has no zeros in $\overline{\mathbb{C}^+}$ outside $\mathcal{J}$. Moreover, since $\|A_1(s) - A_\varepsilon(s)\|_{pm} < \bar{\varepsilon}$, we know from Lemma 5.4 that $p_\varepsilon(z)$ has no zeros on $\mathcal{J}$. Therefore all zeros of $p_\varepsilon(z)$ in $\overline{\mathbb{C}^+}$ are located within $\mathcal{J}$. According to (3), all these zeros are simple and thus we have

$$\forall z \in \overline{\mathbb{C}^+} : \left[p_\varepsilon(z) = 0 \Longrightarrow p_\varepsilon'(z) \neq 0\right]. \tag{41}$$

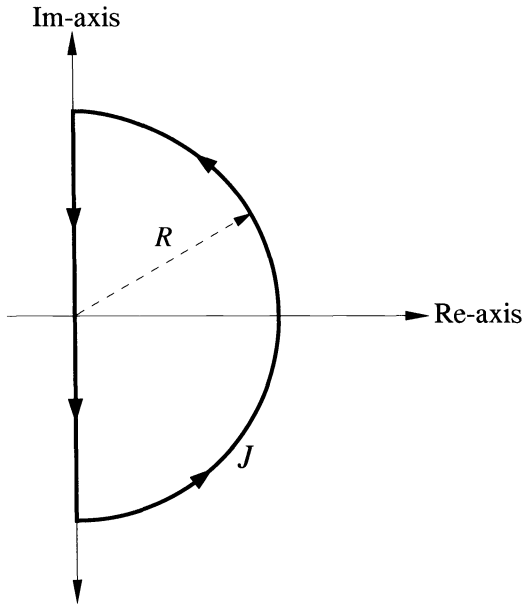


FIG. 5.2. *The Jordan curve* $\mathcal{J}$.

Now, let $\lambda \in \overline{\mathbb{C}^+}$, and assume that $\operatorname{rank}(\lambda I - A_\varepsilon(e^{-\tau\lambda})) < n$. Then $p_\varepsilon(\lambda) = \det(\lambda I - A_\varepsilon(e^{-\tau\lambda})) = 0$, and thus according to (41), $p'_\varepsilon(\lambda) \neq 0$. Applying Corollary 4.3 to the matrix $(zI - A_\varepsilon(e^{-\tau z}))$, we obtain

$$\operatorname{rank}(\lambda I - A_\varepsilon(e^{-\tau\lambda})) = n - 1.$$

This completes the proof. □

In the second part of this section we are concerned with perturbations of the matrix $B(s)$. Suppose that a point $\Sigma = (A(s), B(s), \tau) \in \mathcal{V}$ is given. First perturb $A(s)$ in such a way that for $A_\varepsilon(s)$ conditions (i)-(iii) of Theorem 5.6 are satisfied. From Lemma 5.1 it follows that the analytic function $p_\varepsilon(z) = \det(zI - A_\varepsilon(e^{-\tau z}))$ has only a finite number of zeros in $\overline{\mathbb{C}^+}$, say $\lambda_1, \ldots, \lambda_k$. We know that for each $i \in \{1, \ldots, k\}$, $\operatorname{rank}(\lambda_i I - A_\varepsilon(e^{-\tau\lambda_i})) = n - 1$, and therefore the left-kernel of the matrix $(\lambda_i I - A_\varepsilon(e^{-\tau\lambda_i}))$, i.e., the linear subspace of $\mathbb{C}^n$ consisting of all row vectors x^T such that $x^T \cdot (\lambda_i I - A_\varepsilon(e^{-\tau\lambda_i})) = 0$, is 1-dimensional. So for each $i \in \{1, \ldots, k\}$, this left-kernel is spanned by one row vector $v_i^T \in \mathbb{C}^n$. Now $(\lambda_i I - A_\varepsilon(e^{-\tau\lambda_i}) \mid B(e^{-\tau\lambda_i}))$ has rank n if and only if

$$v_i^T \cdot B(e^{-\tau\lambda_i}) \neq 0. \tag{42}$$

So, to achieve stabilizability, we have to perturb $B(s)$ in such a way that for the perturbed version $B_\varepsilon(s)$ the following holds:

$$\forall i \in \{1, \ldots, k\} : \; v_i^T \cdot B_\varepsilon(e^{-\tau\lambda_i}) \neq 0. \tag{43}$$

To find such a perturbation of $B(s)$, we first look for a vector b that is not perpendicular to a given finite set of vectors.

LEMMA 5.7. *Let the column vectors* $v_1, \ldots, v_k \in \mathbb{C}^n$ *be given, and assume that they are all nonzero. Then there exists a vector* $b \in \mathbb{R}^n$ *such that*

$$\forall i \in \{1, \ldots, k\} : v_i^T \cdot b \neq 0.$$

Proof. First define for all $i = 1, \ldots, k$ the linear spaces

$$V_i := \{x \in \mathbb{R}^n \mid v_i^{\mathrm{T}} \cdot x = 0\}.$$

Since all vectors v_i are nonzero, the sets V_i are linear subspaces of $\mathbb{R}^n$, with dimension smaller than or equal to $n-1$. This implies that each V_i is a nowhere dense subset of $\mathbb{R}^n$. Application of Baire's category theorem (see, for example, [15, Thm. 5.6 and Rem. 5.7]) yields

$$\mathbb{R}^n \neq \bigcup_{i=1}^{k} V_i. \qquad \square$$

Intuitively, the result of Lemma 5.7 is clear. The vectors $v_1, \ldots, v_k$ correspond to linear subspaces $V_1, \ldots, V_k$ in $\mathbb{R}^n$ of dimension smaller than or equal to $n-1$. Now we simply have to pick a vector $b \in \mathbb{R}^n$ that is not an element of one of these subspaces $V_1, \ldots, V_k$. Since we only consider a finite number of subspaces, this is a rather easy task.

Lemma 5.7 makes it possible to find a perturbation of the matrix $B(s)$ that is suitable for our purpose. This result is stated in the next lemma.

LEMMA 5.8. *Let the vectors $v_1, \ldots, v_k \in \mathbb{C}^n$ and $b_1, \ldots, b_k \in \mathbb{C}^n$ be given. Assume that for all $i \in \{1, \ldots, k\} : \|v_i\| = 1$. Then for all $\varepsilon > 0$ there exists a vector $\beta \in \mathbb{R}^n$ such that*

(i) $\|\beta\| < \varepsilon$,

(ii) $\forall i \in \{1, \ldots, k\} : \; v_i^{\mathrm{T}} \cdot (b_i + \beta) \neq 0$.

Proof. Let $\varepsilon > 0$. Choose, according to Lemma 5.7, a vector $\gamma \in \mathbb{R}^n$ such that $v_i^{\mathrm{T}} \cdot \gamma \neq 0$ for all $i \in \{1, \ldots, k\}$. If for all $i \in \{1, \ldots, k\}$ we have $v_i^{\mathrm{T}} \cdot b_i = 0$, then $\beta = \frac{1}{2}\varepsilon \cdot \frac{\gamma}{\|\gamma\|}$ satisfies the claim. Otherwise, choose a $\rho \in (0, \min\{|v_i^{\mathrm{T}} \cdot b_i| \mid v_i^{\mathrm{T}} \cdot b_i \neq 0, \; i = 1, \ldots, k\})$, and define

$$\beta := \frac{1}{2} \cdot \min(\varepsilon, \rho) \cdot \frac{1}{\|\gamma\|} \cdot \gamma.$$

Then (i) is clear: $\|\beta\| \leq \frac{1}{2} \cdot \varepsilon \cdot 1 < \varepsilon$. To prove (ii), let $i \in \{1, \ldots, k\}$. If $v_i^{\mathrm{T}} \cdot b_i = 0$, then

$$v_i^{\mathrm{T}} \cdot (b_i + \beta) = v_i^{\mathrm{T}} \cdot \beta = \frac{1}{2} \cdot (v_i^{\mathrm{T}} \gamma) \cdot \frac{1}{\|\gamma\|} \cdot \min(\varepsilon, \rho) \neq 0.$$

On the other hand, if $v_i^{\mathrm{T}} \cdot b_i \neq 0$, then

$$|v_i^{\mathrm{T}} \cdot (b_i + \beta)| = |v_i^{\mathrm{T}} b_i + v_i^{\mathrm{T}} \beta| \geq |v_i^{\mathrm{T}} b_i| - |v_i^{\mathrm{T}} \beta| \geq \rho - \|v_i\| \cdot \|\beta\| \geq \rho - \frac{1}{2}\rho > 0.$$

So, in either case, $v_i^{\mathrm{T}} \cdot (b_i + \beta) \neq 0$. $\square$

At this point, the proof outlined in the introduction of this section is almost complete. We have only to state and prove the main result.

THEOREM 5.9. *Let $\Sigma = (A(s), B(s), \tau) \in \mathcal{V}$ be given. For all $\varepsilon > 0$ there exists a point $\tilde{\Sigma} = (\tilde{A}(s), \tilde{B}(s), \tilde{\tau}) \in \mathcal{V}$ such that*

(i) $d_{\mathcal{V}}(\Sigma, \tilde{\Sigma}) < \varepsilon$,

(ii) $\deg(\tilde{A}(s)) \leq \max(\deg(A(s)), 2)$ *and* $\deg(\tilde{B}(s)) = \deg(B(s))$,

(iii) *the time-delay system corresponding to $\tilde{\Sigma}$ is stabilizable, i.e.,*

$$\forall z \in \overline{\mathbb{C}^+} : \; \operatorname{rank}(zI - \tilde{A}(e^{-\tilde{\tau}z}) \mid \tilde{B}(e^{-\tilde{\tau}z})) = n.$$

Proof. Let $\varepsilon > 0$. First apply Theorem 5.6 to $A(s)$, and choose a matrix $\tilde{A}(s) \in \mathbb{R}[s]^{n \times n}$ such that

(1) $\|A(s) - \tilde{A}(s)\|_{pm} < \frac{1}{2}\varepsilon$,

(2) $\deg(\tilde{A}(s)) \leq \max(\deg(A(s)), 2)$,

(3) $\forall z \in \overline{\mathbb{C}^+} : \; \operatorname{rank}(zI - \tilde{A}(e^{-\tau z})) \geq n - 1$.

According to Lemma 5.1, the function $\tilde{p}(z) = \det(zI - \tilde{A}(e^{-\tau z}))$ has only a finite number of zeros in $\overline{\mathbb{C}^+}$, say $\lambda_1, \ldots, \lambda_k$. Only in these points, $(zI - \tilde{A}(e^{-\tau z}))$ loses rank, but still $\operatorname{rank}(zI - \tilde{A}(e^{-\tau z})) = n - 1$. So the left-kernel of $(zI - \tilde{A}(e^{-\tau z}))$ is one-dimensional for all $z \in \{\lambda_1, \ldots, \lambda_k\}$. Choose vectors $v_1, \ldots, v_k$ of norm 1 in $\mathbb{C}^n$, spanning these left-kernels:

$$\forall i \in \{1, \ldots, k\} : \operatorname{span}(v_i) = \{x \in \mathbb{C}^n \mid x^{\mathrm{T}} \cdot (\lambda_i I - \tilde{A}(e^{-\tau\lambda_i})) = 0\}.$$

Denote for all $i \in \{1, \ldots, k\}$ the first column of $B(e^{-\tau\lambda_i})$ by b_i. According to Lemma 5.8, there exists a $\beta \in \mathbb{R}^n$ such that $\|\beta\| < \frac{1}{2}\varepsilon$ and $v_i^{\mathrm{T}} \cdot (b_i + \beta) \neq 0$ for all $i = 1, \ldots, k$.

Define $\tilde{B}(s)$ as the sum of $B(s)$ and the $n \times m$ matrix $(\beta \mid 0)$ consisting of the column β, completed with zeros:

$$\tilde{B}(s) := B(s) + (\beta \mid 0).$$

Then (i) and (ii) obviously hold, and we need to show only that $\tilde{\Sigma} = (\tilde{A}(s), \tilde{B}(s), \tau)$ satisfies (iii).

For this, let $z \in \overline{\mathbb{C}^+}$. If $z \notin \{\lambda_1, \ldots, \lambda_k\}$, then $\operatorname{rank}(zI - \tilde{A}(e^{-\tau z})) = n$, so certainly $\operatorname{rank}(zI - \tilde{A}(e^{-\tau z}) \mid \tilde{B}(e^{-\tau z})) = n$.

Otherwise, suppose that $z = \lambda_i$ for certain $i \in \{1, \ldots, k\}$. Let $x \in \mathbb{C}^n$ be such that

$$x^{\mathrm{T}} \cdot (\lambda_i I - \tilde{A}(e^{-\tau\lambda_i}) \mid \tilde{B}(e^{-\tau\lambda_i})) = 0. \tag{44}$$

Hence, x^{T} is an element of the left-kernel of $(\lambda_i I - \tilde{A}(e^{-\tau\lambda_i}))$, and there exists an $\alpha \in \mathbb{C}$ such that $x = \alpha \cdot v_i$. Now the first column of $\tilde{B}(e^{-\tau\lambda_i})$ is $b_i + \beta$, and

$$0 = x^{\mathrm{T}} \cdot (b_i + \beta) = \alpha v_i^{\mathrm{T}} \cdot (b_i + \beta) = \alpha \cdot [v_i^{\mathrm{T}} \cdot (b_i + \beta)].$$

We conclude that $\alpha = 0$. This completes the proof. □

From Theorem 5.9 it follows directly that the subset S of $\mathcal{V}$, consisting of all parametrizations of stabilizable time-delay systems, is a dense subset of $\mathcal{V}$. Note that the conditions on the degrees of $\tilde{A}(s)$ and $\tilde{B}(s)$ are essential. According to Theorem 5.9, it is possible to construct a sequence of time-delay systems $(\Sigma_i)_{i=1}^{\infty} = (A_i(s), B_i(s), \tau_i)_{i=1}^{\infty}$ converging to $\Sigma = (A(s), B(s), \tau)$ (in the sense of §2) with the property that

$$\forall i \in \mathbb{N} : \deg(A_i(s)) \leq \max(\deg(A(s)), 2) \text{ and } \deg(B_i(s)) = \deg(B(s)).$$

This means that to achieve stabilizability, we have to perturb only a finite number of all parameters describing the original system Σ. Construction of a sequence of stabilizable systems converging to Σ, but with an increasing degree in s, is of no use for our genericity result because this requires systems with time-delays of constantly increasing length. Since we can always obtain a stabilizable system using perturbations of an a priori given degree, the result of Theorem 5.9 also holds within the framework of so-called inductive limit topologies, mentioned at the end of §2.

At this stage, our conjecture on the genericity of stabilizability for time-delay systems is reduced to a simple corollary from Theorems 3.1 and 5.9.

THEOREM 5.10. *Time-delay systems of the form* (5) *are generically stabilizable in the following sense: the subset S of the parameter-space $\mathcal{V}$, consisting of all parametrizations $\Sigma = (A(s), B(s), \tau)$ of time-delay systems satisfying*

$$\forall z \in \overline{\mathbb{C}^+} : \operatorname{rank}(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) = n,$$

is an open and dense subset of the metric space $\mathcal{V}$.

6. Generalization to the case of incommensurable time-delays. In §§2-5, a derivation of our genericity result is given for systems with commensurable time-delays. This restriction was made only for notational convenience; the incommensurable delay case is not significantly more difficult. In this section we point out that with exactly the same arguments as before, the genericity result can also be proved for the more general class of systems with incommensurable time-delays.

In the algebraic terminology, systems with k incommensurable time-delays, given by $\tau_1, \ldots, \tau_k$, are modeled as systems over the ring $\mathbb{R}[s_1, \ldots, s_k]$, where the indeterminate s_i corresponds to the delay operator σ_i with time-delay τ_i. To apply a topological approach to our genericity problem, first a parameter-space $\mathcal{W}$ (the incommensurable version of $\mathcal{V}$) has to be introduced. Denoting $\mathbb{R}[s_1, \ldots, s_k]$ by $\mathcal{R}$, $\mathcal{W}$ is defined as

$$\mathcal{W} := \{\Sigma = (A, B, (\tau_1, \ldots, \tau_k)) \mid A \in \mathcal{R}^{n\times n},\ B \in \mathcal{R}^{n\times m},\ \tau_i \in \mathbb{R}^+ (i = 1, \ldots, k)\}.$$

In the same way as in the commensurable delay case, a matrix over $\mathbb{R}[s_1, \ldots, s_k]^{p\times q}$ can be seen as a k-dimensional sequence of $p \times q$ matrices over $\mathbb{R}$, with only a finite number of nonzero elements. So, application of an ℓ_1-norm is possible, and in this way Definition 2.1 may be generalized. In the same way, polynomials in more than two indeterminates can be treated.

With these generalized definitions of the norms, the results of §2 remain valid. Most of these results rely on the fact that for all $z \in \overline{\mathbb{C}^+}$: $|e^{-\tau z}| \leq 1$. Since all time-delays τ_i are strictly larger than zero, we still have

$$\forall i \in \{1, \ldots, k\}\ \forall \tau_i > 0\ \forall z \in \overline{\mathbb{C}^+} :\ |e^{-\tau_i z}| \leq 1, \tag{45}$$

and the same proofs may be applied. The only difficulty left is the result on the continuity of the map χ from a polynomial matrix to its characteristic polynomial. Here exponentials do not play a role, but for this result the number of indeterminates is not significant at all, and therefore it also holds in the incommensurable delay case.

The results of §3 are easily generalized, as far as perturbations of the matrices $A(s_1, \ldots, s_k)$ and $B(s_1, \ldots, s_k)$ are concerned. Perturbations of the lengths of the time-delays are more complicated. However, because of (45), all perturbations of time-delays can be treated successively. In each step $i (i = 1, \ldots, k)$, the exponentials $e^{-\tau_1 z}, \ldots, e^{-\tau_{i-1} z}$ and $e^{-\tau_{i+1} z}, \ldots, e^{-\tau_k z}$, corresponding to all the other time-delays except τ_i, are bounded above by 1 in absolute value because we assume that $z \in \overline{\mathbb{C}^+}$. Therefore exactly the same techniques as in formula (22) may be applied successively for each τ_i separately to arrive at the desired result.

Section 4 is already put in a general context, so here nothing has to be done. Note however that in Proposition 4.6 only one time-delay is required to achieve an appropriate perturbation of the matrix $Q(z)$.

In the first part of §5 we are now dealing with analytic functions of the form

$$p(z) = \det(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})).$$

The assumption on the absence of zeros on the imaginary axis can be removed in almost the same way as stated in Proposition 5.2. Trivially, Rouché's theorem is still valid, and it is also easily seen that all zeros of $p(z) = \det(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}))$ in $\overline{\mathbb{C}^+}$ are contained in a compact subset of $\overline{\mathbb{C}^+}$. Therefore, Lemma 5.4 still holds and the same process of successively reducing the order of the zeros to 1 can be used. Again, Rouché's theorem guarantees that the total number of zeros in $\overline{\mathbb{C}^+}$ remains constant and that simple zeros remain simple. Moreover, the results of §4 imply that the condition on the degree of $A(s_1, \ldots, s_k)$ is satisfied. Perturbations of the matrix $B(s)$ can be obtained in exactly the same way as described for the

commensurable delay case. Therefore Theorem 5.9 may also be generalized to delay systems with incommensurable time-delays.

Summarizing, we conclude that our genericity result for the stabilizability of time-delay systems with commensurable time-delays also holds in the incommensurable delay case. This final conclusion is stated in the last theorem.

THEOREM 6.1. *Time-delay systems with incommensurable time-delays of the form*

$$\dot{x}(t) = A(\sigma_1, \ldots, \sigma_k)x(t) + B(\sigma_1, \ldots, \sigma_k)u(t),$$

where σ_i $(i = 1, \ldots, k)$ *denotes the delay operator corresponding to a time-delay* τ_i, *are generically stabilizable in the following sense: the subset of the parameter-space*

$$\mathcal{W} = \{(A(s_1, \ldots, s_k), B(s_1, \ldots, s_k), (\tau_1, \ldots, \tau_k)) \mid A(s_1, \ldots, s_k) \in \mathbb{R}[s_1, \ldots, s_k]^{n\times n},$$
$$B(s_1, \ldots, s_k) \in \mathbb{R}[s_1, \ldots, s_k]^{n\times m} \text{ and } \forall i \in \{1, \ldots, k\} : \tau_i > 0\},$$

consisting of all parametrizations $\Sigma = (A(s_1, \ldots, s_k), B(s_1, \ldots, s_k), (\tau_1, \ldots, \tau_k))$ *of time-delay systems satisfying*

$$\forall z \in \overline{\mathbb{C}^+} : \operatorname{rank}(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) \mid B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})) = n,$$

is an open and dense subset of the metric space $\mathcal{W}$.

7. Conclusions. In this paper it was shown that time-delay systems with commensurable or incommensurable time-delays are generically stabilizable. First, an algebraic approach was used to model time-delay systems with point delays. For this class of systems, a topological framework was introduced to formalize the concept of genericity. In this setting it was shown that the set of stabilizable time-delay systems is an open and dense subset of the parameter-space describing all time-delay systems. This means that stabilizability is a robust property; it is preserved after small perturbations of the parameters. Moreover, a nonstabilizable time-delay system can be approximated arbitrarily close by a sequence of stabilizable time-delay systems. Therefore the property of stabilizability is very weak; it is generic in the sense described above.

Acknowledgments. I am indebted to Malo Hautus, Henri Huijberts, and Anton Stoorvogel for all their valuable suggestions and help during the writing of this paper. I would especially like to thank Stef van Eijndhoven for his guidance and his enthusiasm for my work. His ideas had a great influence on the contents of this paper.

REFERENCES

[1] J. B. CONWAY, *A Course in Functional Analysis*, Graduate Texts in Mathematics, 96, Springer-Verlag, New York, 1985.

[2] K. B. DATTA AND M. L. J. HAUTUS, *Decoupling of multivariable control systems over unique factorization domains*, SIAM J. Control Optim., 22 (1984), pp. 28-39.

[3] E. EMRE, *On necessary and sufficient conditions for regulation of linear systems over rings*, SIAM J. Control Optim., 20 (1982), pp. 155-160.

[4] E. EMRE AND G. J. KNOWLES, *Control of linear systems with fixed noncommensurate point delays*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 1083-1090.

[5] L. C. G. J. M. HABETS, *Stabilization of Time-Delay Systems: An Overview of the Algebraic Approach*, EUT Report 92-WSK-02, Eindhoven University of Technology, Eindhoven, the Netherlands, 1992.

[6] ———, *Algebraic and Computational Aspects of Time-Delay Systems*, Ph.D. thesis, Eindhoven University of Technology, Eindhoven, the Netherlands, 1994.

[7] J. HALE, *Theory of Functional Differential Equations*, Appl. Math. Sci., 3, Springer-Verlag, New York, 1977.

[8] M. L. J. HAUTUS, *Stabilization controllability and observability of linear autonomous systems*, Nederl. Akad. Wetensch. Proc. Ser. A, 73 (1970), pp. 448-455.

[9] E. W. KAMEN, *Lectures on Algebraic System Theory: Linear Systems Over Rings*, NASA Contractor Report 3016, 1978.
[10] E. KREYSZIG, *Introductory Functional Analysis with Applications*, Wiley, London, 1978.
[11] E. B. LEE AND A. W. OLBROT, *On reachability over polynomial rings and a related genericity problem*, Internat. J. Systems Sci., 13 (1982), pp. 109–113.
[12] L. PANDOLFI, *On feedback stabilization of functional differential equations*, Boll. Un. Mat. Ital., Ser IV, 11, suppl. fasc. 3 (1975), pp. 626–635.
[13] ———, *Controllability properties of perturbed distributed parameter systems*, Linear Algebra Appl., 122/123/124 (1989), pp. 525–538.
[14] Y. ROUCHALEAU, *Régulation statique et dynamique d'un système héréditaire*, in Analysis and Optimization of Systems, Proceedings of the Fifth International Conference on Analysis and Optimization of Systems, Versailles, December 14–17, 1982, Lecture Notes in Control and Inform. Sci., 44, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, Berlin, 1982, pp. 532–547.
[15] W. RUDIN, *Real and Complex Analysis*, 3rd ed., McGraw-Hill, London, 1987.
[16] E. D. SONTAG, *Linear systems over commutative rings: A survey*, Ricerche Automat., 7 (1976), pp. 1–34.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 855-873, May 1996

004

OPTIMAL POSITIONING OF ANODES FOR CATHODIC PROTECTION*

L. STEVEN HOU† AND WEIWEI SUN‡

Abstract. We consider an optimal control problem that arises in cathodic protection. The control is the position of a finite number of anodes, each with a prescribed current density. The goal is to obtain a desired potential distribution on the cathode that will prevent or reduce cathodic corrosion. The existence of an optimal solution is proven. The differentiability of the functional minimized is justified, and the equations satisfied by the derivative are established. Then a numerical algorithm is proposed for computing the optimal position of the anodes. Finally, we present some numerical results obtained by implementing the proposed numerical algorithm with a boundary element method.

Key words. optimal control, cathodic protection, nonlinear boundary value problem, gradient method

AMS subject classifications. 49J20, 49K20, 65K10, 65N30

1. Introduction.

This article is motivated by the desire to reduce corrosion (caused by chemical reactions) of ship propellers surrounded by sea water or of metal containers filled with an electrolyte. The corrosion process can be significantly slowed by maintaining a critical electrical potential on the portion of the structure surface to be protected. This portion of the surface to be protected usually acts as cathodes. The rest part of the surface is either insulated or anodes. The placement of a number of anodes on the structure surface could effectively change the potential distribution on the surface. Thus one could attempt to adjust the current density on the anodes and/or the location of the anodes to best match a desired potential distribution on the cathode. In light of this, mathematicians, scientists, and engineers have been trying to use optimal control theory and techniques to design cathodic protection systems. For instance, [13] contained a survey on this subject. Zamani and Chuang [12] studied a potential matching problem by controlling the electrical current density on the anode. The models used in [12], however, are essentially linear. Hou and Sun [7] discussed, mainly from an algorithmic point of view, several control problems, including adjusting the positions of anodes, in order to achieve a desired potential on the structure; nonlinear models as well as linear ones were employed. Amaya and Aoki [2] used discrete optimization techniques to handle optimal control problems with hybrid controls; i.e., several controls are applied simultaneously. The controls being used in the design of cathodic systems can be loosely divided into two categories. The first is "value controls," i.e., the current density on the anodes or the density of point charges. The second is "location controls," i.e., the location of anodes and/or location of point charges. In this article, we will attempt to mathematically analyze optimal control problems with "location controls" and devise some numerical algorithms that can aid the design of a cathodic control system. The theories presented in this paper can be applied to a variety of cathodic protection problems. Propeller and container protection serves as two concrete examples.

We assume the "corrosive fluid" occupies a physical domain Ω with a boundary Γ. The domain can be finite as in the case of a metal tank containing an electrolyte (see Figure 1) or infinite (but with a bounded boundary) as in the case of a ship body surrounded by sea water (see Figure 2). The "corrosive fluid" in the former case is the electrolyte and in the latter is the sea water. The electrical potential ϕ in Ω is governed by the differential equation

*Received by the editors January 9, 1993; accepted for publication (in revised form) December 28, 1994.

†Department of Mathematics and Statistics, York University, North York, ON M3J 1P3, Canada. The research of this author was supported by Natural Science and Engineering Research Council of Canada grant OGP-0137436 (hou@mathstat.yorku.ca).

‡Department of Mathematics and Statistics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada. Current address: Centre for Mathematics and Its Applications, School of Mathematical Sciences, Australian National University, Canberra, ACT 0200, Australia (weiwei@maths.anu.edu.au). The research of this author was supported in part by the Natural Science and Engineering Research Council of Canada.

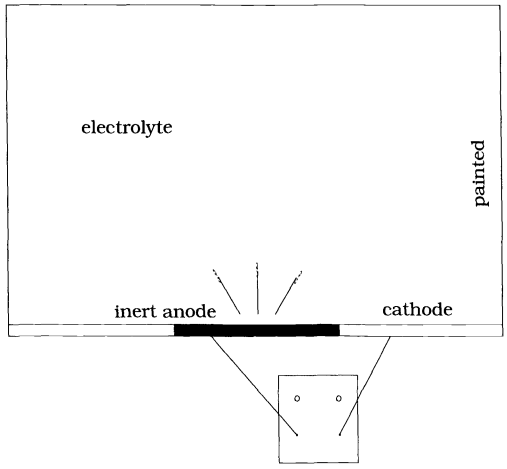


FIG. 1. *An interior domain example: an electrolyte container.*

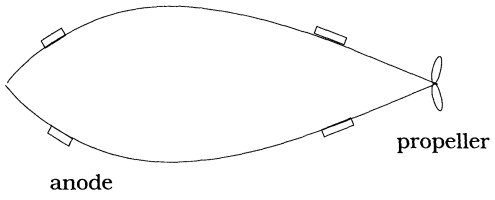


FIG. 2. *An exterior domain example: a ship surrounded by sea water.*

$$-\operatorname{div}(\sigma \operatorname{grad}\phi) = 0 \quad \text{in } \Omega,$$

where the conductivity σ has a positive lower bound.

For simplicity of exposition, we will deal only with interior problems. Since we will use boundary element methods for computations, the treatments for infinite domains and finite domains are essentially the same (see [7], [11] for details). For general theories and applications of boundary element methods, see [4].

The boundary of the domain Ω is divided into three parts: namely, the cathode Γ_C (the surface to be protected), the anode $\cup_{i=1}^{N}\Gamma_i$ (on which electrical current or potential is given), and Γ_0 (the insulated part).

On Γ_C, ϕ satisfies the relation

$$\sigma\frac{\partial\phi}{\partial n} = -f(\phi) \quad \text{on } \Gamma_C,$$

where f is an empirical function that depends on the "corrosive fluid" and the cathode materials (see [3]).

We assume that we have at our disposal a number of given electrical current sources (anodes) that are to be connected to the boundary of Ω, but the location of the anodes can be adjusted, i.e., the location of Γ_i's, acts as control variables. Our task is to determine the optimal location for Γ_i's such that the resulting potential distribution on Γ_C best matches some desired one. On each Γ_i, we have the boundary condition

$$\sigma\frac{\partial\phi}{\partial n} = u_i \quad \text{on } \Gamma_i,$$

which corresponds to the specification of the current density. This is the case for impressed cathodic controls [7], [12] and is the case we will treat in this paper. In some other cases, other types of boundary conditions may be specified instead of the current density, e.g., in the case of a nonpolarizable electrode [13].

A mathematical description proceeds as follows. Let Ω be a bounded open domain in $\mathbb{R}^2$ and Γ be the boundary of Ω. Assume $\Gamma = [\cup_{i=1}^N \Gamma_i] \cup \Gamma_C \cup \Gamma_0$. We assume that all boundary segments are closed and they may intersect with each other only at the end points. We assume that the boundary Γ is parameterized by the arc length coordinate $s \in [0, L]$ and each Γ_i has a fixed arc length $d_i > 0$. Since each point on Γ is uniquely determined by its arc length coordinate s, we will use s to denote, without confusion from the context, both the arc length coordinate for the point and the point itself. Thus each Γ_i is determined by the arc length coordinate t_i of its center point. The location of all the Γ_i's is the control variable, i.e., the center position t_i's constitute the control variable. Each t_i is constrained to range on a finite interval $[a_i, b_i]$. Since the partition of Γ depends on $\mathbf{t} = (t_1, t_2, \ldots, t_N)$, we rewrite $\Gamma = [\cup_{i=1}^N \Gamma_i(\mathbf{t})] \cup \Gamma_C \cup \Gamma_0(\mathbf{t})$. Each segment $\Gamma_i(\mathbf{t})$ is centered at t_i with a fixed arc length d_i. Γ_C is fixed (independent of $\mathbf{t}$) with a positive measure. We assume that each $\Gamma_i(\mathbf{t})$ is parameterized by $s \in [t_i - \frac{d_i}{2}, t_i + \frac{d_i}{2}]$. $\Gamma_0(\mathbf{t}) = \Gamma \setminus \left[\Gamma_C \cup (\cup_{i=1}^N \Gamma_i(\mathbf{t}))\right]$. We assume that $\frac{d_1}{2} \le a_1 < b_1 < a_2 < b_2 < \cdots < a_N < b_N < (L - \frac{d_N}{2})$ and $b_i \le a_{i+1} - \frac{d_i}{2} - \frac{d_{i+1}}{2}$ for $i = 1, \ldots, N-1$ and that $a_{i+1} - b_i \ge \frac{d_i + d_{i+1}}{2}$ for $i = 1, \ldots, N-1$. Recall that each t_i is the center point of the ith anode and the width of the ith anode $d_i > 0$; thus under these assumptions, the situation of two anodes joining together is permitted, but overlapping of two anodes is disallowed. We assume that each u_i is a given function in $C(\mathbb{R})$ with a compact support on the interval $(-\frac{d_i}{2}, \frac{d_i}{2})$. Thus the function $s \mapsto u_i(s - t_i)$ for $s \in [t_i - \frac{d_i}{2}, t_i + \frac{d_i}{2}]$ has a compact support on $\Gamma_i(\mathbf{t})$.

We are concerned with the following optimal control problem: seek a state ϕ and a $\mathbf{t} \in \mathbb{R}^N$ with components $t_i \in [a_i, b_i]$, $i = 1, \ldots, N$, such that the functional

$$\mathcal{J}(\phi, \mathbf{t}) = \frac{1}{2} \int_{\Gamma_C} (\phi - \phi_0)^2 \, d\Gamma \tag{1.1}$$

is minimized subject to the the constraint equations

$$-\operatorname{div}(\sigma \operatorname{grad} \phi) = 0 \quad \text{in } \Omega, \tag{1.2}$$

$$\sigma \frac{\partial \phi}{\partial n} = u_i(s - t_i) \quad \forall\, s \in \left[t_i - \frac{d_i}{2}, t_i + \frac{d_i}{2}\right] \text{ on } \Gamma_i(\mathbf{t}), \quad i = 1, \ldots, N, \tag{1.3}$$

$$\sigma \frac{\partial \phi}{\partial n} = 0 \quad \text{on } \Gamma_0(\mathbf{t}), \tag{1.4}$$

and

$$\sigma \frac{\partial \phi}{\partial n} = -f(\phi) \quad \text{on } \Gamma_C. \tag{1.5}$$

Here ϕ_0 is a desired potential distribution on Γ_C under which the corrosion rate is minimized. The function $f \in C^1(\mathbb{R})$ is given by either

$$f(\phi) = k\phi + l\phi^3, \tag{1.6}$$

where k and l are constants and $k > 0$, $l \ge 0$ (see [7]), or

$$f(\phi) = C_3[e^{C_1\phi} - e^{-C_2\phi}], \tag{1.7}$$

where C_1, C_2, and C_3 are positive constants (see [3]). If $l = 0$ in (1.6), the problem reduces to a linear one. For f given by either (1.6) or (1.7), one can easily verify that there exists an $\alpha > 0$ such that

$$f'(\phi) \geq \alpha \quad \forall\, \phi \qquad \text{and} \qquad f(0) = 0. \tag{1.8}$$

Note that the boundary conditions (1.3)-(1.5) can be written into a unified form

$$\sigma \frac{\partial \phi}{\partial n} = -\tilde{f}(\phi) + \tilde{u}(\cdot, \mathbf{t})$$

with appropriately defined $\tilde{f}$ and $\tilde{u}$.

We will use a variational weak formulation (1.9) of the nonlinear boundary value problem (1.2)-(1.5). We will utilize Sobolev spaces $H^m(\Omega)$, $H^s(\Gamma_i)$, $H^s(\Gamma_C)$, $H^s(\Gamma_0)$, and $H^s(\Gamma)$. The norms on these spaces are denoted by, e.g., $\|\cdot\|_m$, $\|\cdot\|_{s,\Gamma_i}$, and so on. For details, see [1] and [5]. We restate the minimization problem as follows: seek a state $\phi \in H^1(\Omega)$ and a $\mathbf{t} \in \Pi_{i=1}^N [a_i, b_i]$ such that the functional (1.1) is minimized subject to the the constraint equation

$$\begin{aligned} &\int_\Omega \sigma \operatorname{grad}\phi \cdot \operatorname{grad}\psi \, d\Omega + \int_{\Gamma_C} f(\phi)\psi \, d\Gamma \\ &\qquad = \sum_{i=1}^N \int_{\Gamma_i(\mathbf{t})} u_i(s - t_i)\psi \, ds \quad \forall\, \psi \in H^1(\Omega). \end{aligned} \tag{1.9}$$

The nonlinear boundary value problem (1.2)-(1.5) is understood in the sense of (1.9). Of course when the solution is smooth enough, (1.2)-(1.5) can hold in the classical sense.

Now we state a useful fact whose proof can be found in, e.g., [9]. The norm on $H^1(\Omega)$ defined by

$$|||\phi|||_1 = \left\{ \int_\Omega \sigma |\operatorname{grad}\phi|^2 \, d\Omega + \alpha \int_{\Gamma_C} \phi^2 \, d\Gamma \right\}^{1/2} \quad \forall\, \phi \in H^1(\Omega)$$

is equivalent to the usual $H^1(\Omega)$-norm $\|\cdot\|_1$; i.e., there exist constants $\rho > 0$ and $\gamma > 0$ such that

$$\rho \|\phi\|_1^2 \geq \int_\Omega \sigma |\operatorname{grad}\phi|^2 \, d\Omega + \alpha \int_{\Gamma_C} \phi^2 \, d\Gamma \geq \gamma \|\phi\|_1^2 \quad \forall\, \phi \in H^1(\Omega). \tag{1.10}$$

The rest of the paper is organized as follows. In §2 we prove the existence and uniqueness of solutions to (1.9) so that the constraint equation is well posed. In §3 we show the existence of an optimal pair $(\hat{\mathbf{t}}, \hat{\phi})$ that minimizes (1.1) subject to (1.9). In §4 we establish the differentiability of the constraint minimization problem so that a first-order optimality condition is derived. In §5 we propose an algorithm for computing optimal solutions using gradient methods. Finally in §6 we report some numerical results that demonstrate the effectiveness of our theory and algorithm.

2. Existence and uniqueness of solutions to the constraint equations. We first examine the existence of a solution to (1.9), which is a partial differential equation with nonlinear boundary conditions. In this section we assume that $\mathbf{t}$ is given; thus the boundary components Γ_i, $i = 1, \ldots, N$, are fixed. We will show that for f given by either (1.6) or (1.7), equation (1.9) possesses a unique solution $\phi \in H^1(\Omega)$. We point out that [10] studied general nonlinear boundary value problems using a boundary integral method. It is possible to generalize the

ideas of boundary integral methods in [10] to study (1.2)-(1.5) (then, equivalently, (1.9)). For the case in which f is given by (1.7), existence and uniqueness results for (1.9) were established in [8] by a variational method for an energy functional. For completeness we still give a proof of existence and uniqueness, but the approach is different from that of [8] and [10]. Also, although we will restrict our attention to the physical boundary conditions (1.6) and (1.7), it is possible to deal with more general nonlinear boundary conditions with our methods.

LEMMA 2.1. *X is a finite-dimensional Hilbert space whose scalar product is denoted by $(\cdot,\cdot)$ and the corresponding norm by $|\cdot|$. Let F be a continuous mapping form X into X with the following property: there exists an $r > 0$ such that*

$$(F(\phi),\phi) \geq 0 \quad \forall\, \phi \in X \text{ with } |\phi| = r\,.$$

Then there exists a $\phi \in X$ such that

$$F(\phi) = 0 \quad \text{and} \quad |\phi| \leq r\,.$$

Proof. See [5, p. 279]. □

LEMMA 2.2. *Assume $\phi \in H^1(\Omega)$ and $s > 0$. Then $e^{s|\phi|} \in L^1(\Gamma)$. Moreover, there exists a constant κ, independent of ϕ, such that*

$$\int_\Gamma e^{s|\phi|}\,d\Gamma \leq 1 + |\Gamma| + e^{s^2\kappa^2\|\phi\|_1^2}\,|\Gamma| < \infty\,,$$

where $|\Gamma|$ is the measure of Γ.

Proof. The proof is based on an imbedding theorem for Sobolev-Orlicz spaces [1]. See [8] for details. □

LEMMA 2.3. *Assume $\{\phi_n\} \subset L^2(\Gamma_C)$ is a sequence such that $\phi_n \to \phi$ almost everywhere on Γ_C and*

$$\int_{\Gamma_C} f(\phi_n)\phi_n\,d\Gamma \leq B \quad \forall\, n, \tag{2.1}$$

where f is defined by (1.1) and $B > 0$ is a constant independent of n. Then

$$\int_{\Gamma_C} f(\phi)\phi\,d\Gamma \leq \liminf_{n\to\infty} \int_{\Gamma_C} f(\phi_n)\phi_n\,d\Gamma$$

and

$$\lim_{n\to\infty} \int_{\Gamma_C} |f(\phi_n) - f(\phi)|\,d\Gamma = 0\,.$$

Proof. See [8] or [6]. □

THEOREM 2.4. *Assume each $u_i \in C(\Gamma_i)$, $i = 1,\ldots,N$. Then there exists a unique $\phi \in H^1(\Omega)$ that satisfies (1.9). Furthermore, $\phi \in C^2(\Omega) \cap C(\overline{\Omega})$.*

Proof. We first show the existence of a $\phi \in H^1(\Omega)$ that satisfies (1.9). Using (1.8) and the mean value theorem we have

$$\int_{\Gamma_C} f(\phi)\phi\,d\Gamma = \int_{\Gamma_C} [f(\phi) - f(0)]\phi\,d\Gamma = \int_{\Gamma_C} [f'(\tilde{\phi})]\phi^2\,d\Gamma \geq \alpha \int_{\Gamma_C} \phi^2\,d\Gamma,$$

where $\tilde{\phi}$ is between 0 and ϕ. It follows from the last relation and (1.10) that

$$\int_\Omega \sigma\,\operatorname{grad}\phi \cdot \operatorname{grad}\phi\,d\Omega + \int_{\Gamma_C} f(\phi)\phi\,d\Gamma \geq \gamma\|\phi\|_1^2 \quad \forall\, \phi \in H^1(\Omega)\,.$$

Using the Cauchy-Schwarz inequality and trace theorems we obtain the estimate

$$\left|\sum_{i=1}^{N}\int_{\Gamma_i} u_i(s-t_i)\phi\,ds\right| \le \beta\|\phi\|_1 \sum_{i=1}^{N}\|u_i(\cdot-t_i)\|_{0,\Gamma_i} \quad \forall\,\phi\in H^1(\Omega),$$

where β is a positive constant. By combining the last two estimates we deduce that for $r=\frac{\beta}{\gamma}\left\{\sum_{i=1}^{N}\|u_i(\cdot-t_i)\|_{0,\Gamma_i}+1\right\}>0$ we have

$$\int_\Omega \sigma\,\operatorname{grad}\phi\cdot\operatorname{grad}\phi\,d\Omega+\int_{\Gamma_C} f(\phi)\phi\,d\Gamma-\sum_{i=1}^{N}\int_{\Gamma_i} u_i(s-t_i)\phi\,ds \ge 0 \quad \forall\,\phi\in H^1(\Omega) \text{ with } \|\phi\|_1=r. \tag{2.2}$$

Since $H^1(\Omega)$ is separable, we choose a countable basis of $H^1(\Omega)$: $\{\psi_i\}_{i=1}^{\infty}$. We set $X_n=\operatorname{span}\{\psi_1,\dots,\psi_n\}$. The inner product and norm on each X_n are defined by that of $H^1(\Omega)$ restricted to X_n. We introduce the mapping $F_n: X_n\to X_n$ defined by

$$(F_n(\phi),\psi_j)=\int_\Omega \sigma\,\operatorname{grad}\phi\cdot\operatorname{grad}\psi_j\,d\Omega+\int_{\Gamma_C} f(\phi)\psi_j\,d\Gamma - \sum_{i=1}^{N}\int_{\Gamma_i} u_i(s-t_i)\psi_j\,ds, \quad 1\le j\le n.$$

It follows from (2.2) and Lemma 2.1 that the finite-dimensional problem

$$\int_\Omega \operatorname{grad}\phi_n\cdot\operatorname{grad}\psi\,d\Omega+\int_{\Gamma_C} f(\phi_n)\psi\,d\Gamma=\sum_{i=1}^{N}\int_{\Gamma_i} u_i(s-t_i)\psi\,ds \quad \forall\,\psi\in X_n \tag{2.3}$$

has a solution $\phi_n\in X_n$ with a bound

$$\|\phi_n\|_1\le\frac{\beta}{\gamma}\left\{\sum_{i=1}^{N}\|u_i(\cdot-t_i)\|_{0,\Gamma_i}+1\right\}.$$

We can extract a subsequence of $\{\phi_n\}$, still denoted by $\{\phi_n\}$, that converges weakly to some $\phi\in H^1(\Omega)$ as $n\to 0$. Then $\{\phi_n\}$ also converges weakly in $H^{1/2}(\Gamma)$, using a trace theorem. By compact imbedding results, $\{\phi_n\}$ converges strongly in $L^3(\Gamma)$, and by extracting a further subsequence, $\{\phi_n\}$ converges pointwise almost everywhere. Now we examine the two cases (1.6) and (1.7) separately. For f given by (1.6), using the strong convergence of $\{\phi_n\}$ in $L^3(\Gamma)$, the obvious growth condition

$$|f(\phi)|\le C(1+|\phi|^3) \quad \forall\,\phi$$

and the well-known continuity properties of the Němyckiĭ operator on Carathéodory functions, we may pass to the limit in (2.3) to show that

$$\int_\Omega \sigma\,\operatorname{grad}\phi\cdot\operatorname{grad}\psi\,d\Omega+\int_{\Gamma_C} f(\phi)\psi\,d\Gamma=\sum_{i=1}^{N}\int_{\Gamma_i} u_i(s-t_i)\psi\,ds \quad \forall\,\psi\in C^1(\overline{\Omega}).$$

Then, using the denseness of $C^1(\overline{\Omega})$ in $H^1(\Omega)$ and the fact that $\phi|_{\Gamma_C}\in L^4(\Gamma_C)$, we conclude ϕ satisfies (1.9). For f given by (1.7) we see that by setting $\psi=\phi_n$ in (2.3) we obtain

$$\begin{aligned}\int_{\Gamma_C} f(\phi_n)\phi_n\,d\Gamma &\le \|u\|_{0,\Gamma_A}\|\phi_n\|_{0,\Gamma_A}\\ &\le C\|u\|_{0,\Gamma_A}\|\phi_n\|_1\le C\,\|u\|_{0,\Gamma_A}\frac{\beta}{\gamma}\left\{\|u\|_{0,\Gamma_A}+1\right\}.\end{aligned}$$

By Lemma 2.3 we have

$$\int_{\Gamma_C} f(\phi)\phi \, d\Gamma \leq \liminf_{n\to\infty} \int_{\Gamma_C} f(\phi_n)\phi_n \, d\Gamma$$

and

$$\lim_{n\to\infty} \int_{\Gamma_C} |f(\phi_n) - f(\phi)| \, d\Gamma = 0 .$$

Thus for each $\psi \in C^1(\overline{\Omega})$ we may pass to the limit in (2.3) and obtain

$$\int_\Omega \sigma \operatorname{grad}\phi \cdot \operatorname{grad}\psi \, d\Omega + \int_{\Gamma_C} f(\phi)\psi \, d\Gamma = \int_{\Gamma_A} u\psi \, d\Gamma \quad \forall\, \psi \in C^1(\overline{\Omega}) .$$

Using the denseness of $C^1(\overline{\Omega})$ in $H^1(\Omega)$ and the fact that $f(\phi) \in L^2(\Gamma_C)$ (which follows from Lemma 2.2), we conclude ϕ satisfies (1.9).

We briefly examine the regularity of the solution $\phi \in H^1(\Omega)$. If f is given by (1.6), then trace theorems on $\Omega \subset \mathbb{R}^2$ imply that $f(\phi) \in L^r(\Gamma)$ for all $r > 1$. If f is given by (1.7), then Lemma 2.2 implies that $f(\phi) \in L^r(\Gamma)$ for all $r > 1$. Employing elliptic regularity theories, we see that $\phi \in W^{3/2,r}(\Omega)$ for all $r > 1$, which in turn implies $\phi|_\Gamma \in W^{1,r}(\Gamma)$ so that $\phi \in C(\overline{\Omega})$. Interior regularity results for the Laplacian equation implies $\phi \in C^2(\Omega)$. Thus we have shown that there exists a solution $\phi \in H^1(\Omega) \cap C^2(\Omega) \cap C(\overline{\Omega})$.

To answer the question of uniqueness, we assume that ϕ and $\bar{\phi}$ are two solutions to (1.9). Then we have

$$\int_\Omega \sigma \operatorname{grad}(\phi - \bar{\phi}) \cdot \operatorname{grad}\psi \, d\Omega + \int_{\Gamma_C} [f(\phi) - f(\bar{\phi})]\psi \, d\Gamma = 0 \quad \forall\, \psi \in H^1(\Omega) .$$

Setting $\psi = \phi - \bar{\phi}$, we see that

$$\int_\Omega \sigma |\operatorname{grad}(\phi - \bar{\phi})|^2 \, d\Omega + \int_{\Gamma_C} [f(\phi) - f(\bar{\phi})](\phi - \bar{\phi}) \, d\Gamma = 0 .$$

Using the mean value theorem,

$$\int_\Omega \sigma |\operatorname{grad}(\phi - \bar{\phi})|^2 \, d\Omega + \int_{\Gamma_C} f'(\tilde{\phi})(\phi - \bar{\phi})^2 \, d\Gamma = 0$$

for some $\tilde{\phi}$ between ϕ and $\bar{\phi}$. Using (1.8), we see that

$$\int_\Omega \sigma |\operatorname{grad}(\phi - \bar{\phi})|^2 \, d\Omega + \alpha \int_{\Gamma_C} (\phi - \bar{\phi})^2 \, d\Gamma \leq 0 .$$

Hence we deduce that $\operatorname{grad}(\phi - \bar{\phi}) = 0$ in Ω and $(\phi - \bar{\phi}) = 0$ on Γ_C. This in turn implies $(\phi - \bar{\phi}) = 0$ in Ω; i.e., uniqueness holds. □

3. Existence of an optimal solution. We have shown that for each fixed $\mathbf{t}$, (1.9) has a unique solution, which will be denoted by $\phi = \phi(\mathbf{t})$ in subsequent discussion. We are now prepared to study the existence of an optimal $\mathbf{t}$ that minimizes the functional (1.1) subject to (1.9).

THEOREM 3.1. *Assume $u_i \in C^1(\mathbb{R})$ with a compact support on Γ_i, $i = 1, \ldots, N$. Then there exists a $(\hat{\mathbf{t}}, \hat{\phi}) \in \left(\prod_{i=1}^N [a_i, b_i]\right) \times H^1(\Omega)$ that minimizes* (1.1) *subject to* (1.9).

Proof. We first prove that the mapping $\mathbf{t} \mapsto \phi(\mathbf{t})$, where $\phi(\mathbf{t})$ is the solution of (1.9), is continuous from $\prod_{i=1}^{N}[a_i, b_i]$ into $H^1(\Omega)$. Let $\mathbf{t} \in \prod_{i=1}^{N}[a_i, b_i]$ be given. For each $\delta\mathbf{t} \in \mathbb{R}^N$ we introduce

$$\phi_{\delta\mathbf{t}} = \phi(\mathbf{t} + \delta\mathbf{t}) - \phi(\mathbf{t}).$$

Then $\phi_{\delta\mathbf{t}} \in H^1(\Omega)$ satisfies

$$\int_\Omega \operatorname{grad}\phi_{\delta\mathbf{t}} \cdot \operatorname{grad} v \, d\Omega + \int_{\Gamma_C} [f(\phi(\mathbf{t}+\delta\mathbf{t})) - f(\phi(\mathbf{t}))]\, v \, d\Gamma$$
$$= \sum_{i=1}^{N} \int_{\Gamma_i(\mathbf{t})} [u_i(s - t_i - \delta t_i) - u_i(s - t_i)]\, v \, ds \quad \forall\, v \in H^1(\Omega).$$

Here we have used the fact that $u_i(\cdot)$ has a compact support on $\left(-\frac{d_i}{2}, \frac{d_i}{2}\right)$, and thus for $|\delta\mathbf{t}|$ small enough, the support of $u_i(s - t_i - \delta t_i)$ for $s \in \left(t_i + \delta t_i - \frac{d_i}{2}, t_i + \delta t_i + \frac{d_i}{2}\right)$ lies on $\Gamma_i(\mathbf{t})$. Using the mean value theorem we have

$$\int_\Omega \operatorname{grad}\phi_{\delta\mathbf{t}} \cdot \operatorname{grad} v \, d\Omega + \int_{\Gamma_C} f'(\tilde{\phi})\phi_{\delta\mathbf{t}} v \, d\Gamma$$
$$= \sum_{i=1}^{N} \int_{\Gamma_i(\mathbf{t})} [u_i(s - t_i - \delta t_i) - u_i(s - t_i)]\, v \, ds \quad \forall\, v \in H^1(\Omega).$$

By setting $v = \phi_{\delta\mathbf{t}}$ and using (1.8), (1.10), and a trace theorem, we see that

$$\gamma \|\phi_{\delta\mathbf{t}}\|_1^2 \le \|\phi_{\delta\mathbf{t}}\|_{0,\Gamma_C} \sum_{i=1}^{N} \|u_i(\cdot - t_i - \delta t_i) - u_i(\cdot - t_i)\|_{0,\Gamma_i(\mathbf{t})}$$
$$\le C\, \|\phi_{\delta\mathbf{t}}\|_1 \sum_{i=1}^{N} \|u_i(\cdot - t_i - \delta t_i) - u_i(\cdot - t_i)\|_{0,\Gamma_i(\mathbf{t})},$$

so

$$\|\phi_{\delta\mathbf{t}}\|_1 \le C \sum_{i=1}^{N} \|u_i(\cdot - t_i - \delta t_i) - u_i(\cdot - t_i)\|_{0,\Gamma_i(\mathbf{t})}.$$

Since each u_i is continuous, the right-hand side clearly goes to zero as $|\delta\mathbf{t}| \to 0$. Hence,

$$\|\phi_{\delta\mathbf{t}}\|_1 \to 0 \quad \text{as } |\delta\mathbf{t}| \to 0;$$

i.e., we have shown that the mapping $\mathbf{t} \mapsto \phi(\mathbf{t})$ is continuous from $\prod_{i=1}^{N}[a_i, b_i] \to H^1(\Omega)$. This in turn shows that $\mathcal{J}(\phi(\mathbf{t}), \mathbf{t})$ is continuous on the bounded, closed set $\prod_{i=1}^{N}[a_i, b_i]$ so that $\mathcal{J}(\phi(\mathbf{t}), \mathbf{t})$ has a minimum, attained at, say, $\hat{\mathbf{t}}$. Of course $(\hat{\mathbf{t}}, \phi(\hat{\mathbf{t}}))$ minimizes (1.1) subject to (1.9). The proof is completed by setting $\hat{\phi} = \phi(\hat{\mathbf{t}})$. □

4. Differentiability and first-order optimality conditions. In this section, we will attempt to characterize the optimal solution. The main result can be simply stated as follows: the minimum is attained when each $\hat{t}_i$ is either on the boundary of $[a_i, b_i]$ or in the interior of (a_i, b_i) with a vanishing ith partial derivative.

THEOREM 4.1. *Assume $u_i \in C^1(\mathbb{R})$ with a compact support in $\Gamma_i(\mathbf{t})$, $i = 1, \ldots, N$. Then the mapping $\mathbf{t} \mapsto \phi(\mathbf{t})$ is differentiable for $\mathbf{t} \in \prod_{i=1}^{N}(a_i, b_i)$. Furthermore, let $\lambda_i \in H^1(\Omega)$ be*

the solution of the following equation:

$$\int_\Omega \sigma \operatorname{grad}\lambda_i \cdot \operatorname{grad}\psi \, d\Omega + \int_{\Gamma_C} f'(\phi)\lambda_i \psi \, d\Gamma = -\int_{\Gamma_i(\mathbf{t})} u_i'(s - t_i)\psi \, ds \quad \forall\, \psi \in H^1(\Omega)\,. \tag{4.1}$$

Then $\lambda_i = \frac{\partial \phi(\mathbf{t})}{\partial t_i}$.

Proof. Let a $\mathbf{t} \in \prod_{i=1}^N (a_i, b_i)$ be given. Since $\phi(\cdot)$ is continuous on $\prod_{i=1}^N [a_i, b_i]$ and f is C^1, $f'(\phi(\mathbf{t}))$ is continuous on Γ_C. Equations (1.8) and (1.10) imply that (4.1) is a well-posed mixed Neumann–Robin-type boundary value problem, and there exists a unique $\lambda_i = \lambda_i(\mathbf{t}) \in H^1(\Omega)$ satisfying (4.1) and

$$\|\lambda_i\|_1 \le C(\phi, u_1, \ldots, u_N)\,, \tag{4.2}$$

where $C(\phi, u_1, \ldots, u_N)$ is a constant depending on $(\phi, u_1, \ldots, u_N)$.

For each nonzero $\epsilon \in \mathbb{R}$, we introduce

$$\psi_\epsilon = \frac{\phi(\mathbf{t} + \epsilon \mathbf{e}_i) - \phi(\mathbf{t})}{\epsilon} - \lambda_i\,,$$

where $\mathbf{e}_i$ is the ith standard basis vector in $\mathbb{R}^N$. Using the fact that $u_i(\cdot)$ has a compact support on $\left(-\frac{d_i}{2}, \frac{d_i}{2}\right)$ we see that for ϵ small enough, the support of $u_i(s - t_i - \epsilon)$ for $s \in \left(t_i + \epsilon - \frac{d_i}{2}, t_i + \epsilon + \frac{d_i}{2}\right)$ lies on $\Gamma_i(\mathbf{t})$. Thus the defining equations for $\phi(\mathbf{t} + \epsilon \mathbf{e}_i)$ and $\phi(\mathbf{t})$ are given by, respectively,

$$\int_\Omega \sigma \operatorname{grad}\phi(\mathbf{t} + \epsilon \mathbf{e}_i) \cdot \operatorname{grad} v \, d\Omega + \int_{\Gamma_C} f(\phi(\mathbf{t} + \epsilon \mathbf{e}_i)) v \, d\Gamma = \sum_{j=1}^N \int_{\Gamma_j(\mathbf{t})} u_j(s - t_j - \epsilon \delta_j^i)\, v \, ds \quad \forall\, v \in H^1(\Omega)$$

and

$$\int_\Omega \sigma \operatorname{grad}\phi(\mathbf{t}) \cdot \operatorname{grad} v \, d\Omega + \int_{\Gamma_C} f(\phi(\mathbf{t})) v \, d\Gamma = \sum_{j=1}^N \int_{\Gamma_j(\mathbf{t})} u_j(s - t_j)\, v \, ds \quad \forall\, v \in H^1(\Omega)\,.$$

(δ_j^i is the Kronecker delta.) Now we subtract the last two equations, divide by ϵ, and then subtract (4.1). The resulting equation reads as follows:

$$\int_\Omega \sigma \operatorname{grad}\psi_\epsilon \cdot \operatorname{grad} v \, d\Omega + \int_{\Gamma_C} \left[\frac{f(\phi(\mathbf{t}+\epsilon\mathbf{e}_i)) - f(\phi(\mathbf{t}))}{\epsilon} - f'(\phi(\mathbf{t}))\lambda_i\right] v \, d\Gamma = \int_{\Gamma_i(\mathbf{t})} \left[u_i'(s - t_i) - \frac{u_i(s - t_i - \epsilon) - u_i(s - t_i)}{-\epsilon}\right] v \, ds\,. \tag{4.3}$$

Note that

$$\frac{f(\phi(\mathbf{t}+\epsilon\mathbf{e}_i)) - f(\phi(\mathbf{t}))}{\epsilon} = \int_0^1 f'\big((1-s)\phi(\mathbf{t}) + s\phi(\mathbf{t}+\epsilon\mathbf{e}_i)\big)\, ds \, \frac{\phi(\mathbf{t} + \epsilon \mathbf{e}_i) - \phi(\mathbf{t})}{\epsilon}$$

and

$$f'(\phi(\mathbf{t})) - \int_0^1 f'\big((1-s)\phi(\mathbf{t}) + s\phi(\mathbf{t}+\epsilon\mathbf{e}_i)\big)\,ds$$

$$= \int_0^1 \int_0^1 f''\big(\phi(\mathbf{t}) + s(1-r)(\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t}))\big)\,ds\,dr\,[\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t})];$$

thus by adding and subtracting terms in (4.3) we have

$$\begin{aligned}
&\int_\Omega \sigma\,\mathrm{grad}\,\psi_\epsilon \cdot \mathrm{grad}\,v\,d\Omega + \int_{\Gamma_C}\int_0^1 f'\big((1-s)\phi(\mathbf{t}) + s\phi(\mathbf{t}+\epsilon\mathbf{e}_i)\big)\,ds\,\psi_\epsilon v\,d\Gamma \\
&= \int_{\Gamma_i(\mathbf{t})}\left[u_i'(s-t_i) - \frac{u_i(s-t_i-\epsilon) - u_i(s-t_i)}{-\epsilon}\right] v\,ds \\
&\quad + \int_{\Gamma_C}\int_0^1\int_0^1 f''\big(\phi(\mathbf{t}) + s(1-r)(\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t}))\big)\,ds\,dr\,[\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t})]\,\lambda_i\,v\,d\Gamma.
\end{aligned} \tag{4.4}$$

If f is given by (1.6), then $f''(\phi) = 6\phi$ so that

$$\begin{aligned}
&\left\{\int_{\Gamma_C}\int_0^1\int_0^1 s\left|f''\big(\phi(\mathbf{t}) + s(1-r)(\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t}))\big)\,ds\,dr\right|^4 d\Gamma\right\}^{1/4} \\
&\qquad \le C\,\|\phi(\mathbf{t}) + s(1-r)(\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t}))\|_1 \\
&\qquad \le C\,\{\|\phi(\mathbf{t})\|_1 + \|\phi(\mathbf{t}+\epsilon\mathbf{e}_i)\|_1\}.
\end{aligned}$$

If f is given by (1.7), then $f''(\phi) = C_3\big[C_1^2 e^{C_1\phi} - C_2^2 e^{-C_2\phi}\big]$ so that from Lemma 2.2 we infer that

$$\begin{aligned}
&\left\{\int_{\Gamma_C}\int_0^1\int_0^1 s\left|f''\big(\phi(\mathbf{t}) + s(1-r)(\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t}))\big)\,ds\,dr\right|^4 d\Gamma\right\}^{1/4} \\
&\qquad \le C\big[1 + |\Gamma| + e^{C\,\|\phi(\mathbf{t}) + s(1-r)(\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t}))\|_1^2)}\big]^{1/4} \\
&\qquad \le C\left[1 + |\Gamma| + e^{C\,\{\|\phi(\mathbf{t})\|_1^2 + \|\phi(\mathbf{t}+\epsilon\mathbf{e}_i)\|_1^2\}}\right]^{1/4}.
\end{aligned}$$

Thus by setting $v = \psi_\epsilon$ in (4.4) and using (1.8) and (2.1) we obtain

$$\begin{aligned}
\gamma\,\|\psi_\epsilon\|_1^2 \le C &\left\|u_i'(\cdot - t_i) - \frac{u_i(\cdot - t_i - \epsilon) - u_i(\cdot - t_i)}{-\epsilon}\right\|_{0,\Gamma_i(\mathbf{t})} \|\psi_\epsilon\|_{0,\Gamma_i(\mathbf{t})} \\
&+ K(\|\phi(\mathbf{t})\|_1, \|\phi(\mathbf{t}+\epsilon\mathbf{e}_i)\|_1)\,\|\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t})\|_{L^4(\Gamma_C)}\,\|\lambda_i\|_{L^4(\Gamma_C)}\,\|\psi_\epsilon\|_{L^4(\Gamma_C)},
\end{aligned}$$

where

$$\begin{aligned}
&K(\|\phi(\mathbf{t})\|_1, \|\phi(\mathbf{t}+\epsilon\mathbf{e}_i)\|_1) \\
&= \max\left\{C\left[1 + |\Gamma| + e^{C\,(\|\phi(\mathbf{t})\|_1^2 + \|\phi(\mathbf{t}+\epsilon\mathbf{e}_i)\|_1^2)}\right]^{1/4},\ C\,(\|\phi(\mathbf{t})\|_1 + \|\phi(\mathbf{t}+\epsilon\mathbf{e}_i)\|_1)\right\}.
\end{aligned}$$

Trace theorems imply that

$$\|\psi_\epsilon\|_{0,\Gamma_i(\mathbf{t})} \le \|\psi_\epsilon\|_{0,\Gamma} \le C\|\psi_\epsilon\|_1\,,$$

$$\|\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t})\|_{L^4(\Gamma_C)} \le \|\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t})\|_{L^4(\Gamma)} \le C\|\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t})\|_1\,,$$

$$\|\lambda_i\|_{L^4(\Gamma_C)} \le \|\lambda_i\|_{L^4(\Gamma)} \le C\|\lambda_i\|_1 \quad \text{and} \quad \|\psi_\epsilon\|_{L^4(\Gamma_C)} \le \|\psi_\epsilon\|_{L^4(\Gamma)} \le C\|\psi_\epsilon\|_1$$

so that

$$\|\psi_\epsilon\|_1 \le C\left\|u_i'(\cdot - t_i) - \frac{u_i(\cdot - t_i - \epsilon) - u_i(\cdot - t_i)}{-\epsilon}\right\|_{0,\Gamma_i(\mathbf{t})} + C\,K(\|\phi(\mathbf{t})\|_1, \|\phi(\mathbf{t}+\epsilon\mathbf{e}_i)\|_1)\,\|\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t})\|_1\,\|\lambda_i\|_1\,. \tag{4.5}$$

Since each $u_i \in C^1(\mathbb{R})$, the first term on the right-hand side of (4.5) clearly goes to zero as $\epsilon \to 0$. Since the mapping $\mathbf{t} \mapsto \phi(\mathbf{t})$ is continuous from $\prod_{i=1}^N [a_i, b_i] \to H^1(\Omega)$, we see that the quantity $K(\|\phi(\mathbf{t})\|_1, \|\phi(\mathbf{t}+\epsilon\mathbf{e}_i)\|_1)$ is bounded as $\epsilon \to 0$ and $\|\phi(\mathbf{t}+\epsilon\mathbf{e}_i) - \phi(\mathbf{t})\|_1 \to 0$ as $\epsilon \to 0$. We recall (4.2) that $\|\lambda_i\|_1$ is bounded. Hence we conclude that the right-hand side of (4.5) goes to zero as $\epsilon \to 0$, i.e.,

$$\|\psi_\epsilon\|_1 \to 0 \quad \text{as } \epsilon \to 0\,.$$

Thus we have shown that the mapping $\mathbf{t} \mapsto \phi(\mathbf{t})$ is differentiable from $\prod_{i=1}^N (a_i, b_i) \to H^1(\Omega)$ and $\frac{\partial\phi(\mathbf{t})}{\partial t_i} = \lambda_i$. □

Integration by parts yields that λ_i is the solution of the following linearized boundary value problem:

$$-\operatorname{div}(\sigma \operatorname{grad}\lambda_i) = 0 \quad \text{in } \Omega\,,$$

$$\sigma\frac{\partial\lambda_i}{\partial n} = -u_i'(s - t_i) \quad \forall\, s \in \left[t_i - \frac{d_i}{2}, t_i + \frac{d_i}{2}\right] \text{ on } \Gamma_i(\mathbf{t})\,,$$

$$\sigma\frac{\partial\lambda_i}{\partial n} = 0 \quad \forall\, s \in \left[t_j - \frac{d_j}{2}, t_j + \frac{d_j}{2}\right] \text{ on } \Gamma_j(\mathbf{t})\,, \quad j \ne i\,, \quad j = 1, \ldots, N,$$

$$\sigma\frac{\partial\lambda_i}{\partial n} = 0 \quad \text{on } \Gamma_0(\mathbf{t}),$$

and

$$\sigma\frac{\partial\lambda_i}{\partial n} = -f'(\phi(\mathbf{t}))\lambda_i \quad \text{on } \Gamma_C\,.$$

THEOREM 4.2. *Assume that* $(\hat\phi, \hat{\mathbf{t}})$ *minimizes* (1.1) *subject to* (1.9). *Then for each* i, $\hat t_i$ *either is on the boundary of* $[a_i, b_i]$ *or satisfies the equation*

$$\int_{\Gamma_C} (\phi(\hat{\mathbf{t}}) - \phi_0)\hat\lambda_i\, d\Gamma = 0,$$

where $\hat\lambda_i$ *is the solution of* (4.1) *with* $\mathbf{t} = \hat{\mathbf{t}}$.

Proof. Since $(\hat\phi, \hat{\mathbf{t}})$ satisfies (1.9), we have $\hat\phi = \phi(\hat{\mathbf{t}})$. Define $\mathcal{K}(\mathbf{t}) = \mathcal{J}(\phi(\mathbf{t}), \mathbf{t})$. We have shown in previous lemmas that $\phi(\cdot)$ is continuous in $\prod_{i=1}^N [a_i, b_i]$ and differentiable in $\prod_{i=1}^N (a_i, b_i)$. Minimizing (1.1) subject to (1.9) is obviously equivalent to minimizing $\mathcal{K}$ over

$\prod_{i=1}^{N}[a_i, b_i]$. Thus $\hat{\mathbf{t}}$ is a minimum of $\mathcal{K}$. Hence, we have that for each i, $\hat{t}_i$ is either on the boundary of $[a_i, b_i]$ or in the interior of (a_i, b_i) with

$$\frac{\partial \mathcal{K}(\hat{\mathbf{t}})}{\partial t_i} = 0.$$

Differentiating $\mathcal{K}(\cdot)$ by the chain rule yields

$$\frac{\partial \mathcal{K}(\hat{\mathbf{t}})}{\partial t_i} = \int_{\Gamma_C} (\phi(\hat{\mathbf{t}}) - \phi_0)\hat{\lambda}_i \, d\Gamma,$$

where $\hat{\lambda}_i$ satisfies (4.1) with $\mathbf{t} = \hat{\mathbf{t}}$. Thus if $\hat{t}_i$ is in the interior of (a_i, b_i), then

$$\int_{\Gamma_C} (\phi(\hat{\mathbf{t}}) - \phi_0)\hat{\lambda}_i \, d\Gamma = 0. \qquad \square$$

Remark (three-dimensional formulation). The analysis we have done for the two-dimensional case can be generalized to the three-dimensional case in a straightforward manner by replacing the function space $H^1(\Omega)$ with $W^{1,p}(\Omega)$ for some $p \geq 3$. (Then Lemma 2.2 holds.) In the three-dimensional case, each center point on the ith anode Γ_i is parameterized by (x_i, y_i). Let $\mathbf{t} = (x_1, y_1, x_2, y_2, \ldots, x_N, y_N)$. Then we obtain essentially the same results as in the two-dimensional case. $\square$

Remark (exterior problems). For infinite-domain problems, we assume the following condition at infinity:

$$\phi(\mathbf{x}) \to \phi_\infty \quad \text{as } |\mathbf{x}| \to \infty,$$

where ϕ_∞ is a constant. Then the results of this paper hold equally well. $\square$

5. Gradient methods. Grandient methods for minimizing $\mathcal{K}(\mathbf{t})$ with variable step lengths are given as follows:

$$\mathbf{t}^{n+1} = \mathbf{t}^n - \rho_n \nabla \mathcal{K}(\mathbf{t}^n).$$

This algorithm applied to the problem (1.1)-(1.5) takes on the following form:

a) Choose a $\mathbf{t}^{(0)} \in \prod_{i-1}^{N}(a_i, b_i)$.

b) For $n \geq 0$, solve for $\phi^{(n)} = \phi^{(n)}(\mathbf{t}^{(n)})$ from

$$-\operatorname{div}(\sigma \operatorname{grad} \phi^{(n)}) = 0 \quad \text{in } \Omega, \tag{5.1}$$

$$\sigma \frac{\partial \phi^{(n)}}{\partial n} = u_i(s - t_i^{(n)}) \quad \forall\, s \in \left[t_i^{(n)} - \frac{d_i}{2}, t_i^{(n)} + \frac{d_i}{2}\right] \text{ on } \Gamma_i(\mathbf{t}^{(n)}), \ \forall\, i, \tag{5.2}$$

$$\sigma \frac{\partial \phi^{(n)}}{\partial n} = 0 \quad \text{on } \Gamma_0(\mathbf{t}^{(n)}), \tag{5.3}$$

and

$$\sigma \frac{\partial \phi^{(n)}}{\partial n} = -f(\phi^{(n)}) \quad \text{on } \Gamma_C. \tag{5.4}$$

c) Solve for $\lambda_i^{(n)}$ for each $i = 1, \ldots, N$ from

$$-\operatorname{div}(\sigma \operatorname{grad} \lambda_i^{(n)}) = 0 \quad \text{in } \Omega, \tag{5.5}$$

$$\sigma \frac{\partial \lambda_i^{(n)}}{\partial n} = -u_i'(s - t_i^{(n)}) \quad \forall\, s \in \left[t_i^{(n)} - \frac{d_i}{2}, t_i^{(n)} + \frac{d_i}{2}\right] \quad \text{on } \Gamma_i(\mathbf{t}^{(n)}), \tag{5.6}$$

$$\sigma \frac{\partial \lambda_i^{(n)}}{\partial n} = 0 \quad \forall\, s \in \left[t_j^{(n)} - \frac{d_j}{2}, t_j^{(n)} + \frac{d_j}{2}\right] \text{ on } \Gamma_j(\mathbf{t}^{(n)}), \ \forall\, j \neq i, \tag{5.7}$$

$$\sigma \frac{\partial \lambda_i^{(n)}}{\partial n} = 0 \quad \text{on } \Gamma_0(\mathbf{t}^{(n)}), \tag{5.8}$$

and

$$\sigma \frac{\partial \lambda_i^{(n)}}{\partial n} = -f'(\phi^{(n)})\lambda_i^{(n)} \quad \text{on } \Gamma_C. \tag{5.9}$$

d) Update $\mathbf{t}$ by

$$t_i^{(n+1)} = t_i^{(n)} - \rho_n \int_{\Gamma_C} (\phi^{(n)} - \phi_0)\, \lambda_i^{(n)}\, d\Gamma,$$
$$\text{if } t_i^{(n+1)} > b_i, \quad \text{set } t_i^{(n+1)} = b_i,$$
$$\text{if } t_i^{(n+1)} < a_i, \quad \text{set } t_i^{(n+1)} = a_i.$$

e) $n + 1 \to n$; go to b).

Remark. A stopping criterion has to be used in actual implementation.

Remark. Standard convergence results for the constrained gradient method requires that the step length ρ_n be sufficiently small. However, the convergence will be very slow if the step length is chosen to be too small. In our implementation, we use the test-and-trial strategy to ensure the locations will improve by at least a fixed tolerant distance per iteration so that the number of iterations needed to obtain an optimal solution is typically small. Effective choices of ρ_n also depend on the dimensions of all the variables involved. Boundary element methods are employed to implement the above algorithm, and if necessary, the mesh is adjusted at each iteration so that each set of new locations form part of the set of all grid points.

Remark. In step d) in the algorithm, $t^{(n+1)}$ is calculated and then projected into the admissible set. In our experience with several examples, such a projection is not needed when the step size and initial guesses are chosen properly so that our constrained optimization problem can often be treated as an unconstrained optimization problem.

6. Computational examples. We tested our algorithm with an electrolyte container problem. The metal container is assumed to occupy a rectangular domain $\Omega : 0 \leq x \leq 3,\ 0 \leq y \leq 2$ (see Figures 3, 6, and 9), and the container is filled with an electrolyte. The cathode Γ_C is on the bottom boundary described by the coordinates $1 \leq x \leq 2$ and $y = 0$. We investigated several cases: single-anode control (Example 1), double-anode control with different densities on the top and right segments (Examples 2a and 2b), and double-anode control on the top segment (Example 3).

Example 1 (single-anode control on the top segment). We want to place an anode of width 0.4 on the top boundary. The center point $(x_c, 2)$ of the anode segment is constrained to the interval $0.2 \leq x_c \leq 2.8$. The current density on the anode is chosen to be $u = -20\cos^2(2.5(x - x_c))$ for $x \in (x_c - 0.2, x_c + 0.2)$. We use the empirical function $f(\phi) = 4\phi + 0.4\phi^3$. The desired potential distribution on Γ_C is $\phi_0 = -1$.

Our initial guess of x_c is $x_c^{(0)} = 2.6$ (Figure 3). After four iterations using the gradient algorithm, we arrive at $x_c^{(4)} = 1.5$, which corresponds to the center point on the top boundary

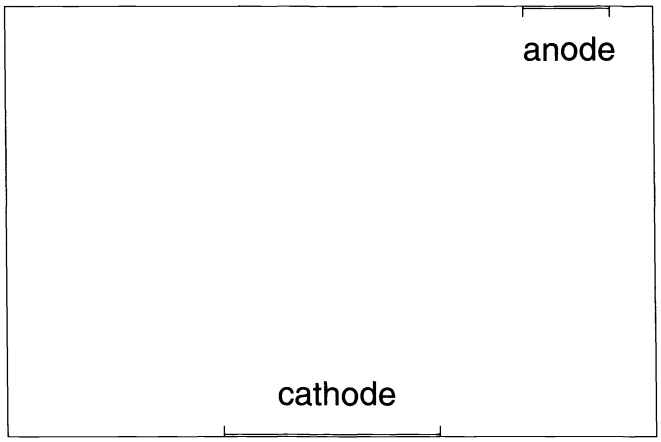


FIG. 3. *Initial guess of the anode position on the container* (*Example* 1).

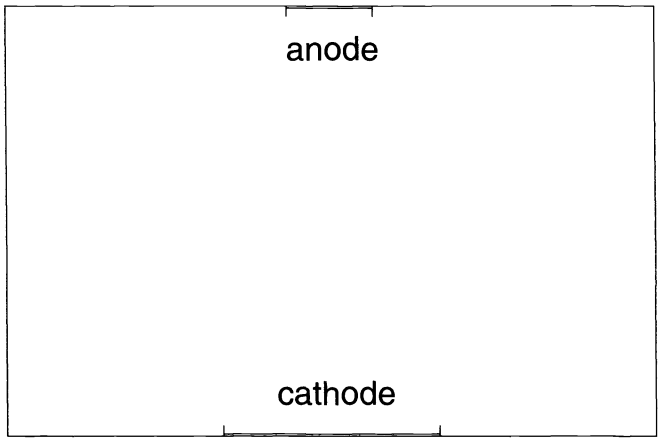


FIG. 4. *Optimal anode position* (*Example* 1).

(Figure 4). The step lengths in the gradient method had to be adjusted properly to ensure fast convergence. In Figure 5 we plotted the initial and final potential distributions on Γ_C. We could verify that $x_c = 1.5$ is indeed the minimum. This result is also consistent with intuitive explanations by chemists.

Example 2a (double-anode control on the top and right segments). We want to place two anodes of width 0.4 on the boundary, one on the top boundary with a center point $(x_c, 2)$, where $0.2 \leq x_c \leq 2.8$, and another on the right boundary with a center point $(3, y_c)$, where $0.2 \leq y_c \leq 1.8$. The current densities on the two anodes are chosen to be $u_1 = -16\cos^2(2.5(x-x_c))$ for $x \in (x_c - 0.2, x_c + 0.2)$ and $u_2 = -4\cos^2(2.5(y - y_c))$ for $y \in (y_c - 0.2, y_c + 0.2)$. The function f and desired ϕ_0 are chosen to be the same as in Example 1.

Our initial guesses for the two anodes are $x_c^{(0)} = 2.6$ on top boundary and $y_c^{(0)} = 0.4$ on the right boundary (Figure 6). After four iterations using the gradient algorithm, we arrive at $x_c^{(4)} = 1.1$ on the top boundary and $y_c^{(4)} = 0.6$ on the right boundary (Figure 7). Again, the step lengths in the gradient method had to be adjusted properly to ensure fast convergence. In Figure 8 we plotted the initial and final potential distributions on Γ_C. We verified by small perturbations of the anode locations that we indeed obtained a minimum with $x_c = 1.1$ on

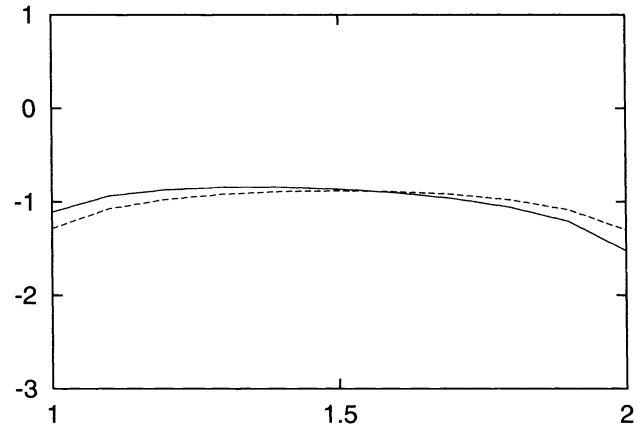


FIG. 5. *Initial and optimal potential distribution on the cathode* Γ_C (*Example* 1) (*initial:* ——, *optimal:* ·····).

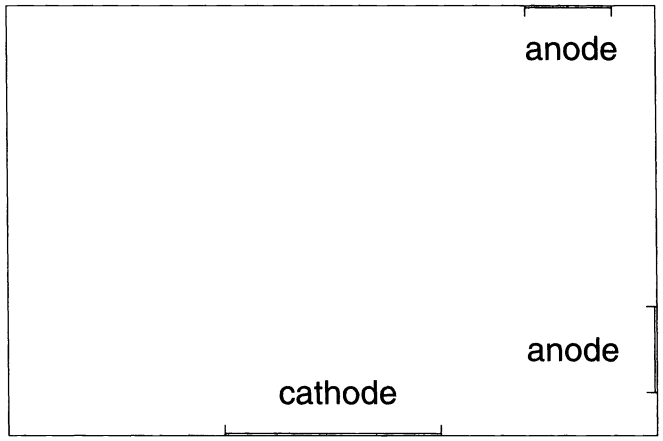


FIG. 6. *Initial guess of the anode positions on the container* (*Examples* 2a *and* 2b).

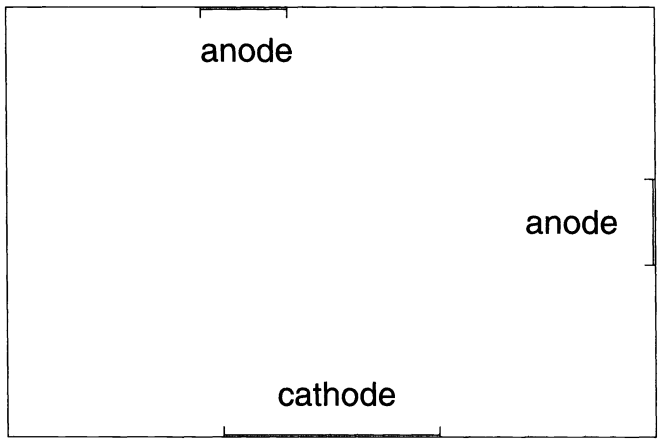


FIG. 7. *Optimal anode positions on the top and right boundary* (*Example* 2a).

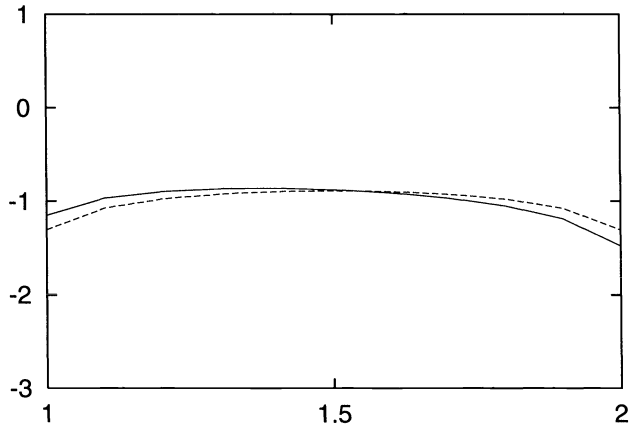


FIG. 8. *Initial and optimal potential distribution on the cathode* Γ_C (*Example* 2a) (*initial:* ——, *optimal:* ·····).

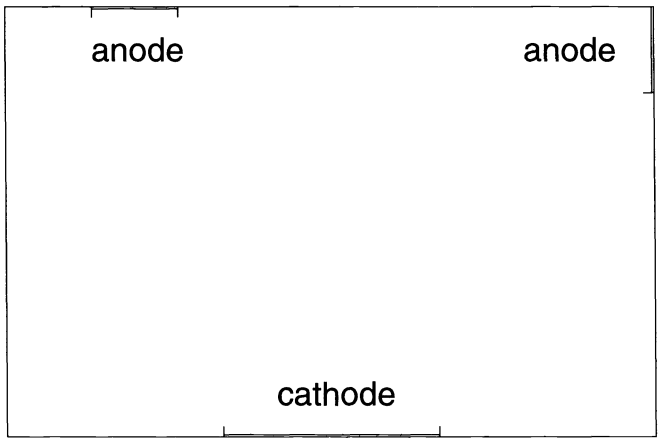


FIG. 9. *Optimal anode positions on the top and right boundary* (*Example* 2b).

the top boundary and $y_c = 0.6$ on the right boundary. This result is not an obvious fact. Our algorithm and calculation proved to be useful in providing the optimal locations for the placement of the double anodes.

Example 2b (double-anode control on the top and right segments—densities interchanged). All the data are the same as in Example 2a, except that the two density functions are interchanged, i.e., $u_1 = -4\cos^2(2.5(x-x_c))$ for $x \in (x_c-0.2, x_c+0.2)$ and $u_2 = -16\cos^2(2.5(y-y_c))$ for $y \in (y_c - 0.2, y_c + 0.2)$.

We use the same initial guess as in Example 2a (Figure 6). After nine iterations using the gradient algorithm, we arrive at $x_c^{(9)} = 2.6$ on the top boundary and $y_c^{(9)} = 1.8$ on the right boundary (Figure 9), which are different from the optimal positions that we obtained in Example 2a. In Figure 10 we plotted the initial and final potential distributions on Γ_C.

Remark. The results of Examples 2a and 2b indicate that after we have found the optimal locations for each anode, it is still important not to misplace the anodes with different current density. □

Example 3 (double-anode control on the top segment). We want to place two anodes of the width 0.4 on the top boundary, one with a center point $(x_{1c}, 2)$, where $0.2 \le x_{1c} \le 1.3$,

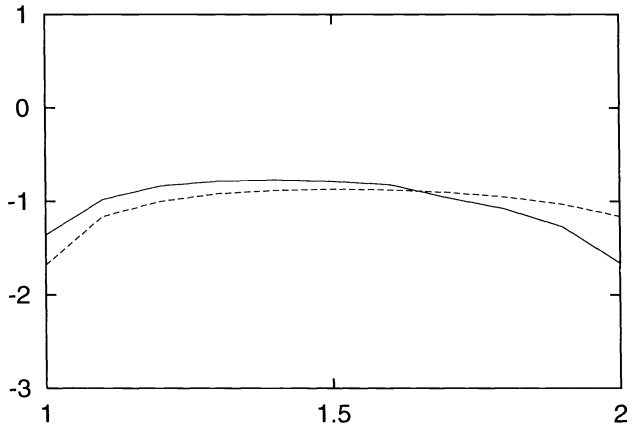


FIG. 10. *Initial and optimal potential distribution on the cathode* Γ_C (*Example* 2b) (*initial:* ——, *optimal:* ·····).

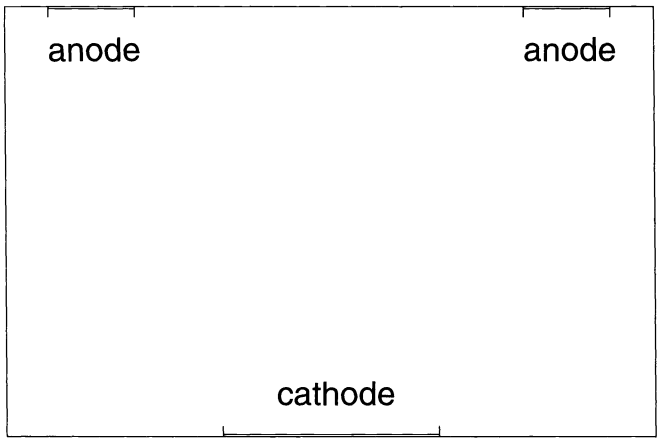


FIG. 11. *Initial guess of the anode positions on the container* (*Example* 3).

and another with a center point $(x_{2c}, 2)$, where $1.7 \leq x_{2c} \leq 2.8$. The current densities on the two anodes are chosen to be $u_1 = -16\cos^2(2.5(x - x_{1c}))$ for $x \in (x_{1c} - 0.2, x_{1c} + 0.2)$ and $u_2 = -4\cos^2(2.5(x - x_{2c}))$ for $x \in (x_{2c} - 0.2, x_{2c} + 0.2)$. The function f and desired ϕ_0 are chosen to be the same as in Example 1.

Our initial guesses for the two anodes are $x_{1c}^{(0)} = 0.4$ and $x_{2c}^{(0)} = 2.6$ (Figure 11). After seven iterations using the gradient algorithm, we arrive at $x_{1c}^{(7)} = 1.2$ and $x_{2c}^{(7)} = 1.6$ (Figure 12). Again, the step lengths in the gradient method had to be adjusted properly to ensure fast convergence. In Figure 13 we plotted the initial and final potential distributions on Γ_C. Clearly, the two anodes at their optimal positions join together at the common end point (1.4, 2) on the top boundary. The common end point of the two anodes has to be insulated in order to maintain the two anode system.

Remark. In Examples 2a and 3 we allow the anodes on different target segments to join together. In the case of Example 2a the optimal solution yields two disjoint anodes. In the case of Example 3 the optimal solution yields two anodes that join together at a point on the top boundary. When this happens, we can simply combine the two anodes to form a larger anode with different density profiles on the two sections of the combined anode. However, to

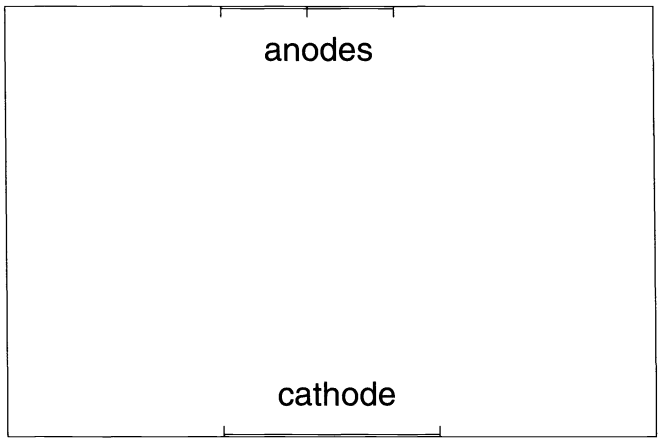


FIG. 12. *Optimal anode positions on the top and right boundary (Example 3).*

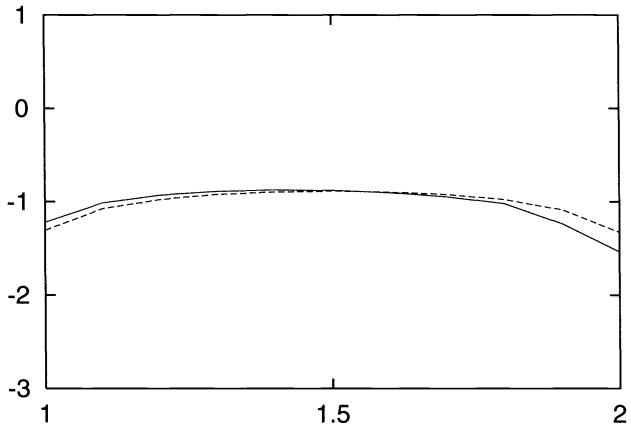


FIG. 13. *Initial and optimal potential distribution on the cathode Γ_C (Example 3) (initial: ——, optimal: ·····).*

maintain the combined (piecewise) density function, one will still need two separate electric sources. Inserting a thin insulator makes it easier to generate the required electric density profile on each anode. In most practical applications, one usually looks for optimal locations for anodes in isolated segments and thus will not encounter the situation of joining anodes. For instance, in the case of ship propeller protection, one wishes to place one anode near the front and two near the rear end; thus the target segments for these anodes are far apart from each other. Returning to Example 3, we may restrict the first anode to near the top left corner (say, $x_{1c} \leq 0.4$) and restrict the second anode to the right half of the top boundary. With such restrictions, the two anodes will never join each other. □

In all examples we have used boundary element methods (see [4], [7], [10]) to carry out the algorithm proposed in §5, i.e., to solve the systems (5.1)-(5.4) and (5.5)-(5.9) repeatedly. The calculations were performed on a SUN Sparc 2. Details for the boundary element method and numerical examples for propeller protection and/or for hybrid controls can be found in [7].

Acknowledgments. The authors thank the anonymous referees for many useful comments that have helped improve the results of this paper.

REFERENCES

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

[2] K. AMAYA AND S. AOKI, *Optimum design of cathodic protection system by 3-D BEM*, Boundary Element Technology, VII (1992), pp. 375-388.

[3] J. O. M. BOCKRIS AND A. K. M. REDDY, *Modern Electrochemistry*, Plenum Press, New York, 1974.

[4] G. CHEN AND J. ZHOU, *Boundary Element Methods*, Academic Press, London, 1992.

[5] V. GIRAULT AND P. RAVIART, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.

[6] P. HESS, *On nonlinear mappings of monotone type with respect to two Banach spaces*, J. Math. Pures Appl., 52 (1973), pp. 13-26.

[7] L. S. HOU AND W. SUN, *Numerical methods for optimal control of impressed cathodic protection systems*, Int. J. Numer. Meth. Engrg., 37 (1994), pp. 2779-2797.

[8] L. S. HOU AND J. C. TURNER, *Analysis and finite element approximation of an optimal control problem in electrochemistry using current density controls*, Numer. Math., 71 (1995), pp. 289-315.

[9] J. NEČAS, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Academia, Prague, 1967.

[10] K. RUOTSALAINEN AND W. WENDLAND, *On the boundary element method for some nonlinear boundary value problems*, Numer. Math., 53 (1988), pp. 299-314.

[11] N. G. ZAMANI AND F. H. PETERS, *Boundary element solution of 2-D corrosion problems*, Technical Report, CADCAMTR-85-03, Technical University of Nova Scotia, 1985.

[12] N. G. ZAMANI AND J. M. CHUANG, *Optimal control of current in a cathodic protection system: A numerical investigation*, Optimal Control Appl. Methods, 8 (1987), pp. 339-350.

[13] N. G. ZAMANI, J. F. PORTER, AND A. A. MUFTI, *A survey of computational efforts in the field of corrosion engineering*, Int. J. Numer. Meth. Engrg., 23 (1986), pp. 1295-1311.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 874-891, May 1996

AUGMENTED LAGRANGIAN-SQP METHODS FOR NONLINEAR OPTIMAL CONTROL PROBLEMS OF TRACKING TYPE*

KAZUFUMI ITO† AND KARL KUNISCH‡

Abstract. An augmented Lagrangian method with second-order update is developed and its relationship to the sequential quadratic programming method is described. The rate of convergence proof depends on a second-order sufficient optimality condition, which is shown to be satisfied for a class of nonlinear optimal control problems of tracking type. Numerical examples are included which demonstrate the globalizing effect of the augmented Lagrangian method.

Key words. Lagrangian methods, SQP methods, nonlinear optimal programming

AMS subject classifications. 49D, 65K

1. Introduction.

The mathematical and, more specifically, the numerical treatment of optimal control problems for nonlinear partial differential equations arising in diverse areas of science has received an increasing amount of attention in the recent past. We mention optimal control problems in phase field modeling [CH, H], in superconductivity [GHS], in combustion, and, of course, in fluid dynamics. The numerical treatment of such problems offers a multitude of open problems and the present paper addresses one such problem.

We study a class of optimal control problems for nonlinear partial differential equations where the cost is of tracking type. The technique that we propose and analyze is the augmented Lagrangian-SQP (sequential quadratic programming) algorithm as developed in [IK]. In this method the differential equation is treated as a constraint which is realized by a Lagrangian term together with a penalty functional. The resulting augmented Lagrangian functional allows a rather straightforward characterization of the second derivative. We prove second-order convergence rate of the algorithm and we also demonstrate this rate with numerical examples. The second-order convergence rate depends upon a second-order sufficient optimality condition. We verify this second-order condition under variants of assumptions on the smallness of the cost functional at the solution. While this analysis is to a certain degree specific for the class of problems under consideration, the general principle becomes apparent: For optimal control problems with cost functional of tracking type, smallness of the cost or the residue helps (guarantees for the problems in this paper) the second-order sufficient optimality condition to hold.

Another important aspect of this research is the global behavior of the algorithm in numerical experiments. We did not implement any globalization strategy and yet convergence was observed even for very unfavorable start-up values. The cost functional for the augmented Lagrangian-SQP algorithm involves a penalty term. We point out that the specific size of the penalty parameter is not too significant (details are given in §5) unless it is chosen to be zero, in which case convergence generally fails. Let us also mention that the approach of this paper is quite different from that chosen in [CH, GHS, H], for example, which is based on iterative techniques for solving the necessary optimality system.

This paper is organized in the following manner. In §2 relevant results on the augmented Lagrangian-SQP technique are summarized. The optimal control problems and the associated

*Received by the editors January 14, 1994; accepted for publication (in revised form) December 28, 1994.

†Center for Research in Scientific Computing, North Carolina State University, Box 8205, Raleigh, NC 27695-8205. The research of this author was supported in part by National Science Foundation grant UINT-8521208.

‡Fachbreich Mathematik, Technische Universität Berlin, Strasse des 17, Juni 136, D-10623 Berlin, Germany. The research of this author was supported in part by the Fonds zur Förderung der wissenschaftlichen Forschung grant P-8146-PHY and the Christian Doppler Laboratory on Parameter Identification and Inverse Problems.

Lagrangian framework are developed in §3. The second-order sufficient optimality condition is analyzed in §4 and numerical test examples are given in §5.

2. Review of augmented Lagrangian–SQP methods. In this section results on augmented Lagrangian–SQP methods which are relevant in the following sections are summarized. We refer to [Be] for further details concerning such methods for finite-dimensional problems and to [IK] for infinite-dimensional problems.

We consider

$$(\mathcal{P}) \qquad \min F(x) \text{ subject to } e(x) = 0,$$

where $F : X \to \mathbb{R}$, $e : X \to Y$, with X and Y Hilbert spaces. We make the following three assumptions.

(H1) There exists a (local) solution x^* of $(\mathcal{P})$, F and e are twice continuously Fréchet differentiable, and the second Fréchet derivatives are Lipschitz continuous in a neighborhood $\tilde{V}(x^*)$ of x^*. The Fréchet derivative of any function with respect to x is denoted by a prime. The Lagrangian $\mathcal{L} : X \times Y \to \mathbb{R}$ associated with $(\mathcal{P})$ is defined by

$$\mathcal{L}(x, \lambda) = F(x) + \langle \lambda, e(x) \rangle_Y,$$

where $\langle \cdot, \cdot \rangle_Y$ stands for the inner product in Y. An element $\lambda^* \in Y$ is called Lagrange multiplier for $(\mathcal{P})$ if

$$\mathcal{L}'(x^*, \lambda^*) = F'(x^*) + e'(x^*)^* \lambda^* = 0. \tag{2.1}$$

Here $e'(x^*)^*$ denotes the adjoint operator of $e'(x^*)$. In (2.1) and also below, we frequently do not distinguish between $F'(x^*) \in \mathcal{L}(X; \mathbb{R})$ and its Riesz representation in X.

(H2) $e'(x^*)$ is surjective.

(H3) There exists $\kappa > 0$ such that

$$\mathcal{L}''(x^*, \lambda^*)(h, h) \geq \kappa |h|_X^2 \quad \text{for all } h \in \ker e'(x^*).$$

Under these assumptions there exists a neighborhood $V(x^*)$ of x^* and constants $\bar{\sigma} > 0$ and $\bar{c} > 0$ such that

$$\mathcal{L}_c(x, \lambda^*) \geq \mathcal{L}_c(x^*, \lambda^*) + \sigma |x - x^*|_X^2 \quad \text{for all } x \in V(x^*) \text{ and } c \geq \bar{c},$$

where $\mathcal{L}_c$ is the augmented Lagrangian defined by

$$\mathcal{L}_c(x, \lambda) = \mathcal{L}(x, \lambda) + \frac{c}{2} |e(x)|_Y^2.$$

We shall describe two algorithms and introduce

$$M(x, \lambda) = \begin{pmatrix} \mathcal{L}''(x, \lambda) & e'(x)^* \\ e'(x) & 0 \end{pmatrix}.$$

ALGORITHM 1.

(i) Choose $\lambda_0 \in Y$, $c \in (\bar{c}, \infty)$, and set $\sigma = c - \bar{c}$, $n = 0$.

(ii) Determine $\tilde{x}$ as solution of

$$(P_{aux}) \qquad \min \mathcal{L}_c(x, \lambda_n) \text{ subject to } x \in \overline{V(x^*)}.$$

(iii) Set $\tilde{\lambda} = \lambda_n + \sigma e(\tilde{x})$.

(iv) Solve for $(\hat{x}, \hat{\lambda})$

$$M(\tilde{x}, \tilde{\lambda}) \begin{pmatrix} \hat{x} - \tilde{x} \\ \hat{\lambda} - \tilde{\lambda} \end{pmatrix} = - \begin{pmatrix} \mathcal{L}'(\tilde{x}, \tilde{\lambda}) \\ e(\tilde{x}) \end{pmatrix}.$$

(v) Set $\lambda_{n+1} = \hat{\lambda}$, $n = n + 1$ goto (ii).

Existence of a solution to (P_{aux}) is guaranteed if $f : X \to \mathbb{R}$ is weakly lower semicontinuous and $e : X \to Y$ maps weakly convergent sequences to weakly convergent sequences. The following algorithm differs from the first one in that (P_{aux}) is eliminated.

ALGORITHM 2.

(i) Choose $(x_0, \lambda_0) \in X \times Y$, $c \geq 0$, and set $n = 0$.

(ii) Set $\tilde{\lambda} = \lambda_n + ce(x_n)$.

(iii) Solve for $(\hat{x}, \hat{\lambda})$

$$M(x_n, \tilde{\lambda}) \begin{pmatrix} \hat{x} - x_n \\ \hat{\lambda} - \tilde{\lambda} \end{pmatrix} = - \begin{pmatrix} \mathcal{L}'(x_n, \tilde{\lambda}) \\ e(x_n) \end{pmatrix}.$$

(iv) Set $(x_{n+1}, \lambda_{n+1}) = (\hat{x}, \hat{\lambda})$, $n = n + 1$ and goto (ii).

PROPOSITION 2.1. *Let* (H1) *and* (H2) *hold, and in the case of Algorithm* 1 *assume that* (P_{aux}) *admits a solution for all* n.

(i) *If* $\frac{1}{c - \bar{c}} |\lambda_0 - \lambda^*|^2$ *is sufficiently small, then Algorithm* 1 *is well defined and its iterates satisfy*

$$|(x_{n+1}, \lambda_{n+1}) - (x^*, \lambda^*)|_{X \times Y} \leq \frac{\hat{K}}{c - \bar{c}} |\lambda_n - \lambda^*|_Y^2,$$

for a constant $\hat{K}$ *independent of* c *and* $n = 0, 1, \ldots$.

(ii) *If* $c|(x_0, \lambda_0) - (x^*, \lambda^*)|_{X \times Y}$ *is sufficiently small, then Algorithm* 2 *is well defined and its iterates satisfy*

$$|(x_{n+1}, \lambda_{n+1}) - (x^*, \lambda^*)|_{X \times Y} \leq \tilde{K} |(x_n, \lambda_n) - (x^*, \lambda^*)|_{X \times Y}^2,$$

for a constant $\tilde{K}$ *independent of* $n = 0, 1, \ldots$.

For a proof we refer to [IK].

3. A class of nonlinear optimal control problems. The general framework of the previous section will be applied to optimal control problems governed by partial differential equations of the type

$$\text{(3.1)} \qquad \begin{cases} -\Delta y + f(y) = \tilde{h} & \text{in} \quad \Omega, \\ \dfrac{\partial y}{\partial n} = g & \text{on} \quad \Gamma_1, \\ \dfrac{\partial y}{\partial n} = g_2 & \text{on} \quad \Gamma_2, \end{cases}$$

where $\tilde{h} \in L^2(\Omega)$ and $g_2 \in L^2(\Gamma_2)$ are fixed and $g \in L^2(\Gamma_1)$ is the control variable. Here Ω is a bounded domain in $\mathbb{R}^n$ with $C^{1,1}$ boundary or Ω is convex. The boundary Γ is assumed to consist of two disjoint sets Γ_1, Γ_2, each of which is connected (or possibly consisting of finitely many connected components), with $\Gamma = \Gamma_1 \cup \Gamma_2$ and Γ_2 possibly empty. Further it is assumed

that $f \in C^2(\mathbb{R})$, $f(H^1(\Omega)) \subset L^{1+\varepsilon}(\Omega)$ for some $\varepsilon > 0$ if $n = 2$, and $f(H^1(\Omega)) \subset L^{\frac{2n}{n+2}}(\Omega)$ if $n \geq 3$. Equation (3.1) is understood in the variational sense, i.e.,

$$\langle \nabla y, \nabla \varphi \rangle_\Omega + \langle f(y), \varphi \rangle_\Omega = \langle \tilde{h}, \varphi \rangle_\Omega + \langle \tilde{g}, \varphi \rangle_\Gamma \quad \text{for all } \varphi \in H^1(\Omega), \tag{3.2}$$

where

$$\tilde{g} = \begin{cases} g & \text{on } \Gamma_1, \\ g_2 & \text{on } \Gamma_2, \end{cases}$$

$\langle \cdot, \cdot \rangle_\Gamma$ denotes the L^2-inner product on Γ, and $\langle \cdot, \cdot \rangle_\Omega$ stands for duality pairing between functions in $L^p(\Omega)$ and $L^q(\Omega)$ with $p^{-1} + q^{-1} = 1$. In (3.2) we should more precisely write $\langle \tilde{g}, \tau_\Gamma \varphi \rangle_\Gamma$ instead of $\langle \tilde{g}, \varphi \rangle_\Gamma$, with τ_Γ the zero-order trace operator on Γ. However, we shall frequently suppress this notation. We refer to (y, g) as a solution of (3.1) if (3.2) holds. The optimal control problem is given by

$$(P) \qquad \begin{cases} \min \dfrac{1}{2}|Cy - y_d|_Z^2 + \dfrac{\alpha}{2}|g|^2_{L^2(\Gamma_1)}, \\ \text{subject to } (y, g) \in H^1(\Omega) \times L^2(\Gamma_1) \text{ a solution of (3.1).} \end{cases}$$

Here C is a bounded linear (observation) operator from $H^1(\Omega)$ to a Hilbert space Z, and $y_d \in Z$ and $\alpha > 0$ are fixed.

To express (P) in the form $(\mathcal{P})$ of §2 we introduce

$$\tilde{e} : H^1(\Omega) \times L^2(\Gamma_1) \to H^1(\Omega)^*$$

with

$$\langle \tilde{e}(y, g), \varphi \rangle_{(H^1)^*, H^1} = \langle \nabla y, \nabla \varphi \rangle_\Omega + \langle f(y) - \tilde{h}, \varphi \rangle_\Omega - \langle \tilde{g}, \varphi \rangle_\Gamma,$$

and

$$e : H^1(\Omega) \times L^2(\Gamma) \to H^1(\Omega)$$

by

$$e = \mathcal{N} \tilde{e},$$

where $\mathcal{N} : H^1(\Omega)^* \to H^1(\Omega)$ is the Neumann solution operator associated with

$$\langle \nabla v, \nabla \varphi \rangle_\Omega + \langle v, \varphi \rangle = \langle h, \varphi \rangle_\Omega \quad \text{for all } \varphi \in H^1(\Omega),$$

where $h \in H^1(\Omega)^*$. In the context of §2 we set

$$X = H^1(\Omega) \times L^2(\Gamma_1), \qquad Y = H^1(\Omega),$$

with $x = (y, g) \in X$, and

$$F(x) = F(y, g) = \frac{1}{2}|Cy - y_d|_Z^2 + \frac{\alpha}{2}|g|^2_{L^2(\Gamma_1)}.$$

We assume that (H1) holds, i.e., that (P) has a solution $x^* = (y^*, g^*)$. The regularity requirements of §2 are clearly met by the mapping F. Those for e are implied by

$$\text{(h0)} \qquad \begin{cases} y \to f(y) & \text{is continuous from } H^1(\Omega) \text{ to } L^{1+\varepsilon}(\Omega) \text{ for some } \varepsilon > 0 \text{ if } n = 2, \\ & \qquad\qquad H^1(\Omega) \text{ to } L^{\frac{2n}{n+2}}(\Omega) \text{ if } n \geq 3, \\ y \to f'(y) & \text{is continuous from } H^1(\Omega) \text{ to } L^{1+\varepsilon}(\Omega) \text{ for some } \varepsilon > 0 \text{ if } n = 2, \\ & \qquad\qquad H^1(\Omega) \text{ to } L^{\frac{n}{2}}(\Omega) \text{ if } n \geq 3, \\ y \to f''(y) & \text{is Lipschitz continuous in a neighborhood of } y^* \\ & \qquad\qquad \text{from } H^1(\Omega) \text{ to } L^{1+\varepsilon}(\Omega) \text{ for some } \varepsilon > 0 \text{ if } n = 2, \\ & \qquad\qquad \text{from } H^1(\Omega) \text{ to } L^{\frac{2n}{6-n}}(\Omega) \text{ if } 3 \leq n \leq 6. \end{cases}$$

Here we used the fact that due to Sobolev's embedding theorem $H^1(\Omega)$ is continuously embedded into $L^p(\Omega)$ for $p \leq \frac{2n}{n-2}$ if $n \geq 3$, and into L^p for every $p \geq 1$ if $n = 2$. We note that (h0) is satisfied by $f(y) = y^k$ for $k \leq 5$, for example.

In the remainder of this section we focus on the Lagrangian functional and on properties of the Lagrange multiplier associated with (P). We require the following hypothesis:

$$\text{(h1)} \qquad \begin{cases} f'(y^*) \in L^{2+\varepsilon}(\Omega) \text{ for some } \varepsilon > 0 \text{ if } n = 2, \\ f'(y^*) \in L^n(\Omega) \text{ if } n \geq 3. \end{cases}$$

With (h1) holding, $f'(y^*)\varphi \in L^2(\Omega)$ for every $\varphi \in H^1(\Omega)$. It is simple to argue that

$$\ker e'(x^*) = \{(v, h) : \langle \nabla v, \nabla \varphi \rangle_\Omega + \langle f'(y^*)v, \varphi \rangle_\Omega = \langle h, \varphi \rangle_{\Gamma_1} \text{ for all } \varphi \in H^1(\Omega)\},$$

i.e., $(v, h) \in \ker e'(x^*)$ if and only if (v, h) is a variational solution of

$$\text{(3.3)} \qquad \begin{cases} -\Delta v + f'(y^*)v = 0 & \text{in} \quad \Omega, \\ \dfrac{\partial v}{\partial n} = h & \text{on} \quad \Gamma_1, \\ \dfrac{\partial v}{\partial n} = 0 & \text{on} \quad \Gamma_2. \end{cases}$$

We also require (H2), assuring existence of a unique Lagrange multiplier $\lambda^* \in H^1(\Omega)$ satisfying

$$\text{(3.4)} \qquad e'(x^*)^*\lambda^* + (\mathcal{N}C^*(Cy^* - y_d), \alpha g^*) = 0 \quad \text{in } H^1(\Omega) \times L^2(\Gamma_1),$$

where $e'(x^*)^* : H^1(\Omega) \to H^1(\Omega) \times L^2(\Gamma_1)$ denotes the adjoint of $e'(x^*)$ and $C^* : Z \to H^1(\Omega)^*$ stands for the adjoint of $C : H^1(\Omega) \to Z$ with Z its pivot space. The Lagrange multiplier satisfies the following proposition.

PROPOSITION 3.1 (Necessary condition). *Let* (H1), (H2), *and* (h1) *hold. Then* λ^* *is a variational solution of*

$$\text{(3.5)} \qquad \begin{cases} -\Delta\lambda^* + f'(y^*)\lambda^* = -C^*(Cy^* - y_d) & \textit{in} \quad \Omega, \\ \dfrac{\partial \lambda^*}{\partial n} = 0 & \textit{on} \quad \Gamma, \end{cases}$$

i.e., $(\nabla\lambda^*, \nabla\varphi) + (f'(y^*)\lambda^*, \varphi) + (Cy^* - y_d, C\varphi)_Z = 0$ *for all* $\varphi \in H^1(\Omega)$ *and*

$$\text{(3.6)} \qquad \tau_{\Gamma_1}\lambda^* = \alpha g^* \quad \textit{on } \Gamma_1.$$

Proof. The Lagrangian associated with (P) can be expressed by

$$\mathcal{L}(y, g, \lambda) = \frac{1}{2}|Cy - y_d|_Z^2 + \frac{\alpha}{2}|g|_{L^2(\Gamma_1)}^2 + \langle \nabla\lambda, \nabla y \rangle_\Omega + \langle \lambda, f(y) - \tilde{h} \rangle_\Omega - \langle \lambda, \tilde{g} \rangle_\Gamma.$$

For every $(v, h) \in H^1(\Omega) \times L^2(\Gamma_1)$ we find

$$\mathcal{L}_y(y^*, g^*, \lambda^*)(v) = \langle Cy^* - y_d, Cv \rangle_Z + \langle \nabla\lambda^*, \nabla v \rangle_\Omega + \langle \lambda^*, f'(y^*)v \rangle_\Omega$$

and

$$\mathcal{L}_g(y^*, g^*, \lambda^*)(h) = \alpha\langle g^*, h \rangle_{\Gamma_1} - \langle \lambda^*, h \rangle_{\Gamma_1}.$$

Thus the claim follows.

In the following regularity result we let $\tilde{g}^*$ denote the function

$$\tilde{g}^* = \begin{cases} g^* & \text{on} \quad \Gamma_1, \\ g_2 & \text{on} \quad \Gamma_2. \end{cases}$$

COROLLARY 3.2. *Under the assumptions of Proposition* 3.1 *and if* $C^*(Cy^* - y_d) \in L^2(\Omega)$, *then* $\lambda^* \in H^2(\Omega)$ *and* $g^* \in H^{3/2}(\Gamma_1)$. *If* $n = 2$ *and* Ω *is a convex curvilinear polygon of class* C^1, $g_2 \in H^{3/2}(\Gamma_2)$, $g^* \in H^{3/2}(\Gamma_1)$, *and* $\tilde{g}^*$ *is "sufficiently smooth" at the endpoints of* Γ_1, *then* $y^* \in H^2(\Omega)$.

These regularity properties follow from well-known results on elliptic boundary value problems [G, pp. 44, 126, 149].

Let $B : H^1(\Omega) \to H^1(\Omega)^*$ be the differential operator given by the left-hand side of (3.5), i.e., $Bv = \varphi$ is characterized as the solution to

$$\langle \nabla v, \nabla \psi \rangle_\Omega + \langle f'(y^*)v, \psi \rangle_\Omega = \langle \varphi, \psi \rangle_{(H^1)^*, H^1} \quad \text{for all } \psi \in H^1(\Omega).$$

We shall use the following hypothesis:

(h2) 0 is not an eigenvalue of B.

Note that (h2) holds, for example, if

$$f'(y^*) \geq \underline{\beta} \quad \text{a.e. on } \Omega,$$

for some $\underline{\beta} > 0$. With (h2) holding, B is an isomorphism from $H^1(\Omega)$ onto $H^1(\Omega)^*$. Moreover, (h2) implies (H2).

COROLLARY 3.3. *Let* (H1), (H2), *and* (h1) *hold.*

(i) *There exists a constant* $K(x^*)$ *such that*

$$|\lambda^*|_{H^1} \leq K(x^*)|(\mathcal{N}C^*(Cy^* - y_d), \alpha g^*)|_X.$$

(ii) *If moreover* (h2) *is satisfied and* $C^*(Cy^* - y_d) \in L^2(\Omega)$, *then there exists a constant* $K(y^*)$ *such that*

$$|\lambda^*|_{H^2} \leq K(y^*)|C^*(Cy^* - y_d)|_{L^2(\Omega)}.$$

Proof. Due to (H2) we have $(e'(x^*)e'(x^*)^*)^{-1} \in \mathcal{L}(H^1(\Omega))$ and thus (i) follows from (3.4). Let us turn to (ii). Due to (h2) and (3.5) there exists a constant K_{y^*} such that

$$|\lambda^*|_{H^1} \leq K_{y^*}|C^*(Cy^* - y_d)|_{(H^1)^*}. \tag{3.7}$$

To obtain the desired $H^2(\Omega)$-estimate for λ^* we apply the well-known H^2 a priori estimate for Neumann problems to

$$\begin{aligned} -\Delta\lambda^* + \lambda^* &= \tilde{f}, \\ \frac{\partial \lambda^*}{\partial n} &= 0, \end{aligned}$$

with $\tilde{f} = \lambda^* - f'(y^*)\lambda^* - C^*(Cy^* - y_d)$. This gives

$$|\lambda^*|_{H^2} \leq K\left(|\lambda^*|_{L^2} + |f'(y^*)\lambda^*|_{L^2} + |C^*(Cy^* - y_d)|_{L^2}\right)$$

for a constant K (depending on Ω but independent of y^*). Since

$$|f'(y^*)\lambda^*|_{L^2} \leq |f'(y^*)|_{L^s}|\lambda^*|_{H^1},$$

where

$$s = \begin{cases} 2+\varepsilon & \text{if } n = 2, \\ n & \text{if } n \geq 3, \end{cases}$$

the desired result follows from (3.7).

To calculate the second Fréchet derivative we shall use

$$\text{(h3)} \qquad \begin{cases} f''(y^*) \in L^{1+\varepsilon}(\Omega) & \text{for some } \varepsilon > 0 \text{ if } n = 2, \\ f''(y^*) \in L^{\frac{2n}{6-n}}(\Omega) & \text{for } 3 \leq n \leq 6. \end{cases}$$

PROPOSITION 3.4. *Let* (H1), (H2) *and* (h1), (h3) *hold. Then*

$$\mathcal{L}''(y^*, g^*, \lambda^*)((v,h),(v,h)) = |Cv|_Z^2 + \alpha |h|_{L^2(\Gamma_1)}^2 + \langle \lambda^*, f''(y^*)v^2 \rangle_\Omega, \tag{3.8}$$

for all $(v,h) \in X$.

Proof. By Sobolev's embedding theorem there exists a constant K_e such that

$$|\langle \lambda^*, f''(y^*)v^2 \rangle_\Omega| \leq K_e |\lambda^*|_{H^1} |f''(y^*)|_{L^q} |v|_{H^1}^2 \tag{3.9}$$

for all $v \in H^1(\Omega)$, where

$$q = \begin{cases} 1+\varepsilon, & \varepsilon > 0, \text{ if } n = 2, \\ \dfrac{2n}{6-n} & \text{if } 3 \leq n \leq 6. \end{cases} \tag{3.10}$$

Referring back to the proof of Proposition 3.1 we see that the claim easily follows.

4. The second-order sufficient optimality condition. We turn now to an analysis of the second-order sufficient optimality condition (H3) for the optimal control problem (P). In view of (3.8) the crucial term is given by $\langle \lambda^*, f''(y^*)v^2 \rangle_\Omega$. Two types of results will be given. The first class of results will guarantee that $|\langle \lambda^*, f''(y^*)v^2 \rangle_\Omega|$ is small. This can be achieved by guaranteeing that λ^*, or, in view of (2.1), that $F'(x^*)$ is small. We may refer to this type of assumption as a small residual problem. The second class of assumptions rests on guaranteeing that $\lambda^* f''(y^*) \geq 0$ on Ω.

In the statement of Theorems 4.1 and 4.2 we use K_e and q which are defined in (3.9), (3.10). Further $\|B^{-1}\|$ denotes the norm of B^{-1} as operator from $H^1(\Omega)^*$ to $H^1(\Omega)$.

THEOREM 4.1. *Let* (H1), (H2), (h1), (h3) *hold.*

(i) *If* $Z = H^1(\Omega)$, $C = \mathrm{id}$, *and*

$$K_e K(x^*) |(y^* - y_d, \alpha g^*)|_X |f''(y^*)|_{L^q} < 1, \tag{4.1}$$

then the second-order sufficient optimality condition (H3) *holds.*

(ii) *If* $Z = L^2(\Omega)$, C *is the injection of* $H^1(\Omega)$ *into* $L^2(\Omega)$, $n \leq 3$, *and if in addition* (h2) *holds and*

$$\tilde{k}_e K(y^*) |y^* - y_d|_{L^2(\Omega)} |f''(y^*)|_{L^\infty(\Omega)} \leq 1, \tag{4.2}$$

where $\tilde{k}_e$ *is the embedding constant of* $H^2(\Omega)$ *into* $L^\infty(\Omega)$, *then* (H3) *is satisfied.*

Proof. (i) By (3.8) and (3.9) we have for every $(v,h) \in X$

$$\begin{aligned} &\mathcal{L}''(y^*, g^*, \lambda^*)((v,h),(v,h)) \\ &\quad \geq |v|_{H^1(\Omega)}^2 + \alpha |h|_{L^2(\Gamma_1)}^2 - K_e |\lambda^*|_{H^1(\Omega)} |f''(y^*)|_{L^q(\Omega)} |v|_{H^1(\Omega)}^2 \\ &\quad \geq \left(1 - K_e K(x^*) |(y^* - y_d, \alpha g^*)|_X |f''(y^*)|_{L^q(\Omega)}\right) |v|_{H^1}^2 + \alpha |h|_{L^2(\Gamma_1)}^2, \end{aligned}$$

where in the last estimate we used Corollary 3.3(i). The claim now follows from (4.1). We observe that in this case $\mathcal{L}''(y^*, g^*, \lambda^*)$ is positive definite on all X, not only on $\ker e'(x^*)$.

(ii) By (3.3) and (h2) we obtain

$$|v|_{H^1(\Omega)} \le \|B^{-1}\| \|\tau_{\Gamma_1}\| |h|_{L^2(\Gamma_1)} \quad \text{for all } (v, h) \in \ker e'(x^*). \tag{4.3}$$

Here $\|\tau_{\Gamma_1}\|$ denotes the norm of the trace operator from $H^1(\Omega)$ onto $L^2(\Gamma_1)$. Hence by Corollary 3.3(ii) and (4.3), we find for every $(v, h) \in \ker e'(x^*)$

$$\begin{aligned}
\mathcal{L}''(v^*, g^*, \lambda^*)((v, h), (v, h)) &= |v|^2_{L^2(\Omega)} + \alpha |h|^2_{L^2(\Gamma_1)} - \langle \lambda^*, f''(y^*)v^2 \rangle_\Omega \\
&\ge |v|^2_{L^2(\Omega)} + \alpha |h|^2_{L^2(\Gamma_1)} - |\lambda^*|_{L^\infty(\Omega)} |f''(y^*)|_{L^\infty(\Omega)} |v|^2_{L^2(\Omega)} \\
&\ge \left[1 - \tilde{k}_e K(y^*) |f''(y^*)|_{L^\infty(\Omega)} |y^* - y_d|_{L^2(\Omega)}\right] |v|^2_{L^2(\Omega)} \\
&\quad + \frac{\alpha}{2} \left(|h|^2_{L^2(\Gamma_1)} + \frac{1}{\|B^{-1}\|^2 \|\tau_{\Gamma_1}\|^2} |v|^2_{H^1(\Omega)} \right).
\end{aligned}$$

Due to (4.2) the expression in brackets is nonnegative and the result follows.

In the following result, C can be a boundary observation operator, for example.

THEOREM 4.2. *Let* (H1), (h1)–(h3) *hold, and let* $\|\tau_{\Gamma_1}\|$ *be the norm of the trace operator from* $H^1(\Omega)$ *onto* $L^2(\Gamma_1)$. *Then*

$$2\|B^{-1}\| \|\tau_{\Gamma_1}\| K_{y^*} |C^*(Cy^* - y_d)|_{(H^1)^*} < \alpha \tag{4.4}$$

implies that the second-order sufficient optimality condition (H3) *is satisfied, where* K_{y^*} *is the constant appearing in* (3.7).

Proof. As in the proof of Theorem 4.1(i) we find

$$\begin{aligned}
\mathcal{L}''(y^*, g^*, \lambda^*)((v, h), (v, h)) &\ge |Cv|^2_Z + \alpha |h|^2_{L^2(\Gamma_1)} - K_e |\lambda^*|_{H^1(\Omega)} |f''|_{L^q(\Omega)} |v|^2_{H^1(\Omega)} \\
&\ge |Cv|^2_Z + \alpha |h|^2_{L^2(\Gamma_1)} - K_{y^*} |C^*(Cy^* - y_d)|_{(H^1)^*} |f''|_{L^q(\Omega)} |v|^2_{H^1(\Omega)},
\end{aligned}$$

where (3.7) was used. For $(v, h) \in \ker e'(x^*)$ this implies by (4.3)

$$\begin{aligned}
\mathcal{L}''(y^*, g^*, \lambda^*)((v, h), (v, h)) &\ge |Cv|^2_Z + \frac{\alpha}{2} |h|^2_{L^2(\Gamma_1)} \\
&\quad + \left[\frac{\alpha}{2\|B^{-1}\| \|\tau_{\Gamma_1}\|} - K_{y^*} |C^*(Cy^* - y_d)|_{(H^1)^*} |f''|_{L^q(\Omega)} \right] |v|^2_{H^1(\Omega)}.
\end{aligned}$$

The desired result follows from (4.4).

In view of (4.1), (4.2), (4.4), and the fact that $|y^* - y_d|$ is decreasing with $\alpha \to 0^+$, the question arises as to whether by decreasing α we can always verify that the second-order sufficient optimality condition holds. The answer to this question is not obvious since the term $|y^* - y_d|$ in (4.1), (4.2), (4.4) is multiplied by factors which depend on x^* and hence on α themselves.

We next analyze one specific situation for which the second-order sufficient optimality condition holds for all α which are sufficiently small. In continuation of Theorem 4.1(i) we consider the specific situation where

$$Z = L^2(\Omega), \ C \text{ is the embedding of } H^1(\Omega) \text{ into } L^2(\Omega), \text{ and } n = 2. \tag{4.5}$$

In the following Lemma 4.3 and Theorem 4.4 we shall write (P_α) instead of (P) and denote by $x^\alpha = (y^\alpha, g^\alpha)$ a solution to

$$(P_\alpha) \qquad \begin{cases} \min \dfrac{1}{2} |y - y_d|^2_{L^2(\Omega)} + \dfrac{\alpha}{2} |g|^2_{L^2(\Gamma_1)}, \\ \text{subject to } (y, g) \in H^1(\Omega) \times L^2(\Gamma_1) \text{ a solution of (3.1).} \end{cases}$$

We assume

$$\text{(h4)} \quad \begin{cases} \text{(i) there exists } a \in \mathbb{R} \text{ such that } f(t)t \geq at^2 \text{ for all } t \in \mathbb{R} \text{ and} \\ \quad y_n \rightharpoonup y \text{ in } H^1(\Omega) \text{ implies } f(y_n) \rightharpoonup f(y) \text{ in } L^2(\Omega); \\ \text{(ii) } f \text{ maps bounded set in } H^1(\Omega) \text{ into bounded sets in } L^{4/3}(\Omega). \end{cases}$$

LEMMA 4.3. *If* (3.1) *has a solution* $(\bar{y}, \bar{g})$ *and if* (h4(i)) *holds, then there exists a solution* $x^\alpha = (y^\alpha, g^\alpha)$ *of* (P_α) *for every* $\alpha > 0$.

Proof. Let $\alpha > 0$ and let (y_n, g_n) be a minimizing sequence of (P_α). Then

$$\lim_{n\to\infty} \left(|y_n - y_\alpha|^2_{L^2(\Omega)} + \alpha |g_n|^2_{L^2(\Gamma_1)} \right) \leq |\bar{y} - y_d|^2_{L^2(\Omega)} + \alpha |\bar{g}|^2_{L^2(\Gamma_1)},$$

and hence $\{(y_n, g_n)\}_{n=1}^\infty$ is bounded in $L^2(\Omega) \times L^2(\Gamma_1)$. By (3.2) and (h4(i)) we find that $\{(y_n, g_n)\}_{n=1}^\infty$ is bounded in $H^1(\Omega) \times L^2(\Gamma_1)$. Consequently there exists a subsequence of $\{(y_n, g_n)\}_{n=1}^\infty$, denoted by the same symbol, and (y^α, g^α) with $(y_n, g_n) \to (y^\alpha, g^\alpha)$ weakly in $H^1(\Omega) \times L^2(\Gamma_1)$. Due to (h4(i)) and weak lower semicontinuity of norms, (y^α, g^α) is a solution of (P_α).

For the next theorem we require additional hypotheses.

$$\text{(h5)} \quad \begin{cases} \text{There exists a solution } (y^0, g^0) \text{ of (3.1) with the property that} \\ |y^0 - y_d|_{L^2(\Omega)} \leq |y - y_d|_{L^2(\Omega)}, \\ \text{where } y \text{ is the first coordinate of any solution } (y, g) \text{ of (3.1).} \end{cases}$$

Under the conditions of Lemma 4.3 let $B_\alpha : H^1(\Omega) \to H^1(\Omega)^*$ denote the differential operators characterzied by $B_\alpha v = \varphi$ where

$$\langle \nabla v, \nabla \psi \rangle_\Omega + \langle f'(y^\alpha)v, \psi \rangle_\Omega = \langle \varphi, \psi \rangle_{(H^1)^*, H^1} \quad \text{for all } \psi \in H^1(\Omega).$$

$$\text{(h6)} \quad \begin{cases} \text{There exists } \bar{\alpha} > 0 \text{ and } \bar{b} \text{ such that } B_\alpha \text{ is surjective} \\ \text{and } \|B_\alpha^{-1}\| \leq \bar{b} \text{ for every } \alpha \in [0, \bar{\alpha}]. \end{cases}$$

A sufficient condition that implies (h6) is given in Remark 4.5 below.

THEOREM 4.4. *Let* (4.5), (h4), (h5), *and* (h6) *hold, and assume that* $n = 2$ *and* Γ *is* $C^{1,1}$ *smooth. Then there exists* $\hat{\alpha} \in (0, \bar{\alpha}]$ *such that the second-order sufficient optimality condition holds for* (P_α) *for all* $\alpha \in (0, \hat{\alpha}]$.

Proof. Due to Lemma 4.3 and (h5) there exists a solution $x^\alpha = (y^\alpha, g^\alpha)$ of (P_α) for every $\alpha > 0$. By (h5)

$$\begin{aligned} |y^0 - y_d|^2_{L^2(\Omega)} + \alpha |g^\alpha|^2_{L^2(\Gamma_1)} &\leq |y^\alpha - y_d|^2_{L^2(\Omega)} + \alpha |g^\alpha|^2_{L^2(\Gamma_1)} \\ &\leq |y^0 - y_d|^2_{L^2(\Omega)} + \alpha |g^0|^2_{L^2(\Gamma_1)}, \end{aligned}$$

which implies

$$|g^\alpha|_{L^2(\Gamma_1)} \leq |g^0|_{L^2(\Gamma_1)}$$

and

$$|y^\alpha - y_d|^2_{L^2(\Omega)} \leq |y^0 - y_d|^2_{L^2(\Omega)} + \alpha \left(|g^0|^2_{L^2(\Gamma_1)} - |g^\alpha|^2_{L^2(\Gamma_1)} \right).$$

Hence $\{(y^\alpha, g^\alpha)\}_{\alpha>0}$ is bounded in $L^2(\Omega) \times L^2(\Gamma_1)$. From (h4(i)) and (3.2) it follows that $\{(y^\alpha, g^\alpha)\}_{\alpha>0}$ is also bounded in $H^1(\Omega) \times L^2(\Gamma_1)$.

Next we argue that there exists a constant K_1 such that

$$|y^\alpha|_{W^{1,4}(\Omega)} \leq K_1 \quad \text{for all } \alpha \in [0, \bar{\alpha}]. \tag{4.6}$$

In fact, since $n = 2$, the functionals $\varphi \to \tilde{G}_\alpha(\varphi) = \int_\Gamma \tilde{g}_\alpha \varphi ds$, with

$$\tilde{g}_\alpha = \begin{cases} g_\alpha & \text{on } \Gamma_1, \\ g_2 & \text{on } \Gamma_2, \end{cases}$$

are elements of $W^{1,4/3}(\Omega)^*$ [Tr, p. 72], and there exist constants K_2 and K_3 such that

$$\|G_\alpha\|_{(W^{1,4/3})^*} \leq K_2 |\tilde{g}_\alpha|_{L^2(\Gamma)} \leq K_3 \quad \text{for all } \alpha \in [0, \bar{\alpha}]. \tag{4.7}$$

The functions y^α satisfy

$$\langle \nabla y^\alpha, \nabla \varphi \rangle_\Omega + \langle y^\alpha, \varphi \rangle_\Omega = \langle \tilde{h}, \varphi \rangle_\Omega + \langle \tilde{g}_\alpha, \varphi \rangle_\Gamma + \langle y_\alpha, \varphi \rangle_\Omega + \langle f(y_\alpha), \varphi \rangle_\Omega. \tag{4.8}$$

Due to (h4(ii)) and (4.7) the right-hand side in (4.8) describes a family (in α) of bounded linear functionals on $W^{1,4/3}(\Omega)$. Hence there exists K_1 such that (4.6) holds [Tr, p. 179]. From continuity of f' from $\mathbb{R}$ to $\mathbb{R}$ and the fact that $W^{1,4}(\Omega)$ is continuously embedded in $C(\Omega)$ for $n = 2$, it follows that

$$|f'(y^\alpha)\lambda|_{L^2(\Omega)} \leq M |\lambda|_{L^2(\Omega)}$$

for a constant M independent of $\alpha \in [0, \bar{\alpha}]$ and $\lambda \in L^2(\Omega)$. Therefore $B_\alpha \in \mathcal{L}(H^1(\Omega), (H^1(\Omega))^*)$ and (h6) is applicable.

Due to (h6) there exists a Lagrange multiplier λ_α satisfying the variational form of

$$\begin{aligned} -\Delta \lambda^\alpha + f'(y^\alpha)\lambda^\alpha &= -(y^\alpha - y_d) \quad \text{in } \Omega, \\ \frac{\partial \lambda^\alpha}{\partial n} &= 0 \quad \text{on } \Gamma \end{aligned}$$

and

$$|\lambda^\alpha|_{H^1(\Omega)} \leq \bar{b} |y^\alpha - y_d|_{(H^1)^*} \tag{4.9}$$

for every $\alpha \in [0, \bar{\alpha}]$. As in the proof of Corollary 3.3 we show that

$$|\lambda^\alpha|_{H^2(\Omega)} \leq K_4 |y^\alpha - y^d|_{L^2(\Omega)}$$

for a constant K_4 independent of $\alpha \in [0, \bar{\alpha}]$.

With these preliminaries we find for the Hessian

$$\begin{aligned} \mathcal{L}''(y^\alpha, g^\alpha, \lambda^\alpha)((v, h), (v, h)) &= |v|^2_{L^2(\Omega)} + \alpha |h|^2_{L^2(\Gamma_1)} + \langle \lambda^\alpha, f''(y^\alpha) v^2 \rangle_\Omega \\ &\geq |v|^2_{L^2(\Omega)} + \alpha |h|^2_{L^2(\Gamma_1)} - |\lambda^\alpha|_{L^\infty(\Omega)} |f''(y^\alpha)|_{L^\infty(\Omega)} |v|^2_{L^2(\Omega)} \\ &\geq (1 - \tilde{k}_e K_4 |y^\alpha - y_d|_{L^2(\Omega)} |f''(y^\alpha)|_{L^\infty(\Omega)}) |v|^2_{L^2(\Omega)} + \alpha |h|^2_{L^2(\Gamma_1)}, \end{aligned}$$

where $\tilde{k}_e$ denotes the embedding constant of $H^2(\Omega)$ into $L^\infty(\Omega)$. Since $f \in C^2(\mathbb{R})$ by assumption, (4.6) implies the existence of K_5 such that $|f''(y^\alpha)|_{L^\infty(\Omega)} \leq K_5$. Combining these estimates with (h6) we find

$$\begin{aligned} \mathcal{L}''(y^\alpha, g^\alpha, \lambda^\alpha)((v, h), (v, h)) \geq (1 - \tilde{k}_e K_4 K_5 |y^\alpha - y_d|_{L^2(\Omega)}) |v|^2_{L^2(\Omega)} \\ + \frac{\alpha}{2} \left(|h|^2_{L^2(\Gamma_1)} + \frac{1}{\bar{b}^2 \|\tau_{\Gamma_1}\|^2} |v|^2_{H^1(\Omega)} \right) \end{aligned}$$

for all $(v, h) \in \ker e'(x^\alpha)$. Finally

$$|y^\alpha - y_d|^2_{L^2(\Omega)} \leq |y^0 - y_d|^2_{L^2(\Omega)} + \alpha(|g^0|^2_{L^2(\Gamma_1)} - |g^\alpha|^2_{L^2(\Gamma_1)})$$

and $|g^\alpha|_{L^2(\Gamma_1)} \leq |g^0|_{L^2(\Gamma_1)}$, and therefore

$$\mathcal{L}''(y^\alpha, g^\alpha, \lambda^\alpha)((v, h), (v, h)) \geq \frac{\alpha}{2}\left(|h|^2_{L^2(\Gamma_1)} + \frac{1}{\tilde{b}^2\|\tau_{\Gamma_1}\|^2}|v|^2_{H^1(\Omega)}\right)$$

$$+\left[1 - \tilde{k}_e K_4 K_5\left(|y^0 - y_d|_{L^2(\Omega)} + \sqrt{\alpha}\sqrt{|g^0|^2_{L^2(\Gamma_1)} - |g^\alpha|^2_{L^2(\Gamma_1)}}\right)\right]|v|^2_{L^2(\Omega)}$$

for all $(v, h) \in \ker e'(x^\alpha)$. The conclusion follows from this estimate.

Remark 4.5. Let the assumptions of Theorem 4.4 except possibly (h6) hold, and assume that (y^0, g^0) in (h5) is unique. Then, using (4.6) we can argue that $(y^\alpha, g^\alpha) \to (y^0, g^0)$ weakly in $W^{1,4}(\Omega) \times L^2(\Gamma_1)$ as $\alpha \to 0^+$. If moreover $f'(y^0) \geq 2\beta > 0$ on Ω, then there exists $\bar{\alpha} > 0$ such that $f'(y^\alpha) \geq \beta$ on Ω for all $\alpha \in [0, \bar{\alpha}]$. By the Lax–Milgram lemma, (h6) follows.

In Theorems 4.1, 4.2, and 4.4 the second-order optimality condition was guaranteed by assuring that the term $\langle \lambda^*, f''(y^*)v^2\rangle_\Omega$ is small when compared to $|Cv|^2_Z + \alpha|h|^2_{L^2(\Gamma_1)}$. Alternatively we can proceed by assuming that

(h7) $$\lambda^* f''(y^*) \geq 0 \quad \text{a.e. on } \Omega.$$

THEOREM 4.6. *Assume that* (H1), (H2), (h1), (h3), *and* (h7) *hold and that*

(a) $Z = H^1(\Omega)$, *and* $C = \mathrm{id}$, *or*

(b) (h2) *is satisfied.*

Then (H3) *holds.*

Proof. By Proposition 3.4 and (h7) we find for all $(v, h) \in X$

$$\mathcal{L}''(y^*, g^*, \lambda^*)((v, h), (v, h)) \geq |Cv|^2_Z + \alpha|h|^2_{L^2(\Gamma_1)}.$$

In the case where (a) holds, the conclusion is obvious and $\mathcal{L}''(v^*, g^*, \lambda^*)$ is positive not only on $\ker e'(x^*)$ but also on all of X. In case (b) we use (4.3) to conclude that

$$\mathcal{L}''(y^*, g^*, \lambda^*)((v, h), (v, h)) \geq |Cv|^2_Z + \frac{\alpha}{2}\left(|h|^2_{L^2(\Gamma_1)} + \frac{1}{\|B^{-1}\|^2\|\tau_{\Gamma_1}\|^2}|v|^2_{H^1(\Omega)}\right)$$

for all $(v, h) \in \ker e'(x^*)$.

Next we give a sufficient condition for (h7).

THEOREM 4.7. *Let* (H1), (H2), (h1) *hold and assume that*

(i) $\langle C^*(y_d - Cy^*), \psi\rangle_{(H^1)^*, H^1} \geq 0$ *for all* $\psi \in H^1(\Omega)$ *with* $\psi \geq 0$, *a.e.*,

(ii) $f'(y^*) \geq 0$ *a.e. on* Ω,

(iii) $f''(y^*) \geq 0$ *a.e. on* Ω.

Then (h7) *holds. The conclusion remains correct if the inequalities in* (i) *and* (iii) *are reversed.*

Proof. Set $\varphi = \inf(0, \lambda^*) \in H^1(\Omega)$ in (3.5). Then we have

$$\int_\Omega |\nabla\varphi|^2 dx + (f'(y^*)\varphi, \varphi) + \langle c^*(Cy^* - y_d), \varphi\rangle_{(H^1)^* \times H^1} = 0.$$

Since $f'(y^*) \geq 0$ it follows from (i) that $|\nabla\varphi|^2_{L^2(\Omega)} = 0$ and $\lambda^* \geq 0$. Together with (iii) we find $\lambda^* f''(y^*) \geq 0$ a.e. on Ω. If the inequalities in (i) and (iii) are reversed, we take $\varphi = \sup(0, \lambda^*)$.

Example 4.8. We consider

$$\begin{aligned} -\Delta y + y^3 - y &= \tilde{h} \quad \text{in } \Omega, \\ \frac{\partial y}{\partial n} &= g \quad \text{on } \Gamma, \end{aligned} \tag{4.10}$$

and the associated optimal control problem

$$\begin{cases} \min \dfrac{1}{2}\displaystyle\int_\Omega |y - y_d|^2 dx + \frac{\alpha}{2}\int_\Gamma g^2 dx, \\ \text{subject to } (y, g) \in H^1(\Omega) \times L^2(\Gamma) \text{ a solution of (4.10).} \end{cases} \tag{4.11}$$

Here Ω is a bounded domain in $\mathbb{R}^n$, $n \leq 3$, satisfying the general requirements specified at the beginning of this section. In the present example $\Gamma_1 = \Gamma = \partial\Omega$, $\Gamma_2 = \emptyset$, and consequently $\tilde{g} = g$. In the context of the general theory we put

$$Z = L^2(\Omega), \quad C : H^1(\Omega) \to L^2(\Omega) \text{ canonical injection}, \qquad f(t) = t^3 - t.$$

Equation (4.10) represents a simplified Ginzburg-Landau model for superconductivity with y denoting the wave function, which is valid in the absence of internal magnetic fields [T, Chaps. 1, 4]. Both (4.10) and

$$-\Delta y + y^3 + y = \tilde{h} \tag{4.12}$$

are of interest for superconductivity, but we prefer to concentrate on (4.10) here rather than on (4.12), since the former has three equilibria while the latter has only one equilibrium.

The optimal control problem (4.11) with a slightly more general version of (4.10) (i.e., (4.12)) was studied in [GHS]. The numerical approach in that paper is based on solving the optimality system associated with (4.11), which involves solving (4.10) and the costate equation.

Let us discuss the validity of some of the conditions (hi) and (Hi) for the present problem. Existence of a minimum $x^* = (y^*, g^*)$ is proved in [GHS] for any $\alpha > 0$. It is also proved there (using $n \leq 3$) that $e'(x^*)$ is surjective and thus (H1) and (H2) hold. Sobolev's embedding theorem and

$$f'(t) = 3t^2 - 1 \quad \text{and} \quad f''(t) = 6t$$

imply that (h1), (h3), and (h4) hold. Let us note that f has the three equilibria ± 1 and 0, of which ± 1 are stable and 0 is unstable. It is therefore quite reasonable to conjecture that in general for $\tilde{h} = 0$, $y_d \geq 1$ implies $1 \leq y^* \leq y_d$ and, similarly, that $y_d \leq -1$ implies $y_d \leq y^* \leq -1$. This was confirmed numerically. In these cases $f'(y^*) \geq \underline{\beta} > 0$ and (i)-(iii) of Theorem 4.7 hold. The situation is more delicate if $-1 \leq y_d \leq 1$ and the (numerical) solution y^* depends qualitatively more significantly on the cost α of the control. We shall elaborate further on this point when discussing numerical examples below.

Example 4.9. Here we investigate

$$\begin{aligned} -\Delta y + y^3 - y &= \tilde{h} \quad \text{in } \Omega, \\ \frac{\partial y}{\partial n} &= g \quad \text{on } \Gamma_1, \\ \frac{\partial y}{\partial n} &= 0 \quad \text{on } \Gamma_2, \end{aligned} \tag{4.13}$$

and the associated optimal control problem is given by

$$\text{(4.14)} \qquad \begin{cases} \min \dfrac{1}{2}\displaystyle\int_{\Gamma_3} |y - y_d|^2 ds + \dfrac{\alpha}{2}\int_{\Gamma_1} g^2 ds, \\ \text{subject to } (y, g) \in H^1(\Omega) \times L^2(\Gamma_1) \text{ a solution of (4.13),} \end{cases}$$

where Γ_3 is a nontrivial measurable subset of Γ_2 and Ω satisfies the properties specified at the beginning of this section. Moreover

$$Z = L^2(\Gamma_3), \qquad C \text{ is the trace operator from } H^1(\Omega) \text{ onto } L^2(\Gamma_3),$$

and f is as in Example 4.8. To argue existence of a solution to (4.14), first note that (4.13) is satisfied for at least one pair $(y, g) \in H^1(\Omega) \times L^2(\Gamma_1)$, which renders the cost functional in (4.14) finite. Hence there exists a minimizing sequence (y_n, g_n) in $H^1(\Omega) \times L^2(\Gamma_1)$. Using (4.13) it is simple to deduce that (y_n, g_n) is bounded in $H^1(\Omega) \times L^2(\Gamma_1)$ and every weak subsequential limit is a solution to (4.14). Thus (H1) holds. Concerning the remaining hypotheses, the same remarks apply to (4.14) as to (4.11).

Example 4.10. This is the singular system

$$\text{(4.15)} \qquad \begin{aligned} -\Delta y - y^3 &= \tilde{h} \quad \text{in } \Omega \\ \frac{\partial y}{\partial n} &= g \quad \text{on } \Gamma, \end{aligned}$$

and the associated optimal control problem

$$\text{(4.16)} \qquad \begin{cases} \min \dfrac{1}{2}|y - y_d|^2_{H^1(\Omega)} + \dfrac{\alpha}{2}\displaystyle\int_{\Gamma} g^2 ds, \\ \text{subject to } (y, g) \in H^1(\Omega) \times L^2(\Gamma_1) \text{ a solution of (4.15),} \end{cases}$$

where $y_d \in H^1(\Omega)$. If (4.15) admits at least one feasible pair (y, g), then it is simple to argue that (4.16) has a solution $x^* = (y^*, g^*)$. (We refer to [L, Chap. 3] for existence results in the case where the cost functional is of the form $|y - y_d|^r_{L^r(\Omega)} + \alpha|g|^2_{L^2(\Gamma)}$ for appropriately chosen $r > 2$.) The existence of a Lagrange multiplier is assured in the same manner as in Example 4.8. Clearly (h1) and (h3) are satisfied. For $y_d = \text{ const } \geq \frac{1}{2}$, we observed that $0 \leq y^* \leq y_d$, $\lambda^* < 0$, which in view of (h7) and Theorem 4.6 explains the second-order convergence rate that is observed numerically.

5. Numerical tests. We carried out numerous tests for Examples 4.8–4.10 and focused our interest on the following aspects:

- convergence for wide ranges of initial values and values for the penalty parameter c (no globalization, e.g., by line search),
- rate of convergence,
- how well the desired state y_d can be reached and how strongly this depends on α,
- influence of equilibria.

All of the tests were carried out with $\Omega = [0, 1] \times [0, 2]$ with a five-point central finite difference discretization of the state equation (appropriately modified at the boundary so that the discretized system equation for (4.10) is symmetric) on the grid $\{(\frac{i}{N}, \frac{j}{N})\}_{i=0,\ldots,N}^{j=0,\ldots,2N}$. The programs were written in MATLAB. Convergence was achieved for all test examples for a wide range of values for c. For $c \in [10, 1000]$ and $\alpha \in [10^{-7}, 10^{-1}]$, the algorithm was quite insensitive with respect to the specific value of c. For $c \in [10^{-7}, 10]$ the number of iterations increased as c was decreased. For $c = 0$ divergence was observed in all runs. The case $c = 0$

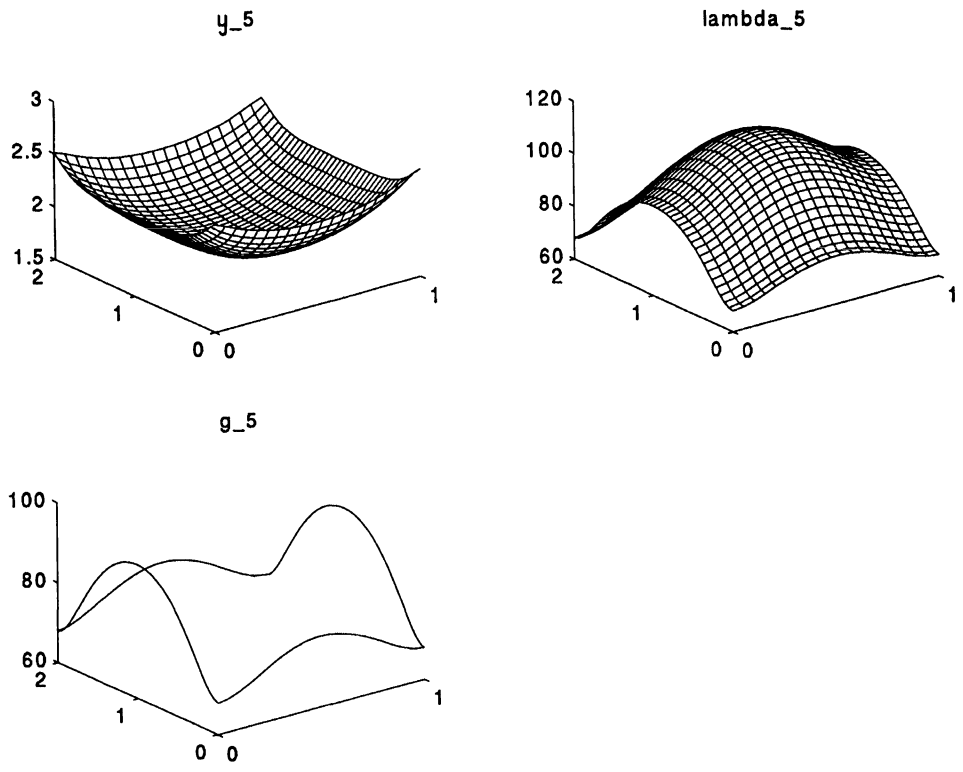


FIG. 5.1.

in Algorithm 2 corresponds to the SQP method without globalization strategy or, equivalently, to the Newton method applied to the first-order necessary optimality condition. Second-order rate of convergence was the typical behavior for all three examples with reasonable choices for α and c.

In all examples below we chose Algorithm 2, and $N = 20$, $\tilde{h} = 0$, $\lambda_0 = 0$ (start-up value for λ^*), $y_0 = y_d$ (start-up value for y^*), $g_0 = 0$ (start-up value for g^*), unless otherwise specified. We denote the converged numerical solution by (y^*, g^*, λ^*).

RUN 1. This is Example 4.8 with

$$c = 1,\ \alpha = 10^{-3},\ y_d = 3.$$

In Figure 5.1 we give the graph of the converged solution after five iterations. The values for

$$e_n = \frac{|y_{n+1} - y_n|_{H^1} + |g_{n+1} - g_n|_{L^2}}{|y_n - y_{n-1}|^2_{H^1} + |g_n - g_{n-1}|^2_{L^2}}$$

are given in Table 5.1. We observe that $1 \leq y^*(= y^5) \leq y_d$, i.e., the solution is attracted by the stable equilibrium. As a consequence, Theorem 3.4 applies. Moreover $\tau_\Gamma \lambda^5 = g^5$, which confirms (3.6) of Proposition 3.1.

Let us make some additional comments on further runs. If the desired state was chosen between the two equilibria 0 and 1 so that $0 \leq y_d \leq 1$, then the numerical solution satisfied $0 \leq y^* \leq 1$, as well. In these cases $\lambda^* < 0$ and Theorem 4.6 is not applicable. Nevertheless

TABLE 5.1.

n	1	2	3	4
e_n	.0080	.0019	.0001	.0181

TABLE 5.2.

n	1	2	3	4	5	6	7	8	9
$\hat{e}_n$	13.4	.023	.023	.023	.040	.029	.080	.020	4.25

very good convergence could always be achieved. For $y_d(x) = x, c = 10, \alpha = 1$, for example, we found $\{e_n\}_{n=1}^4 = (.0032, .015, 0.04, 6 * 10^{14})$. While typically convergence was achieved within the first ten iterations, we should also mention our worst result with

$$\tilde{h} = 100^3 - 100, \quad c = 1, \quad \alpha = 10^{-3}, \quad y_0 = 0, \quad g_0 = 1, \quad \lambda_0 = 0, \quad N = 10.$$

The exact solution is $(y^*, g^*, \lambda^*) = (100, 0, 0)$. Convergence was achieved after 294 iterations, and for

$$\tilde{e}_n = \frac{|y_n - y^*|_{H^1}}{|y_{n-1} - y^*|_{H^1}^2},$$

we found $(\tilde{e}_{291}, \ldots, \tilde{e}_{294}) = (.028, .011, .011, .10)$.

RUN 2. This is Example 4.9 with

$$c = 100, \ \alpha = 10^{-5}, \ y_d = 1, \ g_0 = 1, \ \Gamma_1 = (0, 1) \times \{0\}, \ \Gamma_3 = (0, 1) \times \{2\}. \tag{5.1}$$

The rates for

$$\hat{e}_n = \frac{|y_n - y^*|_{H^1} + |g_n - g^*|_{L^2}}{|y_{n-1} - y^*|_{H^1}^2 + |g_{n-1} - g^*|_{L^2}^2}$$

are given in Table 5.2. The exact solution is $(y^*, g^*) = (1, 0)$. We also give graphs for the evolution of y_i. Convergence is a little slower for small c and thus the graphs are more interesting. Figure 5.2 gives, therefore, the results for $c = 1$, with the remaining specifications being those of (5.1).

RUN 3. This is Example 4.9 with (5.1). Here, unlike in the previous run, we chose the value $y_0 = -2$ as the initial guess for y^*. In the course of the iteration y_i has to "pass through" the equilibria -1 and 0 to reach $y^* = y_d = 1$. Selected graphs of the sequence y_i are shown in Figure 5.3. Between iterations 9 and 10 the step over the unstable equilibrium occurs. This step was too large and due to the weak influence of the control which was applied at the ordinate value $y = 0$ on the state at ordinate value $y = 2$, it took about 16 additional iterations until the desired state y_d was reached ($|y_{36} - y^*|_{H^1(\Omega)} = 7 * 10^{-5}$). In Figure 5.4 we give a plot for the evolution of $|g_i|_{L^2(\Gamma_1)}$. We observe that recovery after passing through the unstable equilibrium is only possible with large values of the cost (for $i = 11, \ldots, 14$).

RUN 4. This is Example 4.9 with

$$c = 100, \quad \alpha = 10^{-5} (\alpha = 10^{-7}), \quad y_d = \frac{1}{2} + x, \quad y_0 = 1, \quad \Gamma_1, \Gamma_3 \text{ as in (5.1)}. \tag{5.2}$$

In Figure 5.5 we give the plots for y^6 with $\alpha = 10^{-5}$ and $\alpha = 10^{-7}$, respectively. Clearly the cheaper cost allows us to reach the desired state y_d better than the more expensive cost.

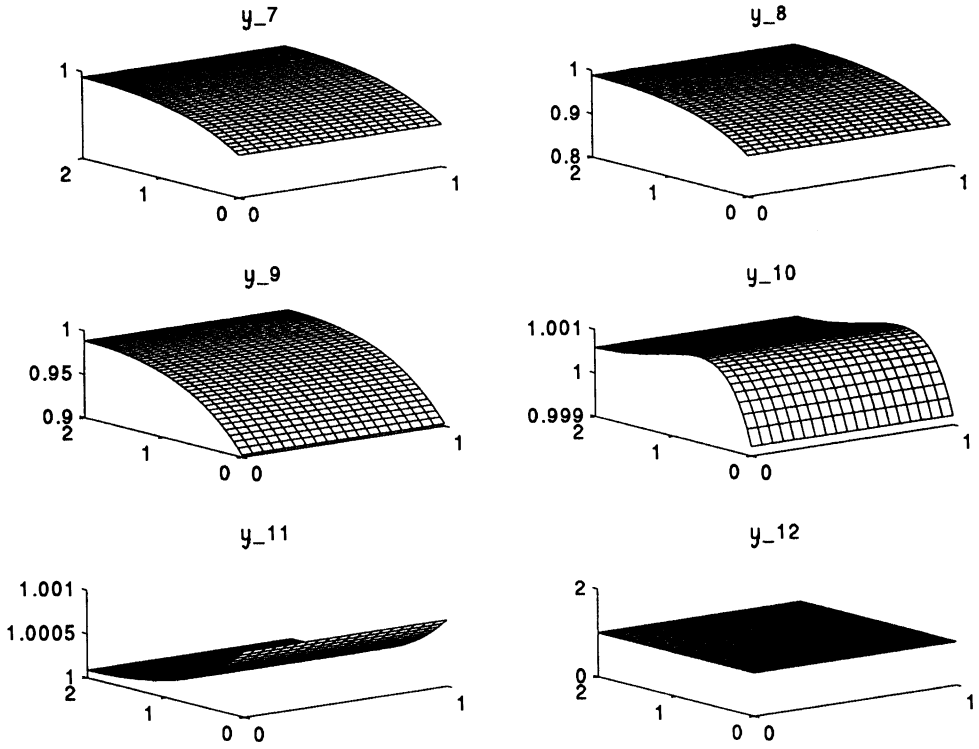


FIG. 5.2.

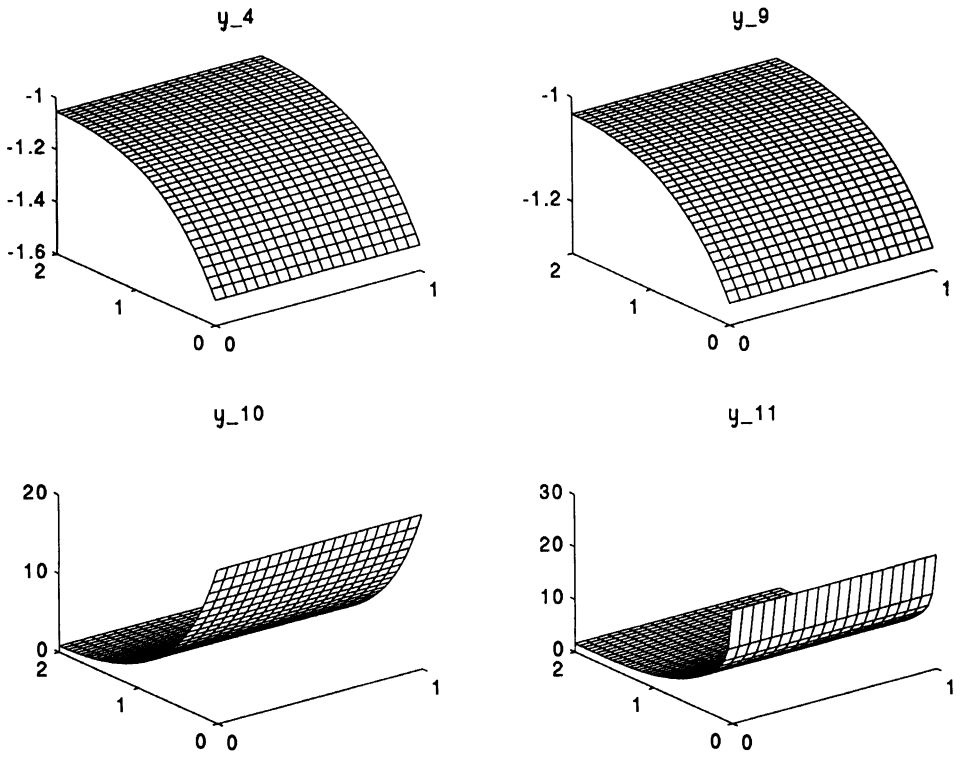


FIG. 5.3.

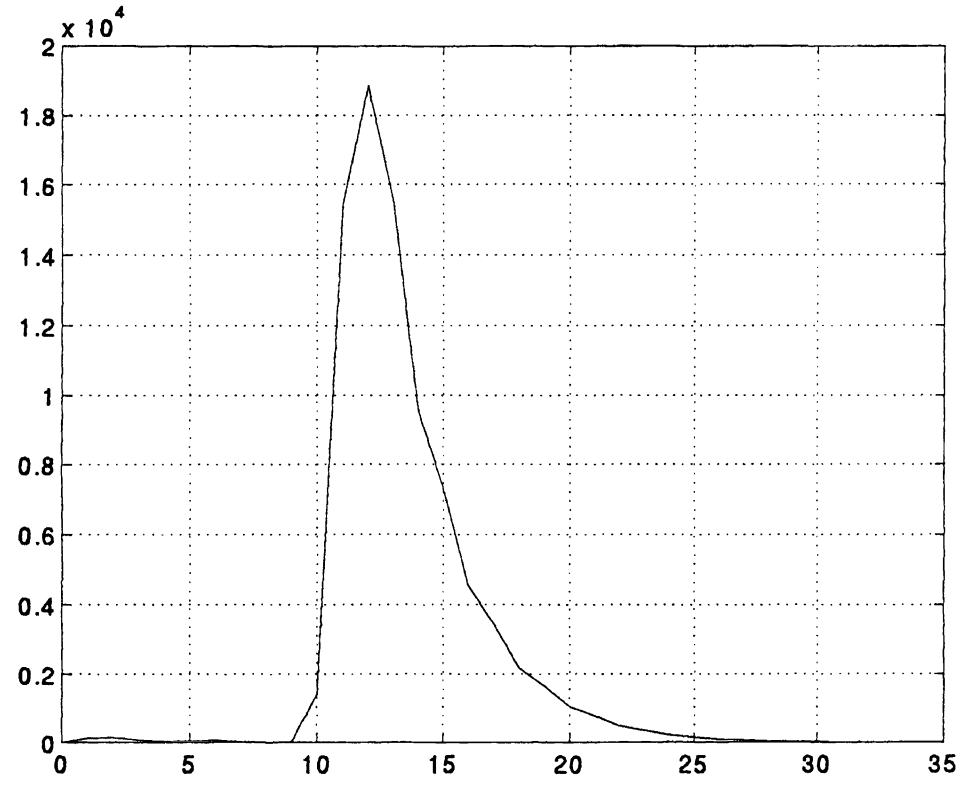


FIG. 5.4.

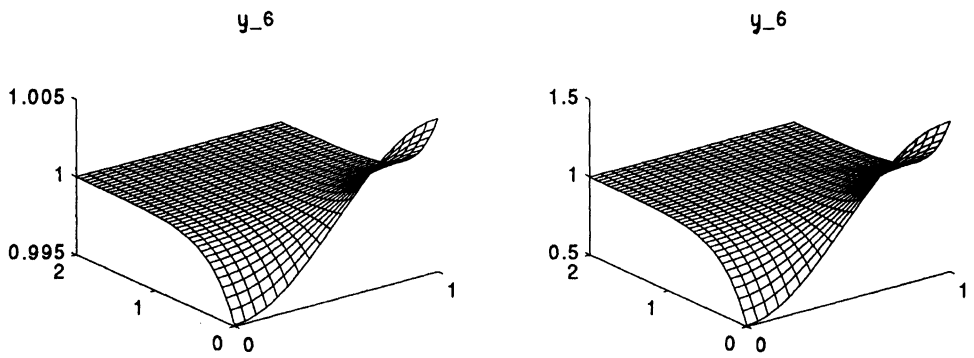


FIG. 5.5.

RUN 5. Here we consider Example 4.9 with a more challenging desired state:

$$N = 40,\ c = 20,\ \alpha = 10^{-7},\ y_d = \frac{1}{2}\sin\left(\frac{3\pi}{2}x\right). \tag{5.3}$$

For the convergence rate we found the results of Table 5.3.

RUN 6. This is the singular control system of Example 4.10 with

$$c = 10,\ \alpha = 10^{-2},\ y_d = 1 + xy,\ y_0 = 0. \tag{5.4}$$

The results after 10 iterations are given in Figure 5.6. We ran other tests with $y_d = \text{const}$ (and, e.g., $c = 10$, $\alpha = .1$) and observed that for const $> 0 (< 0)$ the resulting Lagrange parameter

TABLE 5.3.

n	1	2	3	4	5
$e(n)$	.0029	.0041	.0051	.0050	.0046

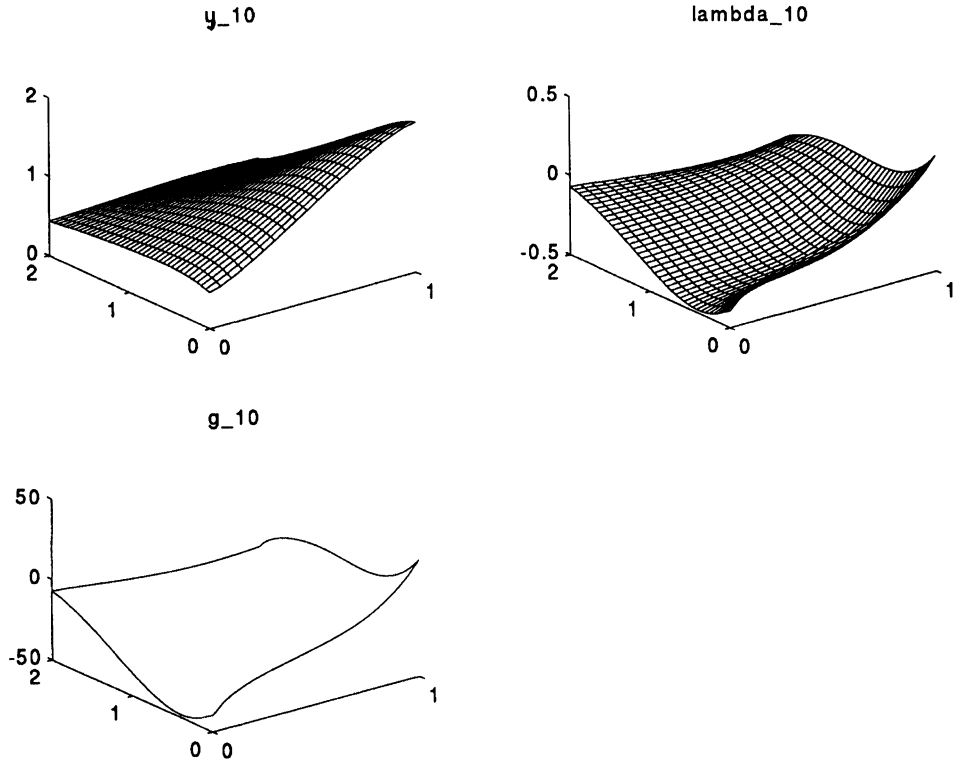


FIG. 5.6.

satisfies $\lambda^* < 0 (> 0)$. If $y_d \geq .5 (\leq .5)$, then $0 \leq y^* \leq y_d$, $(y_d \leq y^* \leq 0)$, so Theorem 4.6 is applicable.

REFERENCES

[Be] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[CH] Z. CHEN AND K. H. HOFFMANN, *Numerical solutions of the optimal control problem governed by a phase field model*, in Control and Estimation of Distributed Parameter Systems, Internat. Series of Numer. Math. 100, Birkhäuser, Boston, 1991, pp. 79-97.

[G] P. GRIVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[GHS] M. D. GUNZBURGER, L. HOU, AND T. P. SVOBODNY, *Finite element approximations of an optimal control problem associated with the scalar Ginzburg-Landau equation*, Comput. Math. Appl., (1991), pp. 123-131.

[H] M. HEINKENSCHLOSS, *Numerical Solution of a Semilinear Parabolic Control Problem*, Virginia Poly. Inst. State University.

[IK] K. ITO AND K. KUNISCH, *Augmented Lagrangian-SQP-methods in Hilbert spaces and application to control in the coefficients problems*, SIAM J. Optim., 6 (1996), pp. 96-125.

[L] J. L. LIONS, *Control of Distributed Singular Systems*, Gauthier-Villars, Paris, 1985.

[T] M. TINKHAM, *Introduction to Superconductivity*, McGraw-Hill, New York, 1975.

[Tr] G. M. TROIANELLO, *Elliptic Differential Equations and Obstacle Problems*, Plenum Press, New York, 1987.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 892-912, May 1996

006

THE KORTEWEG-DE VRIES EQUATION ON A PERIODIC DOMAIN WITH SINGULAR-POINT DISSIPATION*

S. M. SUN†

Abstract. This paper considers the Korteweg-de Vries (KdV) equation

$$u_t + uu_x + u_{xxx} = 0, \quad 0 < x < 1, \ t > 0, \quad u(0, x) = u_0(x),$$

and the periodic boundary conditions $u(t, 1) = u(t, 0), u_{xx}(t, 0) = u_{xx}(t, 1)$ with an L^2-stabilizing control input implemented by a feedback mechanism $u_x(t, 1) = \alpha u_x(t, 0)$ and $|\alpha| < 1$. It can be shown that the solutions conserve the volume $[u] = \int_0^1 u(t, x)dx$ and the constant state $[u_0]$ possesses the smallest energy among solutions with same volume. It has been proved that the solution of the system exists and approaches $[u_0]$ as $t \to +\infty$ when $\alpha \neq -1/2$. This paper studies the case for $\alpha = -1/2$ and gives a proof of the existence and exponential decay of the solutions by deriving estimates of the corresponding Green's function and using semigroup theory. The method used here also works for the other cases with $|\alpha| < 1$.

Key words. KdV equation, point dissipation, stabilization

AMS subject classifications. 93D15, 93C20, 35Q53

1. Introduction.

Recently the Korteweg-de Vries (KdV) equation

$$u_t + \gamma uu_x + u_{xxx} = 0, \tag{1.1}$$

has been studied intensively in many papers. For $\gamma = 0$, this is a third-order linear dispersion equation and has been studied in [12]. The cases with $\gamma \neq 0$ are essentially equivalent and can be covered by letting $\gamma = 1$.

The literature for (1.1) both for x on a periodic domain and for a domain $-\infty < x < \infty$ is enormous, and the reader may be referred to [1, 2, 4, 5, 9, 15] for more details and to [7, 8, 12-14] for related control problems. Here we are considering the following initial boundary value problem:

$$\begin{aligned} &u_t + uu_x + u_{xxx} = 0, \quad 0 < x < 1, \quad t > 0, \\ &u(0, x) = \phi(x), \\ &u(t, 1) = u(t, 0), \quad u_x(t, 1) = \alpha u_x(t, 0), \quad u_{xx}(t, 1) = u_{xx}(t, 0), \end{aligned} \tag{1.2}$$

where $|\alpha| < 1$. By integrating the equation in (1.2) from zero to one and using the boundary conditions, it can be easily obtained that the volume $[u] = \int_0^1 u(t, x)dx$ of the solution $u(t, x)$ is conserved for $t \geq 0$, which is physically reasonable.

It is well known that the KdV equation (1.1) is a model equation for water waves in a channel using long-wave approximations. The equation is usually considered for $x \in (-\infty, +\infty)$. However, the fluid regions are always finite in physical applications and fluid-dynamical experiments as well as numerical computations. Therefore, it is more realistic to consider the KdV equation in a bounded region. The KdV equations with periodic boundary conditions are relevant to wave motions in a circular channel with a great radius. Also, the periodic boundary conditions are mostly used in experiments and numerical computations to find the properties of long waves on the free surfaces of channel flows, which are governed by the KdV equation. Furthermore, it is much easier to control the boundary conditions of the flows, such as by using wave-makers at boundaries, so that certain states of the flows can

*Received by the editors June 13, 1994; accepted for publication (in revised form) December 28, 1994.
†Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.

be reached. The boundary conditions in (1.2) can be viewed as a control mechanism of the vertical velocity of the fluid at boundary, which is intended to stabilize the flow.

In control theory, the system (1.2) is described as a closed-loop point dissipation process. Closed-loop control generally refers to control synthesis via state feedback of some sort and is predominantly concerned with achieving asymptotic stability of an equilibrium or invariant set. From an identity

$$\int_0^1 |u(t,x)|^2 dx = \int_0^1 |u(t,x) - [u]|^2 dx + \int_0^1 |[u]|^2 dx,$$

it shows that among all functions $u \in L^2(0,1)$ for which $[u]$ is a fixed constant, $\|u\|_{L^2(0,1)}$ is uniquely minimized by the constant function $u \equiv c = [u]$, which is an equilibrium state of (1.2). Therefore, a solution of (1.2) may be caused to approach the constant state $[u] = [\phi]$ as $t \to +\infty$ through use of a control process designed to make $\|u(t,\cdot)\|_{L^2(0,1)}$ nonincreasing. By multiplying the equation in (1.2) by $u(t,x)$, integrating it from zero to one, and using the boundary conditions again, we can obtain an identity for the solution of (1.2),

$$\frac{d}{dt}\left(\int_0^1 |u(t,x)|^2 dx\right) = (1/2)(\alpha^2 - 1)|u_x(t,0)|^2. \tag{1.3}$$

When $|\alpha| < 1$, the energy is decreasing unless $u_x(t,0) = u_x(t,1) \equiv 0$. Therefore, the boundary condition $u_x(t,1) = \alpha u_x(t,0)$ for $|\alpha| < 1$ can be viewed as a dissipation mechanism for the system (1.2). This condition was obtained by Russell and Zhang [13] using an L^2-stabilizing control input, implemented by a feedback mechanism. Because of the dissipation mechanism at the boundaries, it is reasonable to expect that the solution $u(t,x)$ of (1.2) approaches $[\phi(x)]$ as $t \to +\infty$.

The system (1.2) was first derived in [13] to study the smoothing properties of the solution of (1.2) with asymptotic decay properties in $L^2(0,1)$. Their results can be summarized as follows. Let

$$(A_\alpha u)(x) = -u'''(x) \tag{1.4}$$

with

$$\mathcal{D}(A_\alpha) = \left\{w \in H^3(0,1) \mid w(1) = w(0),\ w'(1) = \alpha w'(0),\ w''(1) = w''(0)\right\},$$

and let A_α^* be its adjoint operator in $L^2 = L^2(0,1)$. If $\alpha \neq -1/2$, then A_α and A_α^* have complete sets of eigenvectors—respectively, $\{\phi_k \mid -\infty < k < \infty\}$ and $\{\psi_k \mid -\infty < k < \infty\}$—which are normalized so that $\psi_j^* \phi_k = \delta_{kj}$ and form dual Riesz bases for L^2. Define

$$H_\alpha^{s,p} = \left\{ w = \sum_{k=-\infty}^{\infty} c_k \phi_k ; \quad \sum_{k=-\infty}^{\infty} (1 + |k|^{ps})|c_k|^p < \infty \right\}$$

with the norm

$$\|w\|_{H_\alpha^{s,p}}^p = \sum_{k=-\infty}^{\infty} |c_k|^p (1 + |k|^{ps}).$$

If we denote $H_\alpha^{s,2}$ by H_α^s, then the main results in [13] can be stated as follows.

THEOREM 1.1. *Assume $|\alpha| < 1$ and $\alpha \neq -1/2$. Then there exists a $\beta > 0$ such that for any $\phi \in H_\alpha^1$ with $\|\phi\|_{H_\alpha^1} \leq \beta$, (1.2) has a unique solution $u \in C([0,\infty); H_\alpha^1)$ and*

$$\|u(t,\cdot) - [\phi]\|_{L^2} \leq K e^{-\rho t} \|\phi - [\phi]\|$$

for $t \geq 0$, where $K > 0$ and $\rho > 0$ are independent of ϕ and $[\phi] = \int_0^1 \phi dx$.

The existence of solutions in Theorem 1.1 was proven by using an integral equation based on the "variation of parameters" formula and explicit representation of the semigroup associated with the operator A_α obtained from the dual Riesz bases $\{\phi_k\}$ and $\{\psi_k\}$. Then the asymptotic decay of the solutions was obtained by use of Lyapounov techniques based on properties of a linear equation $u_t + u_{xxx} = 0$. However, when $\alpha = -1/2$, the eigenvectors of A_α and A_α^* do not form dual Riesz bases for L^2 and the proof in [13] cannot be carried over. Therefore, a new method must be introduced for the proof when $\alpha = -1/2$.

The objective of this paper is to give a proof of Theorem 1.1 for $\alpha = -1/2$. We consider this singular case in Theorem 1.1 with $\alpha = -1/2$ and construct the semigroup corresponding to A_α using the Green's function to study the existence of solutions for (1.2) when $\alpha = -1/2$. After we derive the estimates for the linear operator, we can invert (1.2) into an integral equation and show that the corresponding integral operator is a contraction. Therefore, the contraction mapping theorem implies the existence of the solutions of (1.2). Finally we use an inequality similar to the Gronwall inequality to prove the exponential decay of the solutions. We note that the method developed here can also be used to prove Theorem 1.1. It is interesting to note that all the large eigenvalues of $A_{-1/2}$ are on the negative real axis. However, Komornik showed that $A_{-1/2}$ does not generate an analytic semigroup [6], which will be given in Appendix 3.

The paper is organized as follows. In §2, the exact form of the semigroup $S(t)$ corresponding to $A_{-1/2}$ is constructed by using Green's function, and some properties of the Green's function are given. In §3, various properties and estimates for $S(t)$ applying to functions in $L^2(0, 1)$ and $H^1(0, 1)$ are derived. In §4, the local and global existence of solutions of (1.2) is proven and the exponential decay of the solutions to their mean values is obtained.

2. Representation of the semigroup for $A_{-1/2}$. Let A_α be an operator defined in (1.4). If $|\alpha| \leq 1$, it is easy to check that A_α is dissipative in $L^2 = L^2(0, 1)$. By using elementary theory of ordinary differential equations, we can show that the range $\mathcal{R}(\lambda_0 I - A)$ is the whole space L^2 if $\lambda_0 > 0$ is large enough. Thus by Lumer-Phillip's theorem [10], A_α is the infinitesimal generator of a C_0-semigroup $S_\alpha(t)$ of contractions in L^2; that is, for $f \in L^2$ and $|\alpha| \leq 1$,

$$\|S_\alpha(t) f\|_{L^2} \leq \|f\|_{L^2}. \tag{2.1}$$

Since $\alpha \neq -1/2$ was studied in [13], here we assume that $\alpha = -1/2$ and let $A = A_{-1/2}$. Now we determine the discrete spectrum of A. Assume that μ_1, μ_2, and μ_3 are the three complex cubic roots of $-\lambda$. Then after some tedious computation, we can see that the eigenvalue λ of $Au = \lambda u$ satisfies either $\lambda = 0$ or

$$3 - (e^{-\mu_1} + e^{-\mu_2} + e^{-\mu_3}) = 0. \tag{2.2}$$

Since A is dissipative, $\operatorname{Re} \lambda \leq 0$ if λ is an eigenvalue. Also, every eigenvalue has multiplicity one. Next we obtain the asymptotic forms of the eigenvalues λ.

LEMMA 2.1. *If λ is an eigenvalue and $|\lambda|$ is large, then λ is real and $\lambda \sim -(2k + 1)\pi/2$ as $k \to +\infty$. Also, there is no eigenvalue on the imaginary axis except $\lambda = 0$.*

Proof. If $\lambda = ir$ with r real, then there exists one cubic root of $-ir$—say, μ_1—such that $\mu_1 = ir^{1/3}$. Therefore, by (2.2),

$$3 - (e^{-ir^{1/3}} + e^{(\sqrt{3}+i)r^{1/3}/2} + e^{(-\sqrt{3}+i)r^{1/3}/2}) = 0.$$

By separating the real part and the imaginary part, it is straightforward to show that the only root is $r = 0$. Thus, no eigenvalues are on the imaginary axis except zero. Next we show that λ is real for large eigenvalues λ. Since all eigenvalues have negative real parts except zero, $\operatorname{Re}(-\lambda) > 0$. Let μ_1 be the cubic root of $-\lambda$ with $|\arg \mu_1| \leq \pi/6$. Then $\mu_1 = \mu_r + i\mu_i$

with $\mu_r > 0$ and $|\mu_i| \le \mu_r/\sqrt{3}$. Substitution of $\mu_1 = \mu_r + i\mu_i$, $\mu_2 = \omega\mu_1$, $\mu_3 = \omega^2\mu_1$ with $\omega = (-1 + i\sqrt{3})/2$ into (2.2) and separating the real and imaginary parts yield

$$\begin{aligned} 3 - e^{-\mu_r}\cos\mu_i - e^{(\mu_r+\sqrt{3}\mu_i)/2}\cos((\sqrt{3}\mu_r - \mu_i)/2) \\ -e^{(\mu_r-\sqrt{3}\mu_i)/2}\cos((\sqrt{3}\mu_r + \mu_i)/2) = 0, \\ e^{-\mu_r}\sin\mu_i + e^{(\mu_r+\sqrt{3}\mu_i)/2}\sin((\sqrt{3}\mu_r - \mu_i)/2) \\ -e^{(\mu_r-\sqrt{3}\mu_i)/2}\sin((\sqrt{3}\mu_r + \mu_i)/2) = 0. \end{aligned}$$

Rewrite these two equations as

$$\begin{aligned} 2\cosh(\sqrt{3}\mu_i/2)\cos(\sqrt{3}\mu_r/2)\cos(\mu_i/2) + 2\sinh(\sqrt{3}\mu_i/2)\sin(\sqrt{3}\mu_r/2)\sin(\mu_i/2) \\ = 3e^{-\mu_r/2} - e^{-(\sqrt{3}\mu_r/2)}\cos\mu_i\,, \\ 2\sinh(\sqrt{3}\mu_i/2)\sin(\sqrt{3}\mu_r/2)\cos(\mu_i/2) - 2\cosh(\sqrt{3}\mu_i/2)\cos(\sqrt{3}\mu_r/2)\sin(\mu_i/2) \\ = e^{-(\sqrt{3}\mu_r/2)}\sin\mu_i\,. \end{aligned}$$

Multiply the first equation by $\cos(\mu_i/2)$ and the second by $\sin(\mu_i/2)$, and subtract the resulting equations to get

$$2\cosh(\sqrt{3}\mu_i/2)\cos(\sqrt{3}\mu_r/2) = \cos(\mu_i/2)(3e^{-\mu_r/2} - e^{-\sqrt{3}\mu_r/2}). \tag{2.3}$$

Then multiply the first by $\sin(\mu_i/2)$ and the second by $\cos(\mu_i/2)$, and add them together to obtain

$$2\sinh(\sqrt{3}\mu_i/2)\sin(\sqrt{3}\mu_r/2) = \sin(\mu_i/2)(3e^{-\mu_r/2} + e^{-\sqrt{3}\mu_r/2}). \tag{2.4}$$

Since $|\mu_i| \le \mu_r/\sqrt{3}$, $\mu_r \to +\infty$ if $\lambda \to \infty$. But $\cosh(\sqrt{3}\mu_i/2) \ge 1$. Thus as $\lambda \to \infty$, by (2.3) $|\cos(\sqrt{3}\mu_r/2)| \le K\exp(-\mu_r/2)$, where K is a fixed constant, which implies

$$\mu_r = (2k+1)\pi/\sqrt{3} + O(e^{-k\pi/\sqrt{3}}) \qquad \text{for } k \text{ large}.$$

Therefore, $|\sin(\sqrt{3}\mu_r/2)| \to 1$, and from (2.4) we have that $\mu_i = o(\exp(-k\pi/\sqrt{3}))$, which yields that for λ large, the cubic root μ_1 of $-\lambda$ must be

$$\mu_1 = (2k+1)\pi/\sqrt{3} + O(e^{-k\pi/\sqrt{3}}). \tag{2.5}$$

However, $2\cos(\sqrt{3}\mu/2) = 3e^{-\mu/2} - e^{-\sqrt{3}\mu/2}$ has infinitely many real roots $\tilde{\mu}_k = (2k+1)\pi/\sqrt{3} + O(\exp(-k\pi/\sqrt{3}))$, which satisfy (2.3) and (2.4) with $\mu_i = 0$. Thus $-\tilde{\mu}_k^3$, $k = 1, 2, \ldots$, are also the eigenvalues of A. Now let us consider the analytic function

$$f(z) = e^{-z/2}(3 - e^{-z} - e^{-\omega z} - e^{-\omega^2 z}).$$

If $\operatorname{Re} z \ge 0$ and $|\operatorname{Im} z| \le 1$, then $|f(z)| + |f'(z)| + |f''(z)| \le K$, where K is a fixed constant. $f(z)$ has infinitely many real zeros $\tilde{\mu}_k$, and at $\tilde{\mu}_k$, $|f'(\tilde{\mu}_k)| = \sqrt{3} + O(\exp(-k\pi/\sqrt{3}))$. For z in a disk centered at $\tilde{\mu}_k$ with radius $1/2$, we have

$$f(z) = f(\tilde{\mu}_k) + f'(\tilde{\mu}_k)(z - \tilde{\mu}_k) + (1/2)f''(\tilde{z})(z - \tilde{\mu}_k)^2,$$

where $\tilde{z}$ is in the disk. Thus

$$|f(z)| \ge |z - \tilde{\mu}_k|\left(|f'(\tilde{\mu}_k)| - K|z - \tilde{\mu}_k|\right),$$

where K is independent of z and k. Therefore, there is a $\tilde{\delta} > 0$ such that for $|z - \tilde{\mu}_k| \leq \tilde{\delta}$ and $z \neq \tilde{\mu}_k$, $f(z) \neq 0$. Thus μ_1 in (2.5) must be real also, which implies that the large eigenvalues are real, completing the proof of Lemma 2.1.

We note here that A does not generate an analytic semigroup, which will be shown in Appendix 3. In the following, we shall give a representation of the semigroup $S(t) = S_{-1/2}(t)$, where $S_\alpha(t)$ is the corresponding C_0-semigroup of A_α. Let λ be in either the upper half or the lower half plane. Without loss of generality, we assume that λ is in the upper half plane since the discussion is similar for λ in the lower half plane. Let $\mu_1^3 = -\lambda$, $\mu_2 = \omega\mu_1$, $\mu_3 = \omega^2\mu_1$ with $\omega = (-1 + i\sqrt{3})/2$. Since $\operatorname{Im}\lambda \geq 0$, $\operatorname{Re}(-i\lambda) \geq 0$. There is a μ^* such that $(\mu^*)^3 = -i\lambda$ with $|\arg\mu^*| \leq (\pi/6)$ and $(i\mu^*)^3 = -\lambda$. Let $\mu_1 = i\mu^*$, which implies $|(\pi/2) - \arg\mu_1| \leq (\pi/6)$. Now let $G(\lambda, x, \xi)$ be the Green's function of $(\lambda I - A)^{-1}$ in L^2. Then for $f \in L^2$,

$$(\lambda I - A)^{-1} f = \int_0^1 G(\lambda, x, \xi) f(\xi) d\xi \in \mathcal{D}(A)\,, \tag{2.6}$$

where $G(\lambda, x, \xi)$ is given in Appendix 1. Since A is an infinitesimal generator of a C_0-semigroup of contractions, by a formula in [10, Cor. 7.5, p. 28], we have

$$S(t) f = (1/2\pi i) \int_{\gamma - i\infty}^{\gamma + i\infty} e^{\lambda t} (\lambda I - A)^{-1} f d\lambda \tag{2.7}$$

for $t > 0, \gamma > 0$ and $f \in \mathcal{D}(A^2)$. It is well known that $(\lambda I - A)^{-1} f$ is analytic in the complex λ-plane except poles, which are the eigenvalues of A. However, by Lemma 2.1 all the eigenvalues except zero are in the left side of the complex plane. By using the spectral decomposition theorem and the residue theorem [3] for the pole of $(\lambda I - A)^{-1}$ at zero, we have

$$S(t) f = (1/2\pi i) \int_{\Gamma^*} e^{\lambda t} (\lambda I - A)^{-1} f d\lambda + c_0 \int_0^1 f(x) dx,$$

where c_0 is a fixed constant and the last term is the projection of $f(x)$ onto 1 in L^2, since 1 is the eigenvector for the eigenvalue zero. Here Γ^* is a contour with only the eigenvalue zero of A on its right side and asymptotic to $\gamma \pm i\infty$ as $|\lambda| \to +\infty$. Next we deform the contour Γ^* further into $\tilde{\Gamma} = \tilde{\Gamma}_+ \cup \tilde{\Gamma}_-$, where, for $\lambda = \lambda_r + i\lambda_i$,

$$\begin{aligned} \tilde{\Gamma}_+ = \Big\{ & F(\lambda_i) + i\lambda_i, 0 \leq \lambda_i \to +\infty \text{ and } F(\lambda_i) \\ & \text{is a smooth function of } \lambda_i \text{ with } |F(\lambda_i)|^{3/2}/\lambda_i \to \delta > 0 \text{ and } F(\lambda_i) < 0 \Big\}, \\ \tilde{\Gamma}_- = \Big\{ & F(-\lambda_i) + i\lambda_i, 0 \geq \lambda_i \to -\infty \Big\}, \end{aligned} \tag{2.8}$$

$F(0) = \Gamma^* \cap \{\operatorname{Im}\lambda = 0\} < 0$, and all the eigenvalues of A except zero are on the left side of $\tilde{\Gamma}$. Using Cauchy's theorem, we have

$$\begin{aligned} S(t) f &= (1/2\pi i) \left(\int_{\Gamma^* \cap \{\operatorname{Im}\lambda \geq 0\}} + \int_{\Gamma^* \cap \{\operatorname{Im}\lambda \leq 0\}} \right) e^{\lambda t} (\lambda I - A)^{-1} f d\lambda + c_0 \int_0^1 f(x) dx \\ &= (1/2\pi i) \left(\int_{\tilde{\Gamma}_+} + \int_{\tilde{\Gamma}_-} \right) e^{\lambda t} (\lambda I - A)^{-1} f d\lambda + c_0 \int_0^1 f(x) dx \end{aligned} \tag{2.9}$$

for $t > 0$ and $f \in \mathcal{D}(A^2)$ since the integrals over contours at infinity are seen to be zero by using the exponential decay of $e^{\lambda t}$ and the explicit form of $G(\lambda, x, \xi)$ for $(\lambda I - A)^{-1}$.

However, for $t > 0$, $\operatorname{Re}(\lambda t) \to -\infty$ with order of $-|\lambda t|^{2/3}$ if $\lambda \in \tilde{\Gamma}$. Thus the right-hand side of (2.9) is well defined for all $f \in L^2$ if $t > 0$. Since $S(t)$ is equal to the right-hand side of (2.9) for $f \in \mathcal{D}(A^2)$ and $\mathcal{D}(A^2)$ is dense in L^2,

$$S(t)f = \frac{1}{2\pi i}\int_{\tilde{\Gamma}_+} e^{\lambda t}(\lambda I - A)^{-1} f\, d\lambda + \frac{1}{2\pi i}\int_{\tilde{\Gamma}_-} e^{\lambda t}(\lambda I - A)^{-1} f\, d\lambda + c_0 \int_0^1 f(x)dx \tag{2.10}$$

for $f \in L^2$ and $t > 0$. Now we rewrite (2.10) in terms of $\mu_1 = \mu$

$$S(t) = \frac{1}{2\pi i}\left(\int_{\Gamma_+} + \int_{\Gamma_-}\right)(3\mu^2)e^{-\mu^3 t}(\mu^3 I + A)^{-1} f\, d\mu + c_0 \int_0^1 f(x)dx, \tag{2.11}$$

where Γ_+ is in $\{\mu : |\arg \mu - (\pi/2)| \le (\pi/6),\ \operatorname{Im} \mu > 0\}$ and is from $\mu_i = b > 0$ to $\mu_i = +\infty$, and Γ_- is in $\{\mu : |\arg \mu - (3\pi/2)| \le (\pi/6),\ \operatorname{Im} \mu < 0\}$ and is from $\mu_i = -b < 0$ to $\mu_i = -\infty$. Also, Γ_+ and Γ_- are disconnected. For $\mu = \mu_r + i\mu_i$ and $\lambda = \lambda_r + i\lambda_i$ with $\mu^3 = -\lambda$, we have $\lambda_r = 3\mu_r\mu_i^2 - \mu_r^3$ and $\lambda_i = \mu_i^3 - 3\mu_r^2\mu_i$. Therefore, on $\Gamma_\pm$, $\mu_r \le c_1 < 0$, $3\mu_r\mu_i^2 - \mu_r^3 \le c_1 < 0$ for a small fixed negative number c_1, and $\mu_r \to -\delta^{2/3}/3$ as $\mu \to \infty$ (i.e., $\mu_i \to \pm\infty$). From the formula of the Green's function in Appendix 1, we have

$$G(-\mu^3, x, \xi) = \begin{cases} -(1/3\mu^2)e^{\mu(x-\xi)} + R(\mu, x, \xi), & 0 < x < \xi < 1, \\ R(\mu, x, \xi), & 0 < \xi < x < 1. \end{cases} \tag{2.12}$$

Since $G(-\mu^3, x, \xi)$ is analytic in μ when μ is in the upper or lower μ-plane, we assume that for μ_i large—say, $|\mu_i| \ge r$—$\mu_r = -\delta^{2/3}/3 = \delta_0$ and $3\mu_r\mu_i^2 - \mu_r^3 \le -2\beta_1 < 0$ on $\Gamma_\pm$ for two small fixed positive numbers δ and β_1. Also, by checking the terms in $G(-\mu^3, x, \xi)$, we can see that

$$\sup_{0\le x,\xi\le 1} |G(-\mu^3, x, \xi)| \le K/|\mu|^2 \tag{2.13}$$

if $|\mu_r| \le \epsilon_0$ for large μ_i and a fixed small constant $\epsilon_0 > 0$, where μ is on $\Gamma_\pm$ and K is independent of μ. Let δ_0 be fixed with $|\delta_0| \le \epsilon_0$. We then have the following properties of the C_0-semigroup $S(t)$.

3. Properties of the C_0-semigroup $S(t)$. Since we use the contraction mapping theorem to obtain the existence of solutions, we first need the following estimates.

PROPOSITION 3.1. *Let $T > 0$ be given. Then for $f \in L^\infty(0, T; L^2)$,*

$$\sup_{0\le t\le T}\left\|\int_0^t S(\tau)f(t-\tau, \cdot)d\tau\right\|_{H^1} \le K\left(T^{1/4}\sup_{0\le t\le T}\|f(t,\cdot)\|_{L^2} + T\sup_{0\le t\le T}\left|\int_0^1 f(t,x)dx\right|\right),$$

where $H^1 = H^1(0, 1)$ and K is independent of T.

Proof. By the form of $S(t)$ in (2.11), we have

$$\left\|\int_0^t S(\tau)f(t-\tau, \cdot)d\tau\right\|_{L^2} \le I_1 + I_2 + K\int_0^t \left|\int_0^1 f(\tau, x)dx\right| d\tau,$$

where

$$I_{1,2} = \left\|\int_0^t \frac{1}{2\pi i}\int_{\Gamma_\pm} 3\mu^2 e^{-\mu^3\tau}\int_0^1 G(-\tau^3, x, \xi)f(t-\tau, \xi)d\xi\, d\mu\, d\tau\right\|_{L^2}. \tag{3.1}$$

By (2.13), $0 < \delta_0 \le \epsilon_0$, and Hölder's inequality, we have

$$
\begin{aligned}
I_1^2 &\le K \int_0^1 \left(\int_0^t \int_b^{+\infty} e^{-\mu_i^2 \tau \delta^{2/3}} d\mu_i d\tau \right)^2 dx \sup_{0<\tau\le t} \|f(\tau,\cdot)\|_{L^2}^2 \\
&\le K t^{1/2} \sup_{0<\tau\le t} \|f(\tau,\cdot)\|_{L^2}^2.
\end{aligned} \tag{3.2}
$$

By a similar calculation, we have $I_2^2 \le K t^{1/2} \sup_{0<\tau\le t} \|f(\tau,\cdot)\|_{L^2}^2$ and

$$
\left\| \int_0^t S(\tau) f(t-\tau,\cdot) d\tau \right\|_{L^2} \le K \left(T^{1/4} \sup_{0\le t\le T} \|f(t,\cdot)\|_{L^2} + T \sup_{0\le t\le T} \left| \int_0^1 f(t,x) dx \right| \right).
$$

In order to estimate the first-order derivative, from (2.12) we first consider

$$
\begin{aligned}
\|II(x)\|_{L^2} &\overset{\text{def}}{=} \left\| (1/2\pi i) \int_0^1 \int_0^t \int_{\Gamma_+} 3\mu^2 e^{-\mu^3\tau} G_x(-\mu^3, x, \xi) f(t-\tau,\xi) d\mu d\tau d\xi \right\|_{L^2} \\
&= (1/2\pi) \Bigg(\int_0^1 \Bigg| \int_x^1 \int_0^t \int_{\Gamma_+} 3\mu^2 e^{-\mu^3\tau} (-1/3\mu) e^{\mu(x-\xi)} f(t-\tau,\xi) d\mu d\tau d\xi \\
&\qquad + \int_0^1 \int_0^t \int_{\Gamma_+} 3\mu^2 e^{-\mu^3\tau} R_x(\mu,x,\xi) f(t-\tau,\xi) d\mu d\tau d\xi \Bigg|^2 dx \Bigg)^{1/2} \\
&\le (1/2\pi) \left(\int_0^1 \left| \int_x^1 \int_0^t \int_{\Gamma_+} 3\mu^2 e^{-\mu^3\tau} (-1/3\mu) e^{\mu(x-\xi)} f(t-\tau,\xi) d\mu d\tau d\xi \right|^2 dx \right)^{1/2} \\
&\qquad + (1/2\pi) \left(\int_0^1 \left| \int_0^1 \int_0^t \int_{\Gamma_+} 3\mu^2 e^{-\mu^3\tau} R_x(\mu,x,\xi) f(t-\tau,\xi) d\mu d\tau d\xi \right|^2 dx \right)^{1/2} \\
&\overset{\text{def}}{=} II_1 + II_2.
\end{aligned}
$$

By using Fubini's theorem and Minkowski's inequality in integral form,

$$
\begin{aligned}
II_2 &\le (1/2\pi) \left(\int_0^1 \left(\int_0^t \int_{\Gamma_+} \int_0^1 \left| 3\mu^2 e^{-\mu^3\tau} R_x(\mu,x,\xi) f(t-\tau,\xi) \right| d\xi |d\mu| d\tau \right)^2 dx \right)^{1/2} \\
&\le (1/2\pi) \int_0^t \int_{\Gamma_+} \left(\int_0^1 \left(\int_0^1 \left| 3\mu^2 e^{-\mu^3\tau} R_x(\mu,x,\xi) f(t-\tau,\xi) \right| d\xi \right)^2 dx \right)^{1/2} |d\mu| d\tau \\
&\le K \sup_{0<\tau\le t} \|f(\tau,\cdot)\|_{L^2} \int_0^t \int_{\Gamma_+} \left(\int_0^1 \int_0^1 \left| 3\mu^2 e^{-\mu^3\tau} R_x(\mu,x,\xi) \right|^2 d\xi dx \right)^{1/2} |d\mu| d\tau.
\end{aligned}
$$

In Appendix 1, we shall show that

$$
\left(\int_0^1 \int_0^1 \left| R_x(\mu,x,\xi) \right|^2 d\xi dx \right)^{1/2} \le K/|\mu|^{3/2}.
$$

Therefore,

$$
\begin{aligned}
II_2 &\le K \sup_{0<\tau\le t} \|f(\tau,\cdot)\|_{L^2} \int_0^t \int_b^{+\infty} |\mu_i|^{1/2} e^{-\mu_i^2 \delta^{2/3} \tau} d\mu_i d\tau \\
&\le K t^{1/4} \sup_{0<\tau\le t} \|f(\tau,\cdot)\|_{L^2}.
\end{aligned}
$$

For II_1 and $x-\xi<0$,

$$\int_{\Gamma_+}\mu e^{-\mu^3\tau+\mu(x-\xi)}d\mu=\left(\int_{\Gamma_+^*}+\int_{ir}^{+\infty i}\right)\mu e^{-\mu^3\tau+(x-\xi)\mu}d\mu,$$

where Γ_+^* is a finite smooth curve on Γ_+ with $b\le\mu_i\le r$. Then

$$\left|\int_{\Gamma_+^*}\mu e^{-\mu^3\tau+(x-\xi)\mu}d\mu\right|\le K<+\infty,$$

and

$$\int_{ir}^{+\infty i}\mu e^{-\mu^3\tau+(x-\xi)\mu}d\mu=-\int_r^{+\infty}\mu e^{i(\mu^3\tau+(x-\xi)\mu)}d\mu.$$

In Appendix 2, we show that if $x-\xi$ is bounded, then

$$\left|\int_r^{+\infty}\mu e^{i(\mu^3\tau+\mu(x-\xi))}d\mu\right|\le K\tau^{-3/4}.$$

Therefore,

$$\begin{aligned}II_1&\le K\left(\int_0^1\left(\int_x^1\int_0^t|f(t-\tau,\xi)|\tau^{-3/4}d\xi d\tau\right)^2dx\right)^{1/2}\\&\le K\left(\int_0^1\left(\int_0^t\tau^{-3/4}\sup_{0<\tau\le t}\|f(\tau,\cdot)\|_{L^2}d\tau\right)^2dx\right)^{1/2}\\&\le Kt^{1/4}\sup_{0<\tau\le t}\|f(\tau,\cdot)\|_{L^2}.\end{aligned}$$

Thus $\|II(x)\|_{L^2}\le Kt^{1/4}\sup_{0<\tau\le t}\|f(\tau,\cdot)\|_{L^2}$. However, by (3.1) and (3.2) we know that the multiple integrals in $\int_0^t S(\tau)f(t-\tau,\cdot)d\tau$ are absolutely integrable and the order of the integration can be interchanged by Fubini's theorem. Also, since $G(-\mu^3,x,\xi)$ is continuous at $x=\xi$ and each term in $G(-\mu^3,x,\xi)$ has a form $X(x)Y(\xi)$, then

$$\begin{aligned}III(x)&\overset{\text{def}}{=}\frac{\partial}{\partial x}\int_0^t\frac{1}{2\pi i}\int_{\Gamma_+}3\mu^2e^{-\mu^3\tau}\left(\int_0^x+\int_x^1\right)G(-\mu^3,x,\xi)f(t-\tau,\xi)d\xi d\mu d\tau\\&=\frac{\partial}{\partial x}\left(\int_0^x+\int_x^1\right)\int_0^t\frac{1}{2\pi i}\int_{\Gamma_+}3\mu^2e^{-\mu^3\tau}G(-\mu^3,x,\xi)f(t-\tau,\xi)d\mu d\tau d\xi\\&=\int_0^1\int_0^t\frac{1}{2\pi i}\int_{\Gamma_+}3\mu^2e^{-\mu^3\tau}G_x(-\mu^3,x,\xi)f(t-\tau,\xi)d\mu d\tau d\xi=II(x).\end{aligned}$$

Therefore

$$\|III(x)\|_{L^2}\le Kt^{1/4}\sup_{0<\tau\le t}\|f(\tau,\cdot)\|_{L^2}.$$

By a similar proof, we have

$$\begin{aligned}&\left\|\frac{\partial}{\partial x}\int_0^t\frac{1}{2\pi i}\int_{\Gamma_-}3\mu^2e^{-\mu^3\tau}\left(\int_0^x+\int_x^1\right)G(-\mu^3,x,\xi)f(t-\tau,\xi)d\xi d\mu d\tau\right\|_{L^2}\\&\qquad\le Kt^{1/4}\sup_{0<\tau\le t}\|f(\tau,\cdot)\|_{L^2}.\end{aligned}$$

Combining the above estimates, we have

$$\left\| \frac{\partial}{\partial x} \int_0^t S(\tau) f(t-\tau, \cdot) d\tau \right\|_{L^2} \leq K t^{1/4} \sup_{0<\tau\leq t} \|f(\tau, \cdot)\|_{L^2}.$$

Finally we have

$$\sup_{0<t\leq T} \left\| \int_0^t S(\tau) f(t-\tau, \cdot) d\tau \right\|_{H^1} \leq K \left(T^{1/4} \sup_{0<t\leq T} \|f(\tau, \cdot)\|_{L^2} + T \sup_{0<\tau\leq T} \left| \int_0^1 f(\tau, x) dx \right| \right).$$

Proposition 3.1 is a local estimate. Next we give a global estimate under some condition on $f(t, x)$.

PROPOSITION 3.2. *If $f(t, x) \in L^\infty([0, \infty); L^2)$ and $\int_0^1 f(t, x) dx \equiv 0$ for all $t \geq 0$, then there exists a K such that*

$$\sup_{0<\tau<+\infty} \left\| \int_0^t S(\tau) f(t-\tau, \cdot) d\tau \right\|_{H^1} \leq K \sup_{0<\tau<+\infty} \|f(\tau, \cdot)\|_{L^2}.$$

Proof. By Proposition 3.1, we need only to prove the inequality for $t \geq 1$. If $t \geq 1$, then

$$\begin{aligned} \|I\|_{H^1} &\overset{\text{def}}{=} \left\| \int_0^t S(\tau) f(t-\tau, \cdot) d\tau \right\|_{H^1} \\ &\leq \left\| \int_0^1 S(\tau) f(t-\tau, \cdot) d\tau \right\|_{H^1} + \left\| \int_1^t S(\tau) f(t-\tau, \cdot) d\tau \right\|_{H^1} \\ &\overset{\text{def}}{=} II_1 + II_2. \end{aligned}$$

But $II_1 = \| \int_0^1 S(\tau) g_t(1-\tau, \cdot) d\tau \|_{H^1}$, where $g_t(s, x) = f(s+t-1, x)$ for $s \geq 0$. By Proposition 3.1,

$$\begin{aligned} II_1 &\leq K \sup_{0<\tau\leq 1} \|g_t(\tau, \cdot)\|_{L^2} \leq K \sup_{0<\tau\leq 1} \|f(\tau+t-1, \cdot)\|_{L^2} \\ &\leq K \sup_{0<\tau<+\infty} \|f(\tau, \cdot)\|_{L^2}. \end{aligned}$$

By using (2.11), (2.13), and the uniform convergence of the integral and its derivatives for $\tau \geq 1$, we have

$$\begin{aligned} II_2 &\leq K \left\| \int_1^t \left(\int_{\Gamma_+} + \int_{\Gamma_-} \right) |\mu|^2 |e^{-\mu^3 \tau}| \left[\left(\int_0^1 |G(-\mu^3, x, \xi)|^2 d\xi \right)^{1/2} \right. \right. \\ &\qquad \left. \left. + \left(\int_0^1 |G_x(-\mu^3, x, \xi)|^2 d\xi \right)^{1/2} \right] \left(\int_0^1 |f(t-\tau, \xi)|^2 d\xi \right)^{1/2} |d\mu| d\tau \right\|_{L^2} \\ &\leq K \sup_{0<\tau<\infty} \|f(\tau, \cdot)\|_{L^2} \int_1^t \left(\int_{\Gamma_+} + \int_{\Gamma_-} \right) |\mu^2| |e^{-\mu^3 \tau}| (|\mu^{-2}| + |\mu^{-1}|) |d\mu| d\tau \\ &\leq K \sup_{0<\tau<\infty} \|f(\tau, \cdot)\|_{L^2} \int_1^t \left(\int_b^{+\infty} + \int_{-\infty}^{-b} \right) |\mu_i^2| |e^{-\mu_i^2 \tau \delta^{2/3}} (|\mu_i|^{-2} + |\mu_i|^{-1}) d\mu_i d\tau \\ &\leq K \sup_{0<\tau<\infty} \|f(\tau, \cdot)\|_{L^2} \left(\int_b^{+\infty} + \int_{-\infty}^{-b} \right) e^{-\mu_i^2 \delta^{2/3}} (|\mu_i|^{-2} + |\mu_i|^{-1}) d\mu_i \\ &\leq K \sup_{0<\tau<\infty} \|f(\tau, \cdot)\|_{L^2}, \end{aligned}$$

where K is independent of t. Thus $\sup_{t\geq 1} II_2 \leq K \sup_{0<\tau<\infty} \|f(\tau,\cdot)\|_{L^2}$. Combining the estimate for $0 \leq t \leq 1$, we obtain

$$\sup_{0<t<+\infty} \left\| \int_0^t S(\tau) f(t-\tau,\cdot) d\tau \right\|_{H^1} \leq K \sup_{0<\tau<\infty} \|f(\tau,\cdot)\|_{L^2}.$$

Finally we derive the estimates for the C_0-semigroup $S(t)$ applying to an H^1-function.

PROPOSITION 3.3. *For $w_0(x) \in H^1 = H^1(0,1)$, $S(t)w_0 \in H^1$ and $\|S(t)w_0\|_{H^1} \leq K\|w_0\|_{H^1}$ for all $t \in [0,\infty)$, where K is independent of t.*

Proof. Since $S(t)$ is a C_0-semigroup of contraction in L^2,

$$\|S(t)w_0\|_{L^2} \leq \|w_0\|_{L^2}, \quad \|S(t)w_{0x}\|_{L^2} \leq \|w_{0x}\|_{L^2},$$

and $S(0)w_0 = w_0$. For $t > 0$,

$$\begin{aligned} \frac{\partial}{\partial x} S(t)w_0 &= \frac{1}{2\pi i}\left(\int_{\Gamma_+} + \int_{\Gamma_-}\right) \frac{\partial}{\partial x} \int_0^1 3\mu^2 e^{-\mu^3 t} G(-\mu^3, x, \xi) w_0(\xi) d\xi d\mu \\ &= \frac{1}{2\pi i}\left(\int_{\Gamma_+} + \int_{\Gamma_-}\right) \int_0^1 3\mu^2 e^{-\mu^3 t} G_x(-\mu^3, x, \xi) w_0(\xi) d\xi d\mu \,, \end{aligned} \tag{3.3}$$

since the integrals are uniformly convergent for $t > 0$ and $G(-\mu^3, x, \xi)$ is continuous at $x = \xi$. By the form of $G(-\mu^3, x, \xi)$ and integration by parts, we obtain

(3.4)

$$\begin{aligned} &\int_0^1 G_x(-\mu^3, x, \xi) w_0(\xi) d\xi = \int_0^1 G(-\mu^3, x, \xi) w_{0\xi}(\xi) d\xi \\ &+ \int_0^1 [3(-\mu_1^3)(3 - e^{-\mu_1} - e^{-\mu_2} - e^{-\mu_3})]^{-1} \Big[(e^{\mu_2} - e^{-\mu_1}) \mu_3 \mu_2^{-1} (\mu_1 - \mu_2) e^{\mu_1 x - \mu_2 \xi} \\ &+ (e^{\mu_3} - e^{-\mu_1}) \mu_2 \mu_3^{-1} (\mu_1 - \mu_3) e^{\mu_1 x - \mu_3 \xi} + (e^{\mu_3} - e^{-\mu_2}) \mu_1 \mu_3^{-1} (\mu_2 - \mu_3) e^{\mu_2 x - \mu_3 \xi} \\ &+ (e^{\mu_1} - e^{-\mu_2}) \mu_3 \mu_1^{-1} (\mu_2 - \mu_1) e^{\mu_2 x - \mu_1 \xi} + (e^{\mu_1} - e^{-\mu_3}) \mu_2 \mu_1^{-1} (\mu_3 - \mu_1) e^{\mu_3 x - \mu_1 \xi} \\ &+ (e^{\mu_2} - e^{-\mu_3}) \mu_1 \mu_2^{-1} (\mu_3 - \mu_2) e^{\mu_3 x - \mu_2 \xi} \Big] w_{0\xi}(\xi) d\xi \\ &\stackrel{\text{def}}{=} I(\mu, x) + II(\mu, x), \end{aligned}$$

where $\mu_1 = \mu$, $\mu_2 = \omega\mu$, $\mu_3 = \omega^3\mu$ with $\omega = (-1 + i\sqrt{3})/2$. Therefore, for $t > 0$,

$$\left\| \frac{1}{2\pi i} \int_{\Gamma_\pm} 3\mu^2 e^{-\mu^3 t} I(\mu, x) d\mu \right\|_{L^2} \leq \|S(t)w_{0x}\|_{L^2} + |c_0| \int_0^1 |w_{0\xi}| d\xi \leq K\|w_{0x}\|_{L^2}.$$

In order to estimate $II(\mu, x)$, we need the following lemma.

LEMMA 3.1. *If $f(y) \in L^2[0,\infty)$ and $\operatorname{Re}\alpha \leq 0$ with $\alpha \neq 0$, then $\int_0^\infty \exp(\alpha x y) f(y) dy \in L^2[0,\infty)$ and*

$$\left\| \int_0^{+\infty} e^{\alpha x y} f(y) dy \right\|^2_{L^2[0,,\infty)} \leq K \|f(y)\|^2_{L^2[0,,\infty)}.$$

The proof of the lemma can be easily obtained by generalizing a lemma in [3, p. 2332]. Now we rewrite $II(\mu, x)$ as

$$II(\mu, x) = II_1(\mu, x) + II_2(\mu, x) + \cdots + II_6(\mu, x). \tag{3.5}$$

Then

$$III \overset{\text{def}}{=} \left\| \frac{1}{2\pi i} \int_{\Gamma_+} 3\mu^2 e^{-\mu^3 t} II_1(\mu, x) d\mu \right\|_{L^2}^2$$

$$= (1/4\pi^2) \int_0^1 \left| \int_{\Gamma_+} \mu^2 e^{-\mu^3 t + \mu_1 x} (\mu^3 (3 - e^{-\mu_1} - e^{-\mu_2} - e^{-\mu_3}))^{-1} \right.$$

$$\left. \times (e^{\mu_2} - e^{-\mu_1})(\mu_3/\mu_2)(\mu_1 - \mu_2) \int_0^1 e^{-\mu_2 \xi} w_{0\xi} d\xi d\mu \right|^2 dx.$$

Since $\mu = \mu_r + i\mu_i$ and μ_r is a function of μ_i for μ on Γ_+, i.e., $\mu_r = s(\mu_i)$ with $s(\mu_i) = -\delta^{2/3}/3 = \delta_0$ for $\mu_i \geq r$,

$$III = \frac{|1-\omega|}{4\pi^2} \int_0^1 \left| \int_b^{+\infty} e^{-(s(\mu_i)+i\mu_i)^3 t + (s(\mu_i)+i\mu_i)x} \right.$$

$$\times (3 - e^{-(s(\mu_i)+i\mu_i)} - e^{-\omega(s(\mu_i)+i\mu_i)} - e^{-\omega^2(s(\mu_i)+i\mu_i)})^{-1}$$

$$\left. \times (e^{\omega(s(\mu_i)+i\mu_i)} - e^{-(s(\mu_i)+i\mu_i)})(s'(\mu_i) + i) \int_0^1 e^{-\omega(s(\mu_i)+i\mu_i)\xi} w_{0\xi}(\xi) d\xi d\mu_i \right|^2 dx$$

$$= \frac{|1-\omega|}{4\pi^2} \int_0^1 \left| \left(\int_b^r + \int_r^\infty \right) (\cdots) d\mu_i \right|^2 dx$$

$$\overset{\text{def}}{=} III_1 + III_2.$$

Since $-s(\mu_i)^3 + 3s(\mu_i)\mu_i^2 < 0$ for $\mu_i \in [b, r]$ and every term inside the integral of III_1 is bounded, we have

$$III_1 = \frac{|1-\omega|}{4\pi^2} \int_0^1 \left| \int_b^r (\cdots) d\mu_i \right| dx \leq K \left(\int_0^1 |w_{0\xi}(\xi)| d\xi \right)^2 \leq K \|w_{0x}\|_{L^2}^2.$$

However, by Lemma 3.1, we have

$$III_2 = \frac{|1-\omega|}{4\pi^2} \int_0^1 e^{\delta_0 x} \left| \int_r^{+\infty} e^{-(\delta_0 + i\mu_i)^3 t + i\mu_i x} \right.$$

$$\times (3 - e^{-(\delta_0+i\mu_i)} - e^{-\omega(\delta_0+i\mu_i)} - e^{-\omega^2(\delta_0+i\mu_i)})^{-1} (e^{\omega(\delta_0+i\mu_i)} - e^{-(\delta_0+i\mu_i)})$$

$$\left. \times \int_0^1 e^{-\omega(\delta_0+i\mu_i)\xi} w_{0\xi}(\xi) d\xi d\mu_i \right|^2 dx$$

$$\leq K \int_r^{+\infty} \left| e^{-(\delta_0+i\mu_i)^3 t} (3 - e^{-(\delta_0+i\mu_i)} - e^{-\omega(\delta_0+i\mu_i)} - e^{-\omega^2(\delta_0+i\mu_i)})^{-1} \right.$$

$$\left. \times (e^{\omega(\delta_0+i\mu_i)} - e^{-(\delta_0+i\mu_i)}) \int_0^1 e^{-\omega(\delta_0+i\mu_i)\xi} w_{0\xi}(\xi) d\xi \right|^2 d\mu_i$$

$$\leq K \int_r^{+\infty} \left| \int_0^1 e^{i\omega\mu_i - \omega(\delta_0+i\mu_i)\xi} w_{0\xi}(\xi) d\xi \right|^2 d\mu_i$$

$$\leq K \int_r^{+\infty} \left| \int_0^1 e^{i\omega\mu_i(1-\xi)} e^{-\omega\delta_0\xi} w_{0\xi}(\xi) d\xi \right|^2 d\mu_i$$

$$\leq K \int_r^{+\infty} \left| \int_0^1 e^{i\omega\mu_i\zeta} (e^{-\omega\delta_0(1-\zeta)} w_{0\xi}(1-\zeta)) d\zeta \right|^2 d\mu_i.$$

Since $\operatorname{Re}(i\omega) < 0$, by Lemma 3.1 again, we obtain

$$III_2 \le K \int_0^1 |e^{-\omega\delta_0(1-\zeta)} w_{0\zeta}(1-\zeta)|^2 d\zeta \le K\|w_{0x}\|_{L^2}^2.$$

By a similar proof, we have

$$\left\| \int_{\Gamma_-} \frac{3\mu^2}{2\pi i} e^{-\mu^3 t} II_1(\mu, x) d\mu \right\|_{L^2}^2 \le K\|w_{0x}\|_{L^2}^2,$$

and therefore

$$\left\| \left(\int_{\Gamma_+} + \int_{\Gamma_-} \right) \frac{3\mu^2}{2\pi i} e^{-\mu^3 t} II_1(\mu, x) d\mu \right\|_{L^2}^2 \le K\|w_{0x}\|_{L^2}^2.$$

Then by applying the same procedure to $II_i(\mu, x)$, $i = 2, 3, \ldots, 6$, we have

$$\left\| \left(\int_{\Gamma_+} + \int_{\Gamma_-} \right) \frac{3\mu^2}{2\pi i} e^{-\mu^3 t} II_i(\mu, x) d\mu \right\|_{L^2}^2 \le K\|w_{0x}\|_{L^2}^2$$

for $i = 2, 3, \ldots, 6$. By (3.3)-(3.5), we have

$$\left\| \left(\int_{\Gamma_+} + \int_{\Gamma_-} \right) \frac{3\mu^2}{2\pi i} e^{-\mu^3 t} II(\mu, x) d\mu \right\|_{L^2}^2 \le K\|w_{0x}\|_{L^2}^2.$$

By the estimate of $I(\mu, x)$, we have

$$\left\| \frac{\partial}{\partial x} S(t) w_0 \right\|_{L^2}^2 \le K\|w_{0x}\|_{L^2}^2.$$

Thus $\|S(t)w_0\|_{H^1} \le K\|w_0\|_{H^1}$. The proof is completed.

Now we are ready to obtain the existence and exponential decay of the solutions.

4. Existence and exponential decay of solutions with small amplitude. We now study the initial value problem for the KdV equation

$$w_t + ww_x + w_{xxx} = 0 \quad \text{for} \quad 0 < x < 1,\ t > 0, \tag{4.1}$$

$$w(0, x) = w_0(x), \tag{4.2}$$

with boundary conditions

$$w(t, 1) = w(t, 0), \quad w_x(t, 1) = (-1/2)w_x(t, 0), \quad w_{xx}(t, 1) = w_{xx}(t, 0). \tag{4.3}$$

If the nonlinear term ww_x is moved to the right-hand side of (4.1), then variation of parameters yields an equivalent integral equation:

$$w(t, \cdot) = S(t)w_0 - \int_0^t S(\tau)(ww_x)(t-\tau, \cdot) d\tau. \tag{4.4}$$

Let

$$Fu \overset{\text{def}}{=} S(t)w_0 - \int_0^t S(\tau)(uu_x)(t-\tau, \cdot) d\tau. \tag{4.5}$$

Then the solution of (4.4) is a fixed point of Fu in $L^2 = L^2(0, 1)$. First we prove the well-posedness of (4.1)-(4.3) in $H^1 = H^1(0, 1)$.

THEOREM 4.1. *For any $w_0 \in H^1$, there exists a $T = T(\|w_0\|_{H^1}) > 0$ such that (4.1)-(4.3) has a unique solution $u \in C(0, T; H^1)$, where $T \to +\infty$ as $\|w_0\|_{H^1} \to 0$. For any $T' < T(\|\phi(x)\|_{H^1})$, there exists a neighborhood U of $\phi(x)$ in H^1 such that the map $\mathcal{G} : w_0 \to u(t, \cdot)$ from U to $C(0, T; H^1)$ is Lipschitz continuous.*

Proof. Let

$$\mathcal{S}_{T,b} = \left\{ v \in C(0, T; H^1) \;\middle|\; \sup_{0\le t\le T} \|v(t, \cdot)\|_{H^1} \le b \right\}$$

for some $b > 0$ and $T > 0$ to be determined. From Propositions 3.1 and 3.3, we have

$$\begin{aligned} \sup_{0\le t\le T} \|Fv\|_{H^1} &\le K_3\|w_0\|_{H^1} + K_1(T^{1/4} + T) \sup_{0\le t\le T} \|(vv_x)(t, \cdot)\|_{L^2} \\ &\le K_3\|w_0\|_{H^1} + K_1(T^{1/4} + T) \left(\sup_{0\le t\le T} \|v(t, \cdot)\|_{H^1} \right)^2, \end{aligned}$$

where K_1 and K_3 are the constants in Propositions 3.1 and 3.3, respectively, and are independent of T and v. If we let $b = 2K_3\|w_0\|_{H^1}$ and $K_1(T^{1/4}+T)b = 2K_1K_3\|w_0\|_{H^1}(T^{1/4}+T) < 1/2$, then $\sup_{0\le t\le T} \|Fv\|_{H^1} \le b$ and F maps $\mathcal{S}_{T,b}$ into itself. For $v_1, v_2 \in \mathcal{S}_{T,b}$, we have

$$Fv_1 - Fv_2 = -\int_0^t S(\tau)(v_1w_x + wv_{2x})(t - \tau, \cdot)d\tau,$$

where $w = v_1 - v_2$. Thus by Proposition 3.1 we have

$$\begin{aligned} \sup_{0\le t\le T} \|Fv_1 - Fv_2\|_{H^1} &\le K_1(T^{1/4} + T) \sup_{0\le t\le T} \|v_1w_x + wv_{2x}\|_{L^2} \\ &\le K_1(T^{1/4} + T) \left(\sup_{0\le t\le T} \|v_1\|_{H^1} + \sup_{0\le t\le T} \|v_2\|_{H^1} \right) \sup_{0\le t\le T} \|w\|_{H^1} \\ &\le 2K_1b(T^{1/4} + T) \sup_{0\le t\le T} \|w\|_{H^1}. \end{aligned}$$

Since $2K_1b(T^{1/4}+T) < 1$ by the choice of b and T, F is a contraction, and by the contraction mapping theorem F has a unique fixed point in $\mathcal{S}_{T,b}$. If we choose $K_1(T^{1/4} + T)b = 1/4 = 2K_1K_3(T^{1/4} + T)\|w_0\|_{H^1}$, then $T \to +\infty$ as $\|w_0\|_{H^1} \to 0$. Finally it is obvious that for any $T' < T(\|\phi(x)\|_{H^1})$, there is a neighborhood U of ϕ in H^1 such that $\mathcal{G}$ is well defined from U to $C(0, T'; H^1)$. For any $w_1, w_2 \in U$, let $u_1 = \mathcal{G}w_1$, $u_2 = \mathcal{G}w_2$, and $w = u_1 - u_2$. Thus

$$w = S(t)(w_1 - w_2) - \int_0^t S(\tau)(u_1w_x + wu_{2x})(t - \tau, \cdot)d\tau.$$

By Propositions 3.1 and 3.3, we have

$$\begin{aligned} \sup_{0\le t\le T} \|w\|_{H^1} &\le K_3\|w_1 - w_2\|_{H^1} + K_1(T^{1/4} + T) \left(\sup_{0\le t\le T} \|u_1\|_{H^1} \right. \\ &\qquad \left. + \sup_{0\le t\le T} \|u_2\|_{H^1} \right) \sup_{0\le t\le T} \|w\|_{H^1} \\ &\le K_3\|w_1 - w_2\|_{H^1} + 2K_1(T^{1/4} + T)b \sup_{0\le t\le T} \|w\|_{H^1}. \end{aligned}$$

Thus $\sup_{0\le t\le T}\|w\|_{H^1}\le K_3(1-2K_1(T^{1/4}+T)b)^{-1}\|w_1-w_2\|_{H^1}$, which implies $\mathcal{G}$ is Lipschitz continuous from U to $C(0,T;H^1)$.

After we have the local existence, we prove the global existence of the solutions of (4.1)-(4.3) with small initial data.

THEOREM 4.2. *There exists a $\beta>0$ such that for any $w_0\in H^1$ with $\|w_0\|_{H^1}\le\beta$, (4.1)-(4.3) have a unique solution $u\in C(\mathbf{R}^+,H^1)$ with $\mathbf{R}^+=[0,+\infty)$.*

Proof. Let

$$\mathcal{S}_b=\left\{v\in C(\mathbf{R}^+;H^1)\;\middle|\;\sup_{0\le t<+\infty}\|v(t,\cdot)\|_{H^1}\le b \text{ and } v(t,0)=v(t,1)\right\}$$

with $b>0$ to be determined. Let

$$Fv=S(t)w_0-\int_0^t S(\tau)(vv_x)(\tau)d\tau.$$

Using Propositions 3.2 and 3.3, we have

$$\begin{aligned}\sup_{0\le t<+\infty}\|Fv\|_{H^1}&\le K_3\|w_0\|_{H^1}+K_2\sup_{0\le t<+\infty}\|vv_x\|_{L^2}\\&\le K_3\|w_0\|_{H^1}+K_2\left(\sup_{0\le t<+\infty}\|v\|_{H^1}\right)^2,\end{aligned}$$

since $\int_0^1 vv_x dx=0$, where K_2 and K_3 are the constants in Propositions 3.2 and 3.3, respectively. Choose $b>0$ and $\beta>0$ such that $K_2b<(1/2)$ and $K_3\beta<(b/2)$. Then F maps $\mathcal{S}_b$ into itself since $\int_0^t S(\tau)(vv_x)(t-\tau)d\tau\in\mathcal{D}(A)$ and $S(t)w_0\in H^1$ with $(S(t)w_0)(0)=(S(t)w_0)(1)$ by the construction of $S(t)$. The contraction property of F can be obtained similarly in Theorem 4.1. Thus F has a unique fixed point in $\mathcal{S}_b$, which is a solution of (4.1)-(4.3).

Finally we show the exponential decay of the global solutions of (4.1)-(4.3) as $t\to+\infty$.

THEOREM 4.3. *There exists δ_1 with $\beta\ge\delta_1>0$ such that for any $w_0\in H^1$ with $\|w_0\|_{H^1}<\delta_1$, the solution $u(t,x)$ of (4.1)-(4.3) satisfies*

$$\|u(t,\cdot)-[w_0]\|_{L^2}\le Ke^{-\rho t}\|w_0-[w_0]\|_{L^2}$$

for all $t\ge0$, where β is defined in Theorem 4.2, $K>0$ and $\rho>0$ are independent of w_0 and t, and $[w_0]=\int_0^1 w_0(x)dx$.

Proof. Since $[u]=[w_0]$ for all $t\ge0$ and the solution is periodic in L^2, i.e., $u(t,0)=u(t,1)$, by substituting $u-[u]$ into the equation and changing the variable, we can assume that $[w_0]=[u]=0$. For $\delta_1<\beta$, (4.1)-(4.3) has a unique global solution u satisfying

$$u(x,t)=S(t)w_0+\int_0^t S(\tau)(uu_x)(t-\tau,\cdot)d\tau.$$

First let us estimate $\|S(t)w_0\|_{L^2}$ for $t\ge1$. In (2.11), we note that if $\mu_r+i\mu_i\in\Gamma_\pm$, then $3\mu_r\mu_i^2-\mu_r^3<0$. However, since $\mu_r\to-\delta^{2/3}/3=\delta_0$ as $\mu_i\to+\infty$, then $3\mu_r\mu_i^2-\mu_r^3\le-2\beta_1$ for some $\beta_1>0$ whenever $\mu_r+i\mu_i\in\Gamma_\pm$. Since $[w_0]=0$, by

(2.11) and (2.13) we have

$$\begin{aligned}\|S(t)w_0\|_{L^2}^2 &= e^{-2\beta_1 t}(2\pi)^{-2}\int_0^1 \left|\left(\int_{\Gamma_+}+\int_{\Gamma_-}\right)3\mu^2 e^{-(\mu^3-\beta_1)t}\right.\\ &\qquad\left.\times\int_0^1 G(-\mu^3,x,\xi)w_0(\xi)d\xi d\mu\right|^2 dx\\ &\le Ke^{-2\beta_1 t}\left[\left(\int_{\Gamma_+}+\int_{\Gamma_-}\right)|e^{-(\mu^3-\beta_1)t}||d\mu|\right]^2\|w_0\|_{L^2}^2\\ &\le Ke^{-2\beta_1 t}\left(\int_{-\infty}^{+\infty}\exp(-\max(\beta_1,\delta^{2/3}\mu_i^2-\beta_1)t)d\mu_i\right)^2\|w_0\|_{L^2}^2\\ &\le Ke^{-2\beta_1 t}\|w_0\|_{L^2}^2.\end{aligned}$$

Thus for $t \ge 1$, $\|S(t)w_0\|_{L^2} \le Ke^{-\beta_1 t}\|w_0\|_{L^2}$. By the semigroup property of $S(t)$, we have $\|S(t)w_0\|_{L^2} \le Ke^{-\beta_1 t}\|w_0\|_{L^2}$ for all $t \ge 0$. Since $u(0) = u(1)$, $\int_0^1 uu_x dx = 0$, and $G(-\mu^3, x, \xi)$ is continuous at $x = \xi$, the integration by parts yields

$$\begin{aligned}S(t-\tau)((u^2/2)_x)(\tau,x) &= \frac{-1}{2\pi i}\int_{\Gamma_+\cup\Gamma_-}3\mu^2 e^{-\mu^3(t-\tau)}\int_0^1 G(-\mu^3,x,\xi)(u^2(\tau,\xi)/2)_\xi d\xi d\mu\\ &= \frac{1}{2\pi i}\int_{\Gamma_+\cup\Gamma_-}3\mu^2 e^{-\mu^3(t-\tau)}\int_0^1 G_\xi(-\mu^3,x,\xi)(u^2(\tau,\xi)/2)d\xi d\mu\\ &= \frac{e^{-\beta_1(t-\tau)}}{2\pi i}\int_{\Gamma_+\cup\Gamma_-}3\mu^2 e^{-(\mu^3-\beta_1)(t-\tau)}\int_0^1 G(-\mu^3,x,\xi)_\xi(u^2(\tau,\xi)/2)d\xi d\mu.\end{aligned}$$

By using the same derivation of the estimate for $(\partial S(t-\tau)f(\tau,\cdot)/\partial x)$ in Proposition 3.1, we can obtain

$$\begin{aligned}\left\|\int_0^t S(\tau)(u^2(t-\tau,\cdot)/2)_x d\tau\right\|_{L^2} &\le K\int_0^t \tau^{-3/4}e^{-\beta_1\tau}\|u^2(t-\tau,\cdot)\|_{L^2}d\tau\\ &\le K\int_0^t (t-\tau)^{-3/4}e^{-\beta_1(t-\tau)}\|u^2(\tau,\cdot)\|_{L^2}d\tau\\ &\le K\sup_{0\le\tau<+\infty}\|u(\tau,\cdot)\|_{H^1}\int_0^t (t-\tau)^{-3/4}e^{-\beta_1(t-\tau)}\|u(\tau,\cdot)\|_{L^2}d\tau.\end{aligned}$$

Thus the solution $u(t,x)$ satisfies

$$\begin{aligned}&\|u(t,\cdot)\|_{L^2}\\ &\le K\left(e^{-\beta_1 t}\|w_0\|_{L^2}+\sup_{0\le\tau<+\infty}\|u(\tau,\cdot)\|_{H^1}\int_0^t (t-\tau)^{-3/4}e^{-\beta_1(t-\tau)}\|u(\tau,\cdot)\|_{L^2}d\tau\right),\end{aligned}$$

where K is independent of t, u and w_0. Let $K\|w_0\|_{L^2} = c_1$, $K\sup_{0\le\tau<+\infty}\|u(\tau,\cdot)\|_{H^1} = c_2$, and $w(t) = e^{\beta_1 t}\|u(t,\cdot)\|_{L^2}$. Then $w(t)$ satisfies

$$w(t) \le c_1 + c_2\int_0^t (t-\tau)^{-3/4}w(\tau)d\tau.$$

Define $g(t) = \sup_{[t]\le\tau<[t]+1} w(\tau)$, where $[t]$ is the largest integer less than or equal to t. Then

$$w(t) \le c_1 + c_2 \sum_{k=0}^{[t]-1} \int_k^{k+1} (t-\tau)^{-3/4} w(\tau)\,d\tau + c_2 \int_{[t]}^{t} (t-\tau)^{-3/4} w(\tau)\,d\tau$$

$$\le c_1 + c_2 \sum_{k=0}^{[t]-1} g(k) 4((t-k)^{1/4} - (t-k-1)^{1/4}) + c_2 g([t]) 4(t-[t])^{1/4}.$$

But

$$(t-k)^{1/4} - (t-k-1)^{1/4} = \left(\sum_{i=0}^{3} (t-k)^{i/4} (t-k-1)^{(3-i)/4} \right)^{-1} \le (t-k)^{-3/4} \le 1$$

for $k \le [t]-1$. Thus

$$w(t) \le c_1 + 4c_2 \sum_{k=0}^{[t]-1} g(k) + 4c_2 g([t])$$

$$= c_1 + 4c_2 \sum_{k=1}^{[t]} g(k),$$

which implies

$$\sup_{[t]\le\tau<[t]+1} w(\tau) \le c_1 + 4c_2 \sum_{k=1}^{[t]} g(k)$$

and

$$g(t) \le c_1 + 4c_2 \int_0^{[t]+1} g(\tau)\,d\tau$$

$$= c_1 + 4c_2 \int_0^t g(\tau)\,d\tau + 4c_2 \int_t^{[t]+1} g(\tau)\,d\tau$$

$$\le c_1 + 4c_2 g(t) + 4c_2 \int_0^t g(\tau)\,d\tau.$$

If $4c_2 \le 1/2$, then

$$g(t) \le \frac{c_1}{1-4c_2} + \frac{4c_1}{1-4c_2} \int_0^t g(\tau)\,d\tau,$$

which yields

$$g(t) \le \frac{c_1}{1-4c_2} \exp\left(\frac{4c_1 t}{1-4c_2} \right) \le 2c_1 e^{8c_2 t}$$

by the Gronwall inequality. Therefore, from the definition of $g(t)$, $w(t)$, c_1, and c_2, we have

$$\|u(t,\cdot)\|_{L^2} \le 2K\|w_0\|_{L^2} \exp\Big(-(\beta_1 - 8K \sup_{0\le\tau<+\infty} \|u(\tau,\cdot)\|_{H^1})t \Big).$$

By the proof of Theorem 4.2, we choose $\|w_0\|_{H^1}$ so small that

$$\beta_1 - 8K \sup_{0\le\tau<+\infty} \|u(\tau,\cdot)\|_{H^1} \ge \beta_2 > 0$$

and $K \sup_{0\le\tau<+\infty} \|u(\tau,\cdot)\|_{H^1} \le (1/8)$. Then $\|u(t,\cdot)\|_{L^2} \le 2K\|w_0\|_{L^2} e^{-\beta_2 t}$, which implies Theorem 4.3.

Appendix 1. The Green's function $G(\lambda, x, \xi)$ is the following: for $0 < x < \xi < 1$,

$$\begin{aligned} G(\lambda, x, \xi) &= (3\lambda)^{-1}\mu_1 e^{\mu_1(x-\xi)} + (3\lambda(3-B))^{-1}\Big[(e^{-\mu_1}-1)\mu_1 e^{\mu_1(x-\xi)} \\ &\quad +(2-e^{-\mu_1}-e^{-\mu_3})\mu_2 e^{\mu_2(x-\xi)} + (2-e^{-\mu_1}-e^{-\mu_2})\mu_3 e^{\mu_3(x-\xi)} \\ &\quad +(e^{\mu_2}-e^{-\mu_1})\mu_3 e^{\mu_1 x-\mu_2\xi} + (e^{\mu_3}-e^{-\mu_1})\mu_2 e^{\mu_1 x-\mu_3\xi} \\ &\quad +(e^{\mu_3}-e^{-\mu_2})\mu_1 e^{\mu_2 x-\mu_3\xi} + (e^{\mu_1}-e^{-\mu_2})\mu_3 e^{\mu_2 x-\mu_1\xi} \\ &\quad +(e^{\mu_1}-e^{-\mu_3})\mu_2 e^{\mu_3 x-\mu_1\xi} + (e^{\mu_2}-e^{-\mu_3})\mu_1 e^{\mu_3 x-\mu_2\xi}\Big] \\ &= (3\lambda)^{-1}\mu_1 e^{\mu_1(x-\xi)} + R(\mu, x, \xi); \end{aligned}$$

for $0 < \xi < x < 1$,

$$\begin{aligned} G(\lambda, x, \xi) &= (3\lambda(3-B))^{-1}\Big[(e^{-\mu_1}-1)\mu_1 e^{\mu_1(x-\xi)} \\ &\quad +(e^{-\mu_2}-1)\mu_2 e^{\mu_2(x-\xi)} + (e^{-\mu_3}-1)\mu_3 e^{\mu_3(x-\xi)} \\ &\quad +(e^{\mu_2}-e^{-\mu_1})\mu_3 e^{\mu_1 x-\mu_2\xi} + (e^{\mu_3}-e^{-\mu_1})\mu_2 e^{\mu_1 x-\mu_3\xi} \\ &\quad +(e^{\mu_3}-e^{-\mu_2})\mu_1 e^{\mu_2 x-\mu_3\xi} + (e^{\mu_1}-e^{-\mu_2})\mu_3 e^{\mu_2 x-\mu_1\xi} \\ &\quad +(e^{\mu_1}-e^{-\mu_3})\mu_2 e^{\mu_3 x-\mu_1\xi} + (e^{\mu_2}-e^{-\mu_3})\mu_1 e^{\mu_3 x-\mu_2\xi}\Big] \\ &= R(\mu, x, \xi), \end{aligned}$$

where $B = e^{-\mu_1} + e^{-\mu_2} + e^{-\mu_3}$. Here $\mu_1^3 = -\lambda$, $\mu_2 = \omega\mu_1$, and $\mu_3 = \omega^2\mu_1$ with $\omega = (-1 + i\sqrt{3})/2$. Next we show that

$$\int_0^1\int_0^1 \left|\frac{\partial}{\partial x}R(\mu, x, \xi)\right|^2 d\xi dx \le K/|\mu|^3,$$

where $\mu \in \Gamma_\pm$ with $|\mu| > 0$. Note that on $\Gamma_\pm$, $\mu_i \to \pm\infty$ and μ_r is negative and bounded with $\mu_r \to -\delta^{2/3}/3$ as $\mu_i \to \pm\infty$. Write $R(\mu, x, \xi)$ as

$$R(\mu, x, \xi) = A_0 + B_1 + B_2 + A_1 + A_2 + \cdots + A_6.$$

Then we derive the following estimates:

$$\int_0^1\int_0^1 \left|\frac{\partial A_i}{\partial x}\right|^2 d\xi dx \le K|\mu|^{-4} \quad \text{for} \quad i = 0, 3, 6,$$

$$\int_0^1\int_0^1 \left|\frac{\partial A_i}{\partial x}\right|^2 d\xi dx \le K|\mu|^{-3} \quad \text{for} \quad i = 1, 2, 4, 5,$$

$$\int_0^1\int_0^1 \left|\frac{\partial B_i}{\partial x}\right|^2 d\xi dx \le K|\mu|^{-3} \quad \text{for} \quad i = 1, 2.$$

Here we check only a few of them for $\mu \in \Gamma_+$. The others and the case for $\mu \in \Gamma_-$ are similar and left to the reader. By noting that $|3 - B| \ge K > 0$ for $\mu \in \Gamma_+$, μ_r is bounded and small, and $|3 - B|^{-1} \le Ke^{-\sqrt{3}\mu_i/2}$ if $\mu \in \Gamma_+$ and $\mu_i \to +\infty$, we have

$$\int_0^1\int_0^1 \left|\frac{\partial A_0}{\partial x}\right|^2 dxd\xi = \int_0^1\int_0^1 |3\lambda(3-B)|^{-2}|e^{-\mu_1}-1|^2|\mu_1|^4|e^{\mu_1(x-\xi)}|^2 d\xi dx$$
$$\le K|\mu|^{-2}e^{-\sqrt{3}|\mu_i|}\int_0^1\int_0^1 |e^{\mu_1(x-\xi)}|^2 d\xi dx \le K|\mu|^{-4},$$
$$\int_0^1\int_0^1 \left|\frac{\partial A_3}{\partial x}\right|^2 dxd\xi = \int_0^1\int_0^1 |3\lambda(3-B)|^{-2}|e^{\mu_3}-e^{-\mu_2}|^2|\mu_1\mu_2|^2|e^{2\mu_2 x-2\mu_3\xi}|dxd\xi$$
$$\le K|\mu|^{-2}e^{-\sqrt{3}|\mu_i|}\int_0^1\int_0^1 |e^{\mu_3}-e^{-\mu_2}|^2|e^{\mu_2 x-\mu_3\xi}|^2 dxd\xi$$
$$\le K|\mu|^{-2}\int_0^1\int_0^1 e^{-\sqrt{3}|\mu_i|x-\sqrt{3}|\mu_i|\xi}dxd\xi \le K|\mu|^{-4},$$
$$\int_0^1\int_0^1 \left|\frac{\partial A_1}{\partial x}\right|^2 dxd\xi = \int_0^1\int_0^1 |3\lambda(3-B)|^{-2}|e^{\mu_2}-e^{-\mu_1}|^2|\mu_1\mu_3|^2|e^{2\mu_1 x-2\mu_2\xi}|dxd\xi$$
$$\le K|\mu|^{-2}e^{-\sqrt{3}|\mu_i|}\int_0^1\int_0^1 e^{\sqrt{3}|\mu_i|\xi}dxd\xi \le K|\mu|^{-3},$$
$$\int_0^1\int_0^1 \left|\frac{\partial B_1}{\partial x}\right|^2 dxd\xi = \int_0^1\int_x^1 \left|\frac{2-e^{-\mu_1}-e^{-\mu_3}}{3\lambda(3-B)}\right|^2 |\mu_2|^4|e^{\mu_2(x-\xi)}|^2 dxd\xi$$
$$+\int_0^1\int_0^x \left|\frac{e^{-\mu_2}-1}{3\lambda(3-B)}\right|^2 |\mu_2|^4|e^{\mu_2(x-\xi)}|^2 dxd\xi$$
$$\le K|\mu|^{-2}\left(e^{-\sqrt{3}|\mu_i|}\int_0^1\int_x^1 e^{-\sqrt{3}|\mu_i|(x-\xi)}d\xi dx + \int_0^1\int_0^x e^{-\sqrt{3}|\mu_i|(x-\xi)}d\xi dx\right)$$
$$\le K|\mu|^{-3}.$$

By these estimates, we have the L^2-estimates for the x-derivative of $R(\mu, x, \xi)$. Here we note that we can obtain the same estimates for the ξ-derivative of $R(\mu, x, \xi)$.

Appendix 2. We prove the following lemma.

LEMMA A.1.

$$I \overset{\text{def}}{=} \left|\int_a^{+\infty} \xi^\alpha e^{i(\xi^3 t+\xi x)}d\xi\right| \le Kt^{-(1+\alpha)/3}(1+|x^3/t|^{(2\alpha-1)/12}),$$

where K is independent of $t > 0$, $x \in \mathbf{R}$, and $0 \le \alpha \le 1$.

Proof. Let $y = \xi t^{1/3}$. Then I becomes

$$I = t^{-(1+\alpha)/3}\left|\int_{at^{1/3}}^{+\infty} y^\alpha \exp(i(y^3+(xt^{-1/3})y))dy\right|.$$

If $\lambda = xt^{-1/3}$, I is

$$I = t^{-(1+\alpha)/3}\left|\int_{at^{1/3}}^{+\infty} y^\alpha e^{i(y^3+\lambda y)}dy\right|.$$

Let $\varphi_0(y)$ be a C^∞-function such that $\varphi_0 \equiv 0$ for $|y| < 1$ and $\varphi_0 \equiv 1$ for $|y| \ge 2$. Thus

$$I \le t^{-(1+\alpha)/3}\left|\int_{at^{1/3}}^{+\infty} y^\alpha e^{i(y^3+\lambda y)}\varphi_0(y)dy\right|$$
$$+ t^{-(1+\alpha)/3}\left|\int_{at^{1/3}}^{+\infty} y^\alpha e^{i(y^3+\lambda y)}(1-\varphi_0(y))dy\right| \overset{\text{def}}{=} I_1 + I_2.$$

Then $|I_2| \le Kt^{-(1+\alpha)/3}$. If $\lambda > -2$, then $3y^2 + \lambda > 1$ for $y \in [at^{1/3}, +\infty) \cap \{|y| \ge 1\}$. Thus, from integration by parts,

$$I_1 \le t^{-(1+\alpha)/3}\left(\left|\int_{\tilde{1}}^{+\infty}(-1)e^{i(y^3+\lambda y)}\left(\frac{y^\alpha\varphi_0(y)}{i(3y^2+\lambda)}\right)_y dy\right| + K\right)$$

$$\le t^{-(1+\alpha)/3}\left(\left|\int_{\tilde{1}}^{+\infty} e^{i(y^3+\lambda y)}\left((i(3y^2+\lambda))^{-1}\left(\frac{y^\alpha\varphi_0(y)}{i(3y^2+\lambda)}\right)_y\right)_y dy\right| + K\right) \le K,$$

where $\tilde{1} = \max(1, at^{1/3})$. In the following we assume $\lambda \le -2$ and $-\lambda = -|\lambda|$. Let $y = |\lambda|^{1/2}\eta$, which yields

$$I_1 = t^{-(1+\alpha)/3}|\lambda|^{(1+\alpha)/2}\left|\int_{\tilde{1}|\lambda|^{-1/2}}^{+\infty}\eta^\alpha e^{i|\lambda|^{3/2}(\eta^3-\eta)}\varphi_0(|\lambda|^{1/2}\eta)d\eta\right|.$$

Let $\psi_1(\eta)$ have a support in $(-\infty, (1/\sqrt{3})-\hat{\delta})$, $\psi_2(\eta)$ have a support in $((1/\sqrt{3})-2\hat{\delta}, (1/\sqrt{3})+2\hat{\delta})$, and $\psi_3(\eta)$ have a support in $((1/\sqrt{3})+\hat{\delta}, +\infty)$ with $\psi_1+\psi_2+\psi_3 = 1$. Thus

$$I_1 = t^{-(1+\alpha)/3}|\lambda|^{(1+\alpha)/2}\left|\int_{\tilde{1}|\lambda|^{-1/2}}^{+\infty}\eta^\alpha e^{i|\lambda|^{3/2}(\eta^3-\eta)}\varphi_0(|\lambda|^{1/2}\eta)(\psi_1(\eta)+\psi_2(\eta)+\psi_3(\eta))d\eta\right|$$

$$\le II_1 + II_2 + II_3.$$

Using integration by parts several times, we have

$$II_3 = t^{-(1+\alpha)/3}|\lambda|^{(1+\alpha)/2}\left|\int_{(1/\sqrt{3})+\hat{\delta}}^{+\infty} e^{i|\lambda|^{3/2}(\xi^3-\xi)}\left(\frac{\xi^\alpha\psi_3(\xi)\varphi_0(|\lambda|^{1/2}\xi)}{|\lambda|^{3/2}(3\xi^2-1)}\right)_\xi d\xi\right|$$

$$\le Kt^{-(1+\alpha)/3}|\lambda|^{(1+\alpha)/2-(3/2)}.$$

By a similar argument, $|II_1| \le Kt^{-(1+\alpha)/3}|\lambda|^{(\alpha-2)/2}$. Finally, using a corollary in [16, p. 311],

$$II_2 = t^{-(1+\alpha)/3}|\lambda|^{(1+\alpha)/2}\left|\int_{(1/\sqrt{3})-2\hat{\delta}}^{(1/\sqrt{3})+2\hat{\delta}}\eta^\alpha e^{i|\lambda|^{3/2}(\eta^3-\eta)}\varphi_0(|\lambda|^{1/2}\eta)\psi_2(\eta)d\eta\right|$$

$$\le Kt^{-(1+\alpha)/3}|\lambda|^{(1+\alpha)/2-(3/4)}$$

since $\lambda \le -2$, and if $\hat{\delta} > 0$ is small enough, $\varphi_0(|\lambda|^{1/2}\eta) = 1$ for $\lambda \le -14$ and $|\lambda|^{1/2}((1/\sqrt{3})-2\hat{\delta}) > 2$. Therefore, for $\lambda \le -2$, $|I_1| \le Kt^{-(1+\alpha)/2}|\lambda|^{(2\alpha-1)/4}$. By the definition of I, we have

$$|I| \le Kt^{-(1+\alpha)/3}\left(1 + |x^3/t|^{(2\alpha-1)/12}\right).$$

Appendix 3. In this appendix, we shall show that A does not generate an analytic semigroup. The proof presented here was given by Komornik [6].

By Lemma 2.1, there are no eigenvalues λ of A on the imaginary axis except $\lambda = 0$. Thus for $\lambda = i\nu$ with either $\nu < 0$ or $\nu > 0$, $(\lambda I - A)^{-1}$ exists. In order to have A generate an analytic semigroup, $(\lambda I - A)^{-1}$ must satisfy

$$\|(\lambda I - A)^{-1}\| \le \frac{C}{|\lambda|} = \frac{C}{|\nu|}$$

for $\lambda = i\nu$ with ν real and large, where C is a fixed constant [10]. However, from (2.6),

$$(\lambda I - A)^{-1}f = \int_0^1 G(\lambda, x, \xi)f(\xi)d\xi$$

for $f \in L^2$, where $G(\lambda, x, \xi)$ is defined in Appendix 1, and in the following, notations from Appendix 1 will be used. Now we estimate only $\|(\lambda T - A)^{-1}\|$ for $\lambda = i\nu$ with $\nu > 0$. Similar estimates can be obtained for $\lambda = i\nu$ with $\nu < 0$. If $\lambda = i\nu$ with $\nu > 0$, then $\mu_1 = \sqrt[3]{\nu}i$, $\mu_2 = ((-\sqrt{3} - i)/2)\sqrt[3]{\nu}$, $\mu_3 = ((\sqrt{3} - i)/2)\sqrt[3]{\nu}$. From Appendix 1,

$$G(\lambda, x, \xi) = (3\lambda)^{-1}\mu_1 e^{\mu_1(x-\xi)}\chi_{(x,1)}(\xi) + R(\mu, x, \xi),$$

where $R(\mu, x, \xi) = A_0 + B_1 + B_2 + A_1 + \cdots + A_6$ and $\chi_{(x,1)}$ is the characteristic function of interval $(x, 1)$. First we show that for $\mu = \mu_1 = \sqrt[3]{\nu}i$ large,

$$\int_0^1 \int_0^1 |R(\mu, x, \xi)|^2 d\xi dx \le K|\nu|^{-5/3},$$

where K is denoted as a generic constant. We need to derive the following estimates:

$$\int_0^1 \int_0^1 |A_i|^2 d\xi dx \le K|\nu|^{-2} \quad \text{for} \quad i = 0, 3, 6,$$

$$\int_0^1 \int_0^1 |A_i|^2 d\xi dx \le K|\nu|^{-5/3} \quad \text{for} \quad i = 1, 2, 4, 5,$$

$$\int_0^1 \int_0^1 |B_i|^2 d\xi dx \le K|\nu|^{-5/3} \quad \text{for} \quad i = 1, 2.$$

Here we check only several of them. The others are similar and left to the reader. Note that $|3 - B| \ge K_1 > 0$ for ν large and $|3 - B|^{-1} \le Ke^{-\sqrt{3}\sqrt[3]{\nu}/2}$. We then have

$$\begin{aligned}
\int_0^1 \int_0^1 |A_0|^2 dx d\xi &= \int_0^1 \int_0^1 |3\lambda(3 - B)|^{-2} |e^{-\mu_1} - 1|^2 |\mu_1|^2 |e^{\mu_1(x-\xi)}|^2 d\xi dx \\
&\le K|\nu|^{-4/3} e^{-\sqrt{3}\sqrt[3]{\nu}} \int_0^1 \int_0^1 |e^{i\sqrt[3]{\nu}(x-\xi)}|^2 d\xi dx \le K|\nu|^{-2}, \\
\int_0^1 \int_0^1 |A_3|^2 dx d\xi &= \int_0^1 \int_0^1 |3\lambda(3 - B)|^{-2} |e^{\mu_3} - e^{-\mu_2}|^2 |\mu_1|^2 |e^{2\mu_2 x - 2\mu_3 \xi}| dx d\xi \\
&\le K|\nu|^{-4/3} e^{-\sqrt{3}\sqrt[3]{\nu}} \int_0^1 \int_0^1 |e^{\mu_3} - e^{-\mu_2}|^2 |e^{-\sqrt{3}\sqrt[3]{\nu}x - \sqrt{3}\sqrt[3]{\nu}\xi}|^2 dx d\xi \le K|\nu|^{-2}, \\
\int_0^1 \int_0^1 |A_1|^2 dx d\xi &= \int_0^1 \int_0^1 |3\lambda(3 - B)|^{-2} |e^{\mu_2} - e^{-\mu_1}|^2 |\mu_3|^2 |e^{2\mu_1 x - 2\mu_2 \xi}| dx d\xi \\
&\le K|\nu|^{-4/3} e^{-\sqrt{3}\sqrt[3]{\nu}} \int_0^1 \int_0^1 e^{\sqrt{3}\sqrt[3]{\nu}\xi} dx d\xi \le K|\nu|^{-5/3}, \\
\int_0^1 \int_0^1 |B_1|^2 dx d\xi &= \int_0^1 \int_x^1 \left| \frac{2 - e^{-\mu_1} - e^{-\mu_3}}{3\lambda(3 - B)} \right|^2 |\mu_2|^2 |e^{\mu_2(x-\xi)}|^2 dx d\xi \\
&\quad + \int_0^1 \int_0^x \left| \frac{e^{-\mu_2} - 1}{3\lambda(3 - B)} \right|^2 |\mu_2|^2 |e^{\mu_2(x-\xi)}|^2 dx d\xi \\
&\le K|\nu|^{-4/3} \left(e^{-\sqrt{3}\sqrt[3]{\nu}} \int_0^1 \int_x^1 e^{-\sqrt{3}\sqrt[3]{\nu}(x-\xi)} d\xi dx + \int_0^1 \int_0^x e^{-\sqrt{3}\sqrt[3]{\nu}(x-\xi)} d\xi dx \right) \\
&\le K|\nu|^{-5/3}.
\end{aligned}$$

Thus from above estimates for $R(\mu, x, \xi)$, we obtain that for $f \in L^2$,

$$\|(\lambda I - A)^{-1} f\|_{L^2} \geq \left\| \int_x^1 (3\lambda)^{-1} \mu_1 e^{\mu_1(x-\xi)} f(\xi) d\xi \right\|_{L^2} - \left\| \int_0^1 R(\mu, x, \xi) f(\xi) d\xi \right\|_{L^2}$$

$$\geq \left\| \int_x^1 (3\lambda)^{-1} \mu_1 e^{\mu_1(x-\xi)} f(\xi) d\xi \right\|_{L^2} - K|\nu|^{-5/6} \|f(\xi)\|_{L^2}.$$

Consider the function $f(x) = e^{ix\sqrt[3]{\nu}} \in L^2$. Then

$$\left\| \int_x^1 (3\lambda)^{-1} \mu_1 e^{\mu_1(x-\xi)} e^{\sqrt[3]{\nu} i\xi} d\xi \right\|_{L^2} = \|(3\lambda)^{-1}(1-x)\mu_1 e^{\mu_1 x}\|_{L^2}$$

$$= 3^{-3/2}|\nu|^{-2/3} \|e^{\mu_1 x}\|_{L^2}.$$

Thus

$$\|(\lambda I - A)^{-1} e^{\mu_1 x}\|_{L^2} \geq (3^{-3/2}|\nu|^{-2/3} - K|\nu|^{-5/6}) \|e^{\mu_1 x}\|_{L^2},$$

which implies $\|(\lambda I - A)^{-1}\| \geq |\nu|^{-2/3}(3^{-3/2} - K|\nu|^{-1/6})$. Therefore, for ν large, we have that $\|(\lambda I - A)^{-1}\| \geq (3^{-3/2}/2)|\nu|^{-2/3}$. Thus, A does not generate an analytic semigroup.

Acknowledgments. The author wishes to thank Professor D. L. Russell for suggesting the problem studied in this paper and for valuable discussions in the course of development of the work.

REFERENCES

[1] J. L. Bona and R. Smith, *The initial value problem for the Korteweg–de Vries equation*, Proc. Roy. Soc. London Ser. A, 278 (1978), pp. 555-601.

[2] J. Bourgain, *Fourier transform restriction phenomena for certain lattice subsets and applications to non-linear evolution equations, part* II*: the KdV equation*, Geom. Funct. Analy., 3 (1993), pp. 209-262.

[3] N. Dunford and J. T. Schwartz, *Linear Operators, Part* III, Wiley-Interscience, New York, 1971.

[4] T. Kato, *On the Cauchy problem for the (generalized) Korteweg–de Vries equations*, Advances in Mathematics supplementary studies, Stud. Appl. Math., 8 (1983), pp. 93-128.

[5] C. E. Kenig, G. Ponce, and L. Vega, *The Cauchy problem for the Korteweg–de Vries equation in Sobolev spaces of negative indices*, Duke Math. J., 71 (1993), pp. 1-21.

[6] V. Komornik, Private communications, 1992.

[7] V. Komornik, D. L. Russell, and B.-Y. Zhang, *Stabilisation de l'equation de Korteweg–de Vries*, C. R. Acad. Sci. Paris, 312 (1991), pp. 841-843.

[8] ———, *Control and stabilization of the Korteweg–de Vries equation on a periodic domain*, J. Differential Equations, to appear.

[9] R. M. Miura, *The Korteweg–de Vries equation: A survey of results*, SIAM Rev., 18 (1976), pp. 412-459.

[10] A. Pazy, *Semigroup of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci., 44, Springer-Verlag, New York, 1983.

[11] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639-739.

[12] D. L. Russell and B.-Y. Zhang, *Controllability and stabilizability of the third order linear dispersion equation on a periodic domain*, SIAM J. Control Optim., 31 (1993), pp. 659-676.

[13] ———, *Smoothing and decay properties of solutions of the Korteweg–de Vries equation on a periodic domain with point dissipation*, J. Math. Anal. Appl., 190 (1995), pp. 449-488.

[14] ———, *Exact Controllability and Uniform Stability of Small Solutions of the Periodic Korteweg–de Vries Equation*, preprint.

[15] J. C. Saut and R. Teman, *Remarks on the Korteweg–de Vries equation*, Israel J. Math., 24 (1976), pp. 78-87.

[16] E. M. Stein, *Oscillatory integrals in Fourier analysis*, in Beijing Lectures in Harmonic Analysis, Ann. of Math. Stud. 112, E. M. Stein, ed., Princeton University Press, Princeton, NJ, 1986, pp. 307-355.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 913-921, May 1996

007

UNIQUE DETERMINATION OF MULTIPLE CRACKS BY TWO MEASUREMENTS*

GIOVANNI ALESSANDRINI† AND ALVARO DIAZ VALENZUELA‡

Abstract. We study the inverse problem of determining multiple cracks in a planar conductor by electrostatic measurements at the boundary. We prove that two measurements at the boundary suffice to identify multiple cracks with any number of components. We treat the problem under no regularity assumptions on the cracks and on the background conductivity.

Key words. inverse problems, cracks, elliptic equations, level curves

AMS subject classifications. 35R30, 78A30, 31A25

1. Introduction.

The inverse problem of determining cracks by boundary measurements consists of finding the shapes and locations of fractures inside a conductor Ω by applying finitely many current fluxes to the boundary of Ω and measuring the induced potentials on the boundary.

Friedman and Vogelius [F-V] proved that if Ω is a planar domain of known analytic conductivity, then a single crack σ—that is, a (possibly empty) smooth simple curve—is uniquely determined by prescribing two appropriate current fluxes. This result was generalized by Bryan and Vogelius to the case of multiple cracks. In [B-V1] they prove that a collection of N pairwise disjoint smooth simple curves σ_j, $j = 1, \ldots, N$, is uniquely determined by prescribing $N + 1$ appropriate current fluxes.

The main purpose of this paper is to show that two appropriate current fluxes suffice to determine multiple cracks with any number of components. See Theorem 1.1.

It has been known since the paper by Friedman and Vogelius that the crucial step toward uniqueness theorems rests on the description of the shape of the level (equipotential) curves of voltage potentials inside Ω, and this in turn depends on information on the critical (stationary) points of the potentials. Here too we shall elaborate this theme. In fact, we shall make use of a technique, developed by Alessandrini and Magnanini (see [A-M1, A-M2]), which yields a precise evaluation of the number of interior critical points of solutions to elliptic equations in the plane in terms of the number of sign changes of boundary data.

The methods in [A-M2] will enable us to generalize the uniqueness result also in other respects. First we shall consider conductivities in Ω that may be discontinuous and anisotropic; this, we think, may be of some interest for applications with composite materials. Moreover, we shall not require that the components of the multiple crack satisfy any smoothness condition, not even that they be curves. Roughly speaking, we shall only require that each component not be an isolated point (which would have zero capacity) nor break off Ω. A detailed list of our assumptions is given below.

We let Ω be a bounded simply connected open set of $\mathbb{R}^2$ with smooth boundary $\partial\Omega$. We define a *multiple crack* σ in Ω as a possibly empty closed set that is the union of finitely many pairwise disjoint, closed continua $\sigma_1, \ldots, \sigma_N \subset \Omega$ such that $\Omega \backslash \sigma_j$ is connected for every $j = 1, \ldots, N$. Recall that a *continuum* is a connected set with at least two points.

Typical examples of the components σ_j are simple arcs, but, by our definition, sets with dendritic shape are also admitted. Components σ_j may also contain a closed curve γ. In such a case, we have that σ_j must also contain the bounded region surrounded by γ.

*Received by the editors February 2, 1994; accepted for publication (in revised form) January 5, 1995.

†Dipartimento di Scienze Matematiche, Universitá degli Studi di Trieste, Piazzale Europa 1, 34100 Trieste, Italy. The research of this author was supported in part by Fondi MURST.

‡Providencia 2457, Dto. 613, Santiago, Chile.

We represent the known conductivity at $x \in \Omega$ as a 2×2 matrix $A = A(x)$. We assume that its entries belong to $L^\infty(\Omega)$ and that, for some $\lambda > 0$, the following ellipticity condition is satisfied:

$$A(x)\xi \cdot \xi \geq \lambda |\xi|^2 \quad \text{for almost every } x \in \Omega \text{ and for every } \xi \in \mathbb{R}^2. \tag{1.1}$$

We prescribe the boundary current fluxes as follows. Let $\partial\Omega$ be decomposed into three internally pairwise disjoint simple arcs $\gamma_0, \gamma_1, \gamma_2$. Consider three functions $\eta_0, \eta_1, \eta_2 \in L^2(\partial\Omega)$ that have the following properties:

$$\eta_j \geq 0 \quad \text{on } \partial\Omega, \qquad \operatorname{supp} \eta_j \subset \gamma_j \quad \text{for every } j = 0, 1, 2. \tag{1.2a}$$

$$\int_{\partial\Omega} \eta_j = 1 \quad \text{for every } j = 0, 1, 2. \tag{1.2b}$$

Here, integration is meant with respect to ds, the arclength element along $\partial\Omega$. Next, let us define the functions ψ_1, ψ_2 as

$$\psi_k = \eta_0 - \eta_k, \qquad k = 1, 2. \tag{1.3}$$

In the sequel, we shall also make use of their antiderivatives along $\partial\Omega$, that is,

$$\Psi_k(s) = \int \psi_k(s) ds, \qquad k = 1, 2, \tag{1.4}$$

where the indefinite integration is considered along $\partial\Omega$ with the counterclockwise orientation, again with respect to the arclength parameter. Notice that, due to (1.2b) and (1.3), Ψ_1, Ψ_2 are continuous on all of $\partial\Omega$ and uniquely defined up to an additive constant.

As is customary in this field, we distinguish the cases when σ is assumed to be perfectly conducting or perfectly insulating. The boundary value problems corresponding to these two settings are as follows. We consider the weak solutions $u_k \in W^{1,2}(\Omega)$, $w_k \in W^{1,2}(\Omega \backslash \sigma)$, $k = 1, 2$, of the following boundary value problems:

$$\operatorname{div}(A \nabla u_k) = 0 \quad \text{in } \Omega \backslash \sigma, \tag{1.5a}$$

$$u_k = c_{k,j} \quad \text{on } \sigma_j, \qquad j = 1, \ldots, N, \tag{1.5b}$$

$$A \nabla u_k \cdot \nu = \psi_k \quad \text{on } \partial\Omega, \tag{1.5c}$$

$$\int_\beta A \nabla u_k \cdot \nu = 0 \quad \text{for every smooth Jordan curve } \beta \subset \Omega \backslash \sigma; \tag{1.5d}$$

$$\operatorname{div}(A \nabla w_k) = 0 \quad \text{in } \Omega \backslash \sigma, \tag{1.6a}$$

$$A \nabla w_k \cdot \nu = 0 \quad \text{on } \partial\sigma_j, \qquad j = 1, \ldots, N, \tag{1.6b}$$

$$A \nabla w_k \cdot \nu = \psi_k \quad \text{on } \partial\Omega, \tag{1.6c}$$

where the numbers $c_{k,j}$ are also unknowns and ν denotes unit normal, with outward orientation when on $\partial\Omega$. Problem (1.5) represents the perfectly conducting case, whereas problem (1.6)

describes the perfectly insulating one. Problems (1.5) and (1.6) are to be interpreted rigorously as follows:

(1.5′) To find $u_k \in W^{1,2}(\Omega)$ satisfying $u_k = \text{const}$ weakly on each σ_j and also $\int_\Omega A\nabla u_k \cdot \nabla\phi = \int_{\partial\Omega} \psi_k\phi$ for every $\phi \in W^{1,2}(\Omega)$ such that $\phi = \text{const}$ weakly on each σ_j;

(1.6′) To find $w_k \in W^{1,2}(\Omega\backslash\sigma)$ satisfying $\int_\Omega A\nabla w_k \cdot \nabla\eta = \int_{\partial\Omega} \psi_k\eta$ for every $\eta \in W^{1,2}(\Omega\backslash\sigma)$.

The existence and the uniqueness, up to an additive constant, of solutions to (1.5), (1.6) is straightforward in the framework of elliptic equations in divergence form. Let us also observe that being the sets σ_j continua they have positive capacity and their boundary is composed of regular points for the classical Dirichlet problem; see for instance, the book by Tsuji [T, Thm. I, 11]. In view of the celebrated result of Littman, Stampacchia, and Weinberger [L-S-W], we have that solutions u_k to (1.5), $k = 1, 2$, are continuous in all of Ω. Consequently, also the N-tuple $\{c_{k,j}\}_{j=1}^N$ in (1.5b) is uniquely determined up to an additive constant. Notice, in connection with this, the role of condition (1.5d), which is implicit in the formulation (1.5′) and represents a no-flux condition around the cracks. A conditions analogous to (1.5d) holds automatically when u_k is replaced by w_k.

Now we are in position to state our main result.

THEOREM 1.1. *Let Γ be a nonempty simple arc in $\partial\Omega$. Let σ, σ' be two multiple cracks in Ω, and let u_k', w_k', $k = 1, 2$, be the solutions to* (1.5), (1.6), *respectively, when σ is replaced by σ'. If either*

$$u_k = u_k' \quad \text{on } \Gamma \text{ for all } k = 1, 2 \tag{1.7}$$

or

$$w_k = w_k' \quad \text{on } \Gamma \text{ for all } k = 1, 2, \tag{1.8}$$

then we have $\sigma = \sigma'$.

Note that in this theorem the prescribed current fluxes belong to $L^2(\partial\Omega)$; however, up to minor technical adjustments, even less regular data could be considered. For instance, the functions ψ_k could be replaced by measures, provided that conditions (1.2), (1.3) are preserved. In particular, data modeling concentrated electrodes as in [B-V1] would serve the purpose.

In §2 we prepare for the proof of Theorem 1.1. We state and prove Proposition 2.1, which allows us to extend the duality arguments already used in [F-V, B-V1, B-V2] to our setting. Next, in Proposition 2.2, we introduce the tools from [A-M2] that will be needed later.

In §3, we state Propositions 3.1 and 3.2, which enable us to complete the proof of Theorem 1.1.

Section 4 contains the proofs of Propositions 3.1 and 3.2.

2. Stream functions and geometric critical points. We define $B = (\det A)^{-1}A^{\mathrm{T}}$, where $(\cdot)^{\mathrm{T}}$ denotes transpose, whereas we denote by $(\cdot)^{\perp}$ the rotation by 90° in the counterclockwise direction.

PROPOSITION 2.1. *For each $k = 1, 2$, there exists, and it is unique up to an additive constant, a function $v_k \in W^{1,2}(\Omega\backslash\sigma)$ that satisfies*

$$\nabla v_k = (A\nabla w_k)^{\perp} \quad \text{almost everywhere in } \Omega\backslash\sigma. \tag{2.1}$$

Furthermore, v_k is a weak solution of the following boundary value problem:

$$\operatorname{div}(B\nabla v_k) = 0 \quad \text{in } \Omega\backslash\sigma, \tag{2.2a}$$

$$v_k = d_{k,j} \quad \text{on } \partial\sigma_j, \qquad j = 1, \dots, N, \tag{2.2b}$$

$$v_k = \Psi_k \quad \text{on } \partial\Omega, \tag{2.2c}$$

$$\int_\beta B\nabla v_k \cdot \nu = 0 \quad \text{for every smooth Jordan curve } \beta \subset \Omega\backslash\sigma, \tag{2.2d}$$

where the numbers $d_{k,j}$ are unknown.

Before giving a proof, we start with some remarks.

The function v_k is the so-called *stream function* associated with w_k, and we have that problems (1.6) and (2.2) are equivalent through (2.1). The construction of the stream function associated to a solution of an elliptic boundary value problem like (1.6) is a generalization of the notion of conjugate harmonic function and is well known in the case when the coefficients in A are smooth and σ_j are smooth curves or have smooth boundaries. See the book by Bergman and Schiffer [B-S] and, more specifically, for the case of cracks [B-V2].

Note that v_k is continuous in $\overline{\Omega\backslash\sigma}$; in fact, as we already observed, every point in $\partial(\Omega\backslash\sigma) = \partial\Omega \cup \partial\sigma$ is regular for the Dirichlet problem. Moreover, v_k can be continued to a $W^{1,2}(\Omega) \cap C(\overline{\Omega})$ function by setting $v_k = d_{k,j}$ in σ_j, $j = 1, \dots, N$.

The rigorous formulation of (2.2) takes the following form:

(2.2′) To find $v_k \in W^{1,2}(\Omega)$ satisfying $v_k = \text{const}$ weakly on each σ_j, $v_k = \Psi_k$ on $\partial\Omega$, and also $\int_\Omega B\nabla v_k \cdot \nabla\phi = 0$ for every $\phi \in W_0^{1,2}(\Omega)$ such that $\phi = \text{const}$ weakly on each σ_j.

Stream functions t_k associated to solutions v_k of (1.6) could be constructed as well, and a result completely analogous to Proposition 2.1 could be stated. We avoid the details since we shall make use of t_k only locally, and we shall not need to specify its boundary conditions.

Proposition 2.1 implies that condition (1.8) in Theorem 1.1 is equivalent to

$$B\nabla v_k \cdot \nu = B\nabla v_k' \cdot \nu \quad \text{on } \Gamma \text{ for all } k = 1, 2, \tag{2.3}$$

where v_k' denotes a solution to (2.2) when σ is replaced with σ'. In fact, we shall prove Theorem 1.1 by treating v_k rather than w_k, the advantage being that v_k, like u_k, satisfies Dirichlet-type conditions on σ rather than Neumann-type conditions.

Proof of Proposition 2.1 (sketch). Let us approximate σ by closed sets σ^n such that $\sigma^{n-1} \supset \sigma^n \to \sigma$ as $n \to \infty$ and $\partial\sigma_n$ are smooth and A by smooth matrices A_n satisfying uniform ellipticity conditions and converging to A as $n \to \infty$ in $L^p(\Omega)$ for all $p < \infty$. With such replacements, our thesis holds and we have uniform $W^{1,2}$ bounds on corresponding regularized solutions to (1.6) and (2.2). Next we let $n \to \infty$ and find subsequences of regularized solutions weakly converging in $W^{1,2}(\Omega\backslash\sigma)$ to solutions of the original problems (1.6) and (2.2) that also satisfy (2.1). □

We shall need some properties about the geometric character of level lines of solutions of two-dimensional elliptic equations in divergence form and discontinuous coefficients. This

issue was treated in [A-M2] for the case when the matrix A is symmetric, but the following results hold also in the nonsymmetric case.

PROPOSITION 2.2. *Let $u \in W^{1,2}(D)$ be a weak solution to $\operatorname{div}(A\nabla u) = 0$ in a simply connected domain D, and let t be its stream function. Then we have the representation $u + it = f \circ \chi$ where χ is a quasi-conformal mapping of D onto the disk $B_1(0)$ and f is a holomorphic function on $B_1(0)$.*

Moreover, if $u = A\nabla u \cdot \nu = 0$ in the weak sense on an arc $\Gamma \subset \partial D$, then we have $u = 0$ everywhere in D.

Proof. See [A-M2, Thm. 2.1 and Cor. 2.2] for a proof. □

This representation gives that the geometrical structure of the level lines of u has the same character as that of the harmonic function $h = \operatorname{Re} f$.

A point $z \in D$ is then called a *geometric critical point* for u if $\chi(z)$ is a critical point for h, that is, $\nabla h|_{\chi(z)} = 0$.

Let F be a smooth vector field in a planar domain G with smooth boundary, and let $F \neq 0$ on ∂G. We define the index of F in G by

$$I(G, F) = -\frac{1}{2\pi} \int_{\partial G} \mathrm{d}(\arg F), \tag{2.4}$$

where arg F denotes the angle made by F with a fixed direction, and the integral is taken in the counterclockwise orientation. If ζ is a point where F vanishes we define the index of F at ζ as

$$I(\zeta, F) = \lim_{r \to 0} I(B_r(\zeta), F).$$

In places we shall also deal with the index of complex-valued functions $g = u + it$; for this purpose, we shall identify the function g with the vector field $F = \binom{u}{t}$.

We denote the *geometric index* of ∇u at the point z as the index of the vector field ∇h at $\chi(z)$, namely,

$$I(z, \nabla u) = I(\chi(z), \nabla h);$$

the geometric index coincides with the index defined above when u is smooth (see [A-M2]). Note that, contrary to the customary definition of index (see, for instance, [M]), we have chosen to place the minus sign in the definition (2.4) (and hence, index= −winding number) in such a way that, for solutions of elliptic equations, the geometric index is always a nonnegative integer and positive only at geometric critical points. In fact, we have the following result.

LEMMA 2.3. *Assume that the hypotheses of Proposition 2.2 hold. For every $z \in D$ there exists a neighbourhood U of z such that the level set $\{\zeta \in U | u(\zeta) = u(z)\}$ is composed of $I(z, \nabla u) + 1$ simple arcs whose pairwise intersection consists of $\{z\}$ only.*

Proof. See [A-M2, Lem. 2.5] for a proof of the lemma. □

Proposition 2.2 also implies that the points z where $u(z) = t(z) = 0$ are isolated, unless u is constant. At such points, we may define the index of the complex-valued function $g = u + it$. It is easily seen that we have

$$I(z, g) = I(\chi(z), f) = I(z, \nabla u) + 1 \geq 1. \tag{2.5}$$

3. Proof of Theorem 1.1. The proof of Theorem 1.1 will be based on the following two propositions, whose proofs are postponed to §4.

PROPOSITION 3.1. *If (1.7) holds, then we have $u_k = u'_k$, $k = 1, 2$, everywhere in Ω. Likewise, if (2.3) holds, then we have $v_k = v'_k$, $k = 1, 2$, everywhere in Ω.*

For any $\alpha, \beta \in \mathbb{R}$ such that $\alpha^2 + \beta^2 = 1$, set $u = \alpha u_1 + \beta u_2$ and $v = \alpha v_1 + \beta v_2$. We have that u and v are solutions to problems (1.5) and (2.2), respectively, when ψ_k and Ψ_k are replaced with $\psi = \alpha\psi_1 + \beta\psi_2$ and $\Psi = \alpha\Psi_1 + \beta\Psi_2$, respectively.

PROPOSITION 3.2. *Neither u nor v have geometric critical points in $\Omega\backslash\sigma$.*

Proof of Theorem 1.1. Suppose that (1.7) holds, and assume by contradiction $\sigma'\backslash\sigma \neq \emptyset$. Hence $\sigma'\backslash\sigma$ must contain a continuum δ. By Proposition 3.1, we have that there exists a constant c_k' such that $u_k = u_k' = c_k'$ on δ, $k = 1, 2$. Since solutions to (1.5) are unique up to additive constants, we may assume with no loss of generality that $c_k' = 0$, $k = 1, 2$. Therefore $u = 0$ on δ. We shall show that there exist $\alpha, \beta \in \mathbb{R}$, $\alpha^2 + \beta^2 = 1$, such that u has at least one geometric critical point in $\Omega\backslash\sigma$, and this contradicts Proposition 3.2. The same type of argument applies when (1.8) or, as is the same, (2.3) holds.

Let P be a fixed point in δ, and let D be a disk centered at P with sufficiently small radius in such a way that $D \subset \Omega\backslash\sigma$. Let $P_n \in \delta \cap D$, $n = 1, 2, \ldots$, be points such that $P_n \neq P$ for all n and $P_n \to P$ as $n \to \infty$. Let t_1, t_2 be stream functions for u_1, u_2, respectively. We may also require $t_1(P) = 0$, $t_2(P) = 0$. Obviously, we have that $t = \alpha t_1 + \beta t_2$ is a stream function for u and that $t(P) = 0$. For every $n = 1, 2, \ldots$ we may find $\alpha_n, \beta_n \in \mathbb{R}$ satisfying $\alpha_n^2 + \beta_n^2 = 1$ such that the complex-valued function

$$g_n = \alpha_n(u_1 + it_1) + \beta_n(u_2 + it_2)$$

vanishes at the points P and P_n. We have $I(P, g_n)$, $I(P_n, g_n) \geq 1$. Possibly passing to subsequences, we may set

$$\alpha_0 = \lim_{n\to\infty} \alpha_n, \quad \beta_0 = \lim_{n\to\infty} \beta_n, \quad \alpha_0^2 + \beta_0^2 = 1.$$

Let us denote

$$g_0 = \alpha_0(u_1 + it_1) + \beta_0(u_2 + it_2).$$

By the continuity property of the index, we obtain

$$I(P, g_0) \geq \liminf_{n\to\infty} (I(P, g_n) + I(P_n, g_n)) \geq 2.$$

Hence, by (2.5), $I(P, \nabla u) \geq 1$ when $\alpha = \alpha_0$ and $\beta = \beta_0$; that is, P is a geometric critical point for u. □

4. Proofs of Propositions 3.1 and 3.2.

Proof of Proposition 3.1. Consider the connected component G of $\Omega\backslash(\sigma \cup \sigma')$ such that $\partial G \subset \partial\Omega$. Functions u_k, u_k' are solutions to $\operatorname{div}(A\nabla u) = 0$ in $\Omega\backslash(\sigma \cup \sigma')$, and they have the same Cauchy data on Γ. By Proposition 2.2 and continuity, we obtain $u_k = u_k'$ in $\overline{G}$. Suppose $\Omega\backslash(\sigma \cup \overline{G})$ is nonempty, and let D be any of its connected components. We have that one connected component Δ_1 of ∂D is contained in ∂G and the remaining Δ_h are some of the components σ_j of σ. On Δ_1, we have $u_k = u_k'$; since $\Delta_1 \subset \sigma \cup \sigma'$, we obtain $u_k = \text{const}$ on Δ_1. By conditions (1.5b), (1.5d) we deduce $u_k = \text{const}$ in D; by Proposition 2.2 this implies $u_k = \text{const}$ in Ω, which contradicts (1.5c). Hence $\Omega\backslash(\sigma \cup \overline{G})$ is empty and $u_k = u_k'$ in $\Omega\backslash\sigma$. By reversing the roles of σ and σ', we arrive at $u_k = u_k'$ in $\Omega\backslash(\sigma \cap \sigma')$. Finally, $u_k = u_k'$ in Ω by continuity and conditions (1.5b). The same argument applies to solutions v_k. v_k'. □

Proof of Proposition 3.2. Consider the Neumann boundary data $\psi = \alpha\psi_1 + \beta\psi_2$ for u on $\partial\Omega$. We have $\psi = (\alpha + \beta)\eta_0 - \alpha\eta_1 - \beta\eta_2$; therefore, $\partial\Omega$ can be decomposed into two arcs δ_2, δ_2—one being the union of two of the arcs $\gamma_0, \gamma_1, \gamma_2$, and the other is the remaining one of the three such that $\psi \geq 0$ on δ_1 and $\psi \leq 0$ on δ_2. Note also that the Dirichlet data

$\Psi = \alpha\Psi_1 + \beta\Psi_2$ for v on $\partial\Omega$ is a primitive of ψ, and therefore Ψ is nondecreasing in δ_1 and Ψ is nonincreasing in δ_2.

By using the regularization procedure outlined in the proof of Proposition 2.1 and by recalling the continuity properties of the geometric index (see [A-M2, Proposition 2.6]), it suffices to prove the thesis when A and $\partial\sigma$ are smooth. Notice also that, in the regularization procedure, we may also approximate the boundary data ψ and Ψ. Hence we shall also assume that ψ is smooth, $\psi > 0$ in the interior of δ_1, $\psi < 0$ in the interior of δ_2, and Ψ is still a primitive of ψ.

Let us incidentally remark that a straightforward application of the methods in [A-M1] would give, for both u and v,

$$(\text{number of critical points in } \Omega\backslash\sigma) + \frac{1}{2}(\text{ number of critical points on } \partial\sigma) \leq N,$$

where N is the number of components of σ. Thus, if σ were empty, our thesis would follow immediately. In the sequel we shall deal with the case when σ is nonempty. Roughly speaking, our argument will consist of showing that conditions (1.5d), (2.2d) imply that there are exactly $2N$ critical points on $\partial\sigma$.

We shall prove that, for every domain $G \subset \Omega\backslash\sigma$ such that no geometric critical point of u nor v belongs to ∂G, we have

$$I(G, \nabla u) = I(G, \nabla v) = 0.$$

With no loss of generality, we may assume $\partial G = \Gamma_0 \cup \Gamma_1 \cup \cdots \cup \Gamma_N$ where $\Gamma_0, \ldots, \Gamma_N$ are smooth closed curves such that Γ_0 surrounds $\Gamma_1 \cup \cdots \cup \Gamma_N$ and each Γ_j, $j = 1, \ldots, N$, surrounds σ_j. We have

$$2\pi I(G, \nabla u) = \sum_{j=1}^{N} \int_{\Gamma_j} \mathrm{d}\arg \nabla u - \int_{\Gamma_0} \mathrm{d}\arg \nabla u,$$

and the analogous formula holds for v. By using the fact that $\partial\Omega$ splits into two parts where the Neumann data for u and the tangential derivative of the Dirichlet data for v have constant sign and by the arguments used in the proof of Theorem 2.2 in [A-M1] we have

$$\int_{\Gamma_0} \mathrm{d}\arg \nabla u, \int_{\Gamma_0} \mathrm{d}\arg \nabla v \geq 0.$$

The proof will be completed by showing

$$\int_{\Gamma_j} \mathrm{d}\arg \nabla u \leq 0 \quad \text{for every } j = 1, \ldots, N. \tag{4.1}$$

In fact, we shall obtain $I(G, \nabla u) = 0$ since we already know that such an index is nonnegative. The same argument will be applicable to v since the boundary conditions for u and v on σ_j are of the same type.

Let Σ be the annular region bounded by Γ_j and $\partial\sigma_j$. We may find $R > 0$ and a conformal change of coordinates from Σ to the annulus $B_R(0)\backslash\overline{B_1(0)}$ that is smooth up to the boundary and transforms Γ_j into $\partial B_R(0)$ and σ_j into $\partial B_1(0)$, (see, for instance, [N]). By conformal invariance we have, in the new coordinates,

$$\operatorname{div}(A\nabla u) = 0 \quad \text{in } B_R(0)\backslash\overline{B_1(0)}, \tag{4.2}$$

$$u = c = \text{const} \quad \text{on } \partial B_1(0), \tag{4.3}$$

$$\int_\beta A\nabla u \cdot \nu = 0 \quad \text{for every smooth Jordan curve } \beta \subset B_R(0)\backslash\overline{B_1(0)},$$

and, by continuity,

$$\int_{\partial B_1(0)} A\nabla u \cdot \nu = 0. \tag{4.4}$$

Condition (4.4) implies that both the level sets $\{u > c\}$, $\{u < c\}$ have zero distance from $\partial B_1(0)$. Moreover, by the maximum principle, they also have zero distance from $\partial B_R(0)$. Hence, the level curve $\{u = c\}$ contains at least two arcs, both joining $\partial B_1(0)$ to $\partial B_R(0)$. We want to continue u to a solution of an elliptic equation in the larger annulus $B_R(0)\backslash\overline{B_{1/R}(0)}$ by the inversion $z \to \overline{z}^{-1}$, where $\overline{(\cdot)}$ denotes the complex conjugate. If A is given by

$$A(z) = \begin{bmatrix} a(z) & b(z) \\ c(z) & d(z) \end{bmatrix} \quad \text{for every } z \in \overline{B_R(0)}\backslash B_1(0),$$

then we set

$$A(z) = \begin{bmatrix} a(\overline{z}^{-1}) & -b(\overline{z}^{-1}) \\ -c(\overline{z}^{-1}) & d(\overline{z}^{-1}) \end{bmatrix} \quad \text{for every } z \in \overline{B_1(0)}\backslash B_{1/R}(0)$$

and

$$u(z) = 2c - u(\overline{z}^{-1}) \quad \text{for every } z \in \overline{B_1(0)}\backslash B_{1/R}(0).$$

We easily see that A is continued in such a way that ellipticity is preserved and u is continued to a C^1 solution of

$$\operatorname{div}(A\nabla u) = 0 \quad \text{in } B_R(0)\backslash\overline{B_{1/R}(0)}.$$

The level curve $\left\{ z \in B_R(0)\backslash\overline{B_{1/R}(0)} \;\middle|\; u(z) = c \right\}$ is now composed by the circle $\partial B_1(0)$ and by at least two arcs, both joining $\partial B_{1/R}(0)$ to $\partial B_R(0)$ and hence both intersecting $\partial B_1(0)$. Therefore, either there exist at least two geometric critical points of u on $\partial B_1(0)$ or there exists only one of geometric index of at least 2. In both cases we obtain by Lemma 2.3

$$2 \le I(B_R(0)\backslash\overline{B_{1/R}(0)}, \nabla u) = \frac{1}{2\pi}\left(\int_{\partial B_{1/R}(0)} \mathrm{d}\arg\nabla u - \int_{\partial B_R(0)} \mathrm{d}\arg\nabla u \right),$$

where use is made of the C^1 smoothness of u. On the other hand, by a straightforward calculation based on the continuation by reflection of u we have

$$\frac{1}{2\pi}\left(\int_{\partial B_{1/R}(0)} \mathrm{d}\arg\nabla u + \int_{\partial B_R(0)} \mathrm{d}\arg\nabla u \right) = 2$$

and thus

$$\int_{\partial B_R(0)} \mathrm{d}\arg\nabla u \le 0,$$

which is equivalent to (4.1) by the invariance of the winding number under change of coordinates. □

Appendix. After this paper was accepted, the authors learned that H. Kim and J. K. Seo [*Unique determination of a collection of finite number of cracks from two boundary measurements*, SIAM J. Math. Anal., 27 (1996), to appear] have obtained a result similar to Theorem 1.1 in the case when the background conductivity is smooth and the unknown multiple crack is composed of finitely many disjoint smooth simple curves.

REFERENCES

[A-M1] G. ALESSANDRINI AND R. MAGNANINI, *The index of isolated critical points and solutions of elliptic equations in the plane*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 19 (1992), pp. 567–589.

[A-M2] ———, *Elliptic equations in divergence form, geometric critical points of solutions, and Stekloff eigenfunctions*, SIAM J. Math. Anal., 25 (1994), pp. 1259–1268.

[B-V1] K. BRYAN AND M. VOGELIUS, *A uniqueness result concerning the identification of a collection of cracks from finitely many electrostatic boundary measurements*, SIAM J. Math. Anal., 23 (1992), pp. 950–958.

[B-V2] ———, *A computational algorithm to determine crack locations from electrostatic boundary measurements. The case of multiple cracks*, Int. J. Engng. Sci., to appear.

[B-S] S. BERGMAN AND M. SCHIFFER, *Kernel Functions and Differential Equations in Mathematical Physics*, Academic Press, New York, 1953.

[F-V] A. FRIEDMAN AND M. VOGELIUS, *Determining cracks by boundary measurements*, Indiana Univ. Math. J., 38 (1989), pp. 527–556.

[L-S-W] W. LITTMAN, G. STAMPACCHIA, AND H. WEINBERGER, *Regular points for elliptic equations with discontinuous coefficients*, Ann. Scuola Norm. Sup. Pisa (3), 17 (1963), pp. 43–77.

[M] J. MILNOR, *Differential Topology*, Princeton University Press, Princeton, NJ, 1958.

[N] Z. NEHARI, *Conformal Mapping*, McGraw-Hill, New York, 1952.

[T] M. TSUJI, *Potential Theory in Modern Function Theory*, Maruzen, Tokyo, 1959.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 922-930, May 1996

008

REGULARITY AND EXACT CONTROLLABILITY FOR A BEAM WITH PIEZOELECTRIC ACTUATOR*

MARIUS TUCSNAK†

Abstract. We consider an initial and boundary value problem modelling the vibrations of a Bernoulli-Euler beam with an attached piezoelectric actuator. We show that the Sobolev regularity of the solution is by $\frac{1}{2}+\epsilon$ higher than that one obtains by simply using the Sobolev regularity of the control term. The main results concern the dependence of the space of exactly controllable initial data on the location of the actuator. Our approach is based on the Hilbert uniqueness method introduced by Lions [*Contrôlabilité exacte des systèmes distribués*, Masson, Paris, 1988] combined with some results from the theory of diophantine approximation.

Key words. exact controllability, Hilbert uniqueness method, piezoelectric actuator, diophantine approximation

AMS subject classifications. 93C20, 35B75, 35B60

1. Introduction. In recent years a lot of papers were devoted to the study of elastic structures with piezoelectric actuators (e.g., [2], [3], [5], [6]). The main topics covered in the papers quoted above are modelling (see [2], [5], [6]), LQR, and identification problems (see [3]). Concerning the controllability problems, as far as we know, all published works consider finite-dimensional approximations of the initial distributed control problems (see [6] and references therein). The main purpose of the present paper is to study the exact controllabillity of a beam with piezoelectric actuator by using the theory of infinite-dimensional control systems as developed in [13]. More precisely we consider the initial and boundary value problem modelling the vibrations of a Bernoulli-Euler beam that is subject to the action of an attached piezoelectric actuator. If we suppose that the beam is hinged at both ends and that the actuator is excited in a manner so as to produce pure bending moments, the model for the controlled beam can be written as (cf. [5])

$$w''(x,t)+\frac{\partial^4 w}{\partial x^4}(x,t)=u(t)\frac{d}{dx}[\delta_\eta(x)-\delta_\xi(x)],\qquad 0<x<\pi,\ t>0, \tag{1.1}$$

$$w(0,t)=w(\pi,t)=\frac{\partial^2 w}{\partial x^2}(0,t)=\frac{\partial^2 w}{\partial x^2}(\pi,t)=0,\qquad t\geq 0, \tag{1.2}$$

$$w(x,0)=w^0(x),\quad w'(x,0)=w^1(x),\quad 0<x<\pi. \tag{1.3}$$

In the equations above w represents the transverse deflection of the beam, $\xi,\eta\in(0,\pi)$ stand for the ends of the actuator, and δ_y is the Dirac mass at the point y. Moreover by w', w'' we denote the time derivatives of w. The control is given by the function $u:[0,T]\to\mathbf{R}$ representing the time variation of the voltage applied to the actuator. Our main purpose is to determine the initial data that can be steered to rest by means of the control function u. More precisely we can give the following definition.

DEFINITION 1.1. *We say that the initial data w^0, w^1 are "exactly L^2-controllable in (ξ,η) at time T" if there exists $u\in L^2(0,T)$ such that the solution w of* (1.1)-(1.3) *satisfies the condition*

$$w(x,T)=w'(x,T)=0,\qquad 0<x<\pi. \tag{1.4}$$

*Received by the editors April 4, 1994; accepted for publication (in revised form) January 5, 1995.

†Ecole Polytechnique, Centre de Mathématiques Appliquées, 91128 Palaiseau Cedex, France, and Université de Versailles (tucsnak@cmapx.polytechnique.fr).

The plan of this paper is as follows. The second section contains some notation and preliminaries. In the third section we prove an existence and regularity result for (1.1)-(1.3). The exact controllability is studied in the fourth section. Our approach was inspired by ideas and methods used in [9] and [10] for pointwise control problems.

2. Notation and preliminaries. To study the wellposedness and controllability for (1.1)-(1.3) we introduce the function spaces $(Y_\alpha)_{\alpha\in\mathbf{R}}$ defined as follows. If $\alpha > 0$, then Y_α is the closure in $H^\alpha(\Omega)$ of the set of all functions $y \in C^\infty(\bar{\Omega})$ satisfying the conditions $\frac{d^{2n}y}{dx^{2n}}(0) = \frac{d^{2n}y}{dx^{2n}}(\pi) = 0$ for all $n \geq 0$; for negative α we define Y_α as the dual space of $Y_{-\alpha}$ constructed by means of the inner product of $Y_0 = L^2(\Omega)$.

Let us now consider the homogenous initial and boundary value problem

$$\phi''(x,t) + \frac{\partial^4\phi}{\partial x^4}(x,t) = 0, \qquad 0 < x < \pi,\ t \in (0,T), \tag{2.1}$$

$$\phi(0,t) = \phi(\pi,t) = \frac{\partial^2\phi}{\partial x^2}(0,t) = \frac{\partial^2\phi}{\partial x^2}(\pi,t) = 0, \qquad t \in (0,T), \tag{2.2}$$

$$\phi(x,0) = \phi^0(x), \quad \phi'(x,0) = \phi^1(x), \quad 0 < x < \pi. \tag{2.3}$$

It is by now well known that the initial and boundary value problem (2.1)-(2.3) is well posed in $Y_{\alpha+2} \times Y_\alpha$ for all $\alpha \geq -2$. Moreover, as a consequence of the Hilbert uniqueness method (HUM), introduced in [13], the following result holds.

PROPOSITION 2.1. *All initial data in $Y_{\alpha+2} \times Y_\alpha$ are "exactly L^2-controllable in (ξ,η) at time T" if and only if there exists a constant $c > 0$ such that*

$$\int_0^T \left[\frac{\partial\phi}{\partial x}(\xi,t) - \frac{\partial\phi}{\partial x}(\eta,t)\right]^2 dt \geq c(\|\phi^0\|^2_{H^{-\alpha}(\Omega)} + \|\phi^1\|^2_{H^{-\alpha-2}(\Omega)}] \tag{2.4}$$

for all

$$(\phi^0,\phi^1) \in Y_3 \times Y_1.$$

We shall also need some results from the theory of diophantine approximation. For a real number ρ, we denote by $|||\rho|||$ the difference, taken positively, between ρ and the nearest integer, i.e.,

$$|||\rho||| = \min_{n\in\mathbf{Z}} |\rho - n|.$$

Let us also denote by A the set of all irrationals $\rho \in]0,1[$ such that if $[0, a_1, \ldots, a_n, \ldots]$ is the expansion of ρ as a continued fraction, then (a_n) is bounded. The set A plays a very important role in our control problem. Let us note that A is obviously uncountable and, by classical results on diophantine approximation (cf. [4, p. 120]), its Lebesgue measure is equal to zero. In particular, by the Euler-Lagrange theorem (cf. [11, p. 57]) A contains the irrational quadratic numbers (i.e., satisfying a second-degree equation with rational coefficients). We shall essentially use the fact that the elements of A are badly approximable by rational numbers. More precisely, the following result holds true (cf. [11, p. 24]).

PROPOSITION 2.2. *A number $\rho \in (0,1)$ is in A if and only if there exists a constant $C > 0$ such that*

$$|||q\rho||| \geq \frac{C}{q} \tag{2.5}$$

for all strictly positive integer q.

We shall also use the following result on simultenous approximation (cf. [4, p. 14]).

PROPOSITION 2.3. *Let $\rho_1, \ldots, \rho_k$ be k irrationals in $(0, 1)$. Then there exists a strictly increasing sequence of natural numbers q_n such that*

$$q_n^{\frac{1}{k}} \max_{i=1,k}(|||q_n\rho_1|||, \ldots, |||q_n\rho_i|||, \ldots, |||q_n\rho_k|||) \leq \frac{k}{k+1}, \quad \forall n \geq 1.$$

The next proposition, which is proved in [4, p. 120], shows that an inequality slightly weaker than (2.5) holds for almost all points in $(0, 1)$.

PROPOSITION 2.4. *For any $\epsilon > 0$ there exists a set $B_\epsilon \subset (0, \pi)$ having the Lebesgue measure equal to π and a constant $C > 0$, such that for any $\rho \in B_\epsilon$,*

$$|||q\rho||| \geq \frac{C}{q^{1+\epsilon}} \tag{2.6}$$

for any strictly positive integer q.

Let us notice that by Roth's theorem (cf. [4, p. 104]), for all $\epsilon > 0$, B_ϵ contains all algebraic irrational numbers in $(0, \pi)$.

3. Existence and regularity of solutions. The regularity of the solutions of (1.1)-(1.3) is a problem that is similar in many respects to the regularity for the wave or Euler-Bernoulli equations with interior point control (see [14], [16]). As in the cases mentioned above one can prove that the regularity of the solutions of (1.1)-(1.3) is higher than that one obtains by simply using the Sobolev regularity of the right-hand side of (1.1). More precisely if we denote by Ω the interval $(0, \pi)$ we notice that if $u \in L^2(0, T)$ the function $u(\cdot)\frac{d}{dx}(\delta_\xi - \delta_\eta)$ is in $L^2(0, T; H^{-\frac{3}{2}-\epsilon}(\Omega))$ for any $\epsilon > 0$. Standard regularity for the Bernoulli-Euler equation (cf. [13]) implies the existence and the uniqueness of a solution of (1.1)-(1.3) in

$$C([0, T], Y_{\frac{1}{2}-\epsilon}) \cap C^1([0, T], Y_{-\frac{3}{2}-\epsilon}). \tag{3.1}$$

The main result of this section, given below, shows that the space regularity in (3.1) can be improved by $\frac{1}{2} + \epsilon$.

THEOREM 3.1. *Suppose that $w^0 \in Y_1$, $w^1 \in Y_{-1}$. Then the initial and boundary value problem* (1.1)-(1.3) *admits a unique solution having the regularity*

$$w \in C([0, T], Y_1) \cap C^1([0, T], Y_{-1}). \tag{3.2}$$

Consider $\tau \in [0, T]$. To prove Theorem 3.1, following [14], we introduce the homogenous initial and boundary value problem

$$v''(x, t) + \frac{\partial^4 v}{\partial x^4}(x, t) = 0, \qquad 0 < x < \pi, \ t \in (0, \tau), \tag{3.3}$$

$$v(0, t) = v(\pi, t) = \frac{\partial^2 v}{\partial x^2}(0, t) = \frac{\partial^2 v}{\partial x^2}(\pi, t) = 0, \qquad t \in (0, \tau), \tag{3.4}$$

$$v(x, \tau) = 0, \quad v'(x, \tau) = g(x), \quad 0 < x < \pi. \tag{3.5}$$

The following lemma shows that the solution of (3.3)-(3.5) has a point regularity property that is essential for the proof of Theorem 3.1.

LEMMA 3.2. *For any $g \in Y_{-1}$ the initial and boundary value problem* (3.3)-(3.5) *admits a unique solution having the regularity*

$$v \in C([0, T], Y_1) \cap C^1([0, T], Y_{-1}). \tag{3.6}$$

Moreover, for any $\rho \in (0, \pi)$ the function $\frac{\partial v}{\partial x}(\rho, \cdot)$ is in $L^2(0, T)$ and there exists a constant $C > 0$ such that

$$\left\| \frac{\partial v}{\partial x}(\rho, \cdot) \right\|_{L^2(0,T)} \leq C \|g\|_{Y_{-1}}. \tag{3.7}$$

Proof. The existence and uniqueness of a solution having the regularity (3.6) are standard results for the Bernoulli-Euler equation. To prove (3.7) we put

$$g(x) = \sum_{n=1}^{\infty} n^2 a_n \sin(nx).$$

By density it is enough to show that (3.7) holds for $g \in C_0^\infty(\Omega)$. Obviously the solution of (3.3)-(3.5) is given by

$$v(x, t) = \sum_{n=1}^{\infty} a_n \sin(n^2 t) \sin(nx),$$

which implies that

$$\frac{\partial v}{\partial x}(\rho, t) = \sum_{n=1}^{\infty} n a_n \sin(n^2 t) \cos(n\rho). \tag{3.8}$$

If we consider the right-hand side of (3.8) as a Fourier series in t (see Theorem 4.1 in [7] for details) we obtain the existence of a constant C depending on T such that

$$\left\| \frac{\partial v}{\partial x}(\rho, \cdot) \right\|_{L^2(0,T)} \leq C \sum_{n=1}^{\infty} n^2 a_n^2 < \infty, \tag{3.9}$$

which is exactly (3.7). □

Remark 3.1. The result of Lemma 3.2 is optimal in the sense that one can easily find $g \in Y_{-1}$, $T > 0$, and $\rho \in (0, \pi)$ such that equality holds in (3.9). We also remark that another possible way by which to prove Lemma 3.2 is a multiplier technique, as used in [7] for the wave equation with point control.

Proof of Theorem 3.1. Due to the linearity of (1.1) and to well-known properties of the Bernoulli-Euler equation, it is enough to consider the case $w^0 = w^1 = 0$. Suppose again $g \in C_0^\infty(\Omega)$, and let v be the solution of (3.3)-(3.5). If we multiply (1.1) by v and integrate by parts we easily obtain

$$\int_0^\pi w(x, \tau) g(x) dx = -\int_0^\tau u(t) \left[\frac{\partial v}{\partial x}(\xi, t) - \frac{\partial v}{\partial x}(\eta, t) \right] dt. \tag{3.10}$$

Lemma 3.2 implies that

$$\left| \int_0^\tau u(t) \left[\frac{\partial v}{\partial x}(\xi, t) - \frac{\partial v}{\partial x}(\eta, t) \right] dt \right| \leq C \|u\|_{L^2(0,T)} \|g\|_{Y_{-1}},$$

so, by (3.10), we obtain that $w(\cdot, \tau) \in Y_1$, for all $\tau \in [0, T]$. By replacing τ by $\tau + h$ in (3.10) we easily get that

$$w \in C([0, T], Y_1), \tag{3.11}$$

which implies that

$$\frac{\partial^4 w}{\partial x^4} \in C([0, T], Y_{-3}). \tag{3.12}$$

As w satisfies satisfies (1.1), from (3.12) we obtain that

$$w'' \in L^2((0, T), Y_{-3}). \tag{3.13}$$

From (3.11) and (3.13), by applying the intermediate derivative theorem (cf. [15, p. 19]) it follows that

$$w' \in L^2((0, T), Y_{-1}). \tag{3.14}$$

The conclusion (3.2) is now a consequence of (3.11)-(3.14) and of the general lifting result from [12].

Remark 3.2. As by (3.10) the interior regularity of w is equivalent to a point regularity property, Remark 3.1 implies that the result in Theorem 3.1 is sharp.

4. The exact controllability results. In this section we shall study the space of initial data that are exactly L^2-controllable in (ξ, η) at time T. Let us put

$$\phi^0(x) = \sum_{n=1}^{\infty} a_n \sin(nx), \qquad \phi^1(x) = \sum_{n=1}^{\infty} n^2 b_n \sin(nx). \tag{4.1}$$

A simple calculation shows that the solution of (2.1)-(2.3) is given by

$$\phi(x, t) = \sum_{n \geq 1} [a_n \cos(n^2 t) \sin(nx) + b_n \sin(n^2 t) \sin(nx)],$$

which implies that

$$\begin{aligned} &\int_0^T \left[\frac{\partial \phi}{\partial x}(\xi, t) - \frac{\partial \phi}{\partial x}(\eta, t)\right]^2 dt \\ &= 4 \int_0^T \sum_{n \geq 1} [n a_n \cos(n^2 t) + n b_n \sin(n^2 t)]^2 \sin^2\left[\frac{n(\eta + \xi)}{2}\right] \sin^2\left[\frac{n(\eta - \xi)}{2}\right] dt. \end{aligned} \tag{4.2}$$

Relation (4.2) implies that (2.4) is false for any α if $\frac{\xi+\eta}{2}$ or $\frac{\eta-\xi}{2}$ is rational, and by Proposition 2.1 it follows that the condition

$$\frac{\xi + \eta}{2\pi}, \ \frac{\eta - \xi}{2\pi} \in \mathbf{R} - \mathbf{Q} \tag{4.3}$$

is necessary to have exact controllability of all initial states in $Y_{\alpha+2} \times Y_\alpha$. The following result shows that the condition above is not sufficient in the sense that there are ξ, η satisfying (4.3) that do not allow the control of arbitrary regular initial data. More precisely we have the following result.

PROPOSITION 4.1. *For any $\alpha > -2$ there exist $\xi,\ \eta \in (0, \pi)$ satisfying (4.3) such that the space $Y_{\alpha+2} \times Y_\alpha$ contains initial data that are not exactly L^2-controllable in (ξ, η) at time T, for any $T > 0$.*

Proof. By Proposition 2.1 it is enough to show that (2.4) is false for any $\alpha,\ c > 0$. Let us fix $\alpha > 0$ and $\nu > \frac{3\alpha+2}{2}$. We choose then

$$\frac{\xi + \eta}{2\pi} = \sum_{n=1}^{\infty} \frac{a_n}{10^{n!}}, \tag{4.4}$$

where $a_n \in \{0, 1, \ldots, 9\}$ for all $n \geq 1$, and a_n is not identically zero for great n. According to [17] the right-hand side of (4.4) is a Liouville number; i.e., it is transcendental and there exists a strictly increasing sequence of integers q_n such that

$$\left|\left|\left| q_n \frac{\xi + \eta}{2\pi} \right|\right|\right| \leq \frac{1}{q_n^\nu}, \quad \forall n \geq 1. \tag{4.5}$$

We note that

$$\left|\sin\left(q_n \frac{\xi + \eta}{2}\right)\right| = \left|\sin\left[\pi\left(q_n \frac{\xi + \eta}{2\pi} - p\right)\right]\right| \leq \pi \left|q_n \frac{\xi + \eta}{2\pi} - p\right|,$$

for any integer p. The relation above and (4.5) imply that

$$\left|\sin\left(q_n \frac{\xi + \eta}{2}\right)\right| \leq \frac{\pi}{q_n^\nu}, \quad \forall n \geq 1. \tag{4.6}$$

Now consider the sequence of initial data

$$\phi_n^0(x) = q_n^{\frac{3\alpha}{2}} \sin(q_n x), \quad \phi_n^1(x) = 0, \qquad \forall x \in (0, \pi). \tag{4.7}$$

A simple calculation shows that $\{\phi_n^0, \phi_n^1\} \in Y_{\alpha+2} \times Y_\alpha$, and

$$||\phi_n^0||_{Y_{-\alpha}}^2 + ||\phi_n^1||_{Y_{-\alpha-2}}^2 \to \infty, \quad \forall \alpha > 0. \tag{4.8}$$

Moreover the solution of (1.1)–(1.3) with the initial data given by (4.7) is

$$\phi_n(x, t) = q_n^{\frac{3\alpha}{2}} \cos(q_n^2 t) \sin(q_n x),$$

so, by (4.6) we have

$$\int_0^T \left[\frac{\partial \phi_n}{\partial x}(\xi, t) - \frac{\partial \phi_n}{\partial x}(\eta, t)\right]^2 dt \tag{4.9}$$

$$= 4q_n^{3\alpha+2} \sin^2\left[\frac{q_n(\eta + \xi)}{2}\right] \sin^2\left[\frac{q_n(\eta - \xi)}{2}\right] \int_0^T \cos^2(q_n^2 t) dt$$

$$\leq 4T q_n^{3\alpha+2} \sin^2\left[\frac{q_n(\eta + \xi)}{2}\right] \to 0, \quad \text{when } n \to \infty.$$

Relations (4.8) and (4.9) imply that (2.4) is false for any $\alpha > -2$ and $c > 0$. □

Remark 4.1. The estimates in Proposition 4.1 are not sharp. In fact, it seems likely that the methods developed in [10] can be used to prove the existence of $\xi,\ \eta$ satisfying (4.3) such that $\bigcap_{\alpha \geq 0} Y_{\alpha+2} \times Y_\alpha$ contains initial data that are not exactly L^2-controllable in (ξ, η) at

time T, for any $T > 0$. We also remark that, by standard duality arguments, condition (4.3) implies approximate controllability in $Y_{\alpha+2} \times Y_\alpha$.

Proposition 4.1 shows that to obtain exact controllability, we need assumptions stronger than the irrationality of $\frac{\xi+\eta}{2}$ and $\frac{\eta-\xi}{2}$. At this point we shall use the number theoretic preliminaries stated in the second section. A first result in this direction is as follows.

THEOREM 4.2. *Suppose that that $\frac{\xi+\eta}{2\pi}$ and $\frac{\eta-\xi}{2\pi}$ belong to the set A (introduced in the second section). Than all initial data in $Y_3 \times Y_1$ are exactly L^2-controllable in (ξ, η) at time T, for any $T > 0$.*

Proof. By Proposition 2.1 the conclusion of the theorem is equivalent to the existence of a constant $c > 0$ such that

$$\int_0^T \left[\frac{\partial \phi}{\partial x}(\xi, t) - \frac{\partial \phi}{\partial x}(\eta, t)\right]^2 dt \geq c(\|\phi^0\|^2_{Y_{-1}} + \|\phi^1\|^2_{Y_{-3}}) \tag{4.10}$$

for all

$$(\phi^0, \phi^1) \in Y_3 \times Y_1.$$

By applying the Ball-Slemrod generalization of Ingham's inequality (cf. [1], [8]), from (4) we obtain that there exists a constant $C > 0$ such that

$$\int_0^T \left[\frac{\partial \phi}{\partial x}(\xi, t) - \frac{\partial \phi}{\partial x}(\eta, t)\right]^2 dt \tag{4.11}$$

$$\geq C \sum_{n\geq 1} (n^2 a_n^2 + n^2 b_n^2) \sin^2\left[\frac{n(\eta+\xi)}{2}\right] \sin^2\left[\frac{n(\eta-\xi)}{2}\right] dt.$$

As $\frac{\xi+\eta}{2\pi}$ and $\frac{\xi-\eta}{2\pi}$ are in A, from (2.5) we see that there exists a constant $C > 0$ such that for n large enough we have

$$\left|\sin\left[\frac{n(\eta \pm \xi)}{2}\right]\right| = \left|\sin\left\{\pi\left[\frac{n(\eta \pm \xi)}{2\pi} - p\right]\right\}\right| \geq \left|\sin\left(\frac{\pi C}{n}\right)\right| \geq \frac{C}{n} \ \forall n \geq 1. \tag{4.12}$$

Inequalities (4.11) and (4.12) imply that

$$\int_0^T \left[\frac{\partial \phi}{\partial x}(\xi, t) - \frac{\partial \phi}{\partial x}(\eta, t)\right]^2 dt \geq c \sum_{n\geq 1} (n^{-2} a_n^2 + n^{-2} b_n^2),$$

which is exactly (4.10). □

The following result shows that, for almost all choices of the ends of the actuator, we have exact controllability in Sobolev spaces more regular than $Y_3 \times Y_1$.

THEOREM 4.3. *Suppose that $\epsilon > 0$ is arbitrary, and consider the set B introduced in Proposition 2.4. Then, for any ξ, $\eta \in (0, \pi)$ such that $\frac{\xi+\eta}{2\pi}$, $\frac{\xi-\eta}{2\pi} \in B$, all initial data in $Y_{3+\epsilon} \times Y_{1+\epsilon}$ are exactly L^2-controllable in (ξ, η) at time T for any $T > 0$.*

Proof. As $\frac{\xi+\eta}{2\pi}$, $\frac{\eta-\xi}{2\pi} \in B$, from (2.6) it follows that

$$\left|\sin\left[\frac{n(\eta+\xi)}{2}\right]\right| \geq \frac{C}{n^{1+\epsilon}}, \quad \left|\sin\left[\frac{n(\eta-\xi)}{2}\right]\right| \geq \frac{C}{n^{1+\epsilon}} \ \forall n \geq 1. \tag{4.13}$$

Consider again the solution ϕ of (2.1)-(2.3) with the initial data given by (4.1). By applying (4.2) and (4.13) we obtain

$$\int_0^T \left[\frac{\partial \phi}{\partial x}(\xi, t) - \frac{\partial \phi}{\partial x}(\eta, t)\right]^2 dt \geq c(\|\phi^0\|^2_{Y_{-1-\epsilon}} + \|\phi^1\|^2_{Y_{-3-\epsilon}})$$

for all

$$(\phi^0, \phi^1) \in Y_3 \times Y_1.$$

By Proposition 2.1 it follows that all initial data in $Y_{3+\epsilon} \times Y_{1+\epsilon}$ are exactly L^2-controllable in (ξ, η) at time T, for any $T > 0$. □

Let $y(x, t)$ be the solution of

$$y''(x,t) + \frac{\partial^4 y}{\partial x^4}(x,t) = h(t)\frac{d}{dx}[\delta_\eta(x) - \delta_\xi(x)], \qquad 0 < x < \pi,\ t > 0, \tag{4.14}$$

$$y(0,t) = y(\pi,t) = \frac{\partial^2 y}{\partial x^2}(0,t) = \frac{\partial^2 y}{\partial x^2}(\pi,t) = 0, \qquad t \geq 0, \tag{4.15}$$

$$y(x,0) = 0, \quad y'(x,0) = 0, \quad 0 < x < \pi. \tag{4.16}$$

By Theorem 3.1 we have

$$y \in C([0,T], Y_1) \cap C^1([0,T], Y_{-1}).$$

By reversing the sense of the time t in (1.1)-(1.3) we easily see that Theorems 4.2 and 4.3 imply that the space regularity $y(\cdot, T)$ and $y'(\cdot, T)$ is higher than that one obtains by simply using Theorem 3.1.

COROLLARY 4.4. *Suppose that y is the solution of* (4.14)-(4.16), *where $h \in L^2(0,T)$ and $\frac{\xi+\eta}{2\pi}$, $\frac{\eta-\xi}{2\pi}$ belong to the set A (respectively, to B). Then, for any $T > 0$, the application $h \to \{y(\cdot, T), y'(\cdot, T)\}$ maps $L^2(0,T)$ onto $Y_3 \times Y_1$ (respectively, onto $Y_{3+\epsilon} \times Y_{1+\epsilon}$, $\forall \epsilon > 0$).*

Theorems 4.2 and 4.3 give no information on the controllability of initial data in $Y_{\alpha+2} \times Y_\alpha$, with $\alpha < 1$. A partial answer is given by the following result.

PROPOSITION 4.5. *Suppose that $\epsilon > 0$. Then the set $Y_{2-\epsilon} \times Y_{-\epsilon}$ contains initial data that are not exactly L^2-controllable in (ξ, η) at time T, for any $T > 0$ and $\xi,\ \eta \in [0, \pi]$.*

Proof. By applying Proposition 2.3 we easily obtain the existence of a strictly increasing sequence of positive integers (q_n) such that

$$\left|\sin\left[\frac{q_n(\eta+\xi)}{2}\right]\right| \leq \frac{\pi}{\sqrt{q_n}}, \quad \left|\sin\left[\frac{q_n(\eta-\xi)}{2}\right]\right| \leq \frac{\pi}{\sqrt{q_n}}, \qquad \forall n \geq 1. \tag{4.17}$$

Consider now the sequence of initial data

$$\phi_n^0(x) = \sin(q_n x), \quad \phi_n^1(x) = 0 \quad \forall x \in (0, \pi).$$

We note that

$$\|\phi_n^0\|_{Y_\epsilon}^2 + \|\phi_n^1\|_{Y_{\epsilon-2}}^2 = q_n^{2\epsilon} \to \infty, \quad \text{when } n \to \infty \tag{4.18}$$

A simple calculation that gives the corresponding sequence of solutions of (2.1)-(2.3) is

$$\phi_n(x,t) = \cos(q_n^2 t)\sin(q_n x),$$

so by (4.17) we have

$$\int_0^T \left[\frac{\partial \phi_n}{\partial x}(\xi, t) - \frac{\partial \phi_n}{\partial x}(\eta, t)\right]^2 dt \tag{4.19}$$

$$= q_n^2 \sin^2\left[\frac{q_n(\eta+\xi)}{2}\right] \sin^2\left[\frac{q_n(\eta-\xi)}{2}\right] \int_0^T \cos^2(q_n^2 t)dt \leq K \quad \forall n \geq 1,$$

where K is a positive constant. In a similar manner we can show that (4.18) and (4.19) hold true for the sequence of initial data

$$\phi_n^0(x) = 0, \quad \phi_n^1(x) = q_n^2 \sin(q_n x) \quad \forall x \in (0, \pi),$$

so (2.4) is false for $\alpha = -\epsilon$ and arbitrary $c > 0$. □

Remark 4.2. The exact controllability of initial data in $Y_{\alpha+2} \times Y_\alpha$, with $0 \leq \alpha < 1$, seems an open question.

REFERENCES

[1] J. M. Ball and M. Slemrod, *Nonharmonic Fourier series and the stabilization pf semilinear control systems*, Comm. Pure Appl. Math., 32 (1979), pp. 555-587.

[2] H. T. Banks and R. C. Smith, *The modelling of piezoceramic patch interactions with shells, plates and beams*, Tech. Report no. 92-66, Institute for Computer Applications in Science and Engineering, Hampton, VA, 1992.

[3] H. T. Banks, W. Fang, R. J. Silcox, and R. C. Smith, *Approximation methods for control of tye acoustic/structure interaction with piezoceramic actuators*, Journal of Intelligent Material Systems and Structures, 4 (1993), pp. 98-116.

[4] J. W. Cassals, *An Introduction to Diophantine Approximation*, Cambridge University Press, Cambridge, 1966.

[5] E. F. Crawley and E. H. Anderson, *Detailed models for piezoceramic actuation of beams*, Journal of Intelligent Material Systems and Structures, 1 (1990), pp. 79-83.

[6] Ph. Destuynder, I. Legrain, L. Castel, and N. Richard, *Theoretical, numerical and experimental discussion of the use of piezoelectric devices for control-structure interaction*, European J. Mech. A Solids, 11 (1992), pp. 181-213.

[7] C. Fabre and J. P. Puel, *Pointwise controllabillity as limit of internal controllability for the wave equation in one space dimension*, Portugal. Math., 51 (1994), pp. 335-350.

[8] A. Haraux, *Quelques proprietes des séries lacunaires utiles dans l'étude des systèmes élastiques*, Publications du Laboratoire d'Analyse Numérique R88011, Paris.

[9] ———, *Remarques sur la contrôlabilité ponctuelle et spectrale de systèmes distribuées*, Publications du Laboratoire d'Analyse Numérique R89017, Paris.

[10] S. Jaffard, *Sur le contrôle ponctuel des cordes vibrantes et des poutres*, preprint.

[11] S. Lang, *Introduction to Diophantine Approximations*, Addison-Wesley, New York, 1966.

[12] I. Lasiecka and R. Triggiani, *A lifting theorem for the time regularity of solutions to abstract equations with unbounded operators and applications to hyperbolic equations*, Proc. Amer. Math. Soc., 10 (1988), pp. 745-755.

[13] J. L. Lions, *Contrôlabilité exacte des systèmes distribués*, Masson, Paris, 1988.

[14] ———, *Pointwise control for distributed systems*, in Control and Estimation in Distributed Parameter Systems, H. T. Banks, ed., SIAM, Philadelphia, PA, 1992, pp. 1-41.

[15] J. L. Lions and E. Magenes, *Nonhomogenous Boundary Value Problems*, Springer, Berlin, 1972.

[16] R. Triggiani, *Interior and boundary regularity of the wave equation with interior point control*, Differential Integral Equations, 6 (1993), pp. 111-129.

[17] G. Valiron, *Théorie des fonctions*, Masson, Paris, 1990.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 931-965, May 1996

009

A DUALITY THEORY FOR SEPARATED CONTINUOUS LINEAR PROGRAMS *

MALCOLM C. PULLAN†

Abstract. This paper presents a detailed duality theory for a class of continuous linear programs called separated continuous linear programs (SCLP), based on a particular dual problem SCLP*. Using weak duality, a notion of complementary slackness is introduced, and several sufficient conditions for optimality of SCLP are derived along with the existence of complementary slack variables for basic feasible solutions for SCLP. Following this, a fairly general condition for the absence of a duality gap between SCLP and SCLP* is given, as are several conditions for the existence of an optimal solution for SCLP*. Finally, using all these ingredients, a strong duality result between SCLP and SCLP* is proven when the problem data are piecewise analytic. A simple counterexample is presented to show that strong duality may not follow if the assumptions of piecewise analyticity do not hold.

Key words. duality, continuous linear programming, linear optimal control

AMS subject classifications. 49N15, 49N05, 49K30, 90C45

1. Introduction.

In 1953, Bellman [7] introduced a class of optimization problems which he called *bottleneck problems*. These problems have now become known as *continuous linear programs* because they can be formulated as linear programs having variables which are functions of time as follows:

$$\begin{aligned} \text{CLP:} \quad & \text{maximize} \quad \int_0^T c(t)^T x(t)\, dt \\ & \text{subject to} \quad B(t)x(t) + \int_0^t K(s,t)x(s)\, ds \le b(t), \\ & \qquad\qquad x(t) \ge 0, \qquad t \in [0,T], \end{aligned}$$

with $x(t)$, $c(t)$, and the elements of $B(t)$ and $K(s,t)$ being bounded measurable functions.

In this paper we will be considering the following subclass of CLP called *separated continuous linear programs*, first introduced by Anderson [1] in an attempt to model job-shop scheduling problems:

$$\begin{aligned} \text{SCLP:} \quad & \text{minimize} \quad \int_0^T c(t)^T x(t)\, dt \\ (1) \qquad & \text{subject to} \quad \int_0^t Gx(s)\, ds + y(t) = a(t), \\ (2) \qquad & \qquad\qquad Hx(t) + z(t) = b(t), \\ & \qquad\qquad x(t), y(t), z(t) \ge 0, \qquad t \in [0,T]. \end{aligned}$$

Here $x(t)$, $z(t)$, $b(t)$, and $c(t)$ are bounded measurable functions and $y(t)$ and $a(t)$ are absolutely continuous functions. The dimensions of $x(t)$, $y(t)$, and $z(t)$ are n_1, n_2, and n_3, respectively. Thus G is an $n_2 \times n_1$ matrix and H is an $n_3 \times n_1$ matrix. We let $\omega(t)$ denote a complete set of variables for SCLP, i.e., $\omega(t)^T = (x(t)^T, y(t)^T, z(t)^T)$. By differentiating the integral constraint (1) we can see that SCLP is a special type of linear optimal control problem with state positivity but without state feedback.

The usual way to solve CLP (or SCLP) is by discretization (see, for example, Buie and Abrham [10]); however, a number of authors have tried to solve this problem by extending

*Received by the editors October 25, 1993; accepted for publication (in revised form) January 10, 1995.

†Judge Institute of Management Studies, Mill Lane, Cambridge CB2 1RX, United Kingdom. Current address: Department of Mathematical Sciences, Loughborough University of Technology, Loughborough, Leicestershire LE11 3TU, United Kingdom.

the simplex method for finite-dimensional linear programming. This was first attempted by Lehman [21] and extended by Drews [12], Hartberger [19], and Segers [35]. The most comprehensive, but a still incomplete, solution method based on the simplex method can be found in Perold [25], later followed up by Perold [26] and Anstreicher [5].

Now the success of the simplex method for finite-dimensional linear programming is due to the existence of a comprehensive duality theory. We can briefly summarise this as follows. A more comprehensive treatment of this may be found in any standard text on linear programming, e.g., Dantzig [11]. Consider the standard finite-dimensional linear program, FLP:

$$\begin{aligned} \text{FLP:} \quad & \text{minimize} && c^T x \\ & \text{subject to} && Ax = b, \\ & && x \geq 0, \end{aligned}$$

where A is an $m \times n$ matrix and $x \in \mathbb{R}^n$. Given FLP, we may define a corresponding finite-dimensional linear program called the *dual problem*, FLP*:

$$\begin{aligned} \text{FLP*:} \quad & \text{maximize} && b^T y \\ & \text{subject to} && A^T y \leq c. \end{aligned}$$

In this context, FLP is often called the *primal problem*. Now FLP and FLP* exhibit two important properties. The first is that if x is any feasible solution for FLP and y is any feasible solution for FLP*, then $b^T y \leq c^T x$ or, more concisely, $V[\text{FLP*}] \leq V[\text{FLP}]$. (Here and throughout the rest of this paper, we use the notation $V[\text{LP}]$ to denote the optimal value of a linear program LP, with the possibility that $V[\text{LP}] = \infty$ if LP is an infeasible minimization problem and $V[\text{LP}] = -\infty$ if LP is an infeasible maximization problem.) This result is known as *weak duality*. The second duality property is crucial to the simplex method, namely, that $V[\text{FLP*}] = V[\text{FLP}]$ (*no duality gap*) and if $V[\text{FLP}]$ is finite, then there exist optimal solutions for both FLP and FLP*. This result is known as *strong duality*.

Now as the concept of duality is at the heart of the simplex method for finite-dimensional linear programming, to extend this algorithm to CLP (or SCLP) it would be necessary to establish a similar duality theory for CLP (or SCLP). Indeed, this concern has been paramount in the development of the partial algorithms in Lehman [21], Drews [12], Hartberger [19], Segers [35], Perold [25], and Anstreicher [5] mentioned above. Moreover, because of this importance of duality results, many papers that deal solely with duality for CLP, and hence for SCLP, have appeared. In fact, the author of the problem CLP, Bellman, introduced in [8] the following dual problem for CLP:

$$\begin{aligned} \text{CLP*}': \quad & \text{minimize} && \int_0^T w(t)^T b(t)\, dt \\ & \text{subject to} && B(t)^T w(t) + \int_t^T K(s,t)^T w(s)\, ds \geq c(t), \\ & && w(t) \geq 0, \qquad t \in [0, T], \end{aligned}$$

where $w(t)$ is in the space of bounded measurable functions. Restricted to SCLP this gives

$$\begin{aligned} \text{SCLP*}': \quad & \text{maximize} && -\int_0^T a(t)^T u(t)\, dt - \int_0^T b(t)^T v(t)\, dt \\ & \text{subject to} && c(t) + \int_t^T G^T u(s)\, ds + H^T v(t) \geq 0, \\ & && u(t), v(t) \geq 0, \qquad t \in [0, T], \end{aligned}$$

with $u(t)$ and $v(t)$ in the space of bounded measurable functions. On introducing the problem, Bellman then readily established the weak duality result $V[\text{CLP}] \leq V[\text{CLP}^{*\prime}]$. The first strong duality results for CLP and CLP$^{*\prime}$ were given in Tyndall [36]. Among other things, the strong duality result required that $B, K, b \geq 0$. Consequently this result is not very useful for SCLP, as many instances of SCLP which are of practical importance (e.g., network problems) give rise to negative entries in G and a and hence also in K and b in the corresponding CLP.

However, Tyndall's work was soon extended by Levinson [23], Tyndall [37] and Grinold [14–16]. Grinold's results were more general than either Levinson's or Tyndall's; however, the problem still remained that many simple instances of CLP, such as network versions of SCLP, were not covered by the results. Using Grinold's results (in particular [15, Thm. 5, p. 42]), however, it is a trivial matter to establish the following result.

THEOREM 1.1 (no duality gap). *Suppose that* SCLP *is feasible and that H is of the form*

$$H = \begin{bmatrix} \bar{H} \\ I \end{bmatrix}.$$

Then SCLP *has an optimal solution and* $V[\text{SCLP}^{*\prime}] = V[\text{SCLP}]$ (*i.e., there is no duality gap between* SCLP *and* SCLP$^{*\prime}$).

Since the work of Grinold, numerous papers on duality for CLP have appeared. These include Schechter [34], Reiland [32], and Levine and Pomerol [22], to name a few; however, in many ways the results of Grinold remain the most general. It is perhaps not surprising that more general results have not been obtained because it is not difficult to construct counterexamples to possible duality results using the dual problem CLP$^{*\prime}$ (see, for example, Grinold [14]). This means that to establish more general strong duality results for CLP or SCLP it is necessary to consider a dual problem different from CLP$^{*\prime}$. This has been noted for a long time by authors trying to develop algorithms for the solution of CLP, e.g., Lehman [21]. In fact, the successive improvements of the algorithms for CLP mentioned above use more general dual problems by allowing δ-functionals in feasible solutions to the dual. Despite this, the need to consider more general dual problems has largely been ignored by authors studying duality for CLP. A notable exception to this, however, is Papageorgiou [24], who poses the dual problem (and the primal problem) in the space of functions of bounded variation.

There have been numerous works on duality theory for general linear programs, and a review of these may be found in Anderson and Nash [2]. The study of duality in the context of optimal control has also attracted many authors. Here it seems that it has been observed for a long time that dual problems such as CLP$^{*\prime}$ for CLP will not suffice. For instance, in Rockafellar [30], the author studied a class of convex optimal control problems and noted that strong duality theorems could be obtained by allowing dual variables to be of bounded variation. This work was later extended to more general problems in Hager and Mitter [18] and Hager [17].

On a more general note, duality theory for general optimization has been a subject of intense study, with a large number of texts and articles devoted solely to this. Two notable general contributions out of many are Rockafellar [31] and Borwein [9].

Now, in Pullan [28], the following dual problem for SCLP was introduced:

$$\begin{aligned} \text{SCLP}^*\text{:} \quad \text{maximize} \quad & -\int_0^T d\pi(t)^T a(t) - \int_0^T \eta(t)^T b(t)\, dt \\ \text{subject to} \quad & c(t) - G^T \pi(t) + H^T \eta(t) \geq 0, \\ & \eta(t) \geq 0, \text{ a.e. on } [0, T], \\ & \pi(t) \text{ monotonic increasing and right continuous} \\ & \text{on } [0, T] \text{ with } \pi(T) = 0, \end{aligned}$$

with the variables $\eta(t)$ Lebesgue-integrable functions on $[0, T]$ and where $\pi(t)$ monotonic increasing means that each component of $\pi(t)$ is monotonic increasing. Here the expression

$$\int_0^T d\pi(t)^T a(t)$$

is understood to be a Lebesgue-Stieltjes integral. We shall make frequent use of such integrals, and the reader is referred to Kolmogorov and Fomin [20] for its definition and important properties.

The dual problem SCLP* is readily seen to be a generalisation of SCLP*′ in that a feasible solution for SCLP*′ generates a feasible solution for SCLP* of the same cost, but not necessarily vice versa (see [28]). The following weak duality result was also established in [28].

LEMMA 1.2 (weak duality). $V[\text{SCLP*}] \leq V[\text{SCLP}]$.

Using this more general dual in [28], Pullan was able to establish a complete algorithm for solving SCLP when $c(t)$ and $a(t)$ were piecewise linear, with $a(t)$ continuous, and $b(t)$ was piecewise constant. As a corollary to the algorithm, the following strong duality result was also established.

THEOREM 1.3 (strong duality). *Suppose that $a(t)$ and $c(t)$ are piecewise linear (with $a(t)$ continuous) and $b(t)$ is piecewise constant on $[0, T]$. Suppose further that the feasible region for* SCLP *is nonempty and bounded; then $V[\text{SCLP}] = V[\text{SCLP*}]$ and there exist optimal solutions for both* SCLP *and* SCLP*.

In fact, although it was not stated in [28], the optimal solutions for SCLP and SCLP* could be chosen with $x(t)$ piecewise constant and with $\pi(t)$ and $\eta(t)$ piecewise linear.

The purpose of this paper is to continue the duality work for SCLP begun in Pullan [28], based on the dual problem SCLP*. In particular, we establish a strong duality result (Theorem 6.9) under the assumption of piecewise analyticity of the problem data. As with the above result, the optimal solutions in this strong duality result are also seen to be of a particularly simple form for both SCLP and SCLP*. A counterexample to a possible strong duality result is also given in §7 in the case when the assumptions of piecewise analyticity do not hold. As well as strong duality results, we also develop a duality theory in three other directions. This is both for its own sake and for its use in proving the strong duality results of §6.

The first of these is the concept of complementary slackness for SCLP (§3). This concept is based on the corresponding one for FLP, which has been extended in Anderson and Nash [2] to more general linear programs. Complementary slackness is a condition that holds at optimality, given that strong duality holds, and thus generates sufficient conditions for optimality of the problem in question. Using these ideas we thus develop several sufficient conditions for optimality of SCLP. Moreover, given mild assumptions on the costs, we show that it is possible to calculate a set of complementary slack variables for any (or at least any sufficiently well behaved) solution for SCLP that is an extreme point of the set of feasible solutions (i.e., a *basic feasible solution*). This is identical to what happens for FLP and forms an important step of the simplex method.

The second preliminary type of duality result we establish is a fairly general condition for the absence of a duality gap between SCLP and SCLP* (§4). Although not as useful as strong duality results, such results can often be used to prove convergence of possible algorithms. Having done this, we note that a fairly general condition also exists to guarantee the existence of an optimal solution for SCLP (Theorem 2.1). Hence, to establish general strong duality results it is necessary only to prove that SCLP* admits an optimal solution. This turns out to be quite difficult for general problem data unless severe restrictions are placed on the problem and in fact may not be true, as shown by Example 7.1. Nevertheless, two general results for

the existence of an optimal solution for SCLP* are established in §5. We conclude this section with a more cryptic result for the existence of an optimal solution for SCLP*. This last result is used in §6 to prove the strong duality.

Before beginning the analysis it is worth noting that the convex optimal control problem considered in Hager and Mitter [18] and Hager [17] can be seen to include SCLP by suitable transformations. Moreover, the Lagrangian dual problem considered in [18] and [17] can be simplified to SCLP*. Using the results in [18], it is possible to derive the following result.

THEOREM 1.4. *Suppose that $a(t)$ is absolutely continuous and that $b(t)$ and $c(t)$ are continuous. Suppose also that there exists a continuous feasible solution $\omega(t)$ for* SCLP *with $\omega_i(t) > 0$ for each i and $t \in [0, T]$. Then V[SCLP] $= V$[SCLP*] and there exists an optimal solution for* SCLP*.

Conditions similar to the existence of a strictly positive solution above are well known to ensure strong duality in various problems and are known as *Slater-type* conditions. However, it is also well known that such conditions often are difficult to verify or fail to hold in practice. As would be expected, by considering only a linear problem, we are able to obtain a much more complete duality theory.

For the purpose of slightly simplifying notation throughout this paper we will rewrite SCLP* in the following equivalent form, obtained by replacing $\eta(t)$ by $-\eta(t)$:

$$\begin{aligned} \text{SCLP*:}\quad & \text{maximize} && \int_0^T \eta(t)^T b(t)\,dt - \int_0^T d\pi(t)^T a(t) \\ (3)\qquad & \text{subject to} && c(t) - G^T\pi(t) - H^T\eta(t) \geq 0, \\ & && \eta(t) \leq 0, \text{ a.e. on } [0, T], \\ & && \pi(t) \text{ monotonic increasing and right continuous} \\ & && \text{on } [0, T] \text{ with } \pi(T) = 0. \end{aligned}$$

We let $\theta(t)$ denote a complete set of variables for SCLP*, i.e., $\theta(t)^T = (\pi(t)^T, \eta(t)^T)$. Also, given $\theta(t)$ we let $\psi(t)$ be the left-hand side of (3). Thus $\psi(t) = c(t) - G^T\pi(t) - H^T\eta(t)$.

We now begin the discussion by introducing some definitions and stating some previous results on SCLP that will be used throughout this paper.

2. Definitions and established results. In this section we introduce some standard definitions and state some previous results on SCLP that will be useful throughout this paper. We begin by introducing some standard notation.

For any $\zeta \in \mathbb{R}^n$, $\zeta \geq 0$, we use the notation

$$\operatorname{supp}(\zeta) = \{ i : \zeta_i > 0 \}$$

to denote the support of ζ.

Let $f : [a, b] \to \mathbb{R}^n$. We shall say that f is *analytic on a neighbourhood of* $[a, b]$ (or $[a, b)$) if there exists $\varepsilon > 0$ and an analytic function $g : (a - \varepsilon, b + \varepsilon) \to \mathbb{R}^n$ such that $f(t) = g(t)$ for all $t \in [a, b]$ (respectively, $[a, b)$). Let $P = \{t_0, t_1, \ldots, t_m\}$ be a partition of $[a, b]$. We say that f is *piecewise analytic on* $[a, b]$ *with breakpoints in* P, or simply, *piecewise analytic on* $[a, b]$, if $f(t)$ is analytic on a neighbourhood of $[t_{i-1}, t_i]$ for $i = 1, \ldots, m$. The smallest such partition P (excluding a and, if f is continuous at b, b) will be called the *breakpoints*. We use similar definitions for piecewise constant, linear, and polynomial. Finally we use the notations $f(t-)$ to denote $\lim_{s \uparrow t} f(s)$ and $f(t+)$ to denote $\lim_{s \downarrow t} f(s)$.

We shall make frequent use of the standard spaces $L_\infty[a, b]$ (bounded measurable functions on $[a, b]$), $L_1[a, b]$ (Lebesgue-integrable functions on $[a, b]$), $C[a, b]$ (continuous functions on $[a, b]$), and $NBV[a, b]$ (functions of bounded variation on $[a, b]$ normalised so that

they are right continuous on $(a, b]$ and have $f(b) = 0$). The way that functions of bounded variation are normalised in $NBV[a, b]$ is slightly nonstandard. We shall also use the notation $X^n[a, b]$ to denote the n-fold product of $X[a, b]$ with itself, where X is any one of L_∞, L_1, C, and NBV.

We shall also make occasional use of dual pairs of vector spaces (see Schaefer [33]). Let (X, Y) be a dual pair of spaces. We shall use the notation $\sigma(X, Y)$ to denote the weak topology on (X, Y).

Finally, by way of preliminary notation, it will be useful to define $F(\text{SCLP})$ to be the feasible region for SCLP, thus

$$F(\text{SCLP}) = \{x(t) \in L_\infty^{n_1}[0, T] : \text{there exists } y(t) \in C^{n_2}[0, T], z(t) \in L_\infty^{n_3}[0, T] \text{ such that } \omega(t)^T = (x(t)^T, y(t)^T, z(t)^T) \text{ is feasible for SCLP}\}.$$

We now summarise previous results relating to SCLP that will be useful in this paper. The first of these may be found in Anderson, Nash, and Perold [3]. As with finite-dimensional linear programming, we use the notation *basic feasible solution* for SCLP to denote an extreme point of the set of feasible solutions for SCLP.

THEOREM 2.1. *If the feasible region for* SCLP *is nonempty and bounded, then there exists an optimal solution for* SCLP *at a basic feasible solution.*

THEOREM 2.2. *A feasible solution, $\omega(t)$, for* SCLP *is basic if and only if the columns of*

$$K = \begin{bmatrix} G & I & 0 \\ H & 0 & I \end{bmatrix}$$

corresponding to the support of $\omega(t)$ are linearly independent for almost all $t \in [0, T]$.

Due to its use in the above theorem, the matrix K will play a significant role throughout this paper. We define K formally as well as some related properties.

DEFINITION 1. *We define*

$$K = \begin{bmatrix} G & I & 0 \\ H & 0 & I \end{bmatrix}.$$

1. *We let L be the number of basis matrices of K.*
2. *Let B be any matrix consisting of columns of K and $\gamma \in \mathbb{R}^{n_1+n_2+n_3}$. We use the notation γ_B to denote those elements of γ corresponding to the columns of K that are in B in the same order. Thus if K_i is the jth column of B, then γ_i is the jth element of γ_B.*

Let B be a basis matrix for K.

3. *Let $\omega(t)^T = (x(t)^T, y(t)^T, z(t)^T)$ be a set of variables for* SCLP. *We let $x_B(t)$ denote $\omega_B(t)$ restricted to $x(t)$. Thus $x_B(t)$ consists of the elements of $x(t)$ corresponding to those of the first n_1 columns of K that are also in B, arranged in the same order as the columns of B.*
4. *Let $\rho(t)$ be a solution to $B\rho(t) = d(t)$ for some $d(t)$ (that is, $\rho(t) = B^{-1}d(t)$). We use the notation $\rho_x(t)$ to denote the elements of $\rho(t)$ corresponding to those of the first n_1 columns of K that are also in B, arranged in the same order as the columns of B.*

In a recent paper [29], a study was made of the possible forms of optimal solutions for SCLP with various types of problem data. The main purpose of this paper was to establish conditions under which an optimal solution existed for SCLP for which $x(t)$ was either piecewise constant or piecewise analytic. We summarise these two main results below. It will be seen that these results coupled with the previous strong duality result (Theorem 1.3) provide

the key to establishing the strong duality results in §6. We first state a preliminary result proven in [29].

LEMMA 2.3. *Let $a(t)$ be any absolutely continuous function and $b(t)$, any bounded measurable function on $[0, T]$. Let $B^{(1)}, \ldots, B^{(L)}$ be the basis matrices for K and let*

$$\rho^{(i)}(t) = B^{(i)^{-1}} \begin{bmatrix} \dot{a}(t) \\ b(t) \end{bmatrix}.$$

Define $x^{(i)}(t)$ by $x^{(i)}_{B^{(i)}}(t) = \rho^{(i)}_x(t)$ with the other components of $x^{(i)}(t)$ set to zero. Let $\omega(t)$ be any basic feasible solution for SCLP; *then for almost all $t \in [0, T]$, $x(t) = x^{(i_t)}(t)$ for some i_t.*

THEOREM 2.4. *Suppose that $a(t)$ is piecewise linear and continuous, $b(t)$ is piecewise constant, and $c(t)$ is piecewise analytic on $[0, T]$. Suppose also that the feasible region for* SCLP *is nonempty and bounded. Then there exists an optimal basic feasible solution for* SCLP *with $x(t)$ piecewise constant on $[0, T]$. Moreover, let $x^{(1)}, \ldots, x^{(L)}$ be given by Lemma 2.3, where L is the number of basis matrices for K, and $P = \{t_0, t_1, \ldots, t_n\}$ be any partition of $[0, T]$ containing the breakpoints of $a(t)$, $b(t)$, and $c(t)$ and with $\dot{c}(t)^T x^{(i)} = \dot{c}(t)^T x^{(j)}$ for all $t \in (t_{m-1}, t_m)$ or $\dot{c}(t)^T x^{(i)} \neq \dot{c}(t)^T x^{(j)}$ for all $t \in (t_{m-1}, t_m)$, for each $i \neq j$ and each m. Then such $x(t)$ may be chosen so that for all m, the maximum number of breakpoints of $x(t)$ in $[t_{m-1}, t_m)$ is L.*

THEOREM 2.5. *Suppose that the costs $c(t)$ and right-hand sides $a(t)$ and $b(t)$ are piecewise analytic on $[0, T]$ (but with $a(t)$ continuous) and that the feasible region for* SCLP *is nonempty and bounded. Then there exists an optimal basic feasible solution for* SCLP *with $x(t)$ piecewise analytic on $[0, T]$.*

3. Complementary slackness results. Consider the finite-dimensional linear program FLP. The statement that strong duality holds between FLP and FLP* can be equivalently written as follows. There exists x feasible for FLP and y feasible for FLP* such that

$$x^T(c - A^T y) = 0, \tag{4}$$

i.e., x and y are *complementary slack*. This statement forms the basis of the simplex method for finite-dimensional linear programming in the following way. Suppose we have a basic feasible solution x for FLP. The next step in the simplex method is to calculate complementary slack variables for this basic feasible solution and test for optimality. In other words, we calculate a set of variables y satisfying (4) and optimality occurs if y is dual feasible.

In this section we mimic these ideas for SCLP. We will define the concept of complementary slackness in an analogous way and show that complementary slack variables can be calculated for any given basic feasible solution for SCLP, given that it and the costs satisfy some mild assumptions. The calculation of complementary slack variables for FLP amounts to solving a system of linear equations involving the costs and a basis matrix. We will see that a similar method applies for SCLP. As strong duality need not necessarily hold between SCLP and SCLP*, the following complementary slackness results can also be seen as sufficient, but not necessary, conditions for optimality.

The basis of our discussion is the following complementary slackness result based on the weak duality result between SCLP and SCLP* proven in Pullan [28].

LEMMA 3.1. *If ω is feasible for* SCLP, *θ is feasible for* SCLP* *and*

$$\int_0^T \psi(t)^T x(t)\, dt = \int_0^T d\pi(t)^T y(t) = \int_0^T \eta(t)^T z(t)\, dt = 0, \tag{5}$$

(where $\psi(t) = c(t) - G^T\pi(t) - H^T\eta(t)$), then ω and θ are optimal for SCLP *and* SCLP*, *respectively. Moreover, strong duality holds between* SCLP *and* SCLP* *if and only if there exists ω feasible for* SCLP *and θ feasible for* SCLP* *such that* (5) *holds.*

Proof. Suppose that ω is feasible for SCLP, θ is feasible for SCLP*, and (5) holds. Then

$$\begin{aligned}
0 &= \int_0^T \psi(t)^T x(t)\,dt + \int_0^T d\pi(t)^T y(t) - \int_0^T \eta(t)^T z(t)\,dt \\
&= \int_0^T c(t)^T x(t)\,dt - \int_0^T \left(G^T\pi(t) + H^T\eta(t)\right)^T x(t)\,dt + \int_0^T d\pi(t)^T y(t) \\
&\quad - \int_0^T \eta(t)^T z(t)\,dt \\
&= \int_0^T c(t)^T x(t)\,dt + \int_0^T d\pi(t)^T \left(\int_0^t Gx(s)\,ds + y(t)\right) - \int_0^T \eta(t)^T (Hx(t) + z(t))\,dt \\
&= \int_0^T c(t)^T x(t)\,dt - \int_0^T \eta(t)^T b(t)\,dt + \int_0^T d\pi(t)^T a(t)
\end{aligned}$$

by integrating by parts (see, for example, Dunford and Schwartz [13, p. 154]) and the feasibility of $\omega(t)$. The first statement of the lemma now follows by weak duality (Lemma 1.2). By reversing this argument we obtain the second statement. □

It would seem that the equation for SCLP corresponding to (4) for FLP is (5). Indeed, it is possible to show that the above result is exactly the restriction to SCLP of the complementary slackness result for general linear programs given by Anderson and Nash [2]. However, the equations (5) by themselves are not easily solved. If, however, feasibility of ω in SCLP and θ in SCLP* is also assumed, then the integrals in (5) may be simplified. With this in mind we now proceed to simplify (5) before defining the notion of complementary slackness. The more difficult expression to simplify is

$$\int_0^T d\pi(t)^T y(t) = 0. \tag{6}$$

To solve this we need to digress briefly and consider the concept of a function increasing at a point.

DEFINITION 2. *Let $f : [a, b] \to \mathbb{R}$ be a monotonic increasing function and $t \in (a, b)$. We say that f is* strictly increasing *at t if for any t_1, $t_2 \in [a, b]$ with $t \in (t_1, t_2)$ we have $f(t_1) < f(t_2)$. For the definition of strictly increasing at $t = a$ or $t = b$ we use the intervals $[a, t_2)$ and $(t_1, b]$, respectively, in place of (t_1, t_2).*

We now derive some useful results based on this definition. It is worth noting the converse of the above definition explicitly, namely, if f is monotonic increasing but *not* strictly increasing at t, then there exists t_1 and t_2 with $t \in (t_1, t_2)$ and $f(t_1) = f(t_2)$; i.e., f is constant on some interval containing t. This statement forms the basis for our next lemma.

LEMMA 3.2. *Suppose $f : [a, b] \to \mathbb{R}$ is monotonic increasing on $[a, b]$. Suppose there is an open interval $(\alpha, \beta) \subseteq [a, b]$ such that f is* not *strictly increasing at any point of (α, β); then f is constant on (α, β).*

Proof. Let $x, y \in (\alpha, \beta)$. Assume $x < y$. Consider the interval $[x, y]$. For any $t \in [x, y]$, f is not strictly increasing at t, so by the remark above, there exists an open interval containing t on which f is constant. By the Heine–Borel theorem (see, for example, Apostol [6]), a finite number of these intervals cover $[x, y]$, and as these intervals are not pairwise disjoint, we see that f is constant on $[x, y]$. In particular $f(x) = f(y)$, and so, as x and y were arbitrary, f is constant on (α, β). □

In most cases f strictly increasing at a point t will imply that either f is discontinuous at t or f is strictly increasing on some interval containing t. In particular, this will be true if f is analytic on some interval containing t (see Pullan [27, Cor. D.2.1]). However, there do exist continuous monotonic increasing functions that have points where they are strictly increasing but which are not strictly increasing on any interval (see [27, Ex. D.1]).

The next result is the main result that will be used concerning the concept of strictly increasing at a point. It is this result that allows us to unpack (6) to give a more satisfactory definition of complementary slackness for SCLP.

LEMMA 3.3. *Let $f : [a, b] \to \mathbb{R}$ be monotonic increasing and $g : [a, b] \to [0, \infty)$ be continuous. Suppose that f is strictly increasing at $t \in [a, b]$ and*

$$\int_a^b g(s)\, df(s) = 0; \tag{7}$$

then $g(t) = 0$. Conversely, if $g(t) = 0$ for every point t at which f is strictly increasing, then (7) holds.

Proof. Suppose that f is monotonic increasing, g is nonnegative on $[a, b]$, and (7) holds. Suppose also that f is strictly increasing for some t but $g(t) > 0$. Then, as g is continuous, there exists an interval I such that $t \in I$ and $g(s) > 0$ for all $s \in I$. Hence, as f is strictly increasing at t,

$$\int_I g(s)\, df(s) > 0.$$

However, as $g(s) \geq 0$ for all $s \in [a, b]$, we have

$$\int_a^b g(s)\, df(s) > 0,$$

which is not possible, so we must have $g(t) = 0$.

Now suppose that $g(t) = 0$ at every point t at which f is strictly increasing. Let $S = \{t : g(t) > 0\}$, then f is not strictly increasing at any point of S. Suppose $S = \bigcup_{n=1}^{N} (\alpha_n, \beta_n)$, $(N \leq \infty)$, where $(\alpha_n, \beta_n) \cap (\alpha_m, \beta_m) = \emptyset$ if $m \neq n$. Then by Lemma 3.2, f is constant on (α_n, β_n). Hence

$$\int_S g(t)\, df(t) = \sum_{n=1}^{N} \int_{(\alpha_n, \beta_n)} g(t)\, df(t) = 0.$$

If either $g(a) > 0$ or $g(b) > 0$, then a similar argument will give that f is constant on some interval containing a if $g(a) > 0$, or b if $g(b) > 0$. Hence

$$\int_{\{a,b\}} g(t)\, df(t) = 0.$$

Finally, $g(t) = 0$ on $(a, b) - S$ and so the result follows. □

We now define the notation of complementary slackness for SCLP.

DEFINITION 3 (complementary slackness). *Let ω be a feasible solution for* SCLP. *We say that $\theta \in NBV^{n_2}[0, T] \times L_1^{n_3}[0, T]$ is* complementary slack with ω *if π is right continuous at zero, and for any α, β, and i,*

1. *if $y_i(t) > 0$ on (α, β), then π_i is constant on $[\alpha, \beta)$;*
2. *if $y_i(T) > 0$, then $\pi_i(T-) = \pi_i(T) = 0$;*
3. *$\psi_i(t) = 0$ a.e. on $\{t : x_i(t) > 0\}$;*
4. *$\eta_i(t) = 0$ a.e. on $\{t : z_i(t) > 0\}$.*

It is worth noting that if θ is also feasible for SCLP*, each of the conditions for complementary slackness has a corresponding contrapositive. For instance, condition (1) can be written equivalently as follows: if π_i is strictly increasing at t, then $y_i(t) = 0$.

Given this definition, we may now write the complementary slackness result, Lemma 3.1 in an analogous manner to that for FLP.

THEOREM 3.4 (complementary slackness). *Suppose that ω is feasible for* SCLP *and that θ is complementary slack with ω. If θ is feasible for* SCLP*, then ω is optimal for* SCLP *and θ is optimal for* SCLP*. Moreover, strong duality holds between* SCLP *and* SCLP* *if and only if there exists ω feasible for* SCLP *and θ feasible for* SCLP* *such that θ is complementary slack with ω.*

Proof. Suppose that strong duality holds between SCLP and SCLP*. Then by Lemma 3.1 there exists ω feasible for SCLP and θ feasible for SCLP* such that

$$\int_0^T \psi(t)^T x(t)\,dt = \int_0^T d\pi(t)^T y(t) = \int_0^T \eta(t)^T z(t)\,dt = 0.$$

Clearly we have $\psi_i(t) = 0$ a.e. on $\{t : x_i(t) > 0\}$ and $\eta_i(t) = 0$ a.e. on $\{t : z_i(t) < 0\}$ for each i. Suppose that $y_i(t) > 0$ on (α, β). Then by Lemma 3.3, π_i is *not* strictly increasing on any point of (α, β). Hence by Lemma 3.2 and as π is right continuous, π_i is constant on $[\alpha, \beta)$. Similarly if $y_i(T) > 0$, then $\pi_i(T-) = \pi_i(T) = 0$. Hence θ is complementary slack with ω.

This establishes the result one way. The proof of the converse is similar. □

Having defined complementary slackness, there still remains the question of how such complementary slack variables should be calculated and, more importantly, if they exist at all. In the finite-dimensional linear program their existence is guaranteed for *basic* feasible solutions. We see that this is essentially true for SCLP as well, in the sense that a θ can be defined satisfying points 1–4 of the definition. This is the content of the next result. However, there is one slight technicality, and that is in ensuring that this θ is in the appropriate space. As the result below shows, this will be true for sufficiently well-behaved costs and a basic feasible solution with sufficiently well-behaved support over time. For example, right continuous costs of bounded variation and a piecewise continuous basic feasible solution would suffice.

THEOREM 3.5. *Let ω be a basic feasible solution for* SCLP *with $B(t)$ the columns of K corresponding to the support of $\omega(t)$. Let $\zeta \in L_\infty^{n_2}[0, T]$ be any arbitrary function such that if $y_j(t) > 0$ on $(\alpha, \beta) \cap [0, T]$ for some α, β, and j, then ζ_j is constant on $[\alpha, \beta) \cap [0, T]$. Let θ be any solution of*

$$\theta(t)^T B(t) = \bar{c}_B(t)^T \tag{8}$$

for almost all $t \in [0, T)$, where

$$\bar{c}(t) = \begin{bmatrix} c(t) \\ \zeta(t) \\ 0 \end{bmatrix}$$

(where 0 is of dimension n_3), with $\pi(t) = \pi(t+)$ for t such that (8) *does not have a solution, if the limit exists, and $\pi(T) = 0$. If this θ is an element of $NBV^{n_2}[0, T] \times L_1^{n_3}[0, T]$, then it is complementary slack with ω. Conversely, any θ complementary slack with ω satisfies* (8) *for some $\zeta(t)$ for almost all $t \in [0, T]$.*

Proof. Let $\zeta(t) : [0, T] \to \mathbb{R}^{n_2}$ be any arbitrary function satisfying ζ_j constant on $[\alpha, \beta)$ if $y_j(t) > 0$ on (α, β). Let $\bar{c}$ be as above. From Theorem 2.2 there exists a solution $\theta(t)$ (although not necessarily unique unless $B(t)$ is square) to (8) for almost all $t \in [0, T]$. To

show that any solution $\theta \in NBV^{n_2}[0, T] \times L_1^{n_3}[0, T]$ is complementary slack with ω we unpack (8) into the three equivalent separate equations to give

$$\left(\pi(t)^T G + \eta(t)^T H\right)_i = c_i(t), \tag{9}$$

$$\pi_j(t) = \zeta_j(t), \tag{10}$$

$$\eta_k(t) = 0. \tag{11}$$

for those indices i, j, and k which correspond to the columns of $B(t)$ (i.e., i such that K_i is in $B(t)$, j such that K_{n_1+j} is in $B(t)$, and k such that $K_{n_1+n_2+k}$ is in $B(t)$). Consider the conditions of Definition 3 in turn.

1. If $y_j(t) > 0$ on (α, β), then by definition of $\zeta(t)$, $\zeta_j(t)$ is constant on $[\alpha, \beta)$. Hence π_j is constant a.e. on $[\alpha, \beta)$ by (10) and hence everywhere as π is right continuous.
2. If $y_j(T) > 0$, then by a similar argument to the above, we have $\pi_j(T-) = \pi_j(T) = 0$.
3. If $x_i(t) > 0$ on a set S, then for almost all $t \in S$ we have by (9)

$$\left(c(t) - G^T \pi(t) - H^T \eta(t)\right)_i = 0,$$

   i.e., $\psi_i(t) = 0$ a.e. on S.
4. If $z_i(t) > 0$ on a set S, then for almost all $t \in S$ we have by (11) that $\eta_i(t) = 0$.

This establishes the result in one direction and the proof of the converse is similar. □

It should be noted that the construction of the complementary slack variables is by no means unique. Even if (8) has a unique solution, in general there will be a countable number of undetermined constants, one for each interval on which $y_j(t) > 0$ for some j. A similar problem frequently occurs in finite-dimensional linear programming. In this case, when there is an arbitrary manner in which to calculate the complementary slack variables, the basic feasible solution is called *degenerate*.

It is worth noting that problems of degeneracy in an infinite-dimensional context have been a major stumbling block in the development of general CLP algorithms. For instance, in Perold [25, 26], the author attempted to give a general pivot operation for CLP. However, the pivot operation only worked under some nondegeneracy assumptions. Similarly in Anderson [1], an initial attempt was made at developing an algorithm for SCLP with the same restrictions on a, b, and c as those considered in Pullan [28] (see §1). Here again, the method failed to work in general due to degeneracy problems. The first algorithm to overcome the degeneracy problem was that for CNP, the single-commodity network version of SCLP given in Anderson and Philpott [4] (under the same assumptions on a, b, and c as those for SCLP in both [1] and [28]). Here the authors encountered and distinguished between the two types of degeneracy we have encountered for SCLP, namely, the possibility of not solving the complementary slackness equations uniquely and, second, of determining the constants associated with the intervals where $y_j(t) > 0$ for some j. The resulting algorithm, however, was very complicated, and great care was needed in the specification of how to handle degeneracy. The success of the more recent algorithm in Pullan [28] for SCLP is due to the fact that for the particular a, b, and c chosen, it is possible to handle the problems of degeneracy by the ordinary simplex method for FLP.

The problems of degeneracy will play a significant role in the development of strong duality results for SCLP. In §6 we will look again at degeneracy and give a concrete definition for a particular SCLP problem to be degenerate for piecewise analytic a and b (Definition 4). We will see, however, that the only type of degeneracy that need concern us here is the possibility of not solving (8) uniquely. In fact, in the absence of such degeneracy, the strong duality results of §6 become much easier. This is not surprising because strong duality is relatively simple to establish between FLP and FLP* in the case where there is no degeneracy present.

Even in general linear programming, the absence of degeneracy, appropriately defined, allows a strong duality result for a general dual problem to be established with ease (see Anderson and Nash [2, Thm. 2.10]).

4. Conditions for the absence of a duality gap. We now consider sufficient conditions to ensure that there is no duality gap between SCLP and SCLP*, i.e., $V[\text{SCLP}] = V[\text{SCLP*}]$. As well as being interesting in its own right, such a result is one of the necessary conditions for strong duality. The main result of this section is Theorem 4.2. The result here is quite general, and so for most practical instances of SCLP we may deduce that there is no duality gap between SCLP and SCLP*. We begin by establishing an important property of $F(\text{SCLP})$, the feasible region for SCLP.

LEMMA 4.1. *The set* $F(\text{SCLP})$ *is closed in the* $\sigma(L_\infty^{n_1}[0, T], L_1^{n_1}[0, T])$ *topology.*

Proof. Suppose $x \notin F$. There are three cases to consider, depending on which constraint is violated. If $x_i(t) < 0$ on some set S of nonzero measure for some i, then define $f \in L_1^{n_1}[0, T]$ by

$$f_j = \begin{cases} 0, & j \neq i, \\ \chi_S, & j = i, \end{cases} \tag{12}$$

where χ_S is the characteristic function of S. Then

$$\int_0^T f(t)^T x(t)\, dt < 0,$$

but for any $\alpha \in F$,

$$\int_0^T f(t)^T \alpha(t)\, dt \geq 0,$$

and so x is contained in some weakly open set that does not intersect with F.

Now suppose that

$$\int_0^t [Gx(s)]_i\, ds > a_i(t)$$

for some index i and $t \in [0, T]$. By the continuity of both the integral and a, this will be true for all t in some open interval $S = (t_1, t_2)$, with equality at the point t_1. Define $f \in L_1^{n_2}[0, T]$ by (12) and set $g = G^T f$. Then

$$\int_0^T g(t)^T x(t)\, dt > \int_S a_i(t)\, dt,$$

and for any $\alpha \in F$,

$$\int_0^T g(t)^T \alpha(t)\, dt \leq \int_S a_i(t)\, dt,$$

so x is again contained in some weakly open set that does not intersect with F.

The remaining case, namely, $[Hx(t)]_i > b_i(t)$ on some set S of nonzero measure for some i, is similar. □

We now prove the main result for the absence of a duality gap between SCLP and SCLP*. The proof involves considering a sequence of linear programs whose optimal values converge to that of SCLP and by using properties of the weak topology.

THEOREM 4.2 (no duality gap). *Suppose that $a(t)$ is absolutely continuous on $[0, T]$ with Riemann-integrable derivative. Suppose also that $c(t)$ and $b(t)$ are Riemann-integrable and that the feasible region for* SCLP *is nonempty. If, furthermore, there exist a continuous piecewise linear function $\bar{a}(t)$ and a piecewise constant function $\bar{b}(t)$ such that $a(t) \leq \bar{a}(t)$ and $b(t) \leq \bar{b}(t)$ on $[0, T]$ and such that the feasible region for the problem* SCLP *with $a(t)$ and $b(t)$ replaced by $\bar{a}(t)$ and $\bar{b}(t)$, respectively, is bounded, then*

$$V[\text{SCLP}] = V[\text{SCLP*}];$$

i.e., there is no duality gap between SCLP *and* SCLP*.

Proof. Let $\{b^{(n)}\}_{n=1}^{\infty}$ and $\{c^{(n)}\}_{n=1}^{\infty}$ be sequences of piecewise constant functions such that $b^{(n)}(t) \leq \bar{b}(t)$, $b^{(n)}(t) \downarrow b(t)$, and $c^{(n)}(t) \uparrow c(t)$ a.e. on $[0, T]$ (see Apostol [6, Thm. 10.11]). By considering a sequence $r^{(n)}(t) \downarrow \dot{a}(t)$ we may also construct a sequence of continuous piecewise linear functions $\{a^{(n)}\}_{n=1}^{\infty}$ such that $a^{(n)}(t) \leq \bar{a}(t)$ and $a^{(n)}(t) \downarrow a(t)$ uniformly on $[0, T]$. Let SCLP_n be the SCLP problem with $a(t)$, $b(t)$, and $c(t)$ replaced by $a^{(n)}(t)$, $b^{(n)}(t)$, and $c^{(n)}(t)$, respectively. Let $F_n = F(\text{SCLP}_n)$ and $F = F(\text{SCLP})$. As $\{a^{(n)}\}_{n=1}^{\infty}$ and $\{b^{(n)}\}_{n=1}^{\infty}$ are monotonic sequences converging to a in $C^{n_2}[0, T]$ and b in $L_{\infty}^{n_3}[0, T]$, respectively, it is not difficult to see that

$$F_{n+1} \subseteq F_n$$

for all n and

$$F = \bigcap_{n=1}^{\infty} F_n.$$

Now by the boundedness assumption, it follows that F_n, for each n, and F are bounded. Moreover F_n, for each n, and F are closed in the $\sigma(L_{\infty}^{n_1}[0, T], L_1^{n_1}[0, T])$ topology by Lemma 4.1.

Let $x^{(n)}(t)$ be an optimal solution for SCLP_n (i.e., the x part of the optimal $\omega^{(n)}(t)$, the existence of which is given by Theorem 2.1). We now construct a subsequence $\{x_{n_k}\}_{k=1}^{\infty}$ and $x \in F$ such that $x_{n_k} \to x$ in the $\sigma(L_{\infty}^{n_1}[0, T], L_1^{n_1}[0, T])$ topology.

To achieve this we recall a result from functional analysis which states that if X is a separable normed linear space (that is, has a countable dense subset), then any norm-bounded sequence in X^* (the dual of X) has a subsequence that converges in the $\sigma(X^*, X)$ topology to an element of X^* (see, for example, Kolmogorov and Fomin [20, Thm. 4, p. 202]). Now it is well known that $L_1[0, T]$, and hence $L_1^{n_1}[0, T]$, is a separable space (see again [20, Thm. 3, p. 382]). Hence, as $L_1^{n_1}[0, T]^* = L_{\infty}^{n_1}[0, T]$, there exists a subsequence, $\{x_{n_k}\}_{k=1}^{\infty}$, of $\{x_n\}_{n=1}^{\infty}$ and $x \in L_{\infty}^{n_1}[a, b]$ such that $x_{n_k} \to x$ in the $\sigma(L_{\infty}^{n_1}[0, T], L_1^{n_1}[0, T])$ topology. Let k be given, and consider the sequence $\{x_{n_l}\}_{l=k}^{\infty} \subseteq F_{n_k}$. This sequence converges to x, and so as F_{n_k} is closed, we have $x \in F_{n_k}$. As k was arbitrary and $F = \bigcap_{n=1}^{\infty} F_n$, $x \in F$. It will be seen in fact that x is optimal for SCLP.

From the construction above we now have $x \in F$ and

$$\lim_{k \to \infty} \int_0^T d(t)^T x^{(n_k)}(t)\, dt = \int_0^T d(t)^T x(t)\, dt$$

for any $d \in L_1^{n_1}[0, T]$. Hence

$$\lim_{k \to \infty} \int_0^T c^{(m)}(t)^T x^{(n_k)}(t)\, dt = \int_0^T c^{(m)}(t)^T x(t)\, dt$$

for any m. Also by Lebesgue's dominated convergence theorem we have

$$\lim_{m \to \infty} \int_0^T c^{(m)}(t)^T x(t)\, dt = \int_0^T c(t)^T x(t)\, dt.$$

Thus, as $c^{(m)}(t) \uparrow c(t)$ a.e. on $[0, T]$, we have

$$\lim_{k\to\infty} \int_0^T c^{(n_k)}(t)^T x^{(n_k)}(t)\, dt = \int_0^T c(t)^T x(t)\, dt. \tag{13}$$

Now by the strong duality result from Pullan [28, Thm. 1.3], there exists $\theta^{(n)}(t)$ feasible for SCLP_n^* such that

$$\int_0^T c^{(n)}(t)^T x^{(n)}(t)\, dt = \int_0^T \eta^{(n)}(t)^T b^{(n)}(t)\, dt - \int_0^T d\pi^{(n)}(t)^T a^{(n)}(t).$$

But $a^{(n)}(t) \geq a(t)$ and $b^{(n)}(t) \geq b(t)$ a.e. on $[0, T]$. Hence

$$\int_0^T c^{(n)}(t)^T x^{(n)}(t)\, dt \leq \int_0^T \eta^{(n)}(t)^T b(t)\, dt - \int_0^T d\pi^{(n)}(t)^T a(t). \tag{14}$$

Now $c^{(n)}(t) \leq c(t)$ a.e. on $[0, T]$ and hence

$$\begin{aligned} c(t) - G^T\pi^{(n)}(t) - H^T\eta^{(n)}(t) &\geq c^{(n)}(t) - G^T\pi^{(n)}(t) - H^T\eta^{(n)}(t) \\ &\geq 0 \end{aligned}$$

a.e. on $[0, T]$ by the feasibility of $\theta^{(n)}$ in SCLP_n^*. Thus $\theta^{(n)}(t)$ is feasible for SCLP*. Hence by weak duality (Lemma 1.2),

$$\int_0^T \eta^{(n)}(t)^T b(t)\, dt - \int_0^T d\pi^{(n)}(t)^T a(t) \leq \int_0^T c(t)^T x(t)\, dt.$$

Combining this with (13) and (14) above we have

$$\lim_{k\to\infty} \left[\int_0^T \eta^{(n_k)}(t)^T b(t)\, dt - \int_0^T d\pi^{(n_k)}(t)^T a(t) \right] = \int_0^T c(t)^T x(t)\, dt,$$

and so the result follows. □

We note that while the boundedness conditions of Theorem 4.2 may look complicated, it is not difficult to show that they will be satisfied if the set $\{\xi : H\xi \leq b(t),\ \xi \geq 0\}$ is bounded for each $t \in [0, T]$. As noted in Pullan [27], this is not that different in practice from insisting that the feasible region be bounded. It is worth noting that the H of Grinold's result in Theorem 1.1 satisfies the condition $\{\xi : H\xi \leq b(t),\ \xi \geq 0\}$ bounded for each $t \in [0, T]$ and $b \in L_\infty^{n_3}[0, T]$.

5. Conditions for the existence of optimal solutions for SCLP*. We now present two general results that establish the existence of an optimal solution for SCLP* and a third, less general result. The first two general results are relatively easy to prove but are not very useful in practice as they tend to be rather restrictive. It is the third result that will be used to prove strong duality in the next section by using the result for the absence of a duality gap (Theorem 4.2) and the result for the existence of a piecewise analytic optimal solution for SCLP (Theorem 2.5).

The first result shows that SCLP* has an optimal solution if its feasible region is a bounded subset of $L_\infty^{n_2+n_3}[0, T]$. As this result is easy to prove and does not form part of the development of the strong duality results in §6, we give only an outline of its proof. To use this result to ensure strong duality, we would need to guarantee the existence of an optimal solution for SCLP. The most readily available condition for this is that the feasible region for SCLP is also

bounded (Theorem 2.1). However, it can be shown (see Pullan [27]), that both the feasible regions for SCLP and SCLP* bounded in the appropriate sense can be very restrictive and thus not a very practical condition to impose on the problem. Note that we make the distinction that the feasible region of SCLP* is a bounded subset of $L_\infty^{n_2+n_3}[0, T]$, because in general, $\eta \in L_1^{n_3}[0, T]$.

THEOREM 5.1. *Suppose that the feasible region of* SCLP* *is a nonempty and bounded subset of* $L_\infty^{n_2+n_3}[0, T]$. *Then there exists an optimal solution for* SCLP*.

Proof. For an explicit proof see Pullan [27]. It involves showing that in this case the feasible region is closed in an appropriate weak topology. The proof is then made complete by using Alaoglu's theorem for weak compactness (see, for example, Dunford and Schwartz [13]). □

The next general result shows that if there exists a sequence of feasible solutions for SCLP* that are of uniformly bounded variation and whose objective function values converge to the optimal value, then SCLP* attains its optimal value.

THEOREM 5.2. *Suppose that there exists a sequence* $\{\theta^{(n)}\}_{n=1}^\infty$ *of feasible solutions for* SCLP* *of uniformly bounded variation; i.e., there exists N such that*

$$\int_0^T \|d\theta^{(n)}(t)\|_\infty \leq N$$

for all n. Suppose also that

$$\lim_{n\to\infty}\left[\int_0^T \eta^{(n)}(t)^T b(t)\,dt - \int_0^T d\pi^{(n)}(t)^T a(t)\right] = V[\text{SCLP*}];$$

then there exists θ *optimal for* SCLP*.

Proof. By the Helly selection principle (see, for example, Kolmogorov and Fomin [20, Thm. 5, p. 372]) there exists a subsequence $\{\theta^{(n_k)}\}_{k=1}^\infty$ and θ of bounded variation on $[0, T]$ such that $\theta^{(n_k)}(t) \to \theta(t)$ for all $t \in [0, T]$. The results now follow by Helly's convergence theorem (see again, Kolmogorov and Fomin, [20, Thm. 4, p. 370]) and Lebesgue's dominated convergence theorem. □

Our final, less general result in this section is the one that we will be using to establish the strong duality results in the following section. It is a specialised version of the previous result and presents a condition that guarantees the existence of a piecewise analytic optimal solution for SCLP* based on a particular sequence of feasible solutions whose objective function values approach the optimal value of SCLP*. Although its statement is more cryptic than the previous two results, it will be made clearer if it is compared with Theorem 3.5 on the calculation of complementary slack variables. It will be in the context of a sequence of suitable complementary slack variables that this result will be used.

LEMMA 5.3. *Suppose that* $a^{(n)} \to a$ *and* $b^{(n)} \to b$ *uniformly. Let* SCLP_n *be the problem* SCLP *with a replaced by* $a^{(n)}$ *and b replaced by* $b^{(n)}$. *Suppose there exists* $\theta^{(n)}$ *optimal for* SCLP_n^* *(the dual of* SCLP_n*) satisfying the following conditions:*

1. *For each n,* $\theta^{(n)}$ *is piecewise analytic on* $[0, T]$ *with breakpoints in the partition* $\{t_0^{(n)}, \ldots, t_{m_n}^{(n)}\}$ *of* $[0, T]$.
2. *For each* $i = 1, \ldots, m_n$ *there exists a basis matrix* $B(t_i^{(n)})$ *for K such that*

$$\theta^{(n)}(t) = \left(B(t_i^{(n)})^{-1}\right)^T \bar{c}^{(n)}_{B(t_i^{(n)})}(t)$$

for $t \in [t_{i-1}^{(n)}, t_i^{(n)})$, *where*

$$\bar{c}^{(n)}(t) = \begin{bmatrix} c(t) \\ \pi^{(n)}(t_i^{(n)}-) \\ 0 \end{bmatrix}$$

for $t \in [t_{i-1}^{(n)}, t_i^{(n)})$.

Suppose also that there exist M *and* N *such that* $m_n \leq M$ *and* $\|\theta^{(n)}\|_\infty \leq N$ *for all* n. *Then there exists a piecewise analytic optimal solution for* SCLP* *with at most* M *breakpoints.*

Proof. By introducing extra artificial breakpoints, if necessary, we can assume that $m_n = M$ for all n, i.e., that each $\theta^{(n)}$ has exactly M breakpoints. Now $\theta^{(n)}(t)$ is uniformly bounded for all n and t, so there exists $\{n_k\}_{k=1}^{\infty}$ and a partition $P = \{t_0, \ldots, t_M\}$ of $[0, T]$ such that $t_i^{(n_k)} \to t_i$ for $i = 0, \ldots, M$ and $\pi^{(n_k)}(t_i^{(n_k)}-)$ converges for $i = 1, \ldots, M$. As there are only a finite number of basis matrices for K, we may also assume that the subsequence $\{n_k\}_{k=1}^{\infty}$ is chosen so that for each i, $B(t_i^{(n_k)}) = B^{(i)}$ for some basis matrix $B^{(i)}$ of K. We now let

$$\bar{c}(t) = \lim_{k\to\infty} \bar{c}^{(n_k)}(t)$$

for $t \in [0, T] - P$ and $\bar{c}(t) = \bar{c}(t+)$ for $t \in P$. These limits exist as $\pi^{(n_k)}(t_i^{(n_k)}-)$ converges for each i. We now define

$$\theta(t) = \left(B^{(i)^{-1}}\right)^T \bar{c}_{B^{(i)}}(t)$$

for $t \in [t_{i-1}, t_i)$ for each $i = 1, \ldots, M$ and $\theta(T) = 0$. We now have $\theta^{(n_k)}(t) \to \theta(t)$ for all $t \in [0, T] - P$ and $\pi^{(n)}(T) = \pi(T)$ for all n. Now by taking a further subsequence, if necessary, we can assume that $\lim_{k\to\infty} \theta^{(n_k)}(t_i)$ exists for $i = 1, \ldots, M-1$. Note that $\theta(t_i)$ may or may not be equal to $\lim_{k\to\infty} \theta^{(n_k)}(t_i)$. Using Helly's convergence theorem and the fact that the Lebesgue–Stieltjes integral is not affected by the value of θ at its breakpoints in $(0, T)$ (see Kolmogorov and Fomin [20, Prob. 7, p. 377]) we have for any continuous function g that

$$\lim_{k\to\infty} \int_0^T d\pi^{(n_k)}(t)^T g(t) = \int_0^T d\pi(t)^T g(t).$$

This will allow us to establish the result that θ is optimal for SCLP*.

Now the feasible region for SCLP$_n^*$ is the same for all n and the same as that for SCLP*. Hence $\theta^{(n_k)}$ is feasible for SCLP*, and thus θ is feasible for SCLP*. We now show that θ is optimal for SCLP*. Now by the remark above and Lebesgue's dominated convergence theorem we have

$$\lim_{k\to\infty} \left[\int_0^T \eta^{(n_k)}(t)^T b^{(m)}(t)\, dt - \int_0^T d\pi^{(n_k)}(t)^T a^{(m)}(t) \right]$$
$$= \int_0^T \eta(t)^T b^{(m)}(t)\, dt - \int_0^T d\pi(t)^T a^{(m)}(t)$$

for all m. Now $a^{(m)} \to a$ in $C^{n_2}[0, T]$, $b^{(m)} \to b$ in $L_1^{n_3}[0, T]$, and θ acts as a continuous linear functional on $C^{n_2}[0, T] \times L_1^{n_3}[0, T]$. Hence we have

$$\lim_{m\to\infty} \left[\int_0^T \eta(t)^T b^{(m)}(t)\, dt - \int_0^T d\pi(t)^T a^{(m)}(t) \right] = \int_0^T \eta(t)^T b(t)\, dt - \int_0^T d\pi(t)^T a(t).$$

Hence, by the uniform convergence of $a^{(m)}$ and $b^{(m)}$, we have

$$\begin{aligned}\lim_{k\to\infty} V[\text{SCLP}^*_{n_k}] &= \lim_{k\to\infty}\left[\int_0^T \eta^{(n_k)}(t)^T b^{(n_k)}(t)\,dt - \int_0^T d\pi^{(n_k)}(t)^T a^{(n_k)}(t)\right]\\ &= \int_0^T \eta(t)^T b(t)\,dt - \int_0^T d\pi(t)^T a(t)\\ &\le V[\text{SCLP}^*]. \qquad (15)\end{aligned}$$

Suppose that $V[\text{SCLP}^*] > \lim_{k\to\infty} V[\text{SCLP}^*_{n_k}]$. Then there exists $\bar{\theta}$ feasible for SCLP* and such that

$$\int_0^T \bar{\eta}(t)^T b(t)\,dt - \int_0^T d\bar{\pi}(t)^T a(t) > \lim_{k\to\infty} V[\text{SCLP}^*_{n_k}].$$

This is a contradiction since, by similar arguments to the above, we would then have that

$$\lim_{m\to\infty}\left[\int_0^T \bar{\eta}(t)^T b^{(m)}(t)\,dt - \int_0^T d\bar{\pi}(t)^T a^{(m)}(t)\right] = \int_0^T \bar{\eta}(t)^T b(t)\,dt - \int_0^T d\bar{\pi}(t)^T a(t),$$

thus giving a feasible solution for $\text{SCLP}^*_{n_k}$ for some n_k with objective function value strictly greater than $V[\text{SCLP}^*_{n_k}]$. Hence (15) is an equality and so θ is optimal for SCLP*. □

6. Strong duality results. We now turn to strong duality results between SCLP and SCLP*. The main result in this section is that strong duality holds between SCLP and SCLP* if $a(t)$, $b(t)$, and $c(t)$ are piecewise analytic, with $a(t)$ also continuous, and the feasible region for SCLP is both nonempty and bounded (Theorem 6.9). It is also shown that in this case optimal solutions exist for both the primal and the dual which are piecewise analytic on $[0, T]$. Before considering the general case, we will first show that strong duality holds between SCLP and SCLP* in the case where $c(t)$ is piecewise analytic, $a(t)$ is piecewise linear and continuous, and $b(t)$ is piecewise constant (Theorem 6.6). Although Theorem 6.6 is just a special case of Theorem 6.9, it is convenient to separate the two in order to bring greater clarity and understanding.

The proofs of the strong duality results in this section are quite long and involved. However, it is possible to break up the proofs into smaller segments. With this in mind we now present several fairly unrelated results that will be used in the proofs of the strong duality results.

6.1. Preliminary results. In Pullan [28], the key to establishing the strong duality result, Theorem 1.3, was a special discretization of SCLP and its properties. Under the appropriate assumptions on the problem data for SCLP it was shown that if we had an optimal solution for SCLP that was piecewise constant, then we could construct an optimal solution for this discretization in a natural way. As the (finite-dimensional) dual of this discretization was a natural discretization of SCLP*, the strong duality theorem of finite-dimensional linear programming allowed the construction of an optimal solution for SCLP*, thus establishing strong duality. The difficult part of the above analysis was showing that an optimal solution for SCLP gave an optimal solution for the discretization. We now extend this result from piecewise linear to piecewise analytic costs. It turns out that the proof of this result is very similar to the corresponding one in [28]. Once established, this result will then be one of the keys to establishing strong duality results. We begin by defining the discretization used in [28].

Let $P = \{t_0, t_1, \ldots, t_m\}$ be a partition of $[0, T]$. Assume that $a(t)$ is piecewise linear and continuous and that $b(t)$ is piecewise constant, both with breakpoints in P. Let

$$u_i = \frac{t_{i-1} + t_i}{2},$$
$$\tau_i = \frac{t_i - t_{i-1}}{2}.$$

The variables in the discretization are $\hat{x}(t_{i-1}+)$, $\hat{x}(t_i-)$, $\hat{y}(t_i)$, $\hat{y}(u_i)$, $\hat{z}(t_{i-1}+)$, and $\hat{z}(t_i-)$, which, as the notation suggests, correspond in some way to the function values of x, y, and z at $t_{i-1}+$, t_i-, and u_i of a feasible solution for SCLP. We define the discretization of SCLP, called AP(P), as follows:

$$\begin{aligned}
\text{AP}(P)\text{: minimize} \quad & \sum_{i=1}^{m} \left(c(t_{i-1}+)^T \hat{x}(t_{i-1}+) + c(t_i-)^T \hat{x}(t_i-) \right) \\
\text{subject to} \quad & G\hat{x}(t_0+) + \hat{y}(u_1) = a(u_1), \\
& G\hat{x}(t_i-) + \hat{y}(t_i) - \hat{y}(u_i) = a(t_i) - a(u_i), \quad i = 1, \ldots, m, \\
& G\hat{x}(t_{i-1}+) + \hat{y}(u_i) - \hat{y}(t_{i-1}) = a(u_i) - a(t_{i-1}), \quad i = 2, \ldots, m, \\
& H\hat{x}(t_{i-1}+) + \hat{z}(t_{i-1}+) = \tau_i b(t_{i-1}+), \quad i = 1, \ldots, m, \\
& H\hat{x}(t_i-) + \hat{z}(t_i-) = \tau_i b(t_i-), \quad i = 1, \ldots, m, \\
& \hat{x}(t_{i-1}+), \hat{x}(t_i-), \hat{y}(t_i), \hat{y}(u_i), \hat{z}(t_{i-1}+), \hat{z}(t_i-) \geq 0, \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, \ldots, m,
\end{aligned}$$

or in matrix form,

$$\begin{aligned}
\text{AP}(P)\text{: minimize} \quad & \hat{c}^T \hat{\omega} \\
\text{subject to} \quad & A\hat{\omega} = \hat{b}, \\
& \hat{\omega} \geq 0,
\end{aligned}$$

where

$$\begin{aligned}
\hat{c}^T &= (c(t_0+)^T, 0^T, c(t_1-)^T, 0^T, c(t_1+)^T, \ldots, c(t_{m-1}+)^T, 0^T, c(t_m-)^T, 0^T), \\
\hat{b}^T &= (a(u_1)^T, \tau_1 b(t_0+)^T, (a(t_1) - a(u_1))^T, \tau_1 b(t_1-)^T, \ldots, (a(t_m) - a(u_m))^T, \\
&\quad \tau_m b(t_m-)^T), \\
\hat{\omega}^T &= (\hat{x}(t_0+)^T, \hat{y}(u_1)^T, \hat{z}(t_0+)^T, \hat{x}(t_1-)^T, \hat{y}(t_1)^T, \hat{z}(t_1-)^T, \ldots, \hat{x}(t_m-)^T, \hat{y}(t_m)^T, \\
&\quad \hat{z}(t_m-)^T),
\end{aligned}$$

where the 0's in $\hat{c}$ are of dimension $n_2 + n_3$ and

$$A = \begin{bmatrix}
G & I & 0 & 0 & 0 & 0 & \ldots & & & & & & \\
H & 0 & I & 0 & 0 & 0 & \ldots & & & & & & \\
0 & -I & 0 & G & I & 0 & \ldots & & & & & & \\
0 & 0 & 0 & H & 0 & I & \ldots & & & & & & \\
0 & 0 & 0 & 0 & -I & 0 & G & \ldots & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & H & \ldots & & & & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & & & & & \\
0 & & & & \ldots & & & & -I & 0 & G & I & 0 \\
0 & & & & \ldots & & & & 0 & 0 & H & 0 & I
\end{bmatrix}.$$

Note that A depends not on the particular partition chosen but only on its size. In fact, in [28], AP(P) was written in a slightly different but equivalent form where A did depend on the partition chosen. We have rewritten AP(P) as above for this reason.

Now AP(P) has the following (finite-dimensional) dual:

$$\begin{aligned} \text{AP}^*(P): \quad & \text{maximize} && \hat{b}^T\hat{\theta} \\ & \text{subject to} && \hat{\theta}^T A \le \hat{c}^T \end{aligned}$$

or, after making a few simplifications,

$$\begin{aligned} \text{AP}^*(P): \quad \text{maximize} \quad & \hat{\pi}(t_0+)^T a(t_0) + \sum_{i=1}^{m} (\hat{\pi}(t_{i-1}+) + \hat{\pi}(t_i-))^T (a(t_i) - a(u_i)) \\ & + \sum_{i=1}^{m} \tau_i (\hat{\eta}(t_{i-1}+) + \hat{\eta}(t_i-))^T b(t_i-) \\ \text{subject to} \quad & c(t_i-) - G^T\hat{\pi}(t_i-) - H^T\hat{\eta}(t_i-) \ge 0, \quad i = 1, \dots, m, \\ & c(t_{i-1}+) - G^T\hat{\pi}(t_{i-1}+) - H^T\hat{\eta}(t_{i-1}+) \ge 0, \quad i = 1, \dots, m, \\ & \hat{\eta}(t_i-), \hat{\eta}(t_{i-1}+) \le 0, \quad i = 1, \dots, m, \\ & \hat{\pi}(t_i-) - \hat{\pi}(t_{i-1}+) \ge 0, \quad i = 1, \dots, m, \\ & \hat{\pi}(t_i+) - \hat{\pi}(t_i-) \ge 0, \quad i = 1, \dots, m-1, \\ & \hat{\pi}(t_m-) \le 0. \end{aligned}$$

The relevant result about AP(P) established in [28] was the following.

THEOREM 6.1. *Suppose that $\omega(t)$ is feasible for* SCLP *with $x(t)$ piecewise constant with breakpoints in $P = \{t_0, t_1, \dots, t_m\}$. Then $\hat{\omega}$ defined by*

$$(16) \qquad \begin{cases} \hat{x}(t_{i-1}+) = \tau_i x(t_{i-1}+), \\ \hat{x}(t_i-) = \tau_i x(t_i-), \\ \hat{y}(t_i) = y(t_i), \\ \hat{y}(u_i) = y(u_i), \\ \hat{z}(t_{i-1}+) = \tau_i z(t_{i-1}+), \\ \hat{z}(t_i-) = \tau_i z(t_i-), \qquad i = 1, \dots, m, \end{cases}$$

is feasible for AP(P). *Moreover, if $c(t)$ is piecewise linear with breakpoints in P and $\omega(t)$ is optimal for* SCLP, *then $\hat{\omega}$ is optimal for* AP(P).

We now proceed to extend Theorem 6.1 to the case of $c(t)$ piecewise analytic.

THEOREM 6.2. *Suppose that $a(t)$ is piecewise linear (and continuous), $b(t)$ is piecewise constant, and $c(t)$ is piecewise analytic, each with breakpoints in the partition $P = \{t_0, t_1, \dots, t_m\}$. Let $\omega(t)$ be feasible for* SCLP *with $x(t)$ piecewise constant with breakpoints in P. Then $\hat{\omega}$, as given by* (16), *is feasible for* AP(P). *Moreover, if $\omega(t)$ is optimal for* SCLP, *then $\hat{\omega}$ is optimal for* AP(P).

Proof. The feasibility of $\hat{\omega}$ follows from Theorem 6.1. Suppose then that $\omega(t)$ is optimal for SCLP with $x(t)$ piecewise constant with breakpoints in P. Let $\hat{\omega}$ be given by (16). Suppose $\hat{\omega}$ is not optimal for AP(P). Then $\hat{\tilde{\omega}}$ exists, feasible for AP(P), with strictly improved objective function, i.e.,

$$\delta \equiv \hat{c}^T\hat{\tilde{\omega}} - \hat{c}^T\hat{\omega} < 0.$$

Following the constructions in [28] (see §4 in [28]), let

$$\tilde{x}(t) = \begin{cases} \dfrac{1}{\tau_i}\hat{\tilde{x}}(t_{i-1}+), & t \in [t_{i-1}, u_i),\ i = 1, \ldots, m, \\ \dfrac{1}{\tau_i}\hat{\tilde{x}}(t_i-), & t \in [u_i, t_i),\ i = 1, \ldots, m, \\ \dfrac{1}{\tau_m}\hat{\tilde{x}}(t_m-), & t = T. \end{cases}$$

Let $\tau = \min\{\, \tau_i : i = 1, \ldots, m \,\}$ and $\varepsilon \in [0, \tau]$. Set $\varepsilon_i = \varepsilon\tau_i/\tau$ and define

$$\bar{x}_\varepsilon(t) = \begin{cases} \tilde{x}(t), & t \in [t_{i-1}, t_{i-1} + \varepsilon_i) \cup [t_i - \varepsilon_i, t_i) \text{ for } i = 1, \ldots, m, \\ x(t) & \text{otherwise.} \end{cases}$$

Let $\bar{y}_\varepsilon(t)$ and $\bar{z}_\varepsilon(t)$ be given by the constraints of SCLP (i.e., so that $\bar{x}_\varepsilon(t)$, $\bar{y}_\varepsilon(t)$, and $\bar{z}_\varepsilon(t)$ satisfy (1) and (2) in place of $x(t)$, $y(t)$, and $z(t)$, respectively). By the same argument as in [28], the resulting $\bar{\omega}_\varepsilon(t)$ is feasible for SCLP.

We now claim that there exists $\varepsilon > 0$ such that $\bar{\omega}_\varepsilon(t)$ is an improved feasible solution for SCLP. For this purpose we use the standard notation $o(h^n)$ for $n \in \mathbb{N}$ to mean a function defined on an interval containing zero such that $\lim_{h \downarrow 0} o(h^n)/h^n = 0$. Now as $c(t)$ is analytic on a neighbourhood of $[t_{i-1}, t_i)$ for $i = 1, \ldots, m$, we have

$$c(t_{i-1} + \rho) = c(t_{i-1}+) + \rho\dot{c}(t_{i-1}+) + o(\rho)$$

for $\rho > 0$. Hence

$$\int_{t_{i-1}}^{t_{i-1}+\varepsilon_i} c(t)\,dt = \varepsilon_i \left(c(t_{i-1}+) + \frac{\varepsilon_i}{2}\dot{c}(t_{i-1}+) + o(\varepsilon_i) \right).$$

Similarly,

$$\int_{t_i-\varepsilon_i}^{t_i} c(t)\,dt = \varepsilon_i \left(c(t_i-) - \frac{\varepsilon_i}{2}\dot{c}(t_i-) + o(\varepsilon_i) \right).$$

Continuing in exactly the same way as in Lemma 4.3 in [28] we establish that

$$\int_0^T c(t)^T \bar{x}_\varepsilon(t)\,dt - \int_0^T c(t)^T x(t)\,dt = \frac{\varepsilon}{\tau}\left(\delta - \frac{\varepsilon\alpha}{\tau} + o(\varepsilon) \right),$$

where

$$\alpha(\omega) = \sum_{i=1}^{m} \frac{\tau_i^2}{2}(x(t_i-) - x(t_{i-1}+))^T \dot{c}(t_i-).$$

Hence

$$\begin{aligned} \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}\left(\int_0^T c(t)^T \bar{x}_\varepsilon(t)\,dt - \int_0^T c(t)^T x(t)\,dt \right) &= \lim_{\varepsilon \downarrow 0} \frac{1}{\tau}\left(\delta - \frac{\varepsilon\alpha}{\tau} + o(\varepsilon) \right) \\ &= \frac{\delta}{\tau} \\ &< 0, \end{aligned}$$

thus contradicting the optimality of $\omega(t)$. □

In §3 we encountered and discussed a notion of degeneracy. This was a problem that arose when trying to calculate complementary slack variables. As with the corresponding notion in FLP, when degeneracy is present, it is not possible to calculate a set of complementary slack variables uniquely. In the conclusion to §3 we noted that there were essentially two types of degeneracy for SCLP. One of these was when there were not enough nonzero variables at any particular time to form a basis matrix for K. We will see that to establish strong duality results, degeneracy is a significant problem. This is not surprising because in the absence of degeneracy in finite-dimensional linear programming, the proof of the strong duality result becomes relatively straightforward. It turns out that only the one type of degeneracy for SCLP mentioned above plays an important part in the strong duality theory for SCLP. As this is the case and as we have not defined degeneracy formally before, we now present the following definition. It will be seen to be very similar to the corresponding one for FLP. As our strong duality results to follow only cover piecewise analytic problem data, we only give a definition of degeneracy for such data.

DEFINITION 4. *Suppose that $a(t)$ and $b(t)$ are piecewise analytic with breakpoints in the partition $P = \{t_0, t_1, \ldots, t_m\}$.*

1. *Let B be any basis matrix of K. Let*

$$\rho(t) = B^{-1}\begin{bmatrix} \dot{a}(t) \\ b(t) \end{bmatrix}.$$

We say that B is a degenerate basis matrix *for* SCLP *if $\rho_i(t) = 0$ for all $t \in [t_{j-1}, t_j)$ for some i and j. Otherwise we say that B is a* nondegenerate basis matrix *for* SCLP.

2. *We say that* SCLP *is* degenerate *if there exists a basis matrix B for K such that B is a degenerate basis matrix for* SCLP. *Otherwise we say that* SCLP *is* nondegenerate.

We now present a simple lemma based on this definition, a result which has an equivalent counterpart in finite-dimensional linear programming.

LEMMA 6.3. *Suppose that* SCLP *is nondegenerate and that $\omega(t)$ is a basic feasible solution for* SCLP. *Suppose there exists $(\alpha, \beta) \subset [0, T]$ such that* $\operatorname{supp}(\omega(t))$ *is constant on (α, β). Then $|\operatorname{supp}(\omega(t))| = n_2 + n_3$ on (α, β), and so there exists a unique basis matrix B of K such that the nonzero variables of $\omega(t)$ are precisely $\omega_B(t)$ on (α, β).*

Proof. Let B be any basis matrix of K that contains the columns of K corresponding to $\operatorname{supp}(\omega(t))$ on (α, β). This exists as $\omega(t)$ is a basic feasible solution. Let

$$\bar{\omega}(t) = \begin{bmatrix} x(t) \\ \dot{y}(t) \\ z(t) \end{bmatrix};$$

then as $\omega(t)$ is feasible for SCLP, we must have

$$\bar{\omega}_B(t) = B^{-1}\begin{bmatrix} \dot{a}(t) \\ b(t) \end{bmatrix}$$

for $t \in (\alpha, \beta)$. However, by the nondegeneracy assumption we have that each component of $\bar{\omega}_B(t)$ is nonzero for all $t \in (\alpha, \beta)$. Hence $\operatorname{supp}(\omega(t))$ has $n_2 + n_3$ elements on (α, β), one for each column of B. □

Our final result in this section is quite unrelated to the previous ones and comes from Pullan [29]. The result shows that the number of zeros in a linear combination of analytic functions is dependent only on the functions involved and not the particular scalars chosen. This is useful in establishing strong duality results for the following reason. As mentioned above, degeneracy will play an important part in the analysis to follow. In fact, as with FLP, we will see that it is relatively easy to establish a strong duality result for nondegenerate problems.

To establish a strong duality result in general we will then approximate a degenerate problem by a sequence of nondegenerate ones. We will then wish to use Lemma 5.3 to produce an optimal solution for SCLP*. To do this we will have to produce a uniform bound on the number of breakpoints in the optimal solutions for the approximating dual problems. The required uniform bound is given by the result below. For this result we use the notation $|S|$ to denote the cardinality of a set S.

LEMMA 6.4. *Let $f : [a, b] \to \mathbb{R}^n$ be a function analytic on a neighbourhood of $[a, b]$. Then there exists $M(f)$ $(< \infty)$ such that for all $\lambda \in \mathbb{R}^n$ if*

$$S(\lambda, f) = \{ t \in [a, b] : \lambda^T f(t) = 0 \},$$

then either $S(\lambda, f) = [a, b]$ or $|S(\lambda, f)| \leq M(f)$.

6.2. Analytic costs. In this section we will be extending the strong duality result in Pullan [28, Thm. 1.3], to SCLP with piecewise analytic costs. We begin by restricting our attention to the case where the problem data contain no breakpoints. We thus introduce the following assumption that we will assume holds for the problem data of SCLP throughout the rest of this section.

Assumption 6.1. The costs $c(t)$ are analytic on a neighbourhood of $[0, T]$, $a(t)$ is linear on $[0, T]$, $b(t)$ is constant on $[0, T]$, and the feasible region for SCLP is nonempty and bounded.

We now proceed to establish a strong duality result under this assumption. This is done in two stages, first for nondegenerate problems and then for degenerate ones (see Definition 4).

We use the following notation. Given a cost $c(t)$ we let

$$\tilde{c}(t) = \begin{bmatrix} c(t) \\ 0 \\ 0 \end{bmatrix},$$

where the 0's are of dimensions n_2 and n_3, respectively. We now define the following sets. Let

$$Q = \left\{ (B^{-1})^T \tilde{\dot{c}}_B : B \text{ is a basis matrix for } K \right\},$$
$$R = \{ \dot{\psi} : \psi = c - G^T \pi - H^T \eta \text{ for some } \theta^T = (\pi^T, \eta^T) \in Q \},$$
$$S = \{ \rho_j : \rho \in Q \cup R \}.$$

Note that S consists of a finite set of analytic functions. Now each of these functions is identically zero or contains a finite number of zeros on $[0, T]$. Let N be the total number of zeros of functions in S that are not identically zero on $[0, T]$. Let $\omega(t)$ be a piecewise constant optimal solution for SCLP with, say, M breakpoints. This exists by Theorem 2.4. We now let $P = \{t_0, t_1, \ldots, t_m\}$ be the partition of $[0, T]$ with at most $M + N + 2$ points that contains all the breakpoints of $\omega(t)$ and all the zeros of functions in S that are not identically zero, as well as the points 0 and T. We now present a lemma showing that such a partition can be used to construct an optimal solution for SCLP* fairly naturally if SCLP is nondegenerate. This lemma makes use of the standard L_∞ matrix norm, which is defined by

$$\|B\|_\infty = \max \left\{ \sum_{j=1}^{n} |b_{i,j}| : i = 1, 2, \ldots, n \right\}$$

for an $n \times n$ matrix B.

LEMMA 6.5. *Suppose that* SCLP *is nondegenerate. Then there exists a piecewise analytic optimal solution θ for* SCLP* *with breakpoints in P (i.e., having at most $M + N + 1$ breakpoints). Moreover, θ satisfies the following:*

1. *For each* $i = 1, \ldots, m$ *there exists a basis matrix* $B^{(i)}$ *for* K *such that*

$$\theta(t) = \left(B^{(i)^{-1}}\right)^T \bar{c}_{B^{(i)}}(t)$$

for $t \in [t_{i-1}, t_i)$, *where*

$$\bar{c}(t) = \begin{bmatrix} c(t) \\ \pi(t_i-) \\ 0 \end{bmatrix}$$

for $t \in [t_{i-1}, t_i)$.

2. $\|\theta\|_\infty \leq \max\{\, \|\hat{B}^{-1}\|_\infty : \hat{B}$ *is a basis matrix for* $\mathrm{AP}(P)\,\}\|c\|_\infty$.

Proof. By Theorem 6.2 we have $\hat{\omega}$, given by (16), optimal for AP(P). Hence, by the strong duality theorem for finite-dimensional linear programming, there exists a basis matrix $\hat{B}$ for AP(P) such that if

$$\hat{\theta} = (\hat{B}^{-1})^T \begin{bmatrix} \tilde{c}(t_0+) \\ \tilde{c}(t_1-) \\ \vdots \\ \tilde{c}(t_m-) \end{bmatrix},$$

then $\hat{\theta}$ is optimal for AP*(P) and complementary slack with $\hat{\omega}$. Note that

$$\|\hat{\theta}\|_\infty \leq \max\{\, \|\hat{B}^{-1}\|_\infty : \hat{B} \text{ is a basis matrix for } \mathrm{AP}(P)\,\}\|c\|_\infty.$$

We will now construct $\theta(t)$ optimal for SCLP* such that

$$\begin{aligned} \theta(t_i+) &= \hat{\theta}(t_i+), \quad i = 0, \ldots, m-1, \\ \theta(t_i-) &= \hat{\theta}(t_i-), \quad i = 1, \ldots, m, \end{aligned}$$

and such that each component of $\theta(t)$ is monotonic on $[t_{i-1}, t_i)$ for $i = 1, \ldots, m$. Such a $\theta(t)$ will then satisfy (2) in the statement of the lemma.

Fix i, and consider the interval (t_{i-1}, t_i). Let B be the basis matrix of K that contains the columns of K corresponding to $\mathrm{supp}(\omega(t))$ on (t_{i-1}, t_i). The existence and uniqueness of this matrix B is given by Lemma 6.3. Let $\bar{c}(t)$ be as given in the statement of the theorem. Then it is clear that $\hat{\theta}$ complementary slack with $\hat{\omega}$ implies

$$\begin{aligned} \hat{\theta}(t_{i-1}+) &= (B^{-1})^T \bar{c}_B(t_{i-1}+), \\ \hat{\theta}(t_i-) &= (B^{-1})^T \bar{c}_B(t_i-). \end{aligned}$$

Now $\dot{\bar{c}}(t) = \dot{\tilde{c}}(t)$ for all t. Hence by the properties of the partition P, for any j, $[(B^{-1})^T \dot{\bar{c}}_B(t)]_j$ contains no isolated zeros in (t_{i-1}, t_i) and so $[(B^{-1})^T \bar{c}_B(t)]_j$ is monotonic on (t_{i-1}, t_i) for each j. Define

$$\theta(t) = (B^{-1})^T \bar{c}_B(t)$$

for $t \in [t_{i-1}, t_i)$. We then have $\theta_j(t)$ monotonic on $[t_{i-1}, t_i)$ for each j. By a similar argument we also have $\psi_j(t) = (c(t) - G^T\pi(t) - H^T\eta(t))_j$ monotonic on $[t_{i-1}, t_i)$ for each j. Hence, by the feasibility of $\hat{\theta}$ for AP*(P), we have

$$\begin{aligned} &c(t) - G^T\pi(t) - H^T\eta(t) \geq 0, \\ &\eta(t) \leq 0, \quad t \in [t_{i-1}, t_i), \\ &\pi(t) \text{ monotonic increasing and right continuous on } [t_{i-1}, t_i). \end{aligned}$$

If this is done for each i and we set $\theta(T) = 0$, we then obtain a feasible solution for SCLP* (note that $\pi(T-) \leq 0$ as $\hat{\pi}(t_m-) \leq 0$). Moreover, by the construction, we have $\theta(t)$ complementary slack with $\omega(t)$ by Theorem 3.5. Hence, by the complementary slackness theorem, Theorem 3.4, $\theta(t)$ is optimal for SCLP*. Finally, by construction, $\theta(t)$ satisfies the remaining requirements of the lemma. □

We now have all the necessary ingredients to establish a strong duality result for SCLP under Assumption 6.1. This is done by approximating SCLP by a sequence of nondegenerate problems and then using Lemma 5.3 to guarantee an optimal solution for SCLP*. The required uniform bounds for the application of this lemma are provided by the result for the existence of a piecewise constant optimal solution for SCLP, Theorem 2.4, and Lemma 6.4. There is one slight problem in establishing this result, however—namely, the need to construct suitable approximating SCLP problems with bounded feasible regions. This is done by adding upper bound constraints to the problem so that any right-hand sides will give a bounded feasible region. The proof is then made complete by showing that the resulting optimal dual variables for the extra constraints are zero by complementary slackness.

THEOREM 6.6 (strong duality). *Suppose that $c(t)$ is piecewise analytic, $a(t)$ is piecewise linear and continuous, $b(t)$ is piecewise constant on $[0, T]$, and the feasible region for* SCLP *is nonempty and bounded. Then V[SCLP] $= V$[SCLP*] and there exist an optimal solution for* SCLP, *with $x(t)$ piecewise constant, and a piecewise analytic optimal solution for* SCLP*.

Proof. It is required only to prove that SCLP* has a piecewise analytic optimal solution. The other statements are given by previous results (Theorem 2.4 and Theorem 4.2; however, the fact that V[SCLP] $= V$[SCLP*] in this case could also be shown directly in the proof to follow). We restrict ourselves to the case where $a(t)$ is linear, $b(t)$ is constant, and $c(t)$ is analytic on a neighbourhood $[0, T]$. The more general case can be dealt with by repeating the argument below a finite number of times.

Assume for the moment that SCLP contains upper bound constraints on $x(t)$, i.e., that H is of the form

$$H = \begin{bmatrix} \bar{H} \\ I \end{bmatrix}. \tag{17}$$

Let $B^{(1)}, \ldots, B^{(L)}$ be all the possible basis matrices for K. Let

$$Q_{i,j} = \{ B^{(j)}\zeta : \zeta \in \mathbb{R}^{n_2+n_3},\ \zeta_i = 0 \}.$$

Then as $B^{(j)}$ has full rank, $Q_{i,j}$ is a $(n_2 + n_3 - 1)$-dimensional subspace of $\mathbb{R}^{n_2+n_3}$ (i.e., a hyperplane). Now $\{ Q_{i,j} : i = 1, \ldots, n_2 + n_3,\ j = 1, \ldots, L \}$ is a finite set of hyperplanes in $\mathbb{R}^{n_2+n_3}$. It is now not difficult to show that for any $\gamma \in \mathbb{R}^{n_2+n_3}$ there exists $\{\gamma^{(n)}\}_{n=1}^{\infty}$ such that $\gamma^{(n)} \downarrow \gamma$ and $\gamma^{(n)} \notin Q_{i,j}$ for each i and j and for all n. In particular there exist $r^{(n)} \downarrow \dot{a}$ and $s^{(n)} \downarrow b$ such that

$$\begin{bmatrix} r^{(n)} \\ s^{(n)} \end{bmatrix} \notin Q_{i,j}$$

for each i and j and for all n. Let

$$a^{(n)}(t) = a(0) + tr^{(n)},$$
$$b^{(n)}(t) = s^{(n)}.$$

Then $\{b^{(n)}\}_{n=1}^{\infty}$ is a set of constant functions and $\{a^{(n)}\}_{n=1}^{\infty}$ a set of linear functions such that $a^{(n)} \downarrow a$ and $b^{(n)} \downarrow b$ uniformly. Let SCLP$_n$ be the problem SCLP with a and b replaced

by $a^{(n)}$ and $b^{(n)}$, respectively. Now as H is of the form given by (17), the feasible region for SCLP_n is both nonempty and bounded. Moreover, by construction, SCLP_n is nondegenerate.

Now by Lemma 6.5 there exists an optimal solution, $\theta^{(n)}(t)$, for SCLP_n^* with at most $M_n + N + 1$ breakpoints, where M_n is the number of breakpoints of a piecewise constant optimal solution for SCLP_n and N is some constant depending only on $c(t)$ and $B^{(1)}, \ldots, B^{(L)}$ and is thus independent of n. We now wish to apply Lemma 5.3 to obtain a piecewise analytic optimal solution for SCLP*. To apply this lemma it will be sufficient to show that there exists Λ, independent of n, such that M_n can be chosen so that $M_n < \Lambda$. This is because

$$\max\{\, \|\hat{B}^{-1}\|_\infty : \hat{B} \text{ is a basis matrix for AP}(P)\,\}$$

depends only on the size of the partition P and not the actual points in the partition (see comment on page 948). Let $x^{(n,i)}$, $i = 1, \ldots, L$ be the possible values of a basic feasible solution for SCLP_n as given by Lemma 2.3. Now by Lemma 6.4 there exists M, independent of n, such that if

$$S_{i,j,n} = \{\, t \in [0, T] : c(t)^T x^{(n,i)} = c(t)^T x^{(n,j)} \,\}$$

for $i \neq j$, then for each $i \neq j$ either $|S_{i,j,n}| \leq M$ or $S_{i,j,n} = [0, T]$. Using the result for the existence of a piecewise constant optimal solution (Theorem 2.4), we may now deduce that an optimal solution $\omega^{(n)}(t)$ exists for SCLP_n for each n with at most $(M+1)L$ breakpoints. This gives the required uniform upper bound on the number of breakpoints for an optimal solution for SCLP_n and establishes the result for H of the form given by (17).

Suppose that H is not of the form given by (17). Let $R > 0$ be such that $\|x\|_\infty \leq R$ for all $x \in F(\text{SCLP})$. Let $\text{SCLP}(R)$ be the problem SCLP with the extra constraints

$$\begin{aligned} x(t) + v(t) &= 2Re, \\ v(t) &\geq 0, \end{aligned}$$

where e is the vector of all ones. By the definition of R, $F(\text{SCLP}) = F(\text{SCLP}(R))$ and $v(t) \geq Re > 0$ for all $t \in [0, T]$ and for any feasible solution for $\text{SCLP}(R)$. Let $\sigma(t)$ denote the dual variables in $\text{SCLP}(R)^*$ corresponding to $v(t)$, and $\eta(t)$ the dual variables in $\text{SCLP}(R)^*$ corresponding to the original constraints $Hx(t) + z(t) = b(t)$ in SCLP, i.e., so that the constraint $\psi(t) \geq 0$ can be written as $c(t) - G^T\pi(t) - H^T\eta(t) - \sigma(t) \geq 0$. By the argument above, there exists $\theta(t)^T = (\pi(t)^T, \eta(t)^T, \sigma(t)^T)$ piecewise analytic on $[0, T]$ and optimal for $\text{SCLP}(R)^*$. Moreover, $\theta(t)$ is complementary slack with some optimal solution $\omega(t)^T = (x(t)^T, y(t)^T, z(t)^T, v(t)^T)$ for $\text{SCLP}(R)$ by the complementary slackness theorem, Theorem 3.4. But $v(t) > 0$ and so we must have $\sigma(t) = 0$ a.e. on $[0, T]$, and hence everywhere as $\sigma(t)$ is piecewise analytic on $[0, T]$. Thus $(\pi(t)^T, \eta(t)^T)$ is an optimal solution for the original SCLP*. This establishes the result. □

It is worth noting as an aside that Lemma 5.3 also ensures that the optimal $\theta(t)$ for SCLP* derived in the proof of the above theorem has at most $(M+1)L + N + 1$ breakpoints, where M is given in the proof of the theorem.

6.3. Analytic right-hand sides. In this section we will be extending the strong duality result in the previous section to SCLP with piecewise analytic costs and right-hand sides. In the result from the previous section, the bulk of the proof was concerned with the case when the problem data had no breakpoints and there were upper bound constraints on $x(t)$. We thus introduce the following assumption that we will assume holds for the problem data of SCLP throughout the rest of this section.

Assumption 6.2. The costs $c(t)$ and the right-hand sides $a(t)$ and $b(t)$ are analytic on a neighbourhood of $[0, T]$, H is of the form

$$H = \begin{bmatrix} \bar{H} \\ I \end{bmatrix},$$

and the feasible region for SCLP is nonempty.

We now concentrate on proving strong duality under Assumption 6.2. The proof involves approximating $a(t)$ and $b(t)$ by sequences of piecewise linear and piecewise constant functions, respectively, and using the ideas from the previous sections. This same general technique was used in Pullan [29] to prove Theorem 2.5 starting from Theorem 2.4. We will thus use some of the ideas from [29]. The first of these is the concept of a change of basis. This concept allows us to distinguish between two types of breakpoints in a basic feasible solution for SCLP when the problem data have breakpoints. One type of breakpoint results from a breakpoint in the problem data. The other results from a change of basis which we now define.

DEFINITION 5. *Let $\omega(t)$ be a piecewise analytic basic feasible solution for* SCLP *and $x^{(1)}(t), \ldots, x^{(L)}(t)$ be given by Lemma 2.3. By a* change of basis *we mean a time $s \in (0, T)$ such that for some $\varepsilon > 0$, $x(t) = x^{(i)}(t)$ for $t \in (s - \varepsilon, s)$ and $x(t) = x^{(j)}(t)$ for $t \in (s, s + \varepsilon)$ for some i and j such that $x^{(i)}(t) \not\equiv x^{(j)}(t)$ on $(s - \varepsilon, s + \varepsilon)$.*

As with [29], we note that if a and b are analytic on a neighbourhood of $[0, T]$, then a change of basis is identical to a breakpoint. However, if a and b are piecewise analytic on $[0, T]$, then in general, the set of changes of basis for a basic feasible solution will be a subset of the set of breakpoints for that solution. This is because some of the breakpoints may be caused by breakpoints in the problem data and not by a change of basis.

We now begin the development of a strong duality result under Assumption 6.2. As with proving Theorem 2.5 in [29] we do this by taking sequences $\{a^{(n)}\}_{n=1}^{\infty}$ and $\{b^{(n)}\}_{n=1}^{\infty}$ of piecewise linear (and continuous) and piecewise constant functions, respectively, so that $a^{(n)} \geq a$, $b^{(n)} \geq b$ with $a^{(n)} \to a$ and $b^{(n)} \to b$ uniformly. The sequences are constructed so that SCLP_n has an optimal basic solution for SCLP_n with a bounded number of changes of basis independent of n, where SCLP_n is the problem SCLP with a and b replaced by $a^{(n)}$ and $b^{(n)}$, respectively. This construction is given in Lemma 4.2 in [29]. To complete the strong duality result under Assumption 6.2 we then construct a corresponding optimal solution for SCLP_n^* (the dual of SCLP_n) and use the bound on the number of changes of basis to apply Lemma 5.3. As might be expected, however, to construct the corresponding optimal solution for SCLP_n^* we need to ensure that SCLP_n is also nondegenerate.

We begin the analysis by establishing the existence of $a^{(n)}$ and $b^{(n)}$ with the required properties.

LEMMA 6.7. *Suppose that Assumption* 6.2 *holds. Then there exist $M \in \mathbb{N}$, a sequence $\{a^{(n)}\}_{n=1}^{\infty}$ of piecewise linear and continuous functions, and a sequence $\{b^{(n)}\}_{n=1}^{\infty}$ of piecewise constant functions with $a^{(n)}(t) \geq a(t)$ and $b^{(n)}(t) \geq b(t)$ on $[0, T]$ for each n, $a^{(n)} \to a$ and $b^{(n)} \to b$ uniformly on $[0, T]$, and such that SCLP_n is nondegenerate and has a piecewise constant basic feasible optimal solution with at most M changes of basis, where SCLP_n is the problem* SCLP *with right-hand sides $a^{(n)}$ and $b^{(n)}$.*

Proof. This lemma, without the nondegeneracy requirement, is precisely Lemma 4.2 in Pullan [29]. Now in the proof of Theorem 6.6 we saw that it was not difficult to ensure that an approximating problem to a given SCLP problem was nondegenerate. This is because any right-hand side can be approximated arbitrarily closely by a right-hand side for a nondegenerate problem. Thus it would be expected that the extra requirement for SCLP_n to be nondegenerate causes no new difficulties. In fact the proof of the current result is virtually identical to the proof of Lemma 4.2 in [29]. As this is the case, and as the proof of the Lemma 4.2 in [29] is also rather long and detailed, we omit the discussion. □

This is actually the bulk of the proof of a strong duality result between SCLP and SCLP* under Assumption 6.2. What remains to do now is prove a result similar to Lemma 6.5 in the previous section that generates an optimal solution, $\theta^{(n)}(t)$, for SCLP_n^* (the dual of SCLP_n as given by the previous lemma) that is uniformly bounded with respect to n. The existence of an optimal $\theta^{(n)}(t)$ for SCLP_n^* follows in a similar way to Lemma 6.5; however, this $\theta^{(n)}(t)$ may have breakpoints at each of the breakpoints of $a^{(n)}(t)$ and $b^{(n)}(t)$. The number of breakpoints in $\theta^{(n)}(t)$ could thus become unbounded and so we would not be able to apply Lemma 5.3 to guarantee an optimal solution for SCLP*. This is where the uniform bound on the number of changes of basis of an optimal solution for SCLP_n is used.

We now set up the same notation to that used in the previous section. Given a cost $c(t)$ we let

$$\tilde{c}(t) = \begin{bmatrix} c(t) \\ 0 \\ 0 \end{bmatrix},$$

where the 0's are of dimensions n_2 and n_3, respectively. We also define the following sets. Let

$$\begin{aligned} Q &= \left\{ (B^{-1})^T \dot{\tilde{c}}_B : B \text{ is a basis matrix for } K \right\}, \\ R &= \{ \dot{\psi} : \psi = c - G^T\pi - H^T\eta \text{ for some } \theta^T = (\pi^T, \eta^T) \in Q \}, \\ S &= \{ \rho_j : \rho \in Q \cup R \}. \end{aligned}$$

As with Lemma 6.5, let N be the total number of zeros of functions in S that are not identically zero on $[0, T]$. Let $\omega^{(n)}(t)$ be a piecewise constant optimal solution for SCLP_n with M changes of basis, where M is given by the previous lemma. We now fix n and define two partitions of $[0, T]$ as follows. Strictly speaking, these partitions depend on n; however, we drop this dependence in the notation for simplicity. Let

$$P = \{t_0, t_1, \ldots, t_m\}$$

be the partition of $[0, T]$ that contains all the changes of basis for $\omega^{(n)}(t)$ and all the zeros of functions in S that are not identically zero. Our next partition of $[0, T]$ is

$$\begin{aligned} \Omega &= \{r_0, r_1, \ldots, r_q\} \\ &= \{s_1^{(0)}, s_1^{(1)}, \ldots, s_{p_1}^{(1)}, s_1^{(2)}, \ldots, s_{p_2}^{(2)}, \ldots, s_{p_m}^{(m)}\}, \end{aligned}$$

which contains P, written in this case as

$$P = \{s_1^{(0)}, s_{p_1}^{(1)}, \ldots, s_{p_m}^{(m)}\},$$

and all the breakpoints of $a^{(n)}(t)$ and $b^{(n)}(t)$. For convenience we set $p_0 = 1$ and $s_0^{(i)} = s_{p_{i-1}}^{(i-1)}$ for $i = 1, \ldots, m$. We now present the result guaranteeing an optimal solution $\theta^{(n)}(t)$ for SCLP_n^* which is uniformly bounded with respect to n.

LEMMA 6.8. *Let* SCLP_n *be given by the previous lemma. Then there exists a piecewise analytic optimal solution* $\theta^{(n)}$ *for* SCLP_n^* *with breakpoints in* P *(i.e., having at most* $M+N+1$ *breakpoints). Moreover,* $\theta^{(n)}$ *satisfies the following:*

1. *For each* $i = 1, \ldots, m$ *there exists a basis matrix* $B^{(i)}$ *for* K *such that*

$$\theta^{(n)}(t) = \left(B^{(i)^{-1}}\right)^T \bar{c}_{B^{(i)}}(t)$$

for $t \in [t_{i-1}, t_i)$, *where*

$$\bar{c}(t) = \begin{bmatrix} c(t) \\ \pi(t_i-) \\ 0 \end{bmatrix}$$

for $t \in [t_{i-1}, t_i)$.

2. $\|\theta^{(n)}\|_\infty \leq \max\{\, \|\hat{B}^{-1}\|_\infty : \hat{B}$ *is a basis matrix for* $\mathrm{AP}(P)\,\}\|c\|_\infty$.

Proof. Using arguments similar to those for the corresponding result in the previous section (Lemma 6.5), we see that there exists an optimal solution $\bar{\omega}$ for AP(Ω) with a corresponding basis matrix $\bar{B}$ for AP(Ω) such that if

$$\bar{\theta} = (\bar{B}^{-1})^T \begin{bmatrix} \tilde{c}(r_0+) \\ \tilde{c}(r_1-) \\ \vdots \\ \tilde{c}(r_q-) \end{bmatrix}, \tag{18}$$

then $\bar{\theta}$ is optimal for AP*(Ω). Moreover, an optimal solution, $\theta^{(n)}(t)$, which is complementary slack with $\omega^{(n)}(t)$, can be defined from $\bar{\theta}$ in the following way. Fix i and consider (t_{i-1}, t_i). As $\mathrm{supp}(\omega^n(t))$ is constant on (t_{i-1}, t_i), there exists a unique basis matrix B of K, given by Lemma 6.3, corresponding to the support of $\omega^{(n)}(t)$ on (t_{i-1}, t_i). The optimal $\theta^{(n)}(t)$ is then defined by

$$\theta^{(n)}(t) = (B^{-1})^T \bar{c}_B(t)$$

for $t \in [t_{i-1}, t_i)$, where

$$\bar{c}(t) = \begin{bmatrix} c(t) \\ \pi(t_i-) \\ 0 \end{bmatrix}$$

for $t \in [t_{i-1}, t_i)$.

Now by the construction of the partition P, the components of $\theta^{(n)}(t)$ are monotonic on $[t_{i-1}, t_i)$. We currently have $\theta^{(n)}$ constructed from (18). This could theoretically produce $\theta^{(n)}$ with a breakpoint at every point in Ω. We now show that this is not possible. In particular, let

$$\hat{\theta} = \begin{bmatrix} \bar{\theta}(t_0+) \\ \bar{\theta}(t_1-) \\ \vdots \\ \bar{\theta}(t_m-) \end{bmatrix}.$$

We will now show that in fact

$$\hat{\theta} = (\hat{B}^{-1})^T \begin{bmatrix} \tilde{c}(t_0+) \\ \tilde{c}(t_1-) \\ \vdots \\ \tilde{c}(t_m-) \end{bmatrix} \tag{19}$$

for some basis matrix $\hat{B}$ for AP(P). This will give us the required bound on the size of $\theta^{(n)}$ stated in the lemma and establish the result. We note that a basis matrix $\hat{B}$ must contain $2m(n_2 + n_3)$ columns of the constraint matrix $\hat{A}$ for AP(P). (Recall that m is the size of the partition P.)

The basic idea of the proof is that as $\text{supp}(\omega^{(n)}(t))$ is constant on (t_{i-1}, t_i) for each i, the essential information for $\bar{B}$ can be compressed into a smaller matrix $\hat{B}$. This idea is quite simple in concept but difficult to notate due to the size of the constraint matrices involved. Recall that the constraint matrix for AP(Ω) is given by

$$\bar{A} = \begin{bmatrix} G & I & 0 & 0 & 0 & 0 & \dots & & & & & \\ H & 0 & I & 0 & 0 & 0 & \dots & & & & & \\ 0 & -I & 0 & G & I & 0 & \dots & & & & & \\ 0 & 0 & 0 & H & 0 & I & \dots & & & & & \\ 0 & 0 & 0 & 0 & -I & 0 & G & \dots & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & H & \dots & & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & & & & \\ 0 & & & & \dots & & & & -I & 0 & G & I & 0 \\ 0 & & & & \dots & & & & 0 & 0 & H & 0 & I \end{bmatrix},$$

i.e., a matrix with a repeating block structure with $2q(n_1 + n_2 + n_3)$ columns. (Recall that q is the size of the partition Ω.) Similarly, the constraint matrix, $\hat{A}$, for AP(P) will be of the same form but with $2m(n_1 + n_2 + n_3)$ columns.

We begin by making a simple observation. Consider the matrix

$$\Lambda = \begin{bmatrix} G & I & 0 & 0 & 0 \\ H & 0 & I & 0 & 0 \\ 0 & -I & 0 & G & 0 \\ 0 & 0 & 0 & H & I \end{bmatrix},$$

made up of five blocks of columns which we call block 1 to block 5 for convenience. Suppose that Π is a matrix of columns of Λ with the following property: column i of block 1 is in Π if and only if column i of block 4 is in Π, and column i of block 3 is in Π if and only if column i of block 5 is in Π. In other words, the columns of the blocks containing G and H are the same, and the columns of the blocks containing a single I (of dimension n_3) are the same. We claim that for such a Π to have full column rank, it must contain no more than $n_2 + n_3$ columns corresponding to the first three blocks. Suppose that Π contains more than $n_2 + n_3$ columns from the first three blocks. Then there exists $\alpha^T = (u^T, v^T, w^T) \in \mathbb{R}^{n_1+n_2+n_3}$ with $\alpha \neq 0$ and with $i \in \text{supp}(\alpha)$ only if Λ_i is a column in Π, such that

$$K\alpha = \begin{bmatrix} G & I & 0 \\ H & 0 & I \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = 0.$$

Define

$$\beta^T = (u^T, v^T, w^T, -u^T, -w^T).$$

Then $\beta \neq 0$, and by the definition of Π, we have $i \in \text{supp}(\beta)$ only if Λ_i is a column in Π. Moreover, $\Lambda\beta = 0$. Thus Π does not have full column rank. Hence, as claimed, if Π has full column rank, then it must contain no more $n_2 + n_3$ columns corresponding to the first three blocks.

With this observation in mind we now return to the basis matrix $\bar{B}$ of the constraint matrix $\bar{A}$ for AP(Ω). Fix i and consider (t_{i-1}, t_i). Consider the columns of $\bar{A}$ corresponding to the variables of $\bar{\omega}$ relating to the times in $\Omega \cap [t_{i-1}, t_i]$ (except for $\bar{y}(s_0^{(i)})$), i.e., corresponding to the variables $\bar{x}(s_0^{(i)}+)$, $\bar{y}((s_0^{(i)} + s_1^{(i)})/2)$, $\bar{z}(s_0^{(i)}+)$, $\bar{x}(s_1^{(i)}-)$, $\bar{y}(s_1^{(i)})$, $\bar{z}(s_1^{(i)}-)$, $\dots$, $\bar{x}(s_{p_i-1}^{(i)}+)$,

$\bar{y}((s^{(i)}_{p_i-1} + s^{(i)}_{p_i})/2)$, $\bar{z}(s^{(i)}_{p_i-1}+)$, $\bar{x}(s^{(i)}_{p_i}-)$, $\bar{y}(s^{(i)}_{p_i})$, and $\bar{z}(s^{(i)}_{p_i}-)$. This can be written as

$$\bar{A}^{(i)} = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & \dots & & 0 & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & & \vdots & \vdots & \vdots & \vdots & \vdots \\
G & I & 0 & 0 & 0 & 0 & \dots & & & & & & \\
H & 0 & I & 0 & 0 & 0 & \dots & & & & & & \\
0 & -I & 0 & G & I & 0 & \dots & & & & & & \\
0 & 0 & 0 & H & 0 & I & \dots & & & & & & \\
0 & 0 & 0 & 0 & -I & 0 & G & \dots & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & H & \dots & & & & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & & & & & \\
0 & & & & \dots & & & & -I & 0 & G & I & 0 \\
0 & & & & \dots & & & & 0 & 0 & H & 0 & I \\
0 & & & & \dots & & & & 0 & 0 & 0 & -I & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & \dots & & 0 & 0 & 0 & 0 & 0
\end{bmatrix}.$$

In this case we refer to each group of $n_1 + n_2 + n_3$ columns as a block. Thus $\bar{A}^{(i)}$ contains $2p_i$ blocks. Now as $\text{supp}(\omega^{(n)}(t))$ is constant on (t_{i-1}, t_i) and has $n_2 + n_3$ elements (as SCLP_n is nondegenerate), $\bar{B}$ contains at least $n_2 + n_3$ columns of $\bar{A}^{(i)}$ from each block, except possibly the last one. (The last block is excluded as we may have $y_j(t_i) = 0$ for some j but $y_j(t) > 0$ for $t \in (t_{i-1}, t_i)$.) Moreover, the columns of $\bar{A}^{(i)}$ that are in $\bar{B}$ are the same for each block in the sense that column j of the first block of $\bar{A}^{(i)}$ is in $\bar{B}$ if and only if column j of each block of $\bar{A}^{(i)}$, except possibly the last one, is in $\bar{B}$. However, by using the observation made above about the matrix Λ (which in some sense is a building block of $\bar{A}^{(i)}$), for $\bar{B}$ to have full column rank it must contain exactly $n_2 + n_3$ columns of $\bar{A}^{(i)}$ from each block, except possibly the last one. This shows that any degenerate columns in $\bar{B}$, i.e., columns corresponding to zero variables in $\bar{\omega}$, must be taken from the last block of $\bar{A}^{(i)}$ for some i. Thus the total number of columns of $\bar{B}$ that are taken from the last block of $\bar{A}^{(i)}$ for some i must be $m(n_2 + n_3)$.

Now consider AP(P). Let $\hat{\omega}$ denote a set of variables for AP(P). Consider the columns of the constraint matrix for AP(P), $\hat{A}$, corresponding to the variables of $\hat{\omega}$ relating to the times t_{i-1} and t_i (except for $\hat{y}(t_{i-1})$), i.e., corresponding to the variables $\hat{x}(t_{i-1}+)$, $\hat{y}((t_{i-1} + t_i)/2)$, $\hat{z}(t_{i-1}+)$, $\hat{x}(t_i-)$, $\hat{y}(t_i)$ and $\hat{z}(t_i-)$. This can be written as

$$\hat{A}^{(i)} = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
G & I & 0 & 0 & 0 & 0 \\
H & 0 & I & 0 & 0 & 0 \\
0 & -I & 0 & G & I & 0 \\
0 & 0 & 0 & H & 0 & I \\
0 & 0 & 0 & 0 & -I & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}.$$

We refer to the first $n_1 + n_2 + n_3$ and last $n_1 + n_2 + n_3$ columns of $\hat{A}^{(i)}$ as the first and second blocks, respectively. Define a matrix $\hat{B}^{(i)}$ consisting of columns of $\hat{A}^{(i)}$ as follows. If column j of the first block of $\bar{A}^{(i)}$ is in $\bar{B}$ (and hence column j of each block of $\bar{A}^{(i)}$, except possibly the last, is in $\bar{B}$), then column j of the first block of $\hat{A}^{(i)}$ is in $\hat{B}^{(i)}$. This generates $n_2 + n_3$ columns of $\hat{B}^{(i)}$. The remaining columns of $\hat{B}$ are chosen in accordance with the columns of $\bar{B}$ from the last block of $\bar{A}^{(i)}$. In particular, if column j of the last block of $\bar{A}^{(i)}$ is in $\bar{B}$, then column j of the second block of $\hat{A}^{(i)}$ is in $\hat{B}^{(i)}$. By the properties of $\bar{B}$ noted above as well

as (18) we now claim that

$$(\hat{B}^{(i)})^T\hat{\theta} = \begin{bmatrix} \tilde{c}(t_{i-1}+) \\ \tilde{c}(t_i-) \end{bmatrix}. \tag{20}$$

In fact, suppose for instance that $[\bar{\pi}(s_j^{(i)}-) - \bar{\pi}(s_{j-1}^{(i)}+)]_k = 0$ is one equation in (18). Then we must have $y_k(t) > 0$ on (t_{i-1}, t_i) and so the equations $[\bar{\pi}(s_l^{(i)}-) - \bar{\pi}(s_{l-1}^{(i)}+)]_k = 0$ for each $l = 1, \ldots, p_i$ and $[\bar{\pi}(s_l^{(i)}+) - \bar{\pi}(s_l^{(i)}-)]_k = 0$ for each $l = 1, \ldots, p_i - 1$ are in (18). Hence we have $[\bar{\pi}(t_i-) - \bar{\pi}(t_{i-1}+)]_k = 0$. But by definition of $\hat{B}^{(i)}$, this equation is in (20). A similar argument follows for the other equations.

Having defined $\hat{B}^{(i)}$ for each i we set

$$\hat{B} = \begin{bmatrix} \hat{B}^{(1)} & \hat{B}^{(2)} & \ldots & \hat{B}^{(m)} \end{bmatrix};$$

then by the observations made above we see that $\hat{B}$ contains $2m(n_2 + n_3)$ columns of $\hat{A}$, $(n_2 + n_3)$ from each of the first blocks of $\hat{A}^{(i)}$ for each i, and $m(n_2 + n_3)$ in total from the second blocks of each $\hat{A}^{(i)}$. Hence $\hat{B}$ is a square matrix. Moreover, from (20) we see that $\hat{\theta}$ satisfies

$$\hat{B}^T\hat{\theta} = \begin{bmatrix} \tilde{c}(t_0+) \\ \tilde{c}(t_1-) \\ \vdots \\ \tilde{c}(t_m-) \end{bmatrix}.$$

All that remains now is to show that $\hat{B}$ has full column rank. Then $\hat{B}$ will be a basis matrix for AP(P) and so (19) will hold.

Suppose that $\hat{B}$ does not have full column rank. Then there exists $\gamma \neq 0$ such that $i \in \text{supp}(\gamma)$ only if $\hat{A}_i$ is a column in $\hat{B}$ and such that $\hat{A}\gamma = 0$. Suppose $\gamma^T = (\alpha^{(1)^T}, \beta^{(1)^T}, \ldots, \alpha^{(m)^T}, \beta^{(m)^T})$, where $\alpha^{(i)}, \beta^{(i)} \in \mathbb{R}^{n_1+n_2+n_3}$ for each i. Define $\rho^{(i)} \in \mathbb{R}^{n_1+n_2+n_3}$ by $\rho_j^{(i)} = \alpha_j^{(i)}$, if $j = n_1 + 1, \ldots, n_1 + n_2$ and $\rho_j^{(i)} = 0$ otherwise. Define

$$\delta^T = (\alpha^{(1)^T}, \rho^{(1)^T}, \ldots, \rho^{(1)^T}, \beta^{(1)^T}, \ldots, \alpha^{(m)^T}, \rho^{(m)^T}, \ldots, \rho^{(m)^T}, \beta^{(m)^T}),$$

where $\rho^{(i)}$ is repeated $2(p_i - 1)$ times for each i. Then it can be observed that $\bar{A}\delta = 0$, thus showing that $\bar{B}$ does not have full column rank. □

The strong duality result under Assumption 6.2 now follows immediately by appealing to Lemma 5.3 to guarantee the existence of an optimal solution for SCLP*. Note that this lemma guarantees a piecewise analytic optimal solution for SCLP* with at most $M + N + 1$ breakpoints, where M is given by Lemma 6.7. The existence of a piecewise analytic optimal solution for SCLP is given by Theorem 2.5 and the absence of a duality gap by Theorem 4.2. By using the same arguments as in the strong duality result for analytic costs, Theorem 6.6, the result can be extended to include general H that does not necessarily contain the upper bound constraints on $x(t)$. Again we may repeat the overall argument a finite number of times to arrive at the main result of this paper.

THEOREM 6.9 (strong duality). *Suppose that the costs $c(t)$ and the right-hand sides $a(t)$ and $b(t)$ are piecewise analytic on $[0, T]$ (with $a(t)$ continuous) and the feasible region for* SCLP *is nonempty and bounded. Then V[SCLP] = V[SCLP*] and there exist piecewise analytic optimal solutions for both* SCLP *and* SCLP*.

Note that the result for the absence of a duality gap contained in the above result has a less restrictive boundedness assumption than Theorem 4.2.

7. Conclusions and counterexamples. The previous section contained a general strong duality result between SCLP and SCLP*. The proof was largely constructive and involved finding an optimal solution for SCLP*. The proof showed us that, at least locally, an optimal solution $\theta(t)$ for SCLP* satisfies

$$\theta(t) = (B^{-1})^T \bar{c}_B(t)$$

for t in some interval $[\alpha, \beta)$, where

$$\bar{c}(t) = \begin{bmatrix} c(t) \\ \pi(\beta-) \\ 0 \end{bmatrix}$$

for $t \in [\alpha, \beta)$. Hence the optimal $\theta(t)$ has the same properties as the costs $c(t)$. Thus if, for instance, $c(t)$ were a polynomial of degree n, then the optimal $\theta(t)$ would be piecewise polynomial of degree n. Such an observation has also been made about optimal solutions for the primal problem SCLP in the conclusion to Pullan [29], namely, that optimal solutions for the primal reflect the nature of the right-hand sides $a(t)$ and $b(t)$. This leads to many possible variations on the strong duality theorem of the previous section. Recalling one result from the conclusion to [29] we state one of these many possible variations.

THEOREM 7.1 (strong duality). *Suppose that the costs $c(t)$ and the right-hand sides $a(t)$ and $b(t)$ are piecewise polynomial on $[0, T]$ (with $a(t)$ continuous) of degrees n, $m+1$, and m, respectively. Suppose also that the feasible region for* SCLP *is nonempty and bounded. Then V[SCLP] $= V$[SCLP*] and there exist piecewise polynomial optimal solutions of degree m for* SCLP *and degree n for* SCLP*.

It is also interesting to speculate whether strong duality holds between SCLP and SCLP* in more general circumstances, for instance, continuously differentiable problem data. To partially answer this question we present a simple counterexample to show that the result in Theorem 6.9 cannot be extended beyond analytic a and b to, say, n-times continuously differentiable a and b for some n, even if the costs are assumed to be constant. In such a case an optimal solution for the primal problem may have an infinite number of breakpoints. Thus, because of the necessity of complementary slackness for strong duality, it is possible that any potential optimal solution for the dual must have an infinite number of breakpoints, which can cause the potential optimal $\pi(t)$ to be unbounded, i.e., not of bounded variation.

Example 7.1. Consider a simple network of two nodes connected by two arcs (see Figure 7.1). We will consider a network problem over the time interval $[0, 1]$ where storage is permitted at the nodes (in other words, a CNP example as discussed by Anderson and Philpott [4]). Let $x_i(t)$ denote the rate of flow in arc i at time t and $y_i(t)$ denote the amount of storage at time t in node i for $i = 1, 2$. The rate of flow in each of the two arcs is subject to an upper bound of 1, and the cost per unit flow in each arc is 1. Choose $n \geq 1$ and define the supplies, $r_i(t)$, in node i by

$$r_1(t) = \begin{cases} (1-t)^n \sin\left(\dfrac{1}{1-t}\right), & t \in [0, 1), \\ 0, & t = 1, \end{cases}$$
$$r_2(t) = -r_1(t)$$

for $t \in [0, 1]$. We then define

$$a_i(t) = \int_0^t r_i(s)\, ds$$

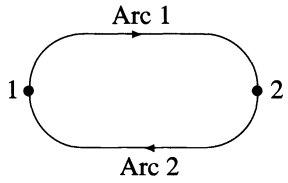


FIG. 7.1. *The network in Example* 7.1.

for $t \in [0, 1]$ and $i = 1, 2$. Then $a_i(t)$ is n-times continuously differentiable. Our SCLP (or CNP) problem is now

$$\begin{aligned}
\text{Ex:}\quad \text{minimize}\quad & \int_0^1 (x_1(t) + x_2(t))\, dt \\
\text{subject to}\quad & \int_0^t (x_1(s) - x_2(s))\, ds + y_1(t) = a_1(t), \\
& \int_0^t (x_2(s) - x_1(s))\, ds + y_2(t) = a_2(t), \\
& x_1(t) + z_1(t) = 1, \\
& x_2(t) + z_2(t) = 1, \\
& x_1(t), x_2(t), y_1(t), y_2(t), z_1(t), z_2(t) \geq 0, \quad t \in [0, 1].
\end{aligned}$$

Let

$$S = \{ t \in [0, 1] : r_1(t) > 0 \};$$

then it is clear that Ex has an optimal solution, $\omega(t)^T = (x_1(t), x_2(t), y_1(t), y_2(t), z_1(t), z_2(t))$, where

$$x_1(t) = \begin{cases} r_1(t), & t \in S, \\ 0, & t \notin S, \end{cases}$$

$$x_2(t) = \begin{cases} 0, & t \in S, \\ r_2(t), & t \notin S. \end{cases}$$

This then gives $y_1(t) = y_2(t) = 0$ and $z_1(t), z_2(t) > 0$ on $[0, 1]$. Now the dual of Ex is given by

$$\begin{aligned}
\text{Ex*:}\quad \text{maximize}\quad & \int_0^1 (\eta_1(t) + \eta_2(t))\, dt - \int_0^1 a_1(t)\, d\pi_1(t) - \int_0^1 a_2(t)\, d\pi_2(t) \\
\text{subject to}\quad & 1 - \pi_1(t) + \pi_2(t) - \eta_1(t) \geq 0, \\
& 1 - \pi_2(t) + \pi_1(t) - \eta_2(t) \geq 0, \\
& \eta_1(t), \eta_2(t) \leq 0, \text{ a.e. on } [0, 1], \\
& \pi_1(t), \pi_2(t) \text{ monotonic increasing and right continuous} \\
& \text{on } [0, 1] \text{ with } \pi_1(1) = \pi_2(1) = 0.
\end{aligned}$$

By a simple application of Theorem 4.2 we see that there is no duality gap between Ex and Ex*. However, as we shall see below, Ex* does not attain its optimal value, and hence strong duality does not hold between Ex and Ex*.

Suppose that strong duality holds, i.e., that Ex* attains its optimal value. Then by the complementary slackness result (Theorem 3.4), there exists $\theta(t)^T = (\pi_1(t), \pi_2(t), \eta_1(t), \eta_2(t))$

optimal for Ex* and complementary slack with $\omega(t)$. Using the definition of complementary slackness, we see that this must imply that

$$1 - \pi_1(t) + \pi_2(t) - \eta_1(t) = 0 \text{ a.e. on } S, \tag{21}$$

$$1 - \pi_2(t) + \pi_1(t) - \eta_2(t) = 0 \text{ a.e. on } [0, 1] - S, \tag{22}$$

$$\eta_1(t) = \eta_2(t) = 0 \text{ a.e. on } [0, 1]. \tag{23}$$

We now proceed to show that these equations cannot be satisfied while maintaining $\theta(t)$ feasible for Ex*.

Suppose that $\pi_2(0) = M$. As $\pi_2(1) = 0$ we must have $M \leq 0$. Now S is an open disconnected set in [0, 1] with an infinite number of components. Suppose then that

$$S = \{0\} \cup \bigcup_{i=1}^{\infty} (t_i, s_i)$$

and that $0 = t_1 < s_1 < t_2 < s_2 \cdots$. Now $(t_1, s_1) \subset S$, so we must have by (21) and (23) that

$$\pi_1(t) = 1 + \pi_2(t) \text{ a.e. on } (t_1, s_1).$$

Hence, as $\pi_2(t) \geq M$ on (t_1, s_1), we must have

$$\pi_1(t) \geq 1 + M \text{ on } (t_1, s_1).$$

Hence $\pi_1(t) \geq 1 + M$ on (t_1, t_2) by the monotonicity of $\pi_1(t)$. Now by repeating the above argument for the interval (s_1, t_2), which is not a component of S, and using (22) and (23) this time, we obtain

$$\pi_2(t) \geq 2 + M \text{ on } (s_1, t_2).$$

Continuing in this manner it is apparent that we may construct $t \in (0, 1)$ such that $\pi_1(t) > 0$. This is a contradiction since $\pi_1(1) = 0$ and $\pi_1(t)$ is monotonic increasing on [0, 1]. Hence no optimal solution for Ex* exists, so strong duality does not hold between Ex and Ex*. □

The question of whether it is possible to extend Theorem 6.6 (i.e., the result with analytic costs, linear $a(t)$, and constant $b(t)$) to include more general costs appears more difficult and remains unresolved.

REFERENCES

[1] E. J. ANDERSON, *A Continuous Model For Job-Shop Scheduling*, Ph.D. thesis, University of Cambridge, UK, 1978.

[2] E. J. ANDERSON AND P. NASH, *Linear Programming in Infinite-Dimensional Spaces*, Wiley-Interscience, Chichester, 1987.

[3] E. J. ANDERSON, P. NASH, AND A. F. PEROLD, *Some properties of a class of continuous linear programs*, SIAM J. Control Optim., 21 (1983), pp. 758–765.

[4] E. J. ANDERSON AND A. B. PHILPOTT, *A continuous-time network simplex algorithm*, Networks, 19 (1989), pp. 395–425.

[5] K. M. ANSTREICHER, *Generation of Feasible Descent Directions in Continuous Time Linear Programming*, Tech. Report SOL 83-18, Department of Operations Research, Stanford University, Stanford, CA, 1983.

[6] T. M. APOSTOL, *Mathematical Analysis*, 2nd ed., Addison-Wesley, Reading, MA, 1974.

[7] R. E. BELLMAN, *Bottleneck problems and dynamic programming*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 947–951.

[8] ———, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

[9] J. M. BORWEIN, *Adjoint process duality*, Math. Oper. Res., 8 (1983), pp. 403–434.

[10] R. N. BUIE AND J. ABRHAM, *Numerical solutions to continuous linear programming problems*, Z. Oper. Res., 17 (1973), pp. 107–117.

[11] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.
[12] W. P. DREWS, *A simplex-like algorithm for continuous-time linear optimal control problems*, in Optimization Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak, New York, 1974, pp. 309-322.
[13] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators Part* I: *General Theory*, Wiley-Interscience, New York, 1988.
[14] R. C. GRINOLD, *Continuous Programming*, Ph.D. thesis, Operations Research Center, University of California, Berkeley, CA, 1968.
[15] ———, *Continuous programming part one: Linear objectives*, J. Math. Anal. Appl., 28 (1969), pp. 32-51.
[16] ———, *Symmetric duality for continuous linear programs*, SIAM J. Appl. Math., 18 (1970), pp. 84-97.
[17] W. H. HAGER, *Lipschitz continuity for constrained problems*, SIAM J. Control Optim., 17 (1979), pp. 321-338.
[18] W. H. HAGER AND S. K. MITTER, *Lagrange duality theory for convex control problems*, SIAM J. Control Optim., 14 (1976), pp. 843-856.
[19] R. J. HARTBERGER, *Representation extended to continuous time*, in Optimization Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak, New York, 1974, pp. 297-307.
[20] A. N. KOLMOGOROV AND S. V. FOMIN, *Introductory Real Analysis*, Dover, New York, 1975.
[21] R. S. LEHMAN, *On the Continuous Simplex Method*, RM-1386, Rand Corporation, Santa Monica, CA, 1954.
[22] P. LEVINE AND J.-CH. POMEROL, *Sufficient conditions for Kuhn-Tucker vectors in convex programming*, SIAM J. Control Optim., 17 (1979), pp. 689-699.
[23] N. LEVINSON, *A class of continuous linear programming problems*, J. Math. Anal. Appl., 16 (1966), pp. 73-83.
[24] N. S. PAPAGEORGIOU, *A class of infinite dimensional linear programming problems*, J. Math. Anal. Appl., 87 (1982), pp. 228-245.
[25] A. F. PEROLD, *Fundamentals of a Continuous Time Simplex Method*, Tech. Report SOL 78-26, Department of Operations Research, Stanford University, Stanford, CA, 1978.
[26] ———, *On a Continuous Time Simplex Method* I: *Local Basis Change*, Tech. Report, Graduate School of Business Administration, Harvard University, Boston, MA., 1982.
[27] M. C. PULLAN, *Separated Continuous Linear Programs: Theory and Algorithms*, Ph.D. thesis, University of Cambridge, UK, 1992.
[28] ———, *An algorithm for a class of continuous linear programs*, SIAM J. Control Optim., 31 (1993), pp. 1558-1577.
[29] ———, *Forms of optimal solutions for separated continuous linear programs*, SIAM J. Control Optim., 33 (1995), pp. 1952-1977.
[30] R. T. ROCKAFELLAR, *State constraints in convex problems of Bolza*, SIAM J. Control, 10 (1972), pp. 691-715.
[31] ———, *Conjugate Duality and Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1974.
[32] T. W. REILAND, *Optimality conditions and duality in continuous programming* II. *The linear problem revisited*, J. Math. Anal. Appl., 77 (1980), pp. 329-343.
[33] H. H. SCHAEFER, *Topological Vector Spaces*, Springer-Verlag, New York, 1971.
[34] M. SCHECHTER, *Duality in continuous linear programming*, J. Math. Anal. Appl., 37 (1972), pp. 130-141.
[35] R. G. SEGERS, *A generalised function setting for dynamic optimal control problems*, in Optimization Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak, New York, 1974, pp. 279-296.
[36] W. F. TYNDALL, *A duality theorem for a class of continuous linear programming problems*, SIAM J. Appl. Math., 13 (1965), pp. 644-666.
[37] ———, *An extended duality theorem for continuous linear programming problems*, SIAM J. Appl. Math., 15 (1967), pp. 1294-1298.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 966-986, May 1996

010

CYCLE DECOMPOSITIONS AND SIMULATED ANNEALING*

ALAIN TROUVÉ†

Abstract. The behavior of simulated annealing algorithms is tightly related to the hierarchical decomposition of their configuration spaces in cycles. In the generalized annealing framework, this decomposition is defined recursively. In this paper, its structure is extensively studied, and it is shown that the decomposition can be achieved through an implementable algorithm which allows exact computation of the fundamental constants underlying the behavior of these algorithms.

Key words. cycle decomposition, large deviations, generalized simulated annealing, rate of convergence

AMS subject classifications. Primary, 60F10; Secondary, 60J10, 93E25

1. Introduction.

Simulated annealing is now used extensively for optimizing large-scale problems. Its main advantages are its wide applicability and its simplicity, allowing a large variability in the choice of the constraints defining the cost function to be minimized.

Let us recall briefly the theoretical scheme. We consider the problem of finding the configurations minimizing a function $U : E \to \mathbb{R}$ on a finite set E, called the configuration space (or the state space). The function U is generally called the cost function or the energy function. One considers an inhomogeneous Markov chain $X = (X_n)_{n\in\mathbb{N}}$ on E with transition kernel at time n given by $Q_{T_{n+1}}$, where for all $T > 0$ (called the temperature)

$$Q_T(i, j) = q(i, j)e^{-(U(j)-U(i))^+/T}, \tag{1}$$

and q is an irreducible Markov kernel satisfying $q(i, j) = q(j, i)$. The sequence $\mathcal{T} = (T_n)_{n\in\mathbb{N}}$ is a sequence of nonnegative numbers called the cooling schedule. For low temperature, the Markov chain preferably follows configuration paths with decreasing energy value, but hill-climbing moves are allowed with a probability controlled by the temperature. The theory of simulated annealing says that if the cooling schedule is sufficiently slowly decreasing, then we have

$$\sup_{i\in E} P(\, U(X_n) \neq \min U \mid X_0 = i \,) \underset{n\to+\infty}{\to} 0 .$$

Using the large deviation approach of Wentzell and Freidlin [6] one can have much more information on the behavior at low temperature. It appears that the Markov chain performs a hierarchical exploration of the configuration space well described by the cycle decomposition of E introduced in [6]. The role of the cycles becomes clear if we consider the Markov chain under a constant cooling schedule at temperature T. When the Markov chain enters into a cycle, the time to exit from the cycle is of order of $e^{H_e/T}$, where H_e called the exit height is a constant depending only on the cycle and not on the entering point. The smallest cycles are the singletons, and the largest one is the whole space E. Since two cycles are either disjoint or included one in the other, the cycles are structured as a tree, and the Markov chain moves from cycle to cycle along characteristic paths. The generalization of this approach to decreasing temperature schedules for the study of simulated annealing algorithms requires some important work, for which the reader can refer to [2-4, 9, 10].

Let $\mathcal{C}(E)$ denote the set of all the cycles, and for every cycle Π, $H_e(\Pi)$ denotes the exit height of Π. Many of the characteristics of the asymptotic behavior of the sequential

*Received by the editors November 16, 1993; accepted for publication (in revised form) January 16, 1995.

†Laboratoire de Mathématiques de l'Ecole Normale Supérieure/Distributed Intelligent Applications and Mathematics, URA 732, Ecole Normale Supérieure, 45 rue d'Ulm, 75230, Paris cedex 05, France (trouve@ens.ens.fr).

annealing under decreasing cooling schedule can be computed. For instance (see [2, 7]), if $\mathcal{T}$ is a decreasing cooling schedule vanishing to zero, then we have

$$\sup_{i\in E} P(\, U(X_n) \neq \min U \mid X_0 = i\,) \underset{n\to+\infty}{\longrightarrow} 0 \text{ iff } \sum_{n\geq 0} e^{-H_1/T_n} = +\infty, \tag{2}$$

where H_1 denotes $\sup\{\, H_e(\Pi) \mid \Pi \in \mathcal{C}(E) \text{ and } U(\Pi) > \min U \,\}$ and $U(\Pi) = \min_{i\in\Pi} U(i)$. The constant H_1 is called the critical height of the cycle decomposition. Moreover, there exists a positive constant K_1 such that for any cooling schedule $\mathcal{T}$ we have

$$\sup_{i\in E} P(\, U(X_n) > \min U \mid X_0 = i\,) \geq K_1/n^{\alpha_{opt}}, \tag{3}$$

where $\alpha_{opt} = \inf\{\, \frac{U(\Pi)-\min U}{H_e(\Pi)} \mid \Pi \in \mathcal{C}(E) \text{ and } U(\Pi) > \min U \,\}$. The optimality of the exponent α_{opt} is proven by a final result established in [2] which says that there exists a positive constant K such that for all $N > 0$ there exists a cooling schedule for which we have

$$\sup_{i\in E} P(\, U(X_N) > \min U \mid X_0 = i\,) \leq K/N^{\alpha_{opt}}. \tag{4}$$

The number α_{opt} is called the optimal convergence exponent.

In order to have a better convergence speed toward the global minima of U, many computer scientists have proposed parallelized schemes of the usual sequential simulated annealing. The underlying idea is to distribute the amount of computation on several processors (see [1, 13]). It appears that the parallel schemes lead to a more general form of the simulated annealing called the generalized simulated annealing (G.S.A.) introduced by Hwang and Sheu in [10] and which corresponds to a discrete-time and finite-space analogue of the general framework introduced by Wentzell and Freidlin in [6] in their study of random perturbation of dynamical systems. Instead of assuming that the transition kernel Q_T satisfies (1), we just assume that there exists $\kappa \geq 1$ such that

$$\frac{1}{\kappa} q(i,j) e^{-V(i,j)/T} \leq Q_T(i,j) \leq \kappa q(i,j) e^{-V(i,j)/T}, \tag{5}$$

where the family V called the communication cost satisfies $V(i,j) \in [0, +\infty]$ and $V(i,j) = +\infty$ iff $q(i,j) = 0$. This new framework is adapted to the study of many of the extensions (parallel or not) of the usual sequential scheme. One gets from Wentzell and Freidlin's theory that there exists a virtual energy W which plays the same role as the energy U for low temperature since if we denote by μ_T the unique equilibrium probability measure of Q_T, then

$$\lim_{T\to 0} T \ln \mu_T(i) = -(W(i) - \min W),$$

and with the help of the Dobrushin theory of inhomogeneous Markov chains we easily prove that if T is sufficiently slowly decreasing, then

$$\sup_{i\in E} P(\, W(X_n) > \min W \mid X_0 = i\,) \underset{n\to+\infty}{\to} 0.$$

However, if we want to establish more precise results for the general framework, like (2)-(4) (with U replaced by W), we have to deal with an extension of the cycle decomposition. This extension has been proposed in the original work of Wentzell and Freidlin, where they again define cycles for which we keep the crucial property that at constant temperature, the time to exit from a cycle Π is of order of $e^{H_e(\Pi)/T}$, where $H_e(\Pi)$ does not depend on the starting point. Using this extended cycle decomposition, Hwang and Sheu in [10] proved a weak version of (2) for cooling schedules $\mathcal{T}$ satisfying $T_n = c/\ln(n+2)$, and in [13, 14] the author proves the exact analogues of (2)-(4) for the G.S.A. (where U is replaced by W). Hence the cycle

decomposition plays a central role in the asymptotic behavior of the G.S.A. as well as in the computation of H_1 and α_{opt}.

We will in this paper perform an extended study of the cycle decomposition in the generalized annealing framework. In §2, starting from the recursive definition of the cycles, which depends only on the communication cost V (and not on W), we will establish the links between the virtual energy W and the exit height $H_e(\Pi)$ and show that these links lead to a decomposition diagram structuring the cycle decomposition in a valued tree. Then we will introduce the altitude of communication by

$$A_c(i, j) = \inf_{n\in\mathbb{N}} \inf_{i=g_0,g_1,\dots,g_n=j} \sup_{k<n} W(g_k) + V(g_k, g_{k+1}),$$

which will allow us to give a new construction of the cycle decomposition very close to the sequential case construction. We will show also that this altitude of communication is symmetric ($A_c(i, j) = A_c(j, i)$) and that this property characterizes (up to a multiplicative constant) the virtual energy. This will allow us to shed a new light on the weak reversibility of Hajek. Finally, in §3, we propose an implementable algorithm to compute automatically the cycle decomposition as well as all the critical constants H_1 and α_{opt}. This may be useful in order to study the *exact* asymptotic behavior of the G.S.A on small state spaces—for example, when evaluating parallel schemes (see [13]).

2. Generalized simulated annealing.

2.1. Definition. Let us first recall some notation. The set E denotes a finite configuration space, and q an irreducible Markov kernel on E called the communication kernel. The irreducibility of Q means that for all distinct configurations i and j there exists a path $(i_k)_{0\le k\le n}$ satisfying

$$i_0 = i,\ i_n = j, \quad \text{and } q(i_k, i_{k+1}) > 0 \text{ for all } k \le n-1.$$

Let us also consider a real-valued number $\kappa \ge 1$. We can now define precisely the families of kernels we will consider.

DEFINITION 2.1. *Let $(Q_T)_{T>0}$ be a family of Markov kernels on E. We say that $(Q_T)_{T>0}$ is admissible for q and κ if there exists a family of positive real-valued numbers $(V(i, j))_{(i,j)\in E\times E}$ (some of them may take the value $+\infty$) such that*

1. *$V(i, j) < +\infty$ iff $q(i, j) > 0$;*
2. *for all $T > 0$, all $i, j \in E$,*

$$\frac{1}{\kappa} q(i, j) e^{-V(i,j)/T} \le Q_T(i, j) \le \kappa q(i, j) e^{-V(i,j)/T}.$$

The function $V : E \times E \to [0, +\infty]$ is called the communication cost function.

NOTATION 2.2. *The set of all the admissible families $(Q_T)_{T>0}$ will be denoted $\mathcal{A}(q, \kappa)$.*

The set $\mathcal{A}(q, \kappa)$ contains all the families of kernels associated with sequential simulated annealing algorithms and also many parallelized versions of the sequential scheme.

We define now the generalized simulated annealing algorithms.

DEFINITION 2.3. *We say that $(X_n)_{n\in\mathbb{N}}$ is a generalized simulated annealing with parameters $(E, V, q, \kappa, \nu_0, \mathcal{T})$, where ν_0 is a probability measure on E and $\mathcal{T} = (T_n)_{n\in\mathbb{N}}$ is a decreasing sequence of strictly positive real-valued numbers called the cooling schedule, if there exists a family of Markov kernel $(Q_T)_{T>0}$ in $\mathcal{A}(q, \kappa)$ with communication cost function V such that $(X_n)_{n\in\mathbb{N}}$ is a Markov chain on E satisfying*

$$P(X_0 = i) = \nu_0(i), \qquad i \in E,$$

$$P(X_{n+1} = j \mid X_n = i) = Q_{T_{n+1}}(i, j), \qquad i, j \in E.$$

2.2. Virtual energy. Let us consider a family $(Q_T)_{T>0}$ of Markov kernels in $\mathcal{A}(q, \kappa)$ and $V : E \times E \to [0, +\infty]$, the associated communication cost function. In the following definition, we recall the notion of A-graphs as defined by Wentzell and Freidlin in [6]. Like them, $i \to j$ will denote the pair (i, j).

DEFINITION 2.4. *Let $A \subset E$. We say that a set g of arrows $i \to j$ in $A^c \times E$ is an A-graph iff*

1. *for each $i \in A^c$, there exists a unique $j \in E$ such that $i \to j \in g$;*
2. *for each $i \in A^c$, there is a path $i = i_0 \to i_1 \to \cdots \to i_n$ such that $i_k \to i_{k+1} \in g$, ending on a configuration in A.*

We denote by $G(A)$ the set of the A-graphs. Furthermore, for each $g \in G(A)$ we denote

$$V(g) = \sum_{i \to j \in g} V(i, j).$$

We can now define the virtual energy.

DEFINITION 2.5. *We say that $W : E \to \mathbb{R}$ is the virtual energy associated with the communication cost function V if*

$$W(i) = \inf_{g \in G(\{i\})} V(g), \qquad i \in E.$$

The virtual energy will play the same role as the energy U for the sequential simulated annealing, as shown in the proposition below.

PROPOSITION 2.6 (Wentzell and Freidlin). *For all $T > 0$, we denote by μ_T the unique invariant probability measure for Q_T (since q irreducible implies Q_T irreducible). Then we have*

$$T \ln(\mu_T(i)) \underset{T \to 0}{\longrightarrow} -(W(i) - W(E)),$$

$$\text{where } W(E) = \inf_{j \in E} W(j).$$

Hence, the generalized simulated annealing can be seen again as an optimization algorithm minimizing the virtual energy. However, the virtual energy is now only implicitly defined by the communication cost function.

2.3. Decomposition in cycles. The study of the virtual energy is not sufficient to understand the behavior of a generalized simulated annealing algorithm. One can have two communication cost functions giving the same virtual energy but leading to very different asymptotic behaviors. This leads us to the computation of a decomposition in cycles as done for the sequential case [2]. However, the cycles cannot be obtained by a simple substitution of U for W. The correct construction has been given by Wentzell and Freidlin [6] and is reported here with slight modifications. For this purpose, let us give the following preliminary definition.

DEFINITION 2.7. *Let F be a finite set, $C : F \times F \to [0, +\infty]$ be a function, and i and j be two distinct configurations of F.*

1. *We denote by $Pth_F(i, j)$ the set of all paths $(g_k)_{0 \le k \le n}$ in F such that $g_0 = i$ and $g_n = j$. The path-dependent integer n will be denoted by n_g and will be called the length of g.*
2. *For g to be in $Pth_F(i, j)$, define $C(g)$ by*

$$C(g) = \sum_{k=0}^{n_g - 1} C(g_k, g_{k+1}).$$

We adopt the convention that $C(g) = +\infty$ if one of the summation terms is $+\infty$.

The decomposition in cycles is defined in an iterative way. First the set E^0 of cycles of order 0 is defined by

$$E^0 = \{ \{i\} \mid i \in E \}.$$

Let us consider the communication cost function V^0 on E^0 defined by

$$V^0(\{i\}, \{j\}) = V(i, j).$$

Assume that the set E^k of the cycles of order k and a communication cost function V^k on E^k have been constructed. The construction of the pair E^{k+1}, V^{k+1} can be split in several steps:

1. From V^k, we define another communication cost V^k_* on E^k by

$$V^k_*(\Pi, \Pi') = \begin{cases} 0 & \text{if } \Pi = \Pi', \\ V^k(\Pi, \Pi') - H^k_e(\Pi) & \text{otherwise,} \end{cases}$$

$$\text{where } H^k_e(\Pi) = \inf\{V^k(\Pi, \Pi'') \mid \Pi'' \in E^k,\ \Pi'' \neq \Pi \}.$$

2. On E^k, we define the relation $\overset{k}{\to}$ by

$$\Pi \overset{k}{\to} \Pi' \text{ if either } \Pi = \Pi' \text{ or there exists } g \in \mathrm{Pth}_{E^k}(\Pi, \Pi') \text{ such that } V^k_*(g) = 0,$$

and the equivalence relation $\mathcal{R}_k$ by

$$\Pi \mathcal{R}_k \Pi' \text{ if either } \Pi = \Pi' \text{ or } \Pi \overset{k}{\to} \Pi' \text{ and } \Pi' \overset{k}{\to} \Pi.$$

3. Define D^{k+1} by

$$D^{k+1} = \{ \bigcup_{\Pi' \mathcal{R}_k \Pi} \Pi' \mid \Pi \in E^k \},$$

and define on D^{k+1} the partial order $\preceq$ by $\Pi^{k+1}_1 \preceq \Pi^{k+1}_2$ if there exists $\Pi^k_i \subset \Pi^{k+1}_i$ for $i = 1, 2$ such that $\Pi^k_2 \overset{k}{\to} \Pi^k_1$. We denote D^{k+1}_* by the set of the minimal elements of D^{k+1} for the order $\preceq$.

4. Define E^{k+1} by

$$E^{k+1} = D^{k+1}_* \cup \{ \Pi^k \in E^k \mid \exists \Pi^{k+1} \in D^{k+1} \setminus D^{k+1}_*,\ \Pi^k \subset \Pi^{k+1} \}.$$

5. We define now a communication cost V^{k+1} on E^{k+1} by

$$(6)\quad V^{k+1}(\Pi^{k+1}_0, \Pi^{k+1}_1) = H^{k+1}_m(\Pi^{k+1}_0) + \inf\{ V^k_*(\Pi^k_0, \Pi^k_1) \mid (\Pi^k_0, \Pi^k_1) \in E^k \times E^k,\ \Pi^k_0 \subset \Pi^{k+1}_0,\ \Pi^k_1 \subset \Pi^{k+1}_1 \},$$

$$\text{where } H^{k+1}_m(\Pi^{k+1}_0) = \sup\{ H^k_e(\Pi^k_0) \mid \Pi^k_0 \subset \Pi^{k+1}_0,\ \Pi^k_0 \in E^k \}.$$

The construction goes on until $E^k = \{E\}$. We denote the order n_E of the decomposition and the set $\mathcal{C}(E)$ of all the cycles by

$$n_E = \inf\{ k \in \mathbb{N} \mid E^{k+1} = \{E\} \}$$

and

$$\mathcal{C}(E) = \cup_{k=0}^{n_E+1} E^k.$$

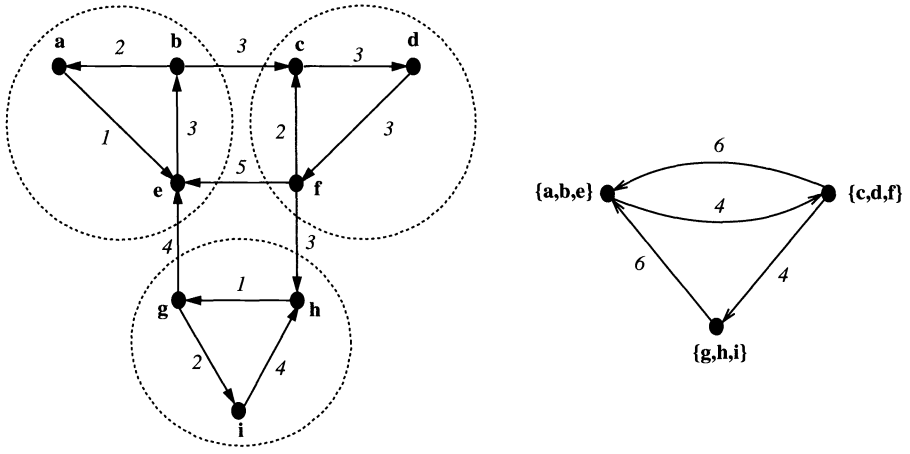


FIG. 1. *Contruction of the cycles.*

TABLE 1.

	H_m^1	H_e^1
$\{a, b, e\}$	3	4
$\{c, d, f\}$	3	4
$\{g, h, i\}$	4	6

TABLE 2.

V^1	$\{a, b, e\}$	$\{c, d, f\}$	$\{g, h, i\}$
$\{a, b, e\}$		4	$+\infty$
$\{c, d, f\}$	6		4
$\{g, h, i\}$	6	$+\infty$	

This procedure gives a hierarchical decomposition of the state space as a tree beginning with the singletons and ending with the whole space.

We shall illustrate this decomposition on a small configuration space:

$$E = \{a, b, c, d, e, f, g, h, i\}.$$

On the left-hand side of Fig. 1, we have represented only edges with finite cost, and their valuation $V(i, j)$ is reported above. The construction of the cycles can be obtained directly on such a small example. We explain here how to obtain the cycles of order 1 since the cycles of greater order can be obtained by iteration of the process. For this purpose, we construct first the *exit graph*, which is a subgraph of the previous graph for which we keep only the edges $i \to j$ satisfying $V(i, j) = \inf_{k \in E, k \neq i} V(i, k)$. The cycles of order 1 are the strongly connected components of the exit graph. We recall that for an oriented graph, a strongly connected component is a family of vertices maximal for the inclusion among all the families which stay connected even if one of their vertices is suppressed. For each cycle Π constructed in this way, we get the quantity $H_m^1(\Pi)$ as the maximal valuation among the edges of the exit graph joining two vertices of Π. Finally, for each pair of distinct cycles Π and Π', we get $V^1(\Pi, \Pi')$ from (6). In our example, we get the following three cycles of order one: $\{a, b, e\}$, $\{c, d, f\}$, and $\{g, h, i\}$. We obtain for H_m^1, H_e^1 and V^1 the values given in Tables 1 and 2. The values of V^1 are reported on the right-hand side of Fig. 1. For the order 2, we have only the whole space E. Hence, we get the decomposition tree in Fig. 2.

2.4. Decomposition diagram. In this section, we will show that we can define as for sequential annealing the bottom of a cycle, its exit height, and any of the quantities defined previously. We will study the links between the exit heights and the potential of the cycles and

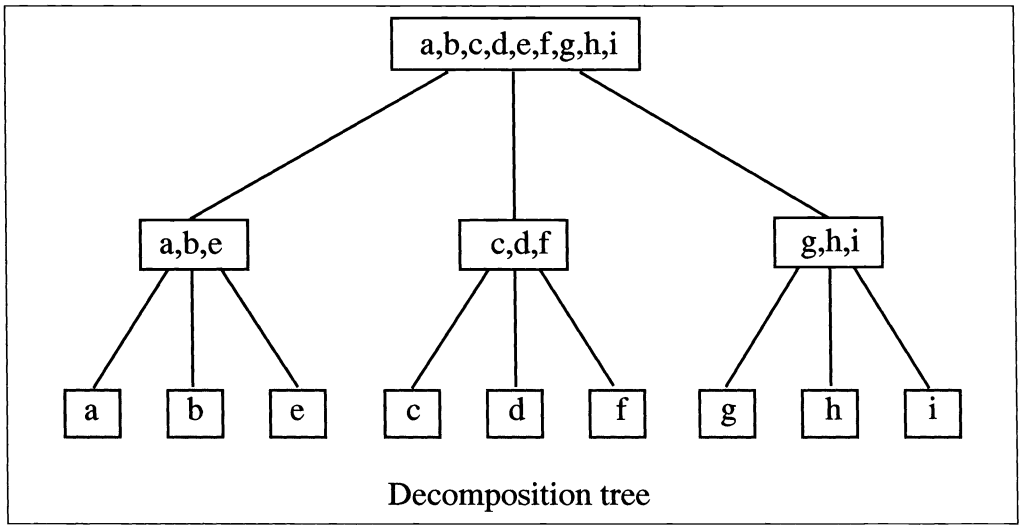


FIG. 2. *Decomposition tree.*

show the existence of an extension of the decomposition diagram introduced in the sequential case.

DEFINITION 2.8. *Let $i \in E$. We define by induction the increasing family of cycles $(i^k)_{0\leq k\leq n_E}$ by $i^0 = \{i\}$ and*

$$i^{k+1} \in E^{k+1}, \quad i^k \subset i^{k+1} \quad \text{for } k \leq n_E.$$

DEFINITION 2.9. *Let $A \subset E$.*

1. *We define the maximal proper partition $\mathcal{M}_*(A)$ of A for $|A| > 1$ by*

$$\mathcal{M}_*(A) = \{\, \Pi \in \mathcal{C}(E) \mid \Pi \text{ is a maximal element in } \mathcal{C}_A^*(E) \,\},$$

$$\text{where } \mathcal{C}_A^*(E) = \{\, \Pi \in \mathcal{C}(E) \mid \Pi \subset A,\ \Pi \neq A \,\}.$$

2. *For all $i \in A$, we call the order of i in A the nonnegative integer $n_{A,i}$ defined by*

$$n_{A,i} = \sup\{\, k \in \mathbb{N} \mid 0 \leq k \leq n_E \text{ and } i^k \subset A \,\}.$$

We will now extend the quantities defined previously on cycles.

DEFINITION 2.10. *Let $\Pi \in \mathcal{C}(E)$. We define*

1. *the order n_Π of Π by*

$$n_\Pi = \inf\{\, k \in \mathbb{N} \mid 0 \leq k \leq n_E \text{ and } \Pi \in E^k \,\};$$

2. *the exit height $H_e(\Pi)$ of Π by*

$$H_e(\Pi) = \begin{cases} \sup\{H_e^k(\Pi) \mid k \leq n_E,\ \Pi \in E^k\} & \text{if } \Pi \neq E, \\ +\infty & \text{otherwise;} \end{cases}$$

3. *the mixing height $H_m(\Pi)$ of Π by*

$$H_m(\Pi) = \begin{cases} \sup\{\, H_e(\Pi') \mid \Pi' \in \mathcal{M}_*(\Pi) \,\} & \text{if } |\Pi| > 1, \\ 0 & \text{if } \Pi \text{ is a singleton;} \end{cases}$$

4. *the potential* $W(\Pi)$ *of* Π *by*

$$W(\Pi) = \inf\{W(i) \mid i \in \Pi\};$$

5. *the bottom* $F(\Pi)$ *of* Π *by*

$$F(\Pi) = \{ i \in \Pi \mid W(i) = W(\Pi) \}.$$

We can make a few comments about these definitions.

Remark 1. The definition of the exit height of a cycle as a *sup* can be surprising. However, we can make the following remark, whose easy verification is left to the reader. Let Π be a cycle of E. If for an integer $k \leq n_E$ we have $\Pi \in E^k \cap E^{k+1}$, then

$$H_m^{k+1}(\Pi) = H_e^{k+1}(\Pi) = H_e^k(\Pi). \tag{7}$$

Hence, $H_e(\Pi) = H_e^k(\Pi)$ as soon as $\Pi \in E^k$.

This remark allows us to use the notation $H_e(i)$ for $H_e^0(\{i\})$.

Remark 2. From the construction of the cycles, for all $\Pi^k \in E^k$ and $\Pi^{k+1} \in E^{k+1}$ such that $\Pi^k \subset \Pi^{k+1}$ and $\Pi^k \neq \Pi^{k+1}$, we get $H_m^{k+1}(\Pi^{k+1}) < H_e^{k+1}(\Pi^{k+1})$. Hence, using Remark 1, we deduce that for each cycle Π such that $|\Pi| > 1$, $H_m(\Pi) < H_e(\Pi)$.

Some parts of the previous definition can be extended to arbitrary subsets of E.

DEFINITION 2.11. *Let* $A \subset E$, $A \neq \emptyset$.

1. *We define the exit height* $H_e(A)$ *of* A *by*

$$H_e(A) = \sup\{ H_e(\Pi) \mid \Pi \in \mathcal{C}(E),\ \Pi \subset A \}.$$

2. *We denote by* $W(A)$ *the real-valued number*

$$W(A) = \inf\{ W(i) \mid i \in A\}.$$

3. *We define the bottom* $F(A)$ *of* A *by*

$$F(A) = \{ i \in A \mid W(i) = W(A) \}.$$

The following proposition reveals the strong link between the decomposition tree and the virtual energy.

PROPOSITION 2.12. *Let* $i \in E$. *Then*

$$W(i) = A_c(E) - \sum_{0 \leq k \leq n_E} (H_e^k(i^k) - H_m^k(i^k)),$$

$$\text{where } A_c(E) = \sum_{k=0}^{n_E} \sum_{\Pi \in E^k} (H_e^k(\Pi) - H_m^k(\Pi)) \tag{8}$$

with the convention $H_m^0(i^0) = 0$.

Proof. The proof is split in four steps.

Step 1: Let $g \in G(\{i\})$. There exists an $\{i\}$-graph g' such that $V(g') \leq V(g)$ and such that for all $\Pi \in E^1 \setminus \{i^1\}$, $\mathcal{H}(\Pi, g')$ is true (recall that i^1 is the cycle of order 1 containing i):

$\mathcal{H}(\Pi, g')$: There exists in g' a unique arrow $e \to f$ going out of Π (e.g., $e \in \Pi$ and $f \notin \Pi$). Moreover if $f' \in \Pi \setminus \{e\}$ and $f' \to f'' \in g'$, then $V(f', f'') - H_e(f') = 0$.

This statement can be proved by induction on the number of $\Pi \in E^1 \setminus \{i^1\}$ such that $\mathcal{H}(\Pi, g)$ is false. Assume that there exists $\Pi \in E^1 \setminus \{i^1\}$ such that $\mathcal{H}(\Pi, g')$ is false. Then consider the following procedure:

1. Define $g_0 = g$ and $A_0 = \{e\}$, where e is an exit point of Π for g for which there exists a path $e = e_0 \to \cdots \to e_n = i$ in g satisfying $e_k \notin \Pi$ for $1 \leq k \leq n$.
2. Assume that (g_k, A_k) has been constructed for $k \leq p$. Then define

$$A_{p+1} = \{ f_1 \in \Pi \setminus \cup_{k \leq p} A_k \mid \exists f_2 \in A_p \text{ such that } V(f_1, f_2)) = H_e(f_1) \}.$$

If $A_{p+1} = \emptyset$, then the construction ends with g_p. Otherwise, for all $f_1 \in A_{p+1}$, we choose $f_2 \in A_p$, denoted $l(f_2)$, such that $V(f_1, f_2) = H_e(f_1)$ and define g_{p+1} by

$$g_{p+1} = \{ j \to k \mid j \notin A_{p+1} \text{ and } j \to k \in g_p \} \cup \{ j \to l(j), \mid j \in A_{p+1} \}.$$

Since Π is a cycle, we verify easily that the procedure ends within at most $|\Pi| - 1$ steps and that at each step, $g_p \in G(\{i\})$ and $V(g_p) \leq V(g_{p-1})$. Moreover, the algorithm ends with an $\{i\}$-graph g_{p_0} for which $\mathcal{H}(\Pi, g_{p_0})$ is true.

Step 2: Let g' be the graph obtained by Step 1. For all $\Pi \in E^1 \setminus \{i^1\}$, we denote by $e(\Pi)$ the unique exit point of Π for g' and by $f(\Pi)$ the unique configuration such that $e(\Pi) \to f(\Pi) \in g'$. Now consider the graph g^1 on E^1 defined by

$$g^1 = \{ e(\Pi)^1 \to f(\Pi)^1 \mid \Pi \in E^1 \setminus \{i^1\} \}.$$

One easily sees that $g^1 \in G^1(\{i^1\})$. We can make $V^1(g^1)$ appear in the computation of $V(g')$:

$$\begin{aligned} V(g') &= \sum_{j \in E} H_e(j) + \sum_{\Pi \in E^1 \setminus \{i^1\}} (V(e(\Pi), f(\Pi)) - H_e(e(\Pi))) - H_e(i) \\ &\geq \sum_{j \in E} H_e(j) + V^1(g^1) - \sum_{\Pi \in E^1} H_m^1(\Pi) - \left[H_e(i) - H_m^1(i^1) \right]. \end{aligned}$$

Step 3: Moreover, if $g^1 \in G^1(\{i^1\})$, there exists $g \in G(\{i\})$ such that

$$V(g) = \sum_{j \in E} H_e(j) + V^1(g^1) - \sum_{\Pi \in E^1} H_m^1(\Pi) - \{H_e(i) - H_m^1(i^1)\}. \tag{9}$$

Such an $\{i\}$-graph g can be obtained by the following procedure: first, for all $\Pi \to \Pi' \in g^1$, we choose $e(\Pi) \in \Pi$ and $f(\Pi) \in \Pi'$ such that

$$V(e(\Pi), f(\Pi)) - H_e(e(\Pi)) = \inf_{j \in \Pi, k \in \Pi'} V(j, k) - H_e(j) = V^1(\Pi, \Pi') - H_m^1(\Pi).$$

1. Define $g_0 = \emptyset$, $A_0 = A_0' = \{i\}$, $B_0 = \{i^1\}$, and $\Pi_0 = i^1$.
2. Assume that g_k, A_k, A_k', B_k, and $\Pi_k \in E^1$ have been constructed for $k \leq p$. Define

$$A_{p+1} = \{ f_1 \in \Pi_k \setminus A_p' \mid \exists f_2 \in A_p \text{ such that } V(f_1, f_2) = H_e(f_1) \}.$$

- If $A_{p+1} = \emptyset$ and $B_p = E^1$, then the algorithm is ended.
- If $A_{p+1} = \emptyset$ and $B_p \neq E^1$, then define $D = \{ \Pi \notin B_p \mid f(\Pi)^1 \in B_p \}$. Choose an element $\Pi_{p+1} \in D$ ($D \neq \emptyset$) and define $A_{p+1} = A_{p+1}' = \{ e(\Pi_{p+1}) \}$, $B_{p+1} = B_p \cup \{\Pi_{p+1}\}$. The graph g_{p+1} is then defined by $g_{p+1} = g_p \cup \{ e(\Pi_{p+1}) \to f(\Pi_{p+1}) \}$.
- If $A_p \neq \emptyset$, define $\Pi_{p+1} = \Pi_p$, $B_{p+1} = B_p$, $A_{p+1}' = A_p' \cup A_{p+1}$, and for all $j \in A_{p+1}$, choose an element $l(j) \in A_p$ such that $V(j, l(j)) = H_e(j)$. The graph g_{p+1} is then defined by $g_{p+1} = g_p \cup \{j \to l(j) \mid j \in A_{p+1}\}$.

One easily verifies that this procedure ends within $|E| - 1$ steps. Moreover, when $A_{p+1} = \emptyset$, we have

$$V(g_p) = \sum_{\Pi \in B_p} \sum_{j \in \Pi} H_e(j) + \sum_{\Pi \in B_p \setminus \{i^1\}, \Pi \to \Pi' \in g^1} (V^1(\Pi, \Pi') - H_m^1(\Pi)) - H_e(i),$$

so (9) is verified for the final graph which belongs to $G(\{i\})$.

Step 4: We deduce from the previous steps that

$$W(i) = \sum_{j \in E} H_e(i) - \sum_{\Pi \in E^1} H_m^1(\Pi) + W^1(i^1) - (H_e(i) - H_m^1(i^1)),$$

where W^k is defined as W with the $\{i^k\}$-graphs on E^k valued by V^k.

From the recursive construction of the family (E^k, V^k), we finally arrive at

$$W(i) = \sum_{k=0}^{n_E} \sum_{\Pi \in E^k} (H_e^k(\Pi) - H_m^k(\Pi)) + W^{n_E}(i^{n_E}) - \sum_{\Pi \in E^{n_E}, \Pi \neq i^{n_E}} H_e^{n_E}(\Pi) - \left(\sum_{k=0}^{n_E} H_e^k(i^k) - H_m^k(i^k) \right).$$

One easily verifies that $W^{n_E}(i^{n_E}) = \sum_{\Pi \in E^{n_E} \setminus \{i^{n_E}\}} H_e^{n_E}(\Pi)$, so the proof of the proposition is ended. □

COROLLARY 2.13. *Let* Π *be a cycle. Then for all* $f \in F(\Pi)$ *we have*

1. $H_e^k(f^k) = H_m^{k+1}(f^{k+1})$ *for* $0 \leq k \leq n_\Pi - 1$,
2. $H_e(\Pi) = H_e(f) + \sum_{k=1}^{n_\Pi} (H_e^k(f^k) - H_m^k(f^k))$.

Proof. From Proposition 2.12, we can write

$$W(j) = A_c(E) - H_e(j) - \sum_{k=1}^{n_E} (H_e^k(j^k) - H_m^k(j^k)).$$

Assume that $j \in \Pi$; then we have

$$W(j) = A_c(E) - H_e(j) - \sum_{k=n_{\Pi,j}+1}^{n_E} (H_e^k(j^k) - H_m^k(j^k)) - \sum_{k=1}^{n_{\Pi,j}} (H_e^k(j^k) - H_m^k(j^k)).$$

One verifies that $n_{\Pi,j} = n_\Pi$ for all $j \in \Pi$. Hence, since $j^{n_\Pi} = \Pi$, the last summation is the only one depending effectively on j for $j \in \Pi$. Thus if $f \in F(\Pi)$, we have

$$H_e(f) + \sum_{k=1}^{n_\Pi} (H_e^k(f^k) - H_m^k(f^k)) = \sup_{j \in \Pi} H_e(j) + \sum_{k=1}^{n_\Pi} (H_e^k(j^k) - H_m^k(j^k)).$$

A straightforward computation gives for $j \in \Pi$

$$H_e(j) + \sum_{k=1}^{n_\Pi} (H_e^k(j^k) - H_m^k(j^k)) = H_e(\Pi) - \sum_{k=1}^{n_\Pi - 1} (H_m^{k+1}(j^{k+1}) - H_e^k(j^k)).$$

Since $H_m^{k+1}(j^{k+1}) - H_e^k(j^k) \geq 0$ we deduce that

$$H_e(j) + \sum_{k=1}^{n_\Pi} (H_e^k(j^k) - H_m^k(j^k)) \leq H_e(\Pi),$$

where the equality is reached for $j_0 \in \Pi$ satisfying

$$H_e^k(j_0^k) = H_m^{k+1}(j_0^{k+1}) \text{ for } 0 \leq k \leq n_\Pi - 1; \tag{10}$$

hence, we obtain points 1. and 2. □

COROLLARY 2.14. *Let* $\Pi \in \mathcal{C}(E)$, $|\Pi| > 1$. *There exists a nonnegative constant* $A_c(\Pi)$ *called the altitude of* Π *such that for all* $\Pi' \in \mathcal{M}_*(\Pi)$, *we have*

$$W(\Pi') + H_e(\Pi') = A_c(\Pi). \tag{11}$$

If Π *is a singleton, we define* $A_c(\Pi) = W(\Pi)$.

Proof. Let Π be a cycle of E which is not a singleton. Consider $\Pi' \in \mathcal{M}_*(\Pi)$ and $f \in F(\Pi')$:

$$W(\Pi') + H_e(\Pi') = W(f) + H_e(f) + \sum_{k=1}^{n_{\Pi',f}} (H_e^k(f^k) - H_m^k(f^k))$$

$$= A_c(E) - \sum_{k=n_{\Pi'}+1}^{n_E} (H_e^k(f^k) - H_m^k(f^k)).$$

However, from Remark 1 it follows that for all $k \in \mathbb{N}$ verifying $n_{\Pi'} + 1 \leq k \leq n_\Pi$ we have $H_e^k(f^k) - H_m^k(f^k) = 0$. Hence, one has

$$W(\Pi') + H_e(\Pi') = A_c(E) - \sum_{k=n_\Pi}^{n_E} (H_e^k(f^k) - H_m^k(f^k)),$$

so we get the result since the summation is now independent of Π'. □

Considering $\Pi' \in \mathcal{M}_*(\Pi)$ such that $H_e(\Pi') = H_m(\Pi)$, we get from equation (11) that $W(\Pi') = W(\Pi)$ and

$$A_c(\Pi) = W(\Pi) + H_m(\Pi). \tag{12}$$

Both equations (11) and (12) reveal that potentials, exit, and mixing heights of the cycles are linked through the altitudes of communication. We will now show that we can deduce from (11) and (12) a top-to-bottom computation of the values of $W(\Pi)$ and $A_c(\Pi)$ for all the cycles as well as the organization of the decomposition tree in a valued graph called the decomposition diagram.

Assume that the quantities $H_e(\Pi)$ and $H_m(\Pi)$ have been computed during the cycle decomposition process.

- We start with the computation of $A_c(E)$ with the help of equation (8):

$$A_c(E) = \sum_{\Pi \in \mathcal{C}(E) \setminus \{E\}} H_e(\Pi) - \sum_{\Pi \in \mathcal{C}(E) \setminus \{E\}} H_m(\Pi)$$

  and obtain $W(E)$ from (12).
- Assume that $A_c(\Pi)$ has been computed for a given cycle Π; then if $\Pi' \in \mathcal{M}_*(\Pi)$, we get $W(\Pi')$ from equation (11) and $A_c(\Pi')$ from equation (12).

Therefore, starting from the cycle E, one can get step by step all the values of $A_c(\Pi)$ and $W(\Pi)$, going from top to bottom in the decomposition tree. This has been done in our example: in Fig. 3, a cycle is represented either with an horizontal bar (and in this case the cycle configurations are given by the usual lines of descent) or with a black point for the cycles reduced to singletons. The mixing height is given by the difference between the horizontal bar associated with the concerned cycle and its lowest configuration. Finally, the exit height is given by the height difference between the horizontal bar immediately above the bar of the concerned cycle and its lowest configuration. We have reported on the figure the values of $H_m(\{g, h, i\})$ and $H_e(\{g, h, i\})$. In our example, we obtain for the points of E the following virtual energies:

Configuration:	a	b	c	d	e	f	g	h	i
$W - W(E)$:	4	3	2	2	2	3	2	3	0

We see that configuration 9 is the global minimum of the virtual energy so that the generalized simulated annealing converges to it for proper cooling schedules.

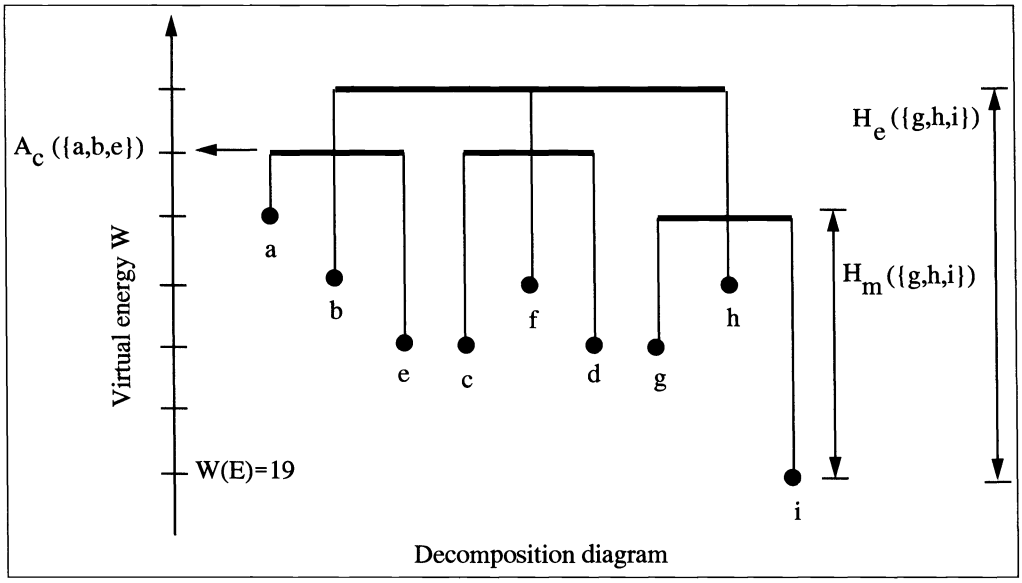


FIG. 3. *Decomposition diagram.*

2.5. Virtual energy and communication altitude. We will extend the notion of altitude introduced in Corollary 2.14 to the definition of the communication altitude between two points. We will see that this function contains in fact all the necessary information to compute the virtual energy (Theorem 2.17) and the decomposition in cycles (Proposition 2.20). Its study will allow us to shed new light on the weak reversibility condition of Hajek (Theorem 2.18).

2.5.1. Characterization of the virtual energy through the communication altitude.

DEFINITION 2.15. *Let $i, j \subset E$. We define the communication altitude from i to j by*

$$A_c(i, j) = \inf_{g \in Pth_E(i,j)} \sup_{0 \le k < n_g} (W(g_k) + V(g_k, g_{k+1})).$$

We adopt the convention that the sup is equal to $W(g_0)$ on a path of length 0.

The link with the definition of the altitude previously introduced in Corollary 2.14 is clarified by the following proposition.

PROPOSITION 2.16.

1. *Let $i, j \in E$; we have*

$$A_c(i, j) = A_c(j, i) = A_c(\Pi_{ij}),$$

where Π_{ij} is the smallest cycle for the inclusion containing i and j.

2. *Let $\Pi \in C(E)$; then*

$$A_c(\Pi) = \sup_{i,j \in \Pi} A_c(i, j).$$

Proof. Point 2 is a straightforward corollary of point 1. We consider now the first point. We prove first the following result.

Let Π be a cycle, $a \in \Pi$, and $b \notin \Pi$. Then we have

$$V(a, b) + W(a) \ge W(\Pi) + H_e(\Pi). \tag{13}$$

We get from Proposition 2.12

$$(14)\quad W(a) - W(\Pi) - H_e(\Pi) + V(a,b) = V(a,b) - \sum_{r=0}^{n_\Pi} H_e^r(a^r) - H_m(a^r)$$
$$= \sum_{r=0}^{n_\Pi - 1} \left(V^r(a^r, b^r) - V^{r+1}(a^{r+1}, b^{r+1}) + H_m^{r+1}(a^{r+1}) - H_e^r(a^r)\right)$$
$$+ \left(V^{n_\Pi}(a^{n_\Pi}, b^{n_\Pi}) - H_e^{n_\Pi}(a^{n_\Pi})\right).$$

We verify easily that each term between parentheses is nonnegative, thus we get (13).

Now consider $i, j \in E$. If $i = j$, then we have $A_c(i,i) = W(i) = A_c(\{i\})$ and thus the result. Assume then that $i \neq j$, and consider $\Pi \in \mathcal{M}_*(\Pi_{ij})$ such that $\Pi \ni i$, where Π_{ij} is the smallest cycle for the inclusion containing i and j. For all paths $g \in \mathrm{Pth}_E(i,j)$, we define $r_g = \inf\{\, 0 \leq k \leq n_g \mid g_k \notin \Pi \,\}$. We have from (14) with $a = i$ and $b = j$

$$\sup_{0 \leq k < n_g} (W(g_k) + V(g_k, g_{k+1})) \geq W(g_{r_g - 1}) + V(g_{r_g - 1}, g_{r_g}) \geq W(\Pi) + H_e(\Pi) = A_c(\Pi_{ij}).$$

Hence we obtain

$$A_c(i,j) \geq A_c(\Pi_{ij}).$$

We will now prove the reverse inequality,

$$(15)\qquad A_c(i,j) \leq A_c(\Pi_{ij}),$$

by induction on $n_{\Pi_{ij}}$. If $n_{\Pi_{ij}} = 0$ then the result is trivial since $i = j$. Now assume that $n_{\Pi_{ij}} = p + 1$ and that (15) is proved for $n_{\Pi_{ij}} \leq p$. There exists $g^p \in \mathrm{Pth}_{E^p}(i^p, j^p)$ such that $V^p(g_k^p, g_{k+1}^p) - H_e^p(g_k^p) = 0$ for all $0 \leq k < n_{g^p}$. Moreover, for all $0 \leq k < n_{g^p}$, we define $a_k \in g_k^p$ and $b_k \in g_{k+1}^p$, satisfying for all $0 \leq r < p - 1$

$$V^{r+1}(a_k^{r+1}, b_k^{r+1}) = H_m^{r+1}(a^{r+1}) + V^r(a_k^r, b_k^r) - H_s^r(a_k^r).$$

Then we get from the equalities (14)

$$V(a_k, b_k) + W(a_k) = H_e(a_k^p) + W(a_k^p) = A_c(\Pi_{ij}).$$

However, defining $b_{-1} = i$ we get from the induction hypothesis that $A_c(b_{k-1}, a_k) \leq A_c(\Pi_{ij})$, so the proposition is proven for $n_{\Pi_{ij}} = p + 1$. □

The symmetry of the communication altitude function ($A_c(i,j) = A_c(j,i)$) proven in Proposition 2.16 is in fact a characterization up to an additive constant of the virtual energy W, as shown in the following theorem.

THEOREM 2.17. *Let $W' : E \to \mathbb{R}$ and $A_{W'} : E \times E \to \mathbb{R}$ such that for all $i, j \in E$, $i \neq j$ we have*

$$A_{W'}(i,j) = \inf_{g \in Pth_E(i,j)} \left(\sup_{0 \leq k < n_g} W'(g_k) + V(g_k, g_{k+1}) \right).$$

If $A_{W'}$ satisfies $A_{W'}(i,j) = A_{W'}(j,i)$ for all $i, j \in E$, $i \neq j$, then there exists $c \in \mathbb{R}$ such that for all $i \in E$

$$W'(i) = W(i) + c.$$

Proof. We will prove by induction on the size of the cycles of E that for all $\Pi \in \mathcal{C}(E)$, there exists $c(\Pi) \in \mathbb{R}$ such that

$$W'(i) = W(i) + c(\Pi), \qquad i \in \Pi.$$

The result is trivial for the singletons. Let $n \geq 1$ and assume that the result is proved for the cycles whose size is smaller or equal to n. Let Π be a cycle of E whose size is $n+1$. We note $(\Pi_l)_{0 \leq l \leq r}$, the family of the elements of $\mathcal{M}_*(\Pi)$, and assume that there exists $c \in \mathbb{R}$ and $0 < l_0 \leq r$ such that $c(\Pi_{l'}) < c < c(\Pi_l)$ for $0 \leq l < l_0 \leq l' \leq r$. Since Π is a cycle, there exists $l' \geq l_0 > l$, $i \in \Pi_{l'}$, and $j \in \Pi_l$ such that $W(i) + V(i, j) = A_c(\Pi)$. Hence

$$A_{W'}(i, j) \leq W(i) + c(\Pi_{l'}) + V(i, j) < A_c(\Pi) + c.$$

However,

$$\begin{aligned} A_{W'}(j, i) &\geq \inf_{a \in \Pi_l,\ b \in \Pi_{l'}} W'(a) + V(a, b) \geq \inf_{a \in \Pi_l,\ b \in \Pi_{l'}} W(a) + c(\Pi_l) + V(a, b) \\ &\geq \inf_{a \in \Pi_l,\ b \in \Pi_{l'}} A_c(a, b) + c(\Pi_l) > A_c(\Pi) + c, \end{aligned}$$

so

$$A_{W'}(j_0, i) - A_{W'}(i, j_0) > 0,$$

which is in contradiction with the symmetry of $A_{W'}$. Hence, all the $c(\Pi_l)$ are equal, and the result is proven for Π. □

2.5.2. Hajek's weak reversibility condition revisited. Theorem 2.17 gives us a new characterization of the virtual energy. In this section, we return to the sequential framework and will deduce from Theorem 2.17 the exact status of the weak reversibility condition of Hajek [7].

In the sequential framework, the communication cost V depends on the energy function U and the communication kernel q and is defined by

$$V(i, j) = \begin{cases} (U(j) - U(i))^+ & \text{if } q(i, j) > 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{16}$$

Moreover, for all $i, j \in E$, $i \neq j$, we define

$$\mathrm{Pth}_q(i, j) = \{g \in \mathrm{Pth}_E(i, j) \mid \forall 0 \leq l < n_g \; q(g_k, g_{k+1}) > 0\}$$

and note

$$D_q(i, j) = \inf_{g \in \mathrm{Pth}_q(i,j)} \sup_{0 \leq k \leq n_g} U(g_k).$$

The weak reversibility condition of Hajek is equivalent to the symmetry of D_q:

$$D_q(i, j) = D_q(j, i), \; i, j \in E. \tag{17}$$

It is well known that for an arbitrary choice of the communication kernel, the energy U is not necessarily the virtual energy associated with V. However, if we assume that the symmetry of the communication kernel q ($q(i, j) = q(j, i)$), then U is the virtual energy up to an additive constant. In [8], Hwang and Sheu prove that the weak reversibility condition of Hajek also is sufficient. In fact, we can go further and deduce from Proposition 2.16 and Theorem 2.17 that Hajek's condition is a necessary and sufficient condition on q to have the equality (up to an additive constant) of U and W.

THEOREM 2.18. *Let V be the communication cost associated with a sequential simulated annealing algorithm (i.e., $V(i,j) = (Uj) - U(i))^+$ for $i \neq j$ and $q(i,j) > 0$), where U is the underlying energy function. Then U is the virtual energy W (up to an additive constant) if and only if the irreducible communication kernel q satisfies the weak reversibility condition* (17) *of Hajek.*

Proof. It is sufficient to note that if A_U is defined by

$$A_U(i,j) = \inf_{g \in \mathrm{Pth}_E(i,j)} \sup_{0 \leq k < n_g} U(g_k) + V(g_k, g_{k+1}), \qquad i, j \in E,$$

then $A_U(i,j) = D_q(i,j)$ since we have

$$U(i') + V(i',j') = \begin{cases} U(i') \vee U(j') & \text{if } q(i',j') > 0, \\ +\infty & \text{if } q(i,j) = 0. \end{cases}$$

Hence, if U is the virtual energy (up to an additive constant) associated with V, then $A_U(i,j) = A_c(i,j) + c$, where $A_c(i,j)$ is the communication altitude from i to j, so that we deduce from point 1 of the Proposition 2.16 that the weak reversibility condition of Hajek (17) is verified. Assume that D_q verifies (17). Then $A_U(i,j) = A_U(j,i)$, so we deduce from Theorem 2.17 that U is the virtual energy up to an additive constant. □

2.5.3. Decomposition in cycle via the communication altitude. We end this section with a proposition which shows that the decomposition in cycles can be computed directly from the values of the communication altitude between configurations. This new characterization of the decomposition in cycles is very close to the definition of the cycles for the sequential case [2] and is its natural extension.

DEFINITION 2.19. *Let $h \in \mathbb{R}$. We note $\mathcal{R}_h$, the equivalence relation defined by*

$$i \mathcal{R}_h j \text{ iff } \begin{cases} i = j \\ \text{or} \\ A_c(i,j) \leq h \quad \text{otherwise.} \end{cases}$$

PROPOSITION 2.20. *We have*

$$\mathcal{C}(E) = \bigcup_{h \in \mathbb{R}} E/\mathcal{R}_h,$$

where $E/\mathcal{R}_h$ denotes the set of the equivalence classes of E for the relation $\mathcal{R}_h$.

Proof. Let $\Pi \in \mathcal{C}(E)$. We want to prove here that $\Pi \in E/\mathcal{R}_h$ for an $h \in \mathbb{R}$. The result is trivial if either Π is a singleton or $\Pi = E$. We consider now the remaining cases and will prove that we can choose $h = A_c(\Pi)$. If $\Pi \notin E/\mathcal{R}_{A_c(\Pi)}$, then there exist $i \in \Pi$ and $j \in \Pi^c$ such that $A_c(i,j) \leq A_c(\Pi)$. Let Π' be the smallest cycle containing i and j. We get from Proposition 2.16 that $A_c(i,j) = A_c(\Pi')$ so that $A_c(\Pi') \leq A_c(\Pi)$ with $\Pi \subset \Pi'$ and $\Pi' \neq \Pi$. However, from (11), (12), and Remark 2 we have

$$A_c(\Pi) = W(\Pi) + H_m(\Pi) < W(\Pi) + H_e(\Pi) = A_c(\Pi'),$$

so we get a contradiction.

Conversely, let $C \in E/\mathcal{R}_h$. If either $C = E$ or C is a singleton, then C is obviously a cycle. We consider the remaining cases. Let Π be the greatest cycle for the inclusion among all the cycles included in C. If $\Pi \neq C$, then we consider $i \in \Pi$ and $j \in C \setminus \Pi$. Since i and j are in C, we have $A_c(i,j) \leq h$. Moreover, $A_c(i,j) = A_c(\Pi_{ij})$, where Π_{ij} is the smallest cycle containing i and j, so $A_c(\Pi_{ij}) \leq h$ and $\Pi_{ij} \subset C$, which is impossible. Hence we get that $\Pi = C$. □

2.6. Optimal exponent. In this part, we will only give the central theorem on the convergence of G.S.A. algorithms, which extends Catoni's result previously stated for the sequential simulated annealing.

DEFINITION 2.21.

1. *We call critical height associated with the decomposition in cycles of E for the communication cost V the real-valued number H_1 defined by*

$$H_1 = H_e(E \setminus F(E)) = \sup_{\Pi \in \mathcal{C}(E), \Pi \cap F(E) = \emptyset} H_e(\Pi).$$

2. *We call optimal exponent for reaching $F(E)$ for the communication cost V the real-valued number α_{opt} defined by*

$$\alpha_{opt} = \inf_{\Pi \in \mathcal{C}(E), \Pi \cap F(E) = \emptyset} \frac{W(\Pi) - W(E)}{H_e(\Pi)}.$$

THEOREM 2.22.

1. *For all decreasing cooling schedules $(T_n)_{n \in \mathbb{N}}$ converging to 0 we have*

$$\sup_{i \in E} P(X_n \notin F(E) \mid X_0 = i) \underset{n \to \infty}{\longrightarrow} 0$$

if and only if

$$\sum_{n=0}^{\infty} e^{-H_1/T_n} = +\infty.$$

2. *We assume that $\alpha_{opt} < +\infty$ and that $Q_T(i, j)\cdot$ is a rational expression of the functions $(e^{a/T})_{a \in \mathbb{R}}$ for any $i, j \in E$. There exist two strictly positive constants R_1 and R_2 such that for all integers $n \geq 1$ we have*

$$\frac{R_1}{n^{\alpha_{opt}}} \leq \inf_{T_1 \geq \cdots \geq T_n \geq 0} \sup_{i \in E} P(W(X_n) \neq W_{\min} \mid X_0 = i) \leq \frac{R_2}{n^{\alpha_{opt}}}.$$

Point 1 of the theorem is the extension of the well-known result of Hajek. It has been proven for the generalized simulated annealing by Hwang and Sheu in [10]. Point 2, which gives the optimal convergence rate toward $F(E)$, has been proven by the author in [13, 14].

3. Energy landscape exploration algorithm. In the previous section we have seen the roles of the critical height H_1 and the optimal exponent α_{opt}. Both depend in a nontrivial way on the communication cost V. Computing these quantities even for very small examples is intractable by hand. However, it can be useful to compute systematically these quantities on medium-size examples in order to be able to test some conjectures on examples and to compute exactly the virtual energy as well as the critical height or the optimal exponent. In this part, we propose a constructive approach to calculate their values on a computer through the recursive construction of the cycle decomposition. We will show that our algorithm has a complexity in $O(|E|^3)$, where $|E|$ is the size of the configuration space.

We call attention here to an alternative approach recently proposed by M. Desai, S. Kumar, and P. R. Kumar in [5] for a direct computation of the virtual energy from the communication cost. Their method is actually not given as a directly implementable algorithm, but this could be done in the spirit of our work with the same complexity. However, since they do not make any explicit reference to the decomposition in cycles, their method is limited to the computation of virtual energy and does not, for instance, provide the critical height H_1 or the optimal exponent α_{opt}.

3.1. Construction of the decomposition tree. The main problem of the effective decomposition in cycles is the construction of a graph isomorphic to the valued graph $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{V})$, which is called a decomposition graph and is defined by the following three points.

1. The set $\mathcal{S}$ of the vertices of $\mathcal{G}$ is defined by

$$\mathcal{S} = \bigcup_{k=0}^{n_E+1} E^k \times \{k\}.$$

The subsets $E^k \times \{k\}$ of $\mathcal{S}$ are called the *levels* of $\mathcal{G}$. If $S = E^k \times \{k\}$ and $S' = E^{k'} \times \{k'\}$ are both levels of $\mathcal{G}$, we will say that S' is the level *above* S if $k' = k + 1$. We note S_0, the level $E^0 \times \{0\}$ called the *base* of $\mathcal{G}$.

2. The edge set $\mathcal{A}$ of $\mathcal{G}$ contains the following three types of edges: for all $v = (\Pi, k)$, $w = (\Pi', k') \in \mathcal{S}$,

$v \rightarrow w \in \mathcal{A}$ if $k = k'$ and $V^k(\Pi, \Pi') < +\infty$.
Then we say that w is *a neighbor* of v.

$v \Rightarrow w \in \mathcal{A}$ if $k = k'$ and $V^k(\Pi, \Pi') = H_e^k(\Pi)$.
Then we say that w is the *exit* out of v.

$v \downarrow w \in \mathcal{A}$ if $k = k' + 1$ and $\Pi' \subset \Pi$.
Then we say that v is *the father* of w and that w is *a son* of v.

3. The set $\mathcal{V}$ of the valuations contains the following valuations of the vertices and on the edges:
   (a) For all $v = (\Pi, k) \in \mathcal{S}$

$$\text{mixing\_height}(v) = H_m^k(\Pi),$$
$$\text{exit\_height}(v) = H_e^k(\Pi).$$

   (b) For all $v = (\Pi, k)$ and $w = (\Pi', k')$ in $\mathcal{S}$ such that $v \rightarrow w \in \mathcal{A}$

$$\text{com\_cost}(v, w) = V^k(\Pi, \Pi').$$

One deduces easily from the recursive decomposition in cycles a construction level by level of the decomposition graph from its base. There exists, however, a step in this algorithm which is not obvious. Assume that the level S has been constructed, as have the edges $u \Rightarrow w$ between vertices of S. In order to construct the level S' above S and the son-father edges $(u \downarrow w)$, we have to identify the equivalence classes on S for the equivalence relation $\mathcal{R}$:

$$u \mathcal{R} v \text{ iff } u \overset{*}{\Rightarrow} v \text{ and } v \overset{*}{\Rightarrow} u,$$

where $w \overset{*}{\Rightarrow} w'$ means that there exists $n \in \mathbb{N}$ and a family $(w_l)_{0 \le l \le n}$ of vertices in S such that $w_0 = w$, $w_n = w'$ and the edges $w_l \Rightarrow w_{l+1}$ exist for $0 \le l < n$. However, this problem is equivalent to the computation of the strongly connected components of $G = (S, A)$, where A is the set of edges $u \Rightarrow v$ between the vertices of S.

DEFINITION 3.1. *We say that $C \subset S$ is a strongly connected component of the oriented graph $G = (S, A)$ if both of the following conditions are verified.*

1. *For all $u, v \in C$, we have*

$$u \overset{*}{\Rightarrow} v \text{ and } v \overset{*}{\Rightarrow} u.$$

2. *The set C is maximal among all the subsets of S verifying* 1.

An easy lemma establishes that the equivalence classes for the relation $\mathcal{R}$ in S are exactly the strongly connected component of $G = (S, A)$. However, Robert Tarjan proposes in [11] an algorithm with complexity $O(|S| + |A|)$ to compute the strongly connected components. Hence, we can now define a recursive algorithm to construct the graph $\mathcal{G}$ from the data given by the base S_0, the edges $u \to v$ between the vertices of S_0, and their valuations com_cost(u, v). We give below a pseudocode version.

```
main()
BEGIN
   load the vertices of S0;
   FOR EACH v in S0
      load the edges v→w;
      FOR EACH neighbor w of v
         load com_cost(v, w);
      mixing_height(v) := 0;
   decompose(S0);
END
```

The function **decompose**, which is the main ingredient of the main program, is defined as follows.

```
FUNCTION decompose(S)
BEGIN
   IF (S is not a singleton)
   THEN
      FOR EACH v in S /* Computation of the valuations exit_height */
         exit_height(v):= min{com_cost(v, w) | v→w};
      FOR EACH v in S /* Creation of the edges v⇒w */
         FOR EACH w in S neighbor of v
            IF (exit_height(v) = com_cost(v, w))
            THEN create the edge v⇒w; /* w is an exit out of v */
      /* Creation of the level S' above S and creation of the edges
         son-father between the vertices of S' and those of S */
      next_level(S);
      /* Creation of the edges v→w between the vertices of S' and
         computation of their valuations */
      FOR EACH v in S'
          FOR EACH v' son of v /* e.g. the edge v↓v' exists */
           FOR EACH w' neighbor of v' /* e.g. the edge v'→w' exists */
              IF (the father w of w' is distinct of v)
              THEN
                 tmp:= mixing_height(v)+com_cost(v', w')−exit_height(v');
                 IF the edge v→w does not exists yet
                 THEN
                    create the edge v→w;
                    com_cost(v, w):= tmp;
                 ELSE com_cost(v, w):= min{ com_cost(v, w) , tmp };
      /* Decomposition of the level S' above S */
      decompose(S');
END
```

We have intentionally isolated the function **next_level** since this function is nothing but the Tarjan's algorithm for the search of the strongly connected components in an oriented graph.

In fact, we have to distinguish the minimal components from the others, but this can be easily achieved since this last problem reduces to finding the leaves in a tree.

Our construction algorithm of the decomposition graph $\mathcal{G}$ is completely recursive, so we obtain a simple implementation. From the complexity point of view, for each level, the complexity of the construction of the next level is dominated by the complexity of the Tarjan's algorithm, that is, $O(|S| + |A|)$, where S is the family of vertices of the current level and A is the family of the edges $u \Rightarrow v$ between the vertices of S. This complexity admits an upper bound in $O(|S| + |B|)$, where B is the family of the edges $u \rightarrow v$ between the vertices of S. However, if S' is the level above S and B' is the family of the edges $u \rightarrow v$ between the vertices of S', we have

$$|S'| \leq |S| - 1 \text{ and } |B'| \leq |B| - 1.$$

Since at each level we have the upper bound $|B| \leq |S|^2$, we deduce that the total complexity of the algorithm is in $O(|S_0|^3)$, where S_0 is the base of $\mathcal{G}$. This value of the complexity shows clearly the difficulty of a computation by hand but also the limitation of this approach even on a computer.

3.2. Derivation of the virtual energy and the critical constants. It is now simple to deduce from the graph $\mathcal{G}$ the value of the interesting quantities. We start here by the computation of the virtual energy W. From Corollary 2.14 we deduce that for all $i \in E$ and all $k \leq n_E$ we have the relation

$$W(i^k) = W(i^{k+1}) + H_m^{k+1}(i^{k+1}) - H_e^k(i^k). \tag{18}$$

Thus, on each vertex $v = (\Pi, k)$ of $\mathcal{S}$, if we define the valuation called virtual_energy(v) by

$$\text{virtual\_energy}(v) = W(\Pi),$$

we deduce from (18) that if w is a son of v (e.g., the edge $v \downarrow w$ exists) then

$$\text{virtual\_energy}(w) = \text{virtual\_energy}(v) + \text{mixing\_height}(v) - \text{exit\_height}(w). \tag{19}$$

The relation (19) allows the computation of virtual_energy(v) for each vertex v of $\mathcal{S}$ recursively from the value of virtual_energy(v_∞), where v_∞ denotes the unique vertex of the last level of $\mathcal{G}$ (e.g., $v_\infty = (E, n_E + 1)$). However, virtual_energy(v_∞) is equal to $W(E) = \inf W$, which can be arbitrarily fixed to the value 0 since it is sufficient to compute the virtual energy up to an additive constant. (One should in fact compute the constant with a recursive algorithm.)

We are now interested in the computation of the bottom $F(E)$ of E (global minima of W). In the framework of $\mathcal{G}$, we have to find the vertices v of the base S_0 of $\mathcal{G}$ for which virtual_energy(v) is minimal. This computation can be achieved without having to compute the virtual energy with a recursive procedure. (We use here the characterization of $F(E)$ given by point 1 of Corollary 2.13.) We will note $F(\mathcal{G})$, the family of the vertices $v = (\{i\}, 0)$ of $\mathcal{S}$ for $i \in F(E)$.

```
bottom_exploration(v∞);
FUNCTION bottom_exploration(v)
BEGIN
   IF (v has no son) /* v ∈ S0  */
   THEN add v to the list;
   ELSE
      FOR EACH son w of v
         IF (exit_height(w) = mixing_height(v))
         THEN bottom_exploration(w);
END
```

For the computation of the critical height $H_1 = H(E \setminus F(E))$ it is sufficient to take the maximal value of exit_height(v) for any vertex v which does not have a descendant in $F(\mathcal{G})$. We obtain the following recursive procedure:

```
critical_height:= 0;
compute_critical_height(v∞, critical_height);
FUNCTION compute_critical_height(v, critical_height);
BEGIN
   IF (v has at least one son)
   THEN
      FOR EACH son w of v
         IF (exit_height(w) < mixing_height(v))
         THEN critical_height:= sup{ critical_height , exit_height(w) };
         ELSE compute_critical_height(w, critical_height);
END
```

Finally, we give here a recursive procedure to compute the optimal exponent α_{opt}. We recall its definition,

$$\alpha_{opt} = \inf_{\Pi \in C(E),\ \Pi \cap F(E)} \frac{W(\Pi) - W(E)}{H_e(\Pi)}.$$

In fact, it is not necessary to consider all the cycles. Indeed, if Π is a cycle such that $\Pi \cap F(E) = \emptyset$, then for any cycle $\Pi' \subset \Pi$, we have $W(\Pi') \geq W(\Pi)$ and $H_e(\Pi') \leq H_e(\Pi)$ so that we get the inequality

$$\frac{W(\Pi') - W(E)}{H_e(\Pi')} \geq \frac{W(\Pi) - W(E)}{H_e(\Pi)}.$$

Now, for any cycle Π distinct of E, we note Π^+, the smallest cycle containing Π and distinct from Π. From the previous remark, it is sufficient for computing α_{opt} to consider the cycles Π such that $\Pi \cap F(E) = \emptyset$ and $\Pi^+ \cap F(E) \neq \emptyset$. For such cycles, we have $W(\Pi) + H_e(\Pi) = W(E) + H_m(\Pi^+)$ so that

$$\frac{W(\Pi) - W(E)}{H_e(\Pi)} = \frac{H_m(\Pi^+) - H_e(\Pi)}{H_e(\Pi)}.$$

Thus we can compute directly from the decomposition in cycles the value of α_{opt} without computing before the virtual energy. We propose the following procedure:

```
alpha_opt:= +∞;
compute_alpha_opt(v∞,alpha_opt);
FUNCTION compute_alpha_opt(v,alpha_opt);
BEGIN
   IF (v has at least one son)
   THEN
      FOR EACH son w of v
         IF (exit_height(w) < mixing_height(v))
         THEN
            tmp:= (mixing_height(v) - exit_height(w))/exit_height(w);
            alpha_opt:= min{ tmp , alpha_opt };
         ELSE compute_alpha_opt(w,alpha_opt);
END
```

3.3. Examples. We have used this decomposition algorithm to evaluate the efficiency of a parallel scheme of simulated annealing on spin-glass energies (see [12] for a presentation of this scheme). Let us briefly present the setting. We consider the configuration space $E = \{-1, 1\}^S$, where S is a finite set called the set of sites, and a spin-glass-like energy

$$U(x) = \sum_{s \neq t \in S} J_{s,t} x_s x_t + \sum_{s \in S} h_s x_s,$$

where the couplings $J_{\{s,t\}}$ are independently and identically distributed (i.i.d.) random variables with normal distribution $\mathcal{N}(0, 1)$ and the h_s are i.i.d. random variables with uniform distribution on [0, 1]. This provides us with a large class of energies with a statistical structure relevant to real minimization problems. On a sample of 200 energies, we have performed the cycle decompositions for the communication costs associated with the sequential simulated annealing and the parallel scheme [13]. The values of H_1 and α_{opt} have been computed also for a comparison. We give below the central processing unit (cpu) time (in seconds) on a Sun 4/65 corresponding to our sample for three values of $|S|$:

$\|S\| = 4$	10 s
$\|S\| = 6$	189 s
$\|S\| = 8$	4778 s

As expected from the complexity study, the cpu time is exponentially increasing with $|S|$, so we have been confined in our experimental work to $|S| \leq 10$, that is, $|E| \leq 1024$. (We see that the average time for a cycle decomposition is about 12 s for $|E| = 2^8 = 256$.) Beyond the attractiveness of an exact calculation of the critical constants, one of the appealing aspects of the explicit construction of the cycle decomposition is to give an efficient numerical tool to progress in the badly known cycle structure of random energies.

REFERENCES

[1] R. AZENCOTT, *Simulated Annealing: Parallelization Techniques*. John Wiley and Sons, New York, 1992.

[2] O. CATONI, *Rough large deviation estimates for simulated annealing: Application to exponential schedules*, Ann. Probab., 20 (1992), pp. 1109-1146.

[3] T.-S. CHIANG AND Y. CHOW, *A limit theorem for a class of inhomogeneous Markov processes*, Ann. Probab., 17 (1989), pp. 1483-1502.

[4] ———, *Asymptotic behavior of eigenvalues and random updating schemes*, Appl. Math. Optim., 28 (1993), pp. 259-275.

[5] M. DESAI, S. KUMAR, AND P. R. KUMAR, *Quasi-statistically Cooled Markov Chains*, preprint, 1992.

[6] M. I. FREIDLIN AND A. D. WENTZELL, *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York, 1984.

[7] B. HAJEK, *Cooling schedule for optimal annealing*, Math. Oper. Res., 13 (1988), pp. 311-329.

[8] C.-R. HWANG AND S.-J. SHEU, *On the Weak Reversibility Condition in Simulated Annealing*, Preprint, Institute of Mathematics, Academia Sinica, Taipei, Taiwan, 1988.

[9] ———, *Large time behavior of perturbated diffusion Markov processes with applications to the second eigenvalue problem for Fokker-Planck operators and simulated annealing*, Acta Appl. Math., 19 (1990), pp. 253-295.

[10] ——— *Singular perturbed Markov chains and exact behaviors of simulated annealing process*, J. Theoret. Probab., 5 (1992), pp. 223-249.

[11] R. TARJAN, *Depth-first search and linear graph algorithms*, SIAM J. Comput., 1 (1972), pp. 146-160.

[12] A. TROUVÉ, *Partially parallel simulated annealing: Low and high temperature approach of the invariant measure*, in I. Karatzas and D. Ocone, eds., Proceedings Volume of the US-French Workshop on Applied Stochastic Analysis (Rutgers University, 29 April-2 May 1991), Lecture Notes in Control and Inform. Sci. 177, Springer-Verlag, New York, 1992.

[13] ———, *Parallélisation massive du recuit simulé*, Ph.D. thesis, Université d'Orsay, Jan. 1993.

[14] ———, *Rough large deviation estimates for the optimal convergence speed exponent of generalized simulated annealing algorithm*, Ann. Inst. H. Poincaré Probab. Statist., 32 (1996).

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 987–1000, May 1996

011

A CHARACTERIZATION OF BOUNDED-INPUT BOUNDED-OUTPUT STABILITY FOR LINEAR TIME-INVARIANT SYSTEMS WITH DISTRIBUTIONAL INPUTS*

CHI-JO WANG† AND J. DANIEL COBB†

Abstract. We consider linear time-invariant operators defined on the space of distributions with left-bounded support. We argue that in this setting the convolution operators constitute the most natural choice of objects for constructing a linear system theory based on the concept of impulse response. We extend the classical notion of bounded-input bounded-output stability to distributional convolution operators and determine precise conditions under which systems characterized by such maps are stable. A variety of expressions for the "gain" of a stable system is derived. We show that every stable system has a natural threefold decomposition based on the classical decomposition of functions of bounded variation. Our analysis involves certain extensions of the Banach spaces L^p in the space of distributions.

Key words. linear systems, stability, distributions

AMS subject classification. 93

1. Introduction. The concept of impulse response has traditionally played a central role in linear system theory. In spite of this fact, certain fundamental system-theoretic ideas have apparently not been developed on a mathematically rigorous level for systems with arbitrary distributional inputs and outputs. In particular, an extract characterization of the impulse and step responses that correspond to bounded-input bounded-output (BIBO) stable systems has not previously appeared in the literature. As an illustration of the problem, recall that if a linear time-invariant system is described by convolution of its inputs with a measurable function h, then the system is BIBO stable if and only if $h \in L^1$. (See, e.g., [1, p. 388].) This characterization is inadequate, however, for studying classes of systems where h may be a distribution since even a simple all-pass system has impulse response $\delta \notin L^1$. Obtaining a complete description of stable distributional systems is the primary goal of this paper.

A somewhat more limited framework than ours that addresses this problem appears in [2, p. 108], where systems are viewed as convolution operators and the impulse response is restricted to be a measurable function plus a linear combination of time shifts of the unit impulse. (In [2] the time-varying case is also included.) Thus, systems that differentiate the input are not included in [2], nor are more exotic cases such as the examples we present in §5. Our framework includes that of [2] (restricted to the time-invariant setting) and gives a more general framework for linear systems and, in particular, BIBO stable systems.

In §2, we consider the problem of meaningfully characterizing linear time-invariant systems in terms of their impulse responses. To set the stage for stability analysis, in §3 we pose and solve the problem of extending the Banach spaces L^p in the space of distributions for $1 \leq p \leq \infty$. In §4 we define BIBO stability for convolution operators on distribution space and obtain exact conditions on the impulse and step responses of a BIBO stable system. Expressions for the induced norm or "gain" of a stable operator with bounded inputs are also established. Section 5 contains a discussion of a three-fold decomposition applicable to all BIBO stable impulse responses. Our results are summarized in §6.

2. Preliminaries. We need a brief introduction to the theory of distributions. (See [3]–[6].) If $\varphi : \mathbb{R} \to \mathbb{R}$, define the support of φ, i.e., supp φ, as the closure of the set $\{t \mid \varphi(t) \neq 0\}$,

*Received by the editors May 3, 1993; accepted for publication (in revised form) January 16, 1995. This research was supported by NSF grant ECS-8920081.

†Department of Electrical and Computer Engineering, University of Wisconsin-Madison, 1415 Johnson Drive, Madison, WI 53706-1691.

and let $\sigma_\tau\varphi$ be the translation of φ defined by $(\sigma_\tau\varphi)(t) = \varphi(t-\tau)$. Let K be the space of C^∞ functions $\varphi : \mathbb{R} \to \mathbb{R}$ with supp φ bounded, and let K' be the dual space of K. (See [6] for an exact description of the topology of K.) A distribution f is an element of K', i.e., a continuous linear functional $\varphi \to \langle f, \varphi\rangle$ on K. For $f \in K'$, supp f is defined to be the complement of the largest open set $U \subset \mathbb{R}$ such that supp $\varphi \subset U$ implies $\langle f, \varphi\rangle = 0$. We may also define the time shift $\sigma_\tau f$ of a distribution f by $\langle \sigma_\tau f, \varphi\rangle = \langle f, \sigma_{-\tau}\varphi\rangle$ and the derivative $\dot{f}$ of f by $\langle \dot{f}, \varphi\rangle = -\langle f, \dot{\varphi}\rangle$. Denote the ith distributional derivative by $f^{(i)}$. It is easy to show that the time shift and differentiation operators commute and that

$$\frac{d}{d\tau}\langle \sigma_\tau f, \varphi\rangle = \langle \sigma_\tau \dot{f}, \varphi\rangle.$$

The unit impulse δ is defined by $\langle \delta, \varphi\rangle = \varphi(0)$. Also, any function f that is locally L^1 determines a distribution according to $\langle f, \varphi\rangle = \int f\varphi$. (Functions that coincide a.e. are identified.) In this way, we may view functions in L^p as distributions for $1 \le p \le \infty$. In particular, the unit step function θ may be considered a distribution. Define $\delta_\tau = \sigma_\tau\delta$ and $\theta_\tau = \sigma_\tau\theta$. If f is locally L^1 and differentiable a.e. in the classical sense, denote this derivative by f'. It is an important fact that there exist f such that $f \ne \dot{f}$. This may occur in trivial ways (e.g., $\dot{\theta} = \delta$, but $\theta' = 0$ a.e.), but such cases also exist where f is continuous. (See §5.) Define $K'_\tau = \{f \in K' \mid \text{supp } f \subset [\tau, \infty)\}$ and

$$K'_+ = \bigcup_{\tau\in\mathbb{R}} K'_\tau.$$

Convergence in K' is defined via its weak* topology, which has a subbasis consisting of all sets of the form

$$U_{\varphi g} = \{f + g \mid |\langle f, \varphi\rangle| < 1\},$$

where $\varphi \in K$ and $g \in K'$. In terms of convergence, this means that a sequence (or net) f_n converges to f iff $\langle f_n, \varphi\rangle \to \langle f, \varphi\rangle$ for all $\varphi \in K$. Thus a linear operator $T : K'_+ \to K'_+$ is weak* continuous iff $\langle f_n, \varphi\rangle \to 0$ implies $\langle T(f_n), \varphi\rangle \to 0$. A linear operator $T : K'_+ \to K'_+$ is causal iff $\inf(\text{supp } T(f)) \ge \inf(\text{supp } f)$ for all $f \in K'_+$.

We are especially interested in convolution operators; the convolution of any pair $f, g \in K'_+$ is defined as follows. It is shown in [3, p. 100] that the map $\overline{\psi}(t) = \langle g, \sigma_{-t}\varphi\rangle$ defines a C^∞ function. Since φ has bounded support, supp $\overline{\psi}$ is bounded above. Choosing $\psi \in K$ to be any function in K such that $\psi(t) = \overline{\psi}(t)$ for all $t \ge \inf(\text{supp } f)$, we define $\langle f^*g, \varphi\rangle = \langle f, \psi\rangle$. This definition is unambiguous, since $\langle f, \gamma\rangle = 0$ whenever supp $\gamma \cap$ supp $\varphi = \phi$. Convolution can be shown to be commutative and to satisfy $f^*\delta^{(i)} = f^{(i)}$. Also, if $h = f^*g$, $\dot{h} = \dot{f}^*g = f^*\dot{g}$. A *convolution operator* $T : K'_+ \to K'_+$ is an operator of the form $T(u) = h^*u$, where $h \in K'_+$. For any convolution operator, $T(\dot{\theta}) = T(\delta)^*\dot{\theta} = T(\delta)$; T is causal iff $h \in K'_0$. We will often refer to the following basic result from [3, p. 105] concerning continuity of convolution operators.

LEMMA 2.1. *Let T be a convolution operator with $T(v) = h^*v$ for every v, and let $u_n \to u$ be a convergent sequence (or net) in K'_+. If there exists a $\tau \in \mathbb{R}$ such that either* supp $h \subset (-\infty, \tau]$ *or* supp $u_n \subset [\tau, \infty)$ $\forall n$, *then $T(u_n) \to T(u)$.*

Let $C_0 = \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ is continuous}, f(-\infty) = f(\infty) = 0\}$ with norm

$$\|f\|_\infty = \sup_t |f(t)|.$$

We denote by BV the space of functions $f : \mathbb{R} \to \mathbb{R}$ with bounded variation. Set $NBV = \{f \in BV \mid f \text{ is left-continuous}, f(-\infty) = 0\}$, and let $\text{Var}(f)$ be the variation of f. From [7,

Ch. 6], *NBV* is the dual of C_0 with induced norm

$$\|f\|_n = \mathrm{Var}(f).$$

In addition, let $DBV = \{\dot{g} \in K' \mid g \in BV\}$ and define

$$\|\dot{g}\|_d = \mathrm{Var}(g).$$

It is easy to verify that $\|\cdot\|_d$ defines a norm on *DBV* and that *NBV* and *DBV* are isometrically isomorphic under the map $g \to \dot{g}$; hence *DBV* may also be viewed as the dual of C_0. It is easy to show that K is dense in C_0 and that the norm $\|\cdot\|_d$ satisfies

$$\|f\|_d = \sup_{\substack{\varphi \in K \\ \|\varphi\|_\infty = 1}} |\langle f, \varphi\rangle| = \sup_{\substack{\varphi \in K \\ \|\varphi\|_\infty = 1}} \left| \int_{-\infty}^{\infty} \varphi(t) dg(t) \right| \tag{1}$$

for any $f = \dot{g} \in DBV$.

Recall that every $f \in BV$ has a decomposition (unique a.e.) of the form $f = f_1 + f_2 + f_3$, where f_1 is bounded and absolutely continuous, f_2 is a bounded saltus function (i.e., $f_2 = \Sigma \alpha_i \theta_{\tau_i}$, where $\Sigma|\alpha_i| < \infty$), and f_3 is a singular function (i.e., f_3 is continuous and nonconstant, $f_3 \in BV$, and $f_3' = 0$ a.e.).

Define L_+^p, BV_+, and DBV_+ to consist of the L^p, BV, and *DBV* distributions $f \in K'$, respectively, with $\inf(\mathrm{supp} f) > -\infty$. For $p < \infty$, L_+^p is a dense subspace of L^p. On the other hand, the closure of L_+^∞ is a proper subspace of L^∞, namely,

$$L_0^\infty = \left\{ f \in L^\infty \mid \operatorname*{ess\,sup}_{t \in (-\infty, -n]} |f(t)| \to 0 \text{ as } n \to \infty \right\}.$$

Let $L_{[0,\infty)}^\infty$ denote the L^∞ functions f with $\mathrm{supp}\, f \subset [0, \infty)$. Note that $L_{[0,\infty)}^\infty$, L_+^p, BV_+, and DBV_+ may be viewed as subspaces of K_+'.

The first question we address is that of determining which operators $T : K_+' \to K_+'$ can be justifiably called "linear systems." Clearly, T should be linear. Also, since we wish to develop a theory based on the concept of impulse response, we need to establish conditions under which $T(\delta)$ uniquely characterizes the operator T. We will limit ourselves to time-invariant operators, although the results of this section can be generalized considerably. As a first step, we might also impose continuity on T, since continuous linear operators are easier to work with. These constraints and the following lemma lead to Theorem 2.2.

LEMMA 2.2. *Let* $\tau \in \mathbb{R}$ *and*

$$I_\tau = \left\{ f \in K' \mid \exists t_i \geq \tau, \beta_i \text{ such that } f = \sum_{i=1}^{k} \beta_i \delta_{t_i} \right\}.$$

Then I_τ *is weak* dense in* K_τ'.

Proof. Using an overbar to denote weak* closure, it is clear that $I_\tau \subset K_\tau' = \overline{K}_\tau'$, so $\bar{I}_\tau \subset K_\tau'$. We will demonstrate that $\bar{I}_\tau \supset \overline{K}_\tau \supset \overline{K}_{\tau b}' \supset K_\tau'$, where

$$K_\tau = \{\varphi \in K \mid \mathrm{supp}\, \varphi \subset [\tau, \infty)\}, \qquad K_{\tau b}' = \{f \in K_\tau' \mid \mathrm{supp}\, f \text{ is bounded}\}.$$

Let $\varphi \in K$ with $\mathrm{supp}\, \varphi \subset [\tau, \tau + \Delta]$ and define

$$\gamma_n = \sum_{k=0}^{n} \frac{\Delta}{n} \varphi\left(\tau + k\frac{\Delta}{n}\right) \delta\left(t - \tau - k\frac{\Delta}{n}\right).$$

For any $\psi \in K$, $\langle \gamma_n, \Psi \rangle$ is a sequence of Riemann sums converging to $\langle \varphi, \psi \rangle$; hence, $\gamma_n \to \varphi$ weak*. Since φ is arbitrary, $\bar{I}_\tau \supset K_\tau$ and $\bar{I}_\tau \supset \overline{K}_\tau$. Let $\varphi_n \in K_0$, $\varphi_n \to \delta$ weak*, $f \in K'_{\tau b}$, and $\psi_n = f^* \varphi_n$. From Lemma 2.1, $\psi_n \in K_\tau$ and $\psi_n \to f$. Hence, $\overline{K}_\tau \supset K'_{\tau b}$ and $\overline{K}_\tau \supset \overline{K}'_{\tau b}$. Finally, let $g \in K'_\tau$ and let $\eta_n \in K$ satisfy $\eta_n(t) = 1$ for $0 \le t \le n$. Then $\eta_n g \in K'_{\tau b}$ and $\eta_n g \to g$ weak*, so $\overline{K}'_{\tau b} \supset K'_\tau$. □

THEOREM 2.3. *Let $T : K'_+ \to K'_+$ be a weak* continuous, linear, time-invariant operator. Then $T(u) = T(\delta)^* u$ for all $u \in K'_+$.*

Proof. Suppose $u \in K'_\tau$. From Lemma 2.2, for any weak* neighborhood U of u, there exists a $v \in U$ of the form

$$v = \sum_{i=1}^{k} \beta_i \delta_{t_i}$$

with $t_i \ge \tau$ for all i. Let $\varphi \in K$. From linearity and time-invariance of T,

$$\langle T(v), \varphi \rangle = \langle T(\delta), \psi \rangle,$$

where

$$\psi = \sum_{i=1}^{k} \beta_i \sigma_{-t_i} \varphi.$$

Note that $\psi \in K$ and $\psi(t) = \langle v, \sigma_{-t} \varphi \rangle$ for all $t \ge \tau$. Thus

$$\langle T(\delta), \psi \rangle = \langle T(\delta)^* v, \varphi \rangle.$$

Since φ is arbitrary, $T(v) = T(\delta)^* v$. From Lemma 2.1, $T(u) = T(\delta)^* u$. □

Unfortunately, the converse to Theorem 2.3 is false; i.e., a convolution operator may fail to be weak* continuous. From Lemma 2.1, boundedness of $T(\delta)$ is sufficient to guarantee continuity of T. The next result establishes the converse.

THEOREM 2.4. *Let $T : K'_+ \to K'_+$ be a weak* continuous convolution operator. Then* supp $T(\delta)$ *is bounded.*

Proof. Suppose supp $T(\delta)$ is unbounded. Then there exist sequences $\varphi_n \in K$ and $\alpha_n \in \mathbb{R}$ such that supp $\varphi_n \subset [\alpha_n, \alpha_n + 1]$, $\alpha_n \to \infty$, and $\langle T(\delta), \varphi_n \rangle = \beta_n \ne 0$. Let

$$\gamma_n = \max_{\substack{i \le n \\ t \in \mathbb{R}}} \left| \frac{d^i \varphi_i(t)}{dt^i} \right|$$

and $\psi_n = \frac{1}{n\gamma_n} \sigma_{-n} \varphi_n$. From [3, p. 2], $\psi_n \to 0$ in K; hence, $\{\psi_n\}$ is a bounded subset of K. Let

$$f_n = \frac{n\gamma_n}{\beta_n} \delta_{-\alpha_n}.$$

Then $f_n \to 0$ weak*, but

$$\begin{aligned} \sup_m \langle T(\delta)^* f_n, \psi_m \rangle &\ge \langle T(\delta)^* f_n, \psi_n \rangle \\ &= \langle \frac{n\gamma_n}{\beta_n} \sigma_{-\alpha_n} T(\delta), \psi_n \rangle \\ &= \frac{1}{\beta_n} \langle T(\delta), n\gamma_n \sigma_{\alpha_n} \psi_n \rangle \\ &= \frac{1}{\beta_n} \langle T(\delta), \varphi_n \rangle \\ &= 1, \end{aligned}$$

so $T(\delta)^* f_n$ does not converge uniformly to 0 on bounded subsets of K. From [4, pp. 55–56], $T(\delta)^* f_n$ does not converge to 0 weak*, so T is not continuous. □

It follows from Theorem 2.4 that there exist many familiar examples of linear systems that are characterized by weak* discontinuous convolution operators. For example, θ has unbounded support, so a simple integrator is discontinuous. In particular, the sequence $\delta_{-n} \to 0$ weak* as $n \to \infty$, but its integrals θ_{-n} converge to the constant distribution 1. In view of such examples, we choose not to restrict ourselves to weak* continuous operators.

Unfortunately, an arbitrary class of discontinuous operators T in general is not uniquely characterized by the values $T(\delta)$, since $T(\delta)$ only determines the action of a linear time-invariant operator on the proper subspace span $\{\delta_\tau \mid \tau \in \mathbb{R}\} \subset K'_+$. On the other hand, the distributions $T(\delta)$ do uniquely characterize the family of convolution operators. In fact, it is easy to show that $h^*u = 0$ for every $u \in K'_+$ implies $h = 0$, so $T \to T(\delta)$ maps the convolution operators one-to-one onto K'_+. Thus any linear time-invariant nonconvolution operator has the same impulse response as some convolution operator.

Based on these observations, we define a *linear time-invariant system* to be a convolution operator $T : K'_+ \to K'_+$. In the next section, we develop the machinery that will enable us to define and characterize BIBO stability for such systems.

3. Extension of normed linear spaces. In this section we examine the problem of extending L^p in K' for arbitrary p. We do this because L^1 is known to play a role in characterizing BIBO stability of an operator and because L^∞ is used in the definition of stability. Values $p \in (1, \infty)$ are not directly related to stability but can be easily handled along with $p = \infty$ and are therefore included. In fact, the problem can be couched in much more general terms without substantially increasing the level of difficulty.

Let X be a Hausdorff topological vector space over $\mathbb{R}$, and let $Y \subset X$ be a normed linear space. Then Y has two topologies: the norm topology and the one inherited from X. Denote the topology of X by $\mathcal{T}$ (i.e., $\mathcal{T}$ is the family of all open subsets U of X), let $\mathcal{T}_Y$ be the relative topology on Y generated by $\mathcal{T}$ (i.e., $\mathcal{T}_Y$ consists of all sets $U \cap Y$). Also, let $B(y, r) \subset Y$ be the norm ball about y with radius r. We make the following assumptions.

A1) $\forall U \in \mathcal{T}, U \cap Y \neq \phi$.
A2) $\forall U \in \mathcal{T}$ and $\forall y \in U \cap Y, \exists \varepsilon > 0$ such that $B(y, \varepsilon) \subset U$.
A3) $\exists U \in \mathcal{T}$ such that $U \cap Y = Y - B(0, 1)$.

Assumption A1) states that Y is dense in X relative to $\mathcal{T}$. The other two assumptions give upper and lower bounds on $\mathcal{T}_Y$. Assumption A2) states that the norm topology on Y is stronger than or equal to $\mathcal{T}_Y$, while A3) says that $B(0, 1)$ is closed in $\mathcal{T}_Y$.

Suppose Y has norm $\|\cdot\|$, let $x \in X$, and let $\{U_\beta\} \subset \mathcal{T}$ be the family of all neighborhoods of x. Define

$$\|x\|^e = \sup_\beta \inf_{y \in U_\beta \cap Y} \|y\|. \tag{2}$$

In view of A1), $\|x\|^e$ is well defined and determines a function $\|\cdot\|^e : X \to [0, \infty]$. The next result establishes that $\|\cdot\|^e$ is a natural extension of $\|\cdot\|$ to all of X.

PROPOSITION 3.1. 1) $\|\cdot\|$ *and* $\|\cdot\|^e$ *coincide on* Y.

2) $\|\cdot\|^e$ *is lower semicontinuous on* X *relative to* $\mathcal{T}$.

3) *If* $\|x\|^e < \infty$, *then for every* $\varepsilon > 0$ *and* $\mathcal{T}$*-neighborhood* U *of* x *there exists a* $y \in U \cap Y$ *such that*

$$\|y\| < \|x\|^e + \varepsilon.$$

4) *If* $\|\cdot\|^f : X \to [0, \infty]$ *satisfies* 1)–3) *(replacing* e *with* f *throughout), then* $\|\cdot\|^f = \|\cdot\|^e$.

Proof. 1) For $x \in Y$,

$$\inf_{y \in U_\beta \cap Y} \|y\| \leq \|x\|$$

for all β, so $\|x\|^e \leq \|x\|$ follows immediately from (2). Suppose $\|x\|^e < a < \|x\|$. Then (2) states that for every β and $\varepsilon > 0$ there exists a $y \in U_\beta \cap Y$ such that $\|y\| < \|x\|^e + \varepsilon$. Setting $\varepsilon = a - \|x\|^e$ yields $\|y\| < a$. Hence $U_\beta \cap B(0, a) \neq \phi$, and $B(0, a)$ is not closed relative to $\mathcal{T}$. Thus $B(0, 1)$ is also not closed, contradicting A3). Therefore $\|x\|^e = \|x\|$.

2) We need to show that

$$\sum_R = \{x \in X \mid \|x\|^e > R\}$$

is $\mathcal{T}$-open for every $R < \infty$. (See [8, p. 84].) From (2) we have

$$\sum_R = \{x \in X \mid \exists \text{ a } \mathcal{T}\text{-neighborhood } U \text{ of } x \text{ and } \varepsilon > 0 \text{ such that } \|y\| > R + \varepsilon \ \forall y \in U \cap Y\}. \tag{3}$$

If $x \in \sum_R$ for some R, then U and ε are determined by (3). In fact, $U \subset \sum_R$, so $\sum_R$ is open.

3) This follows immediately from (2).

4) If $\|x\|^e = \infty$, $\|x\|^f \leq \|x\|^e$. Suppose that $\|x\|^e < \infty$ and $\varepsilon > 0$ are given. From 2), there exists a β such that $\|y\| \geq \|x\|^f - \frac{\varepsilon}{2}$ for every $y \in U_\beta$. Setting $U = U_\beta$ in 3) guarantees the existence of a $z \in U_\beta \cap Y$ such that $\|z\| < \|x\|^e + \frac{\varepsilon}{2}$. Setting $y = z$ yields $\|x\|^f < \|x\|^e + \varepsilon$. Since ε is arbitrary, $\|x\|^f \leq \|x\|^e$. Interchanging the roles of "e" and "f" and applying the same arguments gives $\|x\|^f \geq \|x\|^e$. □

Let $Y_e = \{x \in X \mid \|x\|^e < \infty\}$. From Proposition 3.1, 1), it is obvious that $Y_e \supset Y$. We refer to Y_e as the *X-extension* of Y. The next result further justifies this terminology.

PROPOSITION 3.2. 1) *Y_e is a subspace of X.*

2) *$\|\cdot\|^e$ is a norm on Y_e.*

Proof. 1) If $x \in Y_e$ and $\alpha \in \mathbb{R}$, then

$$\begin{aligned} \|\alpha x\|^e &= \sup_\beta \inf_{y \in \alpha U_\beta \cap Y} \|y\| \\ &= |\alpha| \sup_\beta \inf_{y \in U_\beta \cap Y} \|y\| \\ &= |\alpha| \, \|x\|^e \\ &< \infty. \end{aligned} \tag{4}$$

Furthermore, if $x_1, x_2 \in Y_e$ and U_β is a neighborhood of $x_1 + x_2$, then there exist neighborhoods V_β and W_β of x_1 and x_2, respectively, such that $V_\beta + W_\beta \subset U_\beta$. Hence,

$$\begin{aligned} \|x_1 + x_2\|^e &= \sup_\beta \inf_{y \in U_\beta \cap Y} \|y\| \\ &\leq \sup_\beta \inf_{\substack{y_1 \in V_\beta \cap Y \\ y_2 \in W_\beta \cap Y}} \|y_1 + y_2\| \\ &\leq \|x_1\|^e + \|x_2\|^e \\ &< \infty. \end{aligned} \tag{5}$$

Thus αx and $x_1 + x_2$ belong to Y_e, and 1) follows.

2) In view of (4) and (5), to demonstrate 2) it remains to show that $\|x\|^e = 0$ implies $x = 0$. Indeed, $\|x\|^e = 0$ implies

$$\inf_{y \in U_\beta \cap Y} \|y\| = 0 \tag{6}$$

for every β. Since X is Hausdorff, for $x \neq 0$ there must exist disjoint $\mathcal{T}$-neighborhoods U_β and V of x and 0, respectively. From assumption A2), there exists an $\varepsilon > 0$ such that $B(0, \varepsilon) \subset V$. Thus $B(0, \varepsilon) \cap U_\beta = \phi$, contradicting (6). □

Roughly speaking, Propositions 3.1 and 3.2 say that 1) $\| \cdot \|^e$ is the smallest possible extension of $\| \cdot \|$ such that $\|x\|^e$ is consistent with $\mathcal{T}$-approximations to x from within Y, and 2) Y_e is the largest subspace of X upon which $\| \cdot \|^e$ is a norm.

We may now specialize these ideas to $X = K'_+$ and $Y = L^p_+$. First note that assumptions A1) and A2) follow easily from [4, §II.4.4] and [3, §I.1.8]. The next result verifies assumption A3).

PROPOSITION 3.3. *$B(0, 1) \subset L^p_+$ is weak* closed (relative to K'_+) for $1 \leq p \leq \infty$.*

Proof. Suppose $B(0, 1)$ is not weak* closed. Then there exist $\varepsilon > 0$ and $f \in L^p_+$ such that $\|f\|_p = 1 + \varepsilon$ and such that, for each $\varphi \in K$, there exists a $g \in L^p_+$ with $\|g\|_p \leq 1$ and $|\langle f - g, \varphi\rangle| < \frac{\varepsilon}{4}$. Let q be conjugate to p. Since K is dense in L^q using $\| \cdot \|_q$ for $q < \infty$, we may choose $\varphi \in K$ such that $\|\varphi\|_q = 1$ and $|\langle f, \varphi\rangle| > 1 + \frac{\varepsilon}{2}$. To handle the case $q = \infty$, we note that K is dense in C_0, so [7, Thm. 6.19] guarantees the existence of a $\varphi \in K$ with $\|\varphi\|_\infty = 1$ and $|\langle f, \varphi\rangle > 1 + \frac{\varepsilon}{2}$. Thus, for arbitrary p, $|\langle g, \varphi\rangle| \leq 1$ and

$$\frac{\varepsilon}{2} < |\langle f, \varphi\rangle| - |\langle g, \varphi\rangle| \leq |\langle f - g, \varphi\rangle| < \frac{\varepsilon}{4}.$$

This is a contradiction, so $B(0, 1)$ is closed. □

Since A1)-A3) are satisfied, the K'_+-extension L^p_{+e} of L^p_+ and its norm $\| \cdot \|^e_p$ are well defined. The following two results characterize L^p_{+e} more precisely.

PROPOSITION 3.4. *Let $1 < p \leq \infty$. Then $L^p_{+e} = L^p_+$.*

Proof. Suppose $f \in K' - L^p_+$, let $M < \infty$ be given, and let q be conjugate to p. Since K is dense in L^q relative to $\| \cdot \|_q$, the dual of K with $\| \cdot \|^q$ imposed on it is just L^p. Thus there exists a $\varphi \in K$ with $\|\varphi\|_q = 1$ such that $|\langle f, \varphi\rangle| > M + 1$. Furthermore, there exists a weak* neighborhood U of f such that $\|g\|_p \geq |\langle g, \varphi\rangle| > M$ for all $g \in U \cap L^p_+$; thus $\|f\|^e_p \geq M$. Since M is arbitrary, $\|f\|^e_p = \infty$ and $f \notin L^p_{+e}$. Hence $L^p_{+e} = L^p_+$. □

The case $p = 1$ is somewhat more challenging.

PROPOSITION 3.5. *$L^1_{+e} = DBV_+$ and $\|x\|^e_1 = \|x\|_d$ for all $x \in DBV_+$.*

Proof. Let

$$\|f\|_D = \begin{cases} \|f\|_d, & f \in DBV, \\ \infty, & f \notin DBV. \end{cases}$$

It suffices to verify 1)-3) in Proposition 3.1. If $f \in L^1$, then $f = \dot{g}$ for some absolutely continuous g. Hence

$$\|f\|_D = \mathrm{Var}(g) = \int_{-\infty}^{\infty} 1|dg(t)| = \int_{-\infty}^{\infty} |\dot{g}(t)|\, dt = \|f\|_1,$$

and 1) holds.

To prove 2), let $R < \infty$ and $\sum_R = \{f \in K' \mid \|f\|_D > R\}$. For $f \in \sum_R$, we have from (1) that

$$\sup_{\substack{\varphi \in K \\ \|\varphi\|_\infty = 1}} |\langle f, \varphi\rangle| = \|f\|_D > R.$$

Hence, there exists a $\varphi \in K$ with $\|\varphi\|_\infty = 1$ such that $|\langle f, \varphi\rangle| > R$. Let

$$U = \{g \in K' \mid |\langle f - g, \varphi\rangle| < |\langle f, \varphi\rangle| - R\}.$$

U is a weak* neighborhood of f. If $g \in U$,

$$|\langle f, \varphi\rangle| - |\langle g, \varphi\rangle| \leq |\langle f - g, \varphi\rangle| < |\langle f, \varphi\rangle| - R,$$

so $|\langle g, \varphi\rangle| > R$. Hence,

$$\|g\|_D = \sup_{\substack{\varphi \in K \\ \|\varphi\|_\infty = 1}} |\langle g, \varphi\rangle| > R.$$

Thus $g \in \sum_R$ and $U \subset \sum_R$. Hence $\sum_R$ is weak* open, and $\|\cdot\|_D$ is lower semicontinuous.

Condition 3) can be proven directly using elementary analytic arguments based on the definition of Var($\cdot$), but here we supply a functional analytic proof that is more amenable to generalization. Let U be a weak* neighborhood of f, and let $\varepsilon > 0$. Then there exist $\varphi_1, \ldots, \varphi_n \in K$ such that $h \in U$ whenever $|\langle f - h, \varphi_i\rangle| < 1$ for all i. If $\beta_1, \ldots, \beta_n \in \mathbb{R}$, then

$$\left|\sum \beta_i \langle f, \varphi_i\rangle\right| = \left|\left\langle f, \sum \beta_i \varphi_i\right\rangle\right| \leq \|f\|_d \left\|\sum \beta_i \varphi_i\right\|_\infty.$$

Noting that $K \subset L^\infty$ and that L^∞ is the dual of the Banach space L^1, it follows from [9, Thm. 5, p. 109] that there exists an $h \in L^1$ such that $\langle h, \varphi\rangle = \langle f, \varphi_i\rangle$ for all i and $\|h\|_1 \leq \|f\|_d + \frac{\varepsilon}{2}$. Note that $|\langle f - h, \varphi_i\rangle| = 0$, so $h \in U$. Since L^1_+ is dense in L^1 relative to $\|\cdot\|_1$ (and therefore also weak*), there exists a $g \in U \cap L^1_+$ such that $\|g\|_1 < \|f\|_d + \varepsilon$. □

We can make slight modifications to the arguments of this section and construct an extension L^p_e of L^p in K'. In this way, results similar to Propositions 3.1, 3.4, and 3.5 are obtained; i.e., $L^p_e = L^p$ for $1 < p \leq \infty$ and $L^1_e = DBV$. This construction has the advantage that L^1_e is a Banach space, while L^1_{+e} is not; however, convolution is not defined on all of L^1_e, so we must restrict ourselves to L^1_{+e}.

Besides stability analysis, another important application of our extension theory occurs in treating minimum-norm optimization problems over K'. For example, the issue of extending a quadratic cost functional on L^2 to K' arose naturally in the earlier work of one of us [10]. It is easily seen that [10, Prop. 1] follows immediately from Proposition 3.1, part 2) and Proposition 3.4.

4. BIBO stability. Proposition 3.4 shows that "boundedness" of a distribution $f \in K'_+$ is most naturally interpreted to mean that $f \in L^\infty_+$. Hence, we define a linear operator $T : K'_+ \to K'_+$ to be *BIBO stable* if $T(L^\infty_+) \subset L^\infty_+$. Clearly, this definition extends the classical one, as long as $u, T(u) \in K'_+$.

Since convolution operators satisfying $T(\delta) \in L^1_+$ are known to be BIBO stable, a natural conjecture is that the convolution operators with kernels in the extension space L^1_{+e} described in Proposition 3.5 coincide exactly with the stable operators. This idea is supported by the fact that $\delta \in L^1_{+e}$, since $\delta = \dot{\theta}$ and $\theta \in BV_+$, and $\delta^{(i)} \notin L^1_{+e}$ for $i = 1, 2, 3, \ldots$, since $\delta^{(i-1)} \notin BV$. It is easy to show that $T(u) = \delta^{(i)} * u$ defines a stable operator iff $i = 0$.

Corresponding to each convolution operator T we may associate a $\widetilde{T} : K \to C^\infty$ defined by

$$\widetilde{T}(\varphi)(t) = \langle T(\delta), \sigma_{-t}\varphi\rangle.$$

Indeed, it is established in [3, p. 100] that $\widetilde{T}$ takes values in C^∞ with $\widetilde{T}(\varphi)(t) = 0$ for all t in some interval $[\tau_\varphi, \infty)$. The following result provides preliminary information about BIBO stable operators.

LEMMA 4.1. *Let $T : K'_+ \to K'_+$ be a convolution operator.*

1) *If T is BIBO stable, then*

$$\sup_{\substack{u \in L^\infty_+ \\ \|u\|_\infty = 1}} \|T(u)\|_\infty < \infty.$$

2) *T is BIBO stable iff*

$$\sup_{\substack{\varphi\in K\\ \|\varphi\|_1=1}} \int_{-\infty}^{\infty} |\tilde{T}(\varphi)(t)|\,dt < \infty.$$

Proof. 1) Let P be the restriction of T to $L^\infty_{[0,\infty)}$. We begin by showing that P is continuous relative to $\|\cdot\|_\infty$.

Let $u, u_i \in L^\infty_{[0,\infty)}$ and $\|u_i - u\|_\infty \to 0$. Suppose there exists a $v \in L^\infty$ such that $\|P(u_i) - v\|_\infty \to 0$. Since weak* topology is weaker than norm topology on L^∞, $u_i \to u$ weak* and $P(u_i) \to v$ weak*. From Lemma 2.1, P is weak* continuous, so $P(u_i) \to P(u)$ weak*. Since weak* topology is Hausdorff, $P(u) = v$. The continuity of P follows from the closed graph theorem.

Now let $u_i \in L^\infty_+$ with $\|u_i\|_\infty \to 0$. For each u_i, there exist τ_i such that $\sigma_{\tau_i} u_i \in L^\infty_{[0,\infty)}$. Also, $\|\sigma_{\tau_i} u_i\|_\infty = \|u_i\|_\infty \to 0$. From the continuity of P and time-invariance of T,

$$\|T(u_i)\|_\infty = \|\sigma_{-\tau_i} P(\sigma_{\tau_i} u_i)\|_\infty = \|P(\sigma_{\tau_i} u_i)\|_\infty \to 0.$$

This shows that T is continuous or, equivalently,

$$\sup_{\substack{u\in L^\infty_+\\ \|u\|_\infty=1}} \|T(u)\|_\infty < \infty.$$

2) (sufficient) Let $u \in L^\infty_+$. From the definition of convolution on K', we have

$$\langle T(u), \varphi\rangle = \int_{-\infty}^{\infty} u(t)\tilde{T}(\varphi)(t)\,dt;$$

hence

$$|\langle T(u), \varphi\rangle| \le \|u\|_\infty \int_{-\infty}^{\infty} |\tilde{T}(\varphi)(t)|\,dt$$

and

$$\sup_{\substack{\varphi\in K\\ \|\varphi\|_1=1}} |\langle T(u), \varphi\rangle| < \infty.$$

Since K is dense in L^1, $T(u)$ extends continuously to a unique linear functional on L^1. Hence, $T(u) \in L^\infty_+$.

(necessary) From 1),

$$\begin{aligned}\sup_{\substack{u\in L^\infty_+\\ \|u\|_\infty=1}} \sup_{\substack{\varphi\in K\\ \|\varphi\|_1=1}} \left|\int_{-\infty}^{\infty} u(t)\tilde{T}(\varphi)(t)\,dt\right| &= \sup_{\substack{u\in L^\infty_+\\ \|u\|_\infty=1}} \sup_{\substack{\varphi\in K\\ \|\varphi\|_1=1}} |\langle T(u), \varphi\rangle| \\ &\le \sup_{\substack{u\in L^\infty_+\\ \|u\|_\infty=1}} \|T(u)\|_\infty \\ &< \infty.\end{aligned}$$

Let $\varphi \in K$ and $u_i(t) = \operatorname{sgn}(\tilde{T}(\varphi)(t))\theta_{-i}(t)$. Then

$$\begin{aligned}\int_{-\infty}^{\infty} |\tilde{T}(\varphi)(t)|\,dt &= \lim_{i\to\infty} \int_{-\infty}^{\infty} u_i(t)\tilde{T}(\varphi)(t)\,dt \\ &\le \sup_{\substack{u\in L^\infty_+\\ \|u\|_\infty=1}} \left|\int_{-\infty}^{\infty} u(t)\tilde{T}(\varphi)(t)\,dt\right|,\end{aligned}$$

so

$$\sup_{\substack{\varphi\in K\\ \|\varphi\|_1=1}} \int_{-\infty}^{\infty} |\tilde{T}(\varphi)(t)|\,dt < \infty. \qquad \square$$

Lemma 4.1, part 1) makes the striking statement that every stable convolution operator is also continuous relative to $\|\cdot\|_\infty$. In system-theoretic jargon, this means that, in the time-invariant case, BIBO stability implies that small changes in the system input give rise to only small changes in the output. It is an interesting fact that this statement is demonstrably false in the time-varying case.

We are now in a position to give our main result.

THEOREM 4.2. *Let $T : K'_+ \to K'_+$ be a convolution operator. The following statements are equivalent.*

1) *T is BIBO stable.*
2) $T(\delta) \in L^1_{+e}$.
3) $T(\theta) \in BV_+$.

Proof. The equivalence of 2) and 3) is obvious from Proposition 3.5. To prove that 3) implies 1), let $\varphi \in K$, $u \in L^\infty_+$, and $s = T(\theta)$ and note that

$$\begin{aligned}
\langle T(u), \varphi\rangle &= \int_{-\infty}^{\infty} u(t)\langle \dot{s}, \sigma_{-t}\varphi\rangle\,dt \\
&= -\int_{-\infty}^{\infty} u(t)\langle s, \sigma_{-t}\dot{\varphi}\rangle\,dt \\
&= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(t)s(\tau)\dot{\varphi}(t+\tau)d\tau\,dt \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(t)\varphi(t+\tau)ds(\tau)dt \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} u(t-\tau)ds(\tau)\right)\varphi(t)dt,
\end{aligned}$$

so

$$T(u)(t) = \int_{-\infty}^{\infty} u(t-\tau)ds(\tau) \quad \text{a.e.}$$

and

$$|T(u)(t)| \le \|u\|_\infty \operatorname{Var}(s) \quad \text{a.e.}$$

Thus $T(u) \in L^\infty$.

Finally, we show that 1) implies 2). From [11, Thm. 2.3.9] and Lemma 4.1, part 2), we know that there exists a measurable function $g : \mathbb{R}^2 \to \mathbb{R}$ with $g(t,\cdot) \in L^\infty$ for all t, $\int_{-\infty}^{\infty} g(\cdot,\tau)\varphi(\tau)\,d\tau$ absolutely continuous,

$$\operatorname*{ess\,sup}_{\tau} \operatorname{Var}(g(\cdot,\tau)) < \infty$$

and

$$\tilde{T}(\varphi)(t) = \frac{d}{dt}\int_{-\infty}^{\infty} g(t,\tau)\varphi(\tau)\,d\tau$$

for all $\varphi \in K$.

Since $\tilde{T}$ is time-invariant,

$$\frac{d}{dt}\int_{-\infty}^{\infty} g(t,\tau)\varphi(\tau - t_0)\,d\tau = \frac{d}{dt}\int_{-\infty}^{\infty} g(t - t_0,\tau)\varphi(\tau)\,d\tau$$

for all $t, t_0 \in \mathbb{R}$, $\varphi \in L^1$. Integration and a change of variables yield

$$\int_{-\infty}^{\infty} g(t,\tau + t_0)\varphi(\tau)\,d\tau = \int_{-\infty}^{\infty} g(t - t_0,\tau)\varphi(\tau)\,d\tau.$$

Thus $g(t,\tau + t_0) = g(t - t_0,\tau)$ for all t, t_0, τ. Set $s(t) = g(-t, 0)$. Then $s \in BV$ and

$$g(t,\tau) = g(t - \tau, 0) = s(\tau - t),$$

so

$$\tilde{T}(\varphi)(t) = \frac{d}{dt}\int_{-\infty}^{\infty} s(\tau - t)\varphi(\tau)\,dt = \frac{d}{dt}\langle \sigma_t s, \varphi\rangle = \langle \sigma_t \dot{s}, \varphi\rangle$$

for all $\varphi \in K$. Setting $t = 0$ yields $\langle \dot{s}, \varphi\rangle = \langle T(\delta), \varphi\rangle$; hence $T(\delta) \in DBV \cap K_+' = DBV_+$. □

Our next objective is to obtain a more detailed picture of the additional structure imposed on a linear system by stability. We begin by considering certain extensions of the operators T and $\tilde{T}$.

Since every stable T is continuous on L_+^∞ relative to $\|\cdot\|_\infty$, each such operator may be extended uniquely to a continuous linear operator $T_0 : L_0^\infty \to L_0^\infty$. Similarly, Lemma 4.1, part 2) also states that T is BIBO stable iff $\tilde{T}(K) \subset L^1$ and $\tilde{T}$ is bounded using the L^1 norm throughout. In this case, since K is dense in L^1, $\tilde{T}$ extends uniquely to a continuous linear operator $\tilde{T}_e : L^1 \to L^1$. It is easy to show that T_0 and $\tilde{T}_e$ are time-invariant.

THEOREM 4.3. *Suppose $T : K_+' \to K_+'$ is a BIBO stable convolution operator and $s = T(\theta)$. Let $T_e : L^\infty \to L^\infty$ be defined by*

$$T_e(u)(t) = \int_{-\infty}^{\infty} u(t - \tau)\,ds(\tau).$$

Then

1) $T_e(u) = T_0(u)$ *for all* $u \in L_0^\infty$.

2) $\tilde{T}_e(\varphi)(t) = \int_{-\infty}^{\infty} \varphi(t + \tau)\,ds(\tau)$ for all $\varphi \in L^1$.

3) T_e *is the adjoint of* $\tilde{T}_e$.

Proof. 1) As in the proof of Theorem 4.2,

$$T(u)(t) = \int_{-\infty}^{\infty} u(t - \tau)\,ds(\tau)$$

for any $u \in L_+^\infty$. Since T_e is continuous relative to $\|\cdot\|_\infty$, 1) follows immediately.

2) For any $\varphi \in K$,

$$\begin{aligned}\tilde{T}(\varphi)(t) &= \langle \dot{s}, \sigma_{-t}\varphi\rangle \\ &= -\langle s, \sigma_{-t}\dot{\varphi}\rangle \\ &= \int_{-\infty}^{\infty} s(\tau)\dot{\varphi}(t + \tau)\,d\tau \\ &= \int_{-\infty}^{\infty} \varphi(t + \tau)\,ds(\tau).\end{aligned}$$

Thus $\tilde{T}_e(\varphi) = \tilde{T}(\varphi)$ for all $\varphi \in K$. Since $\tilde{T}_e$ is continuous relative to $\|\cdot\|_1$, 2) follows.

3) Let $s \in BV$, $u \in L^\infty$, and $\varphi \in L^1$. Applying Fubini's theorem and a change of variable, we have

$$\int_{-\infty}^{\infty} u(t)\left(\int_{-\infty}^{\infty} \varphi(t+\tau)\,ds(\tau)\right)dt = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} u(t-\tau)\,ds(\tau)\right)\varphi(t)\,dt$$

or

$$\int_{-\infty}^{\infty} u(t)\widetilde{T}_e(\varphi)(t)\,dt = \int_{-\infty}^{\infty} T_e(u)(t)\varphi(t)\,dt. \qquad \square$$

Note that the proof of Theorem 4.3 applies even if $s \in BV - BV_+$. Hence, the idea of a stable system whose step or impulse response does not have left-bounded support is meaningful; the system may be viewed as an operator on L^∞. However, such an operator does not extend easily to K' or even K'_+.

To conclude this section, we give several equivalent expressions that quantify the "gain" of a stable operator.

THEOREM 4.4. *For any BIBO stable convolution operator* $T : K'_+ \to K'_+$,

$$\sup_{\substack{u\in L^\infty_+ \\ \|u\|_\infty=1}} \|T(u)\|_\infty = \sup_{\substack{u\in L^\infty \\ \|u\|_\infty=1}} \|T_e(u)\|_\infty = \sup_{\substack{\varphi\in L^1 \\ \|\varphi\|_1=1}} \|\widetilde{T}_e(\varphi)\|_1 = \mathrm{Var}(T(\theta)) = \|T(\delta)\|_1^e.$$

Proof. The first equality follows from continuity of T_e. Since the norms of adjoint operators must coincide, the second identity holds. The next equality follows from the representation of $\widetilde{T}_e$ established in Theorem 4.3. The last identity follows immediately from Proposition 3.5. $\square$

5. Additional properties of stable systems. In view of the Theorem 4.2, part 3) the step response of every BIBO system can be decomposed as $T(\theta) = s_1 + s_2 + s_3$, where s_1 is absolutely continuous, s_2 is saltus, and s_3 is singular. (See [12].) Thus, the corresponding impulse response is $T(\delta) = h_1 + h_2 + h_3$, where $h_i = \dot{s}_i$. Since s_1 is bounded and absolutely continuous, $h_1 \in L^1_+$. Also, since s_2 is a saltus function,

$$h_2 = \sum \alpha_i(\delta_{\tau_i}), \tag{7}$$

where $\sum |\alpha_i| < \infty$. Hence the impulse response of any stable system can be uniquely decomposed into the sum of an L^1 function, an impulsive distribution, and the (distributional) derivative of a singular function.

The distribution $h_3 = \dot{s}_3$ is particularly interesting and apparently has not been treated in the literature as a viable impulse response. Distributions of this type illustrate the fact that, for a function $s : \mathbb{R} \to \mathbb{R}$, the operations of differentiation and identification with a distribution do not in general commute, even if s is continuous. Indeed, the derivative s'_3 of s_3 *as a function* vanishes a.e., so s'_3 is identified with the distribution 0. On the other hand, s_3 is by definition not constant a.e., hence its *distributional derivative* $\dot{s}_3$ does not vanish.

A classical example of a singular function on [0, 1] is the Cantor function, which we denote by c_0. (See [12, p. 50].) Define

$$c(t) = \begin{cases} 0, & t < 0, \\ c_0(t), & 0 \le t \le 1, \\ 1, & t > 1. \end{cases}$$

Then c is nondecreasing and singular on $\mathbb{R}$. Since $\mathrm{Var}(c) = 1$, Proposition 3.5 gives $\|\dot{c}\|_1^e = 1$. The support of the distribution $\dot{c}$ is simply the Cantor ternary set, which is uncountable and

has Lebesgue measure zero. (See [12, p. 49].) Note that $\dot{c}$ has a far more elusive structure than a conventional impulsive distribution (7). Nevertheless, Theorem 4.2 guarantees that the system governed by $T(u) = \dot{c} * u$ is BIBO stable.

On the other hand, suppose $T(\delta) = \ddot{c}$. Any attempt to decide the value $\int |T(\delta)|$ by intuitive means would be perilous at best. Using our theory, this case is easily handled by simply noting that $T(\theta) = \dot{c} \notin BV_+$.

Another characterization of stable linear time-invariant systems can be obtained by examining the set $\mathcal{H}$ of Fourier transforms of functions in L_e^1. It follows from [5, p. 189] that the Fourier transform of any $h \in L_e^1$ exists, is a function, and is given by

$$H(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t}\, ds(t), \tag{8}$$

where $s \in BV$ and $\dot{s} = h$. We refer to $\mathcal{H}$ as the set of *BIBO stable transfer functions*. Clearly, a rational function belongs to $\mathcal{H}$ iff it is BIBO stable in the usual sense.

According to Theorem 4.2, the stable transfer functions are generated by letting s vary over BV in (8). Since every function in BV can be written as the difference of two bounded nondecreasing functions, a substantial number of existing results in analysis come into play. For example, working from [13, Ch. VI], we find that all functions in $\mathcal{H}$ are bounded and uniformly continuous. Several complete, albeit abstruse, characterizations of $\mathcal{H}$ are available, perhaps the simplest following from Bochner's theorem: A function H belongs to $\mathcal{H}$ iff H is the difference of two continuous positive semidefinite functions. (See [13, p. 137].)

The Laplace transform

$$H(z) = \int_{-\infty}^{\infty} e^{-zt}\, ds(t)$$

of $h = \dot{s} \in L_{+e}^1$ also exists and is analytic on $\operatorname{Re} z > 0$. It is easy to show that the "boundary function"

$$\omega \to \lim_{\sigma \to 0} H(\sigma + i\omega)$$

is well defined and equals the Fourier transform (8). In fact, if $\operatorname{supp} h \subset [\tau, \infty)$,

$$|H(z)| \leq e^{-\operatorname{Re} z\tau} \operatorname{Var}(s)$$

for all right half-plane z. In particular, if $h \in L_{+e}^1$ with $\tau \geq 0$, then H belongs to the Hardy space $H^\infty(\mathbb{C}^+)$. The converse implication fails, however, since the function

$$H(z) = e^{-\frac{1}{z}}$$

belongs to H^∞, but $H(i\omega)$ is not continuous at $\omega = 0$ and therefore H is not stable. Sufficient conditions on $H(z)$ for stability (other than those on the boundary function) are difficult to obtain. Even analyticity on the whole plane is not sufficient (e.g., let $h = \dot{\delta}$).

Our final comment of this section addresses the issue of linear systems with multiple inputs and outputs. In this case, $T(\delta)$ and its Fourier transform are matrices. Extending the definition of BIBO stability in the obvious way, it is clear that stability simply corresponds to each entry of the matrix being stable in the sense described above.

6. Conclusions. We presented a coherent distributional theory for linear time-invrariant systems based on the concept of impulse response. The property of BIBO stability was shown to be equivalent to a simple condition on either the impulse or step response of the system. We also supplied a somewhat more difficult stability condition related to the system transfer function. The time-varying case is at present under investigation.

REFERENCES

[1] C. T. CHEN, *Linear System Theory and Design*, Holt, Rinehart, and Winston, New York, 1984.
[2] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.
[3] I. M. GELFAND AND G. E. SHILOR, *Generalized Functions, Vol.* I, Academic Press, New York, 1968.
[4] ———, *Generalized Functions, Vol.* II, Academic Press, New York, 1968.
[5] L. SCHWARTZ, *Mathematics for the Physical Sciences*, Addison-Wesley, Reading, MA, 1966.
[6] J. BARROS-NETO, *An Introduction to the Theory of Distributions*, Marcel Dekker, Inc., New York, 1973.
[7] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1987.
[8] J. DUGUNDJI, *Topology*, Allyn and Bacon, Inc., Boston, MA, 1966.
[9] K. YOSIDA, *Functional Analysis*, Springer-Verlag, New York, 1971.
[10] J. D. COBB, *Descriptor variable systems and optimal state regulation*, IEEE Trans. Automat. Control, 28 (1983), pp. 601-611.
[11] N. DUNFORD AND B. J. PETTIS, *Linear operators on summable functions*, Trnas. Amer. Math. Soc., 47 (1940), pp. 323-392.
[12] I. P. NATANSON, *Theory of Functions of a Real Variable, Vol.* I, Fredrick Ungar, New York, 1974.
[13] Y. KATZNELSON, *An Introduction to Harmonic Analysis*, Dover, New York, 1976.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 1001-1043, May 1996

012

FINITE-DIMENSIONAL APPROXIMATION OF A CLASS OF CONSTRAINED NONLINEAR OPTIMAL CONTROL PROBLEMS*

MAX D. GUNZBURGER† AND L. STEVEN HOU‡

Abstract. An abstract framework for the analysis and approximation of a class of nonlinear optimal control and optimization problems is constructed. Nonlinearities occur in both the objective functional and the constraints. The framework includes an abstract nonlinear optimization problem posed on infinite-dimensional spaces, an approximate problem posed on finite-dimensional spaces, together with a number of hypotheses concerning the two problems. The framework is used to show that optimal solutions exist, to show that Lagrange multipliers may be used to enforce the constraints, to derive an optimality system from which optimal states and controls may be deduced, and to derive existence results and error estimates for solutions of the approximate problem. The abstract framework and the results derived from that framework are then applied to three concrete control or optimization problems and their approximation by finite-element methods. The first involves the von Kármán plate equations of nonlinear elasticity, the second the Ginzburg-Landau equations of superconductivity, and the third the Navier-Stokes equations for incompressible, viscous flows.

Key words. optimal control, nonlinear partial differential equations, finite-dimensional approximation, finite-element methods, von Kármán equations, Ginzburg-Landau equations, Navier-Stokes equations

AMS subject classifications. 65J15, 65N30, 49J20, 35J65, 73G05, 76D05, 81Q05

1. Introduction.

The need to solve optimization and control problems arises in many settings. Although in some cases these problems may be easily solved, either analytically or computationally, in many other cases substantial difficulties are encountered. For example, candidate optimal states and controls may belong to infinite-dimensional function spaces and one may have to minimize a nonlinear functional of the state and control variables subject to nonlinear constraints that take the form of a system of partial differential equations whose solutions are in general not unique. In this paper, our goal is to construct and analyze a framework for the approximate solution of many such problems. The setting for our framework is a class of nonlinear control or optimization problems that is general enough to be of use in numerous applications. The major steps in the development and analysis of our framework are as follows:

- define an abstract class of nonlinear control or optimization problems;
- show that, under certain assumptions, optimal solutions exist;
- show that, under certain additional assumptions, Lagrange multipliers exist that may be used to enforce the constraints;
- use the Lagrange multiplier technique to derive an optimality system from which optimal states and controls may be deduced;
- define algorithms for the approximation, in finite-dimensional spaces, of optimal states and controls;
- derive estimates for the error in the approximations to the optimal states and controls.

Two of the key ingredients used to carry out the above plan are a theory given in [21] for showing the existence of Lagrange multipliers and a theory first developed in [6] for the approximation of a class of nonlinear problems. In both of these theories, certain properties of compact operators on Banach spaces play a central role. We point out that the nonuniqueness

*Received by the editors July 13, 1994; accepted for publication (in revised form) January 23, 1995.

†Interdisciplinary Center for Applied Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0531. The research of this author was supported by Office of Naval Research grant N00014-91-J-1493, Air Force Office of Scientific Research grants AFOSR-93-1-0061 and AFOSR-93-1-0280, and National Aeronautics and Space Administration contract NAS1-19840 while he was in residence at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center.

‡Department of Mathematics and Statistics, York University, North York, Ontario M3J 1P3, Canada. The research of this author was supported by Natural Science and Engineering Research Council of Canada grant OGP-0137436.

of solutions of the nonlinear constraint equations deems it appropriate to employ Lagrange multiplier principles.

After having developed and analyzed the abstract framework, we will apply it to some specific, concrete problems. In each case, we use the abstract framework to analyze the concrete problems by merely showing that the latter fit into the former. The particular applications we consider are

- control problems in structural mechanics having geometric nonlinearities that are governed by the von Kármán equations;
- control problems in superconductivity that are governed by the Ginzburg-Landau equations;
- control problems for incompressible, viscous flows that are governed by the Navier-Stokes equations.

In considering these applications, we will purposely choose different types of controls in order to illustrate how these can be accounted for within the abstract framework. In all three cases, approximation will be effected through the use of finite-element methods.

2. The abstract problem and its analysis. In this section we define and analyze an abstract class of constrained nonlinear control problems; an outline of the definitions and results of this section is as follows.

- In §2.1, the abstract class of constrained control problems that we consider is defined.
- In §2.2, a list of assumptions about the class of abstract problems is given.
- In Theorem 2.1 of §2.3, some of the assumptions listed in §2.2 are used to show that optimal solutions of the abstract problem exist.
- In §2.4, some additional assumptions of §2.2 are used to show that Lagrange multipliers exist that may be used to enforce the constraint; also, first-order necessary conditions are given.
- In Theorems 2.5 and 2.6 of §2.4, the first-order necessary conditions for determining optimal states and controls are simplified under additional assumptions about the control set.
- In §2.5, the optimality system from which optimal controls and states can be determined is made more amenable to approximation by simplifying the dependence of the objective functional on the control.

2.1. The abstract setting. We begin with the definition of the abstract class of nonlinear control or optimization problems that we study.

We introduce the spaces and control set as follows. Let G, X, and Y be reflexive Banach spaces whose norms are denoted by $\|\cdot\|_G$, $\|\cdot\|_X$, and $\|\cdot\|_Y$, respectively. Dual spaces will be denoted by $(\cdot)^*$. The duality pairing between X and X^* is denoted by $\langle\cdot,\cdot\rangle_X$; one similarly defines $\langle\cdot,\cdot\rangle_Y$ and $\langle\cdot,\cdot\rangle_G$. The subscripts are often omitted when there is no chance for confusion. Let Θ, the control set, be a closed convex subset of G. Let Z be a subspace of Y with a compact imbedding. Note that the compactness of the imbedding $Z \subset Y$ will play an important role.

We assume that the functional to be minimized takes the form

$$\mathcal{J}(v,z) = \lambda\,\mathcal{F}(v) + \lambda\,\mathcal{E}(z) \quad \forall\,(v,z) \in X \times \Theta\,, \tag{2.1}$$

where $\mathcal{F}$ is a functional on X, $\mathcal{E}$ is a functional on Θ, and λ is a given parameter that is assumed to belong to a compact interval $\Lambda \subset \mathbb{R}_+$.

The constraint equation $M(v,z) = 0$ relating the state variable v and the control variable z is defined as follows. Let N be a differentiable mapping from X to Y, K a continuous linear operator from Θ to Y, and T a continuous linear operator from Y to X. For any $\lambda \in \Lambda$, we

define the mapping M from $X \times \Theta$ to X by

$$M(v, z) = v + \lambda TN(v) + \lambda TK(z) \quad \forall\, (v, z) \in X \times \Theta\,. \tag{2.2}$$

With these definitions we now consider the constrained minimization problem:

$$\min_{(v,z)\in X\times\Theta} \mathcal{J}(v, z) \quad \text{subject to} \quad M(v, z) = 0\,. \tag{2.3}$$

In (2.3), we seek a global minimizer with respect to the set $\{(v, z) \in X \times \Theta : M(v, z) = 0\}$. Although, under suitable hypotheses, we will show that the problem (2.3) has a solution, in practice, one can only characterize local minima, i.e., points $(u, g) \in X \times \Theta$ such that for some $\epsilon > 0$

$$\begin{aligned} &\mathcal{J}(u, g) \le \mathcal{J}(v, z) \qquad \forall\, (v, z) \in X \times \Theta \text{ such that} \\ &\qquad\qquad M(v, z) = 0 \text{ and } \|u - v\|_X \le \epsilon\,. \end{aligned} \tag{2.4}$$

Thus, when we consider algorithms for locating constrained minima of $\mathcal{J}$, we must be content to find local minima in the sense of (2.4).

After showing that optimal solutions exist and that one is justified in using the Lagrange multiplier rule, we will introduce some simplifications in order to render the abstract problem (2.3), or (2.4), more amenable to approximation. The first is only to consider the control set $\Theta = G$. The second is only to consider Fréchet differentiable functionals $\mathcal{E}(\cdot)$ such that the Fréchet derivative $\mathcal{E}'(g) = E^{-1}g$, where E is an invertible linear operator from G^* to G.

2.2. Hypotheses concerning the abstract problem. The first set of hypotheses will be invoked to prove the existence of optimal solutions. It is given by:

(H1) $\inf_{v\in X} \mathcal{F}(v) > -\infty$;

(H2) there exist constants $\alpha, \beta > 0$ such that $\mathcal{E}(z) \ge \alpha\|z\|^\beta \ \ \forall\, z \in \Theta$;

(H3) there exists a $(v, z) \in X \times \Theta$ satisfying $M(v, z) = 0$;

(H4) if $u^{(n)} \rightharpoonup u$ in X and $g^{(n)} \rightharpoonup g$ in G where $\{(u^{(n)}, g^{(n)})\} \subset X \times \Theta$, then $N(u^{(n)}) \rightharpoonup N(u)$ in Y and $K(g^{(n)}) \rightharpoonup K(g)$ in Y;

(H5) $\mathcal{J}(\cdot, \cdot)$ is weakly lower semicontinuous on $X \times \Theta$;

(H6) if $\{(u^{(n)}, g^{(n)})\} \subset X \times \Theta$ is such that $\{\mathcal{F}(u^{(n)})\}$ is a bounded set in $\mathbb{R}$ and $M(u^{(n)}, g^{(n)}) = 0$, then $\{u^{(n)}\}$ is a bounded set in X.

The second set of assumptions will be used to justify the use of the Lagrange multiplier rule and to derive an optimality system from which optimal states and controls can be determined. The second set is given by

(H7) for each $z \in \Theta$, $v \mapsto \mathcal{J}(v, z)$ and $v \mapsto M(v, z)$ are Fréchet differentiable;

(H8) $z \mapsto \mathcal{E}(z)$ is convex, i.e.,

$$\mathcal{E}\big(\gamma z_1 + (1 - \gamma)z_2\big) \le \gamma\, \mathcal{E}(z_1) + (1 - \gamma)\, \mathcal{E}(z_2) \quad \forall\, z_1, z_2 \in \Theta\,,\ \forall\, \gamma \in [0, 1]\,;$$

(H9) for $v \in X$, $N'(v)$ maps X into Z.

In (H9), N' denotes the Fréchet derivative of N.

A simplified optimality system may be obtained if one invokes the additional assumption:

(H10) $\Theta = G$, and the mapping $z \mapsto \mathcal{E}(z)$ is Fréchet differentiable on G.

Hypotheses (H7)-(H10) allow us to obtain a simplified optimality system for almost all values of the parameter $\lambda \in \Lambda$. In many cases, it is possible to show that the same optimality system holds for all values of λ. The following two additional assumptions, which will only be invoked in case $(1/\lambda)$ is an eigenvalue of $-TN'(u)$, each provides a setting in which this last result is valid:

(H11) if $v^* \in X^*$ satisfies $\big(I + \lambda\,[N'(u)]^*T^*\big)v^* = 0$ and $K^*T^*v^* = 0$, then $v^* = 0$;

(H11)′ the mapping $(v, z) \mapsto v + \lambda\,TN'(u)v + \lambda\,TKz$ is onto from $X \times G$ to Y.

In order to make the optimality system more amenable to approximation and computation, we will invoke the following additional assumption:

(H12) $\mathcal{E}'(g) = E^{-1}g$, where E is an invertible linear operator from G^* to G and g is an optimal control for the constrained minimization problem (2.4).

2.3. Existence of an optimal solution. We first use assumptions (H1)-(H6) to establish that optimal solutions exist.

THEOREM 2.1. *Assume that the functional $\mathcal{J}$ and mapping M defined by* (2.1) *and* (2.2), *respectively, satisfy the hypotheses* (H1)-(H6). *Then, there exists a solution to the minimization problem* (2.3).

Proof. Assumption (H3) simply asserts that there is at least one element of $X \times \Theta$ that satisfies the constraint. Thus, we may choose a minimizing sequence $\{(u^{(n)}, g^{(n)})\} \subset X \times \Theta$ such that

$$\lim_{n\to\infty} \mathcal{J}(u^{(n)}, g^{(n)}) = \inf_{(v,z)\in X\times\Theta} \mathcal{J}(v, z)$$

and

$$M(u^{(n)}, g^{(n)}) = 0\,.$$

By (H1) and (H2), the boundedness of $\{\mathcal{J}(u^{(n)}, g^{(n)})\}$ implies the boundedness of the sequences $\{\|g^{(n)}\|_G\}$ and $\{\mathcal{F}(u^{(n)})\}$. Then, by (H6), we deduce that $\{\|u^{(n)}\|_X\}$ is bounded. Thus, we may extract a subsequence $\{(u^{(n)}, g^{(n)})\}$ such that $u^{(n)} \rightharpoonup u$ in X and $g^{(n)} \rightharpoonup g$ in G. Since Θ is closed and convex, we have $g \in \Theta$. Of course, $u \in X$. We next show that (u, g) satisfies the constraint equation. Using (H4), we have that

$$\lim_{n\to\infty} \langle TN(u^{(n)}), f\rangle = \lim_{n\to\infty} \langle N(u^{(n)}), T^*f\rangle = \langle N(u), T^*f\rangle = \langle TN(u), f\rangle \quad \forall\, f \in X^*$$

and

$$\lim_{n\to\infty} \langle TK(g^{(n)}), f\rangle = \lim_{n\to\infty} \langle K(g^{(n)}), T^*f\rangle = \langle K(g), T^*f\rangle = \langle TK(g), f\rangle \quad \forall\, f \in X^*$$

so that

$$0 = \lim_{n\to\infty} \langle M(u^{(n)}, g^{(n)}), f\rangle = \langle u + \lambda\,TN(u) + \lambda\,TK(g), f\rangle \quad \forall\, f \in X^*\,,$$

i.e., $M(u, g) = 0$. Finally, we use (H5), the weak lower semicontinuity of $\mathcal{J}(\cdot, \cdot)$, to conclude that (u, g) is indeed a minimizer in $X \times \Theta$ satisfying the constraint $M(u, g) = 0$. □

Remark. The hypotheses (H1)-(H6) are not sufficient to guarantee that optimal solutions are unique. Indeed, in many applications to nonlinear problems, including the ones we consider in §4, optimal solutions in the sense of (2.4) are in general not uniquely determined. □

2.4. Existence of Lagrange multipliers. We now wish to use the additional assumptions (H7)-(H9) to show that the Lagrange multiplier rule may be used to turn the constrained minimization problem (2.3) into an unconstrained one. Actually, the Lagrange multiplier rule will only enable us to find local minima in the sense of (2.4). We first quote the following abstract Lagrange multiplier rule whose proof can be found in [21].

THEOREM 2.2. *Let X_1 and X_2 be two Banach spaces and Θ an arbitrary set. Suppose $\mathcal{J}$ is a functional on $X_1 \times \Theta$ and M a mapping from $X_1 \times \Theta$ to X_2. Assume that $(u, g) \in X_1 \times \Theta$ is a solution to the following constrained minimization problem:*

$$\begin{aligned} &M(u, g) = 0 \text{ and there exists an } \epsilon > 0 \text{ such that } \mathcal{J}(u, g) \leq \mathcal{J}(v, z) \\ &\text{for all } (v, z) \in X_1 \times \Theta \text{ such that } \|u - v\|_{X_1} \leq \epsilon \text{ and } M(v, z) = 0. \end{aligned} \tag{2.5}$$

Let U be an open neighborhood of u in X_1. Assume further that the following conditions are satisfied:

(2.6) *for each $z \in \Theta$, $v \mapsto \mathcal{J}(v, z)$ and $v \mapsto M(v, z)$ are Fréchet-differentiable at $v = u$;*

(2.7) *for any $v \in U$, $z_1, z_2 \in \Theta$, and $\gamma \in [0, 1]$, there exists a $z_\gamma = z_\gamma(v, z_1, z_2)$ such that*

$$M(v, z_\gamma) = \gamma M(v, z_1) + (1 - \gamma) M(v, z_2)$$

and

$$\mathcal{J}(v, z_\gamma) \leq \gamma \mathcal{J}(v, z_1) + (1 - \gamma) \mathcal{J}(v, z_2);$$

(2.8) $\operatorname{Range}\big(M_u(u, g)\big)$ *is closed with a finite codimension,*

where $M_u(u, g)$ denotes the Fréchet derivative of M with respect to u. Then, there exists a $k \in \mathbb{R}$ and a $\mu \in X_2^*$ that are not both equal to zero such that

$$k \langle \mathcal{J}_u(u, g), v \rangle - \langle \mu, M_u(u, g) v \rangle = 0 \quad \forall\, v \in X_1$$

and

$$\min_{z \in \Theta} \mathcal{L}(u, z, \mu, k) = \mathcal{L}(u, g, \mu, k),$$

where $\mathcal{L}(u, g, \mu, k) = k\, \mathcal{J}(u, g) - \langle \mu, M(u, g) \rangle$ is the Lagrangian for the constrained minimization problem (2.5) *and where $\mathcal{J}_u(u, g)$ denotes the Fréchet derivative of $\mathcal{J}$ with respect to u. Moreover, if*

(2.9) *the algebraic sum $M_u(u, g) X_1 + M(u, \Theta)$ contains $0 \in X_2$ as an interior point, then we may choose $k = 1$, i.e., there exists a $\mu \in X_2^*$ such that*

$$\langle \mathcal{J}_u(u, g), v \rangle - \langle \mu, M_u(u, g) v \rangle = 0 \quad \forall\, v \in X_1$$

and

$$\min_{z \in \Theta} \mathcal{L}(u, z, \mu, 1) = \mathcal{L}(u, g, \mu, 1).$$

Proof. See [21]. □

Next, we apply Theorem 2.2 to the optimization problem (2.4). In doing so, we will need the following result.

LEMMA 2.3. *Let the spaces X, Y, and Z and operators T and N be defined as in* §2.1. *For $v \in X$, assume that $N'(v)$ maps X into Z. Then, $TN'(v)$ is a compact operator from X to X; therefore, $\sigma\big(-TN'(v)\big)$, the spectrum of the operator $\big(-TN'(v)\big)$, is at most countable with zero being the only possible limit point.*

Proof. Since $Z \hookrightarrow\hookrightarrow Y$, we see that $N'(v)$ is a compact linear operator from X to Y. Also, T is a bounded linear operator from Y to X, so $TN'(v)$ is a compact operator from X to X. Hence, $\sigma\big(-TN'(v)\big)$ is at most countable and consists only of 0 and the eigenvalues of $\big(-TN'(v)\big)$. □

Note that in the following result, the existence of at least one pair (u, g) satisfying (2.4) is guaranteed by Theorem 2.1.

THEOREM 2.4. *Let $\lambda \in \Lambda$ be given. Assume that assumptions* (H1)-(H9) *hold. Let $(u, g) \in X \times \Theta$ be an optimal solution satisfying* (2.4). *Then, there exists a $k \in \mathbb{R}$ and a $\mu \in X^*$ that are not both equal to zero such that*

$$k\,\langle \mathcal{J}_u(u, g), w\rangle - \langle \mu, M_u(u, g)\cdot w\rangle = 0 \quad \forall\, w \in X \tag{2.10}$$

and

$$\min_{z\in\Theta} \mathcal{L}(u, z, \mu, k) = \mathcal{L}(u, g, \mu, k)\,. \tag{2.11}$$

Furthermore, if $(1/\lambda) \notin \sigma\big(-TN'(u)\big)$, we may choose $k = 1$; i.e., for almost all λ, there exists a $\mu \in X^*$ such that

$$\langle \mathcal{J}_u(u, g), w\rangle - \langle \mu, M_u(\lambda, u, g)\cdot w\rangle = 0 \quad \forall\, w \in X \tag{2.12}$$

and

$$\min_{z\in\Theta} \mathcal{L}(u, z, \mu, 1) = \mathcal{L}(u, g, \mu, 1)\,. \tag{2.13}$$

Proof. Let $\lambda \in \Lambda$ be given. To show the existence of k and μ such that (2.10) and (2.11) are valid, we only need to verify that the hypotheses (2.6)-(2.8) of Theorem 2.2 hold with $X_1 = X_2 = X$, since in this case (2.5) reduces to (2.4). Obviously, (2.6) is merely a restatement of (H7). Since Θ is convex and since the mappings T and K are linear, we have that if $z_\gamma = \gamma z_1 + (1-\gamma)z_2$, then

$$\begin{aligned} M(v, z_\gamma) &= v + \lambda TN(v) + \lambda TKz_\gamma \\ &= \gamma\big(v + \lambda TN(v)\big) + (1-\gamma)\big(v + \lambda TN(v)\big) + \gamma\lambda\big(TKz_1 + (1-\gamma)TKz_2\big) \\ &= \gamma M(v, z_1) + (1-\gamma)M(v, z_2)\,. \end{aligned}$$

Moreover, (H8) implies that

$$\begin{aligned} \mathcal{J}(v, z_\gamma) &= \lambda\mathcal{F}(v) + \lambda\mathcal{E}(z_\gamma) = \lambda\mathcal{F}(v) + \lambda\mathcal{E}\big(\gamma z_1 + (1-\gamma)z_2\big) \\ &\le \lambda\mathcal{F}(v) + \lambda\big(\gamma\mathcal{E}(z_1) + (1-\gamma)\mathcal{E}(z_2)\big) = \gamma\,\mathcal{J}(v, z_1) + (1-\gamma)\,\mathcal{J}(v, z_2)\,. \end{aligned}$$

Thus, (2.7) holds. The operator $M_u(u, g)$ from X to X is defined by

$$M_u(u, g)\cdot w = w + \lambda TN'(u)\cdot w \quad \forall\, w \in X$$

or, simply,

$$M_u(u, g) = I + \lambda TN'(u)\,.$$

From (H9) and Lemma 2.3, we have that $TN'(u)$ is a compact operator from X to X. As a result, $M_u(u, g) = I + \lambda TN'(u)$ is a Fredholm operator, so it has a closed range with a finite codimension, i.e., (2.8) holds. Thus, by Theorem 2.2, there exists a $k \in \mathbb{R}$ and a $\mu \in X^*$ that are not both equal to zero such that (2.10) and (2.11) hold.

To show the existence of a μ such that (2.12) and (2.13) are valid, we only need to verify that the additional hypothesis (2.9) of Theorem 2.2 holds. In fact, if in addition $(1/\lambda) \notin \sigma\big(-TN'(u)\big)$, then it follows that $X = \text{Range}\big(I + \lambda TN'(u)\big) = \text{Range}\big(M_u(u, g)\big)$, so $\text{Range}\big(M_u(u, g)\big)$ contains $0 \in X$ as an interior point, i.e., (2.9) holds. Hence, by Theorem 2.2 and Lemma 2.3, we conclude that for almost all λ there exists a $\mu \in X^*$ such that (2.12) and (2.13) hold. □

So far Θ has only been assumed to be a closed and convex subset of G. No smoothness condition on the control variable g has been assumed in the functional or in the constraint. Thus, the necessary condition of optimality with respect to variations in the control variable is expressed in the cumbersome relation (2.11). We now turn to the case where Θ contains a neighborhood of g, where (u, g) is an optimal solution. In particular, we assume that $\Theta = G$. In this case, (2.11) can be given a more concrete structure.

THEOREM 2.5. *Let $\lambda \in \Lambda$ be given. Assume that assumptions* (H1)-(H10) *hold. Let $(u, g) \in X \times G$ be a solution of the problem* (2.4). *Then there exist a $k \in \mathbb{R}$ and a $\mu \in X^*$ that are not both equal to zero such that*

$$k\,\langle \mathcal{J}_u(u, g), w\rangle - \langle \mu, \big(I + \lambda TN'(u)\big)w\rangle = 0 \quad \forall\, w \in X \tag{2.14}$$

and

$$k\langle \mathcal{E}'(g), z\rangle - \langle \mu, TKz\rangle = 0 \quad \forall\, z \in G\,. \tag{2.15}$$

Furthermore, if $(1/\lambda) \notin \sigma\big(-TN'(u)\big)$, we may choose $k = 1$; i.e., there exists a $\mu \in X^*$ such that

$$\langle \mathcal{J}_u(u, g), w\rangle - \langle \mu, \big(I + \lambda TN'(u)\big)w\rangle = 0 \quad \forall\, w \in X \tag{2.16}$$

and

$$\langle \mathcal{E}'(g), z\rangle - \langle \mu, TKz\rangle = 0 \quad \forall\, z \in G \tag{2.17}$$

hold.

Proof. Since the hypotheses imply that $\mathcal{J}(v, z)$ is Fréchet differentiable with respect to z, (2.14)-(2.17) follow easily from Theorem 2.4. □

Remark. If $k = 0$, then there exists a $\mu \neq 0$ such that

$$-\langle \mu, M_u(u, g)w\rangle = 0 \quad \forall\, w \in X\,,$$

so the optimality system necessarily has infinitely many solutions. In fact, for any $C \in \mathbb{R}$, $(C\mu)$ is a solution whenever μ is a solution. This creates both theoretical and numerical difficulties. Thus, it is of great interest to try to eliminate this situation. Fortunately, Lemma 2.3 and Theorem 2.4 tell us that we may set $k = 1 \neq 0$ for almost all values of $(1/\lambda)$, i.e., except for the at most countable set of values in $\sigma\big(-TN'(u)\big)$. □

If the control g enters the constraint in a favorable manner, then we may take $k = 1$ even when $(1/\lambda) \in \sigma\big(-TN'(u)\big)$. Specifically, we invoke one of the assumptions (H11) and (H11)′. We then have the following result.

THEOREM 2.6. *Assume that the hypotheses of Theorem* 2.5 *hold. Assume that if $(1/\lambda) \in \sigma\big(-TN'(u)\big)$, then either* (H11) *or* (H11)′ *holds. Then, for all $\lambda \in \Lambda$, there exists a $\mu \in X^*$ such that* (2.16) *and* (2.17) *hold.*

Proof. Because of Theorem 2.5, we need only to examine the case $(1/\lambda) \in \sigma\big(-TN'(u)\big)$ and to show the algebraic sum $M_u(u, g)X + M(u, G) = X$. If (H11)′ holds, the result is a direct application of Theorem 2.2.

If (H11) holds, let $(1/\lambda)$ be a nonzero eigenvalue of $\big(-TN'(u)\big)$. Then, λ is also an eigenvalue of $\big(-N'(u)^*T^*\big)$ with a finite-dimensional eigenspace having the corresponding eigenfunctions $\{v_i^*\}_{i=1}^m \subset X^*$ as a basis. We claim that $\{K^*T^*v_i^*\}_{i=1}^m \subset G^*$ is a linearly independent set. To see this, we assume $\sum_{i=1}^m \alpha_i\, K^*T^*v_i^* = 0$ with $\alpha_i \in \mathbb{R}$; this expression can be rewritten as $K^*T^*\big(\sum_{i=1}^m \alpha_i\, v_i^*\big) = 0$. Because each v_i^* is an eigenvector, we have $\big(I + \lambda\, N'(u)^*T^*\big)\sum_{i=1}^m \alpha_i\, v_i^* = 0$. Thus, the assumption (H11) implies that $\sum_{i=1}^m \alpha_i\, v_i^* = 0$. Since $\{v_i^*\}_{i=1}^m$ is an eigenbasis and, therefore, a linearly independent set, we have each $\alpha_i = 0$. This shows that $\{K^*T^*v_i^*\}_{i=1}^m$ is a linearly independent set in G^*. Hence, we may choose an orthonormal dual basis $\{z_i\}_{i=1}^m \subset G$ such that $\langle z_i, K^*T^*v_j^*\rangle = \delta_{ij}$.

Now, let $w \in X$ be given. We choose $z = \frac{1}{\lambda}\sum_{i=1}^m \langle w, v_i^*\rangle z_i$. Then $\langle w, v_j^*\rangle - \lambda\langle TKz, v_j^*\rangle = \langle w, v_j^*\rangle - \lambda\langle z, K^*T^*v_j^*\rangle = \langle w, v_j^*\rangle - \sum_{i=1}^m \langle w, v_i^*\rangle\delta_{ij} = 0$ for $j = 1, \ldots, m$. Thus, by Fredholm alternatives, there exists a unique $v \in X$ that satisfies $\big(I + \lambda\, TN'(u)\big)v = w - \lambda\, TKz$ or $\big(I + \lambda\, TN'(u)\big)v + \lambda\, TKz = w$; thus, we have shown that $M_u(u, g)X + M(u, G) = X$. Hence, by Theorem 2.2, there exists a $\mu \in X^*$ such that (2.16) and (2.17) hold. □

2.5. The optimality system. Under the assumptions of Theorem 2.6, an optimal state $u \in X$, an optimal control $g \in G$, and the corresponding Lagrange multiplier $\mu \in X^*$ satisfy the optimality system of equations formed by (2.2), (2.16), and (2.17). From (2.1) we have that $\mathcal{J}_u = \lambda\mathcal{F}'$ and $\mathcal{J}_g = \lambda\mathcal{E}'$, where $\mathcal{F}'$ denotes the obvious Fréchet derivative. Then, (2.16)-(2.17) may be rewritten in the form

$$\mu + \lambda\,[N'(u)]^*T^*\mu - \lambda\,\mathcal{F}'(u) = 0 \quad \text{in } X^* \tag{2.18}$$

and

$$\mathcal{E}'(g) - K^*T^*\mu = 0 \quad \text{in } G^*. \tag{2.19}$$

For purposes of numerical approximations, it turns out to be convenient to make the change of variable $\xi = T^*\mu$. Then, the optimality system (2.2), (2.18), and (2.19) for $u \in X$, $g \in G$, and $\xi \in Y^*$ takes the form

$$u + \lambda\, TN(u) + \lambda\, TKg = 0 \quad \text{in } X, \tag{2.20}$$

$$\xi + \lambda\, T^*[N'(u)]^*\xi - \lambda\, T^*\mathcal{F}'(u) = 0 \quad \text{in } Y^*, \tag{2.21}$$

and

$$\mathcal{E}'(g) - K^*\xi = 0 \quad \text{in } G^*. \tag{2.22}$$

It will also be convenient to invoke an additional simplifying assumption concerning the dependence of the objective functional on the control. Specifically, we assume that (H12) holds. Then, (2.20)-(2.22) can be rewritten as

$$u + \lambda\, TN(u) + \lambda\, TKg = 0 \quad \text{in } X, \tag{2.23}$$

$$\xi + \lambda\, T^*[N'(u)]^*\xi - \lambda\, T^*\mathcal{F}'(u) = 0 \quad \text{in } Y^*, \tag{2.24}$$

and

$$g - EK^*\xi = 0 \quad \text{in } G. \tag{2.25}$$

Remark. Note that the optimality systems, e.g., (2.23)-(2.25), are linear in the adjoint variable ξ. Also, note that the control g may be eliminated from the optimality system (2.23)-(2.25). Indeed, the substitution of (2.25) into (2.23) yields

$$u + \lambda TN(u) + \lambda TKEK^*\xi = 0 \quad \text{in } X. \tag{2.26}$$

Thus, (2.24) and (2.26) determine the optimal state u and adjoint state ξ; subsequently, (2.25) may be used to determine the optimal control g from ξ. This observation serves to emphasize the important, direct role that the adjoint state plays in the determination of the optimal control. □

Remark. Given a $\xi \in Y^*$, it is not always possible to evaluate g exactly from (2.25). For example, the application of the operator E may involve the solution of a partial differential equation. Thus, although it is often convenient to devise algorithms for the approximation of optimal control and states based on the simplified optimality system (2.24) and (2.26), in some other cases it is best to deal with the full form (2.23)-(2.25). Thus, when we consider approximations of optimal controls and states, we will deal with the latter. □

Remark. In many applications we have that $X^* = Y$. Since these spaces are assumed to be reflexive, we also have that $Y^* = X$. In this case, we have that both u and ξ belong to X. □

3. Finite-dimensional approximations of the abstract problem. In this section we define and analyze algorithms for the finite-dimensional approximation of solutions of the optimality system (2.23)-(2.25); an outline of the definitions and results of this section is as follows.

- In §3.1, we define the finite-dimensional approximate problems that we consider.
- In §3.2, a list of assumptions about the approximate problems is given.
- In §3.3, we quote a result of [6] that we will use to analyze approximations obtained as solutions of the approximate problems defined in §§3.1 and 3.2.
- In §3.4, we provide error estimates for the approximation of solutions of the optimality system (2.23)-(2.25).

3.1. Formulation of finite-dimensional approximate problems. A finite-dimensional discretization of the optimality system (2.23)-(2.25) is defined as follows. First, one chooses families of finite-dimensional subspaces $X^h \subset X$, $(Y^*)^h \subset Y^*$, and $G^h \subset G$. These families are parameterized by a parameter h that tends to zero. (For example, this parameter can be chosen to be some measure of the grid size in a subdivision of Ω into finite elements.) Next, we define approximate operators $T^h : Y \to X^h$, $E^h : G^* \to G^h$, and $(T^*)^h : X^* \to (Y^*)^h$. Of course, one views T^h, E^h, and $(T^*)^h$ as approximations to the operators T, E, and T^*, respectively. Note that $(T^*)^h$ is not necessarily the same as $(T^h)^*$. The former is a discretization of an adjoint operator, while the latter is the adjoint of a discrete operator.

Once the approximating subspaces and operators have been chosen, an approximate problem is defined as follows. We seek $u^h \in X^h$, $g^h \in G^h$, and $\xi^h \in (Y^*)^h$ such that

$$u^h + \lambda T^h N(u^h) + \lambda T^h K g^h = 0 \quad \text{in } X^h, \tag{3.1}$$

$$\xi^h + \lambda (T^*)^h [N'(u^h)]^* \xi^h - \lambda (T^*)^h \mathcal{F}'(u^h) = 0 \quad \text{in } (Y^*)^h, \tag{3.2}$$

and

$$g^h - E^h K^* \xi^h = 0 \quad \text{in } G^h. \tag{3.3}$$

3.2. Hypotheses concerning the abstract problem and the approximate problem. We make the following hypotheses concerning the approximate operators T^h, $(T^*)^h$, and E^h:

(H13) $$\lim_{h\to 0} \|(T - T^h)y\|_X = 0 \quad \forall\, y \in Y\,,$$

(H14) $$\lim_{h\to 0} \|(T^* - (T^*)^h)v\|_{Y^*} = 0 \quad \forall\, v \in X^*\,,$$

(H15) $$\lim_{h\to 0} \|(E - E^h)s\|_G = 0 \quad \forall\, s \in G^*\,.$$

We also need the following additional hypotheses on the operators appearing in the definition of the abstract problem (2.4):

(H16) $N \in C^3(X; Y)$ and $\mathcal{F} \in C^3(X; \mathbb{R})$;

(H17) N'', N''', $\mathcal{F}''$, and $\mathcal{F}'''$ are locally bounded, i.e., they map bounded sets to bounded sets;

(H18) for $v \in X$, in addition to (H9), i.e., $N'(v) \in \mathcal{L}(X; Z)$ where $Z \hookrightarrow\hookrightarrow Y$, we have that $[N'(v)]^* \in \mathcal{L}(Y^*; \hat{Z})$ where $\hat{Z} \hookrightarrow\hookrightarrow X^*$, that for $\eta \in Y^*$, $[N''(v)]^* \cdot \eta \in \mathcal{L}(Y^*; \hat{Z})$, and that for $w \in X$, $\mathcal{F}''(v) \cdot w \in \mathcal{L}(X; \hat{Z})$;

(H19) K maps G into Z.

Here, $(\cdot)''$ and $(\cdot)'''$ denote second and third Fréchet derivatives, respectively.

3.3. Quotation of results concerning the approximation of a class of nonlinear problems. The error estimate to be derived in §3.4 makes use of results of [6] and [10] (see also [13]) concerning the approximation of a class of nonlinear problems. These results imply that, under certain hypotheses, the error of approximation of solutions of certain nonlinear problems is basically the same as the error of approximation of solutions of related linear problems. Here, for the sake of completeness, we will state the relevant results, specialized to our needs.

The nonlinear problems considered in [6], [10], and [13] are of the following type. For given $\lambda \in \Lambda$, we seek $\psi \in \mathcal{X}$ such that

(3.4) $$\mathcal{H}(\lambda, \psi) \equiv \psi + \mathcal{T}\mathcal{G}(\lambda, \psi) = 0\,,$$

where $\mathcal{T} \in \mathcal{L}(\mathcal{Y}; \mathcal{X})$, $\mathcal{G}$ is a C^2 mapping from $\Lambda \times \mathcal{X}$ into $\mathcal{Y}$, $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces, and Λ is a compact interval of $\mathbb{R}$. We say that $\{(\lambda, \psi(\lambda)) : \lambda \in \Lambda\}$ is a *branch of solutions* of (3.4) if $\lambda \to \psi(\lambda)$ is a continuous function from Λ into $\mathcal{X}$ such that $\mathcal{H}(\lambda, \psi(\lambda)) = 0$. The branch is called a *regular branch* if we also have that $\mathcal{H}_\psi(\lambda, \psi(\lambda))$ is an isomorphism from $\mathcal{X}$ into $\mathcal{X}$ for all $\lambda \in \Lambda$. Here, $\mathcal{H}_\psi(\cdot, \cdot)$ denotes the Fréchet derivative of $\mathcal{H}(\cdot, \cdot)$ with respect to the second argument. We assume that there exists another Banach space $\mathcal{Z}$, contained in $\mathcal{Y}$, with continuous imbedding, such that

(3.5) $$\mathcal{G}_\psi(\lambda, \psi) \in \mathcal{L}(\mathcal{X}; \mathcal{Z}) \quad \forall\, \lambda \in \Lambda \text{ and } \psi \in \mathcal{X}\,,$$

where $\mathcal{G}_\psi(\cdot, \cdot)$ denotes the Fréchet derivative of $\mathcal{G}(\cdot, \cdot)$ with respect to the second argument.

Approximations are defined by introducing a subspace $\mathcal{X}^h \subset \mathcal{X}$ and an approximating operator $\mathcal{T}^h \in \mathcal{L}(\mathcal{Y}; \mathcal{X}^h)$. Then, given $\lambda \in \Lambda$, we seek $\psi^h \in \mathcal{X}^h$ such that

(3.6) $$\mathcal{H}^h(\lambda, \psi^h) \equiv \psi^h + \mathcal{T}^h\mathcal{G}(\lambda, \psi^h) = 0\,.$$

Concerning the operator $\mathcal{T}^h$, we assume the approximation properties

$$\lim_{h\to 0} \|(\mathcal{T}^h - \mathcal{T})\omega\|_{\mathcal{X}} = 0 \quad \forall\, \omega \in \mathcal{Y} \tag{3.7}$$

and

$$\lim_{h\to 0} \|(\mathcal{T}^h - \mathcal{T})\|_{\mathcal{L}(\mathcal{Z};\mathcal{X})} = 0\,. \tag{3.8}$$

Note that whenever the imbedding $\mathcal{Z} \subset \mathcal{Y}$ is compact, (3.8) follows from (3.7) and, moreover, (3.5) implies that the operator $\mathcal{T}\mathcal{G}_\psi(\lambda, \psi) \in \mathcal{L}(\mathcal{X}; \mathcal{X})$ is compact.

We can now state the result of [6] or [10] that will be used in the sequel. In the statement of the theorem, $D^2\mathcal{G}$ represents any and all second Fréchet derivatives of $\mathcal{G}$.

THEOREM 3.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces and Λ a compact subset of $\mathbb{R}$. Assume that $\mathcal{G}$ is a C^2 mapping from $\Lambda \times \mathcal{X}$ into $\mathcal{Y}$ and that $D^2\mathcal{G}$ is bounded on all bounded sets of $\Lambda \times \mathcal{X}$. Assume that* (3.5), (3.7), *and* (3.8) *hold and that $\{(\lambda, \psi(\lambda));\ \lambda \in \Lambda\}$ is a branch of regular solutions of* (3.4). *Then, there exist a neighborhood $\mathcal{O}$ of the origin in $\mathcal{X}$ and, for $h \le h_0$ small enough, a unique C^2 function $\lambda \to \psi^h(\lambda) \in \mathcal{X}^h$ such that $\{(\lambda, \psi^h(\lambda));\ \lambda \in \Lambda\}$ is a branch of regular solutions of* (3.6) *and $\psi^h(\lambda) - \psi(\lambda) \in \mathcal{O}$ for all $\lambda \in \Lambda$. Moreover, there exists a constant $C > 0$, independent of h and λ, such that*

$$\|\psi^h(\lambda) - \psi(\lambda)\|_{\mathcal{X}} \le C\|(\mathcal{T}^h - \mathcal{T})\mathcal{G}(\lambda, \psi(\lambda))\|_{\mathcal{X}} \quad \forall\, \lambda \in \Lambda\,. \tag{3.9}$$ □

3.4. Error estimates for the approximation of solutions of the optimality system. We now apply the result of Theorem 3.1 to study the approximation of solutions of the optimality system. Set $\mathcal{X} = X \times G \times Y^*$, $\mathcal{Y} = Y \times X^*$, $\mathcal{Z} = Z \times \hat{Z}$, and $\mathcal{X}^h = X^h \times G^h \times (Y^*)^h$. (Recall that $\hat{Z}$ was introduced in (H18).) By the hypotheses on Z and $\hat{Z}$, we have that $\mathcal{Z}$ is compactly imbedded into $\mathcal{Y}$. Let $\mathcal{T} \in \mathcal{L}(\mathcal{Y}; \mathcal{X})$ be defined in the following manner: $\mathcal{T}(\tilde{r}, \tilde{\tau}) = (\tilde{u}, \tilde{g}, \tilde{\xi})$ for $(\tilde{r}, \tilde{\tau}) \in \mathcal{Y}$ and $(\tilde{u}, \tilde{g}, \tilde{\xi}) \in \mathcal{X}$ if and only if

$$\tilde{u} + T\tilde{r} = 0\,, \tag{3.10}$$

$$\tilde{\xi} + T^*\tilde{\tau} = 0\,, \tag{3.11}$$

and

$$\tilde{g} - EK^*\tilde{\xi} = 0\,. \tag{3.12}$$

Similarly, the operator $\mathcal{T}^h \in \mathcal{L}(\mathcal{Y}; \mathcal{X}^h)$ is defined as follows: $\mathcal{T}^h(\tilde{r}, \tilde{\tau}) = (\tilde{u}^h, \tilde{g}^h, \tilde{\xi}^h)$ for $(\tilde{r}, \tilde{\tau}) \in \mathcal{Y}$ and $(\tilde{u}^h, \tilde{g}^h, \tilde{\xi}^h) \in \mathcal{X}^h$ if and only if

$$\tilde{u}^h + T^h\tilde{r} = 0\,, \tag{3.13}$$

$$\tilde{\xi}^h + (T^*)^h\tilde{\tau} = 0\,, \tag{3.14}$$

and

$$\tilde{g}^h - E^hK^*\tilde{\xi}^h = 0\,. \tag{3.15}$$

The nonlinear mapping $\mathcal{G} : \Lambda \times \mathcal{X} \to \mathcal{Y}$ is defined as follows: $\mathcal{G}\big(\lambda, (\tilde{u}, \tilde{g}, \tilde{\xi})\big) = (\tilde{r}, \tilde{\tau})$ for $\lambda \in \Lambda$, $(\tilde{u}, \tilde{g}, \tilde{\xi}) \in \mathcal{X}$, and $(\tilde{r}, \tilde{\tau}) \in \mathcal{Y}$ if and only if

$$\tilde{r} = \lambda N(\tilde{u}) + \lambda K\tilde{g} \tag{3.16}$$

and

$$\tilde{\tau} = \lambda\,[N'(\tilde{u})]^*\tilde{\xi} - \lambda\,\mathcal{F}'(\tilde{u})\,. \tag{3.17}$$

It is evident that the optimality system (2.23)-(2.25) and its finite-dimensional counterpart (3.1)-(3.3) can be written as

$$(u, g, \xi) + \mathcal{T}\mathcal{G}\big(\lambda, (u, g, \xi)\big) = 0$$

and

$$(u^h, g^h, \xi^h) + \mathcal{T}^h\mathcal{G}\big(\lambda, (u^h, g^h, \xi^h)\big) = 0\,,$$

respectively, i.e., with $\psi = (u, g, \xi)$ and $\psi^h = (u^h, g^h, \psi^h)$, in the form of (3.4) and (3.6), respectively.

Now we examine the approximation properties of $\mathcal{T}^h$.

LEMMA 3.2. *Let the operators $\mathcal{T}$ and $\mathcal{T}^h$ be defined by* (3.10)-(3.12) *and* (3.13)-(3.15), *respectively. Assume that the hypotheses* (H13)-(H15) *hold. Then*

$$\lim_{h\to 0} \|(\mathcal{T} - \mathcal{T}^h)(r, \tau)\|_{\mathcal{X}} = 0 \quad \forall\,(r, \tau) \in \mathcal{Y}\,. \tag{3.18}$$

Proof. Let $(\tilde{u}, \tilde{g}, \tilde{\xi}) = \mathcal{T}(r, \tau)$; i.e., $(\tilde{u}, \tilde{g}, \tilde{\xi})$ satisfies (3.10)-(3.12). Let $(\tilde{u}^h, \tilde{g}^h, \tilde{\xi}^h) = \mathcal{T}^h(r, \tau)$; i.e., $(\tilde{u}^h, \tilde{g}^h, \tilde{\xi}^h)$ satisfies (3.13)-(3.15). Subtracting the corresponding equations yields that

$$\|\tilde{u} - \tilde{u}^h\|_X = \|(T - T^h)r\|_X\,,$$
$$\big\|\tilde{\xi} - \tilde{\xi}^h\big\|_{Y^*} = \|(T^* - (T^*)^h)\tau\|_{Y^*}\,,$$

and

$$\begin{aligned}\|\tilde{g} - \tilde{g}^h\|_G &= \big\|(E - E^h)K^*\tilde{\xi}^h + EK^*\big(\tilde{\xi} - \tilde{\xi}^h\big)\big\|_G \\ &\le \big\|(E - E^h)K^*\tilde{\xi}^h\big\|_G + \|EK^*\|_{\mathcal{L}(Y^*;G)}\,\big\|\big(\tilde{\xi} - \tilde{\xi}^h\big)\big\|_G\,.\end{aligned}$$

Thus, for some constant $C > 0$,

$$\begin{aligned}&\|(\mathcal{T} - \mathcal{T}^h)(r, \tau)\|_{\mathcal{X}} \\ &\quad\le C\left\{\|(T - T^h)r\|_X + \|(T^* - (T^*)^h)\tau\|_{Y^*} + \|(E - E^h)K^*\tilde{\xi}^h\|_G\right\}.\end{aligned}$$

Then the result of the proposition follows from (H13)-(H15). □

Next, we examine the derivative of the mapping $\mathcal{G}$.

LEMMA 3.3. *Let the mapping $\mathcal{G} : \Lambda \times \mathcal{X} \to \mathcal{Y}$ be defined by* (3.16)-(3.17). *Assume that the hypotheses* (H9), (H16), *and* (H18)-(H19) *hold. Then, for every $\lambda \in \Lambda$ and every $(u, g, \xi) \in \mathcal{X}$, the operator $\mathcal{G}_{(u,g,\xi)}\big(\lambda, (u, g, \xi)\big) \in \mathcal{L}(\mathcal{X}; \mathcal{Z})$.*

Proof. A simple calculation shows that $\mathcal{G}_{(u,g,\xi)}\big(\lambda, (u, g, \xi)\big) \in \mathcal{L}(\mathcal{X}; \mathcal{Y})$ is given by

$$\mathcal{G}_{(u,g,\xi)}\big(\lambda, (u, g, \xi)\big)\cdot\big(\tilde{u}, \tilde{g}, \tilde{\xi}\big) = \lambda\begin{pmatrix} N'(u)\cdot\tilde{u} + K\tilde{g} \\ [N''(u)\cdot\tilde{u}]^*\cdot\xi + [N'(u)]^*\cdot\tilde{\xi} - \mathcal{F}''(u)\cdot\tilde{u}\end{pmatrix}.$$

Then, the result follows from (H9) and (H18)-(H19). □

A solution $\big(u(\lambda), g(\lambda), \xi(\lambda)\big)$ of the optimality system (2.23)–(2.25) is called *regular* if the system (for the unknowns $(\tilde{u}, \tilde{g}, \tilde{\xi})$)

$$\tilde{u} + \lambda T N'(u)\tilde{u} + \lambda T K \tilde{g} = \tilde{x}\,, \tag{3.19}$$

$$\tilde{\xi} + \lambda T^*[N''(u)]^*\tilde{u}\cdot\xi + \lambda T^*[N'(u)]^*\tilde{\xi} - \lambda T^*\mathcal{F}''(u)\tilde{u} = \tilde{y}\,, \tag{3.20}$$

and

$$\tilde{g} - EK^*\tilde{\xi} = \tilde{z} \tag{3.21}$$

is uniquely solvable for any $(\tilde{x}, \tilde{z}, \tilde{y}) \in \mathcal{X} = X \times G \times Y^*$. (Note that the linear operator appearing on the left-hand side of (3.19)–(3.21) is obtained by linearizing the optimality system (2.23)–(2.25) about (u, g, ξ).)

In the following theorem, we will assume that the solution $\big(u(\lambda), g(\lambda), \xi(\lambda)\big)$ of the optimality system (2.23)–(2.25) that we are trying to approximate is a regular solution. The assumptions we have made, in particular (H9), (H18)–(H19), are sufficient to guarantee that for almost all values of λ, this is indeed the case.

LEMMA 3.4. *Assume the hypotheses of Lemma* 3.3. *Then, for almost all* λ, *solutions* $\big(u(\lambda), g(\lambda), \xi(\lambda)\big)$ *of the optimality system* (2.23)–(2.25) *are regular.*

Proof. The system (3.19)–(3.21) is equivalent to

$$\Big(I + \lambda \mathcal{T}\mathcal{S}(u, g, \xi)\Big)(\tilde{u}, \tilde{g}, \tilde{\xi}) = (\tilde{x}, \tilde{z}, \tilde{y})\,, \tag{3.22}$$

where the linear operator $\mathcal{S}(u, g, \xi) : \mathcal{X} \to \mathcal{Y}$ is defined by

$$\begin{aligned}\mathcal{S}(u, g, \xi)\cdot(\tilde{u}, \tilde{g}, \tilde{\xi}) &\equiv \frac{1}{\lambda}\mathcal{G}_{(u,g,\xi)}\big(\lambda, (u, g, \xi)\big)\cdot(\tilde{u}, \tilde{g}, \tilde{\xi}) \\ &= \begin{pmatrix} N'(u)\cdot\tilde{u} + K\tilde{g} \\ [N''(u)\cdot\tilde{u}]^*\cdot\xi + [N'(u)]^*\cdot\tilde{\xi} - \mathcal{F}''(u)\cdot\tilde{u} \end{pmatrix}.\end{aligned}$$

Now, $\mathcal{T} \in \mathcal{L}(\mathcal{Y}; \mathcal{X})$; hence, by Lemma 3.3, $\big(I + \lambda\mathcal{T}\mathcal{S}(u, g, \xi)\big)$ is a compact perturbation of the identity operator from $\mathcal{X}$ to $\mathcal{X}$. Thus, for almost all λ, (3.22), or equivalently (3.19)–(3.21), is uniquely solvable; i.e., for almost all λ, the solution $\big(u(\lambda), g(\lambda), \xi(\lambda)\big)$ of the optimality system (2.23)–(2.25) is regular. □

Using Theorem 3.1, we can now provide an error estimate for approximations of solutions of the abstract problem.

THEOREM 3.5. *Let* $\big(u(\lambda), g(\lambda), \xi(\lambda)\big) \in \mathcal{X}$, *for* $\lambda \in \Lambda$, *be a branch of regular solutions of the optimality system* (2.23)–(2.25). *Assume that the hypotheses* (H13)–(H19) *hold. Then, there exist a* $\delta > 0$ *and an* $h_0 > 0$ *such that for* $h < h_0$ *the discrete optimality system* (3.1)–(3.3) *has a unique solution* $\big(u^h(\lambda), g^h(\lambda), \xi^h(\lambda)\big)$ *satisfying*

$$\big\|\big(u(\lambda), g(\lambda), \xi(\lambda)\big) - \big(u^h(\lambda), g^h(\lambda), \xi^h(\lambda)\big)\big\|_{\mathcal{X}} < \delta\,.$$

Moreover,

$$\lim_{h\to 0}\big\|\big(u(\lambda), g(\lambda), \xi(\lambda)\big) - \big(u^h(\lambda), g^h(\lambda), \xi^h(\lambda)\big)\big\|_{\mathcal{X}} = 0 \tag{3.23}$$

uniformly in $\lambda \in \Lambda$ *and there exists a constant* C, *independent of* h *and* λ, *such that*

$$\begin{aligned}&\lim_{h\to 0}\big\|\big(u(\lambda), g(\lambda), \xi(\lambda)\big) - \big(u^h(\lambda), g^h(\lambda), \xi^h(\lambda)\big)\big\|_{\mathcal{X}} \\ &\quad\le C\lambda\Big\{\big\|(T^h - T)\big(N(u(\lambda)) + Kg(\lambda)\big)\big\|_X + \|(E^h - E)K^*\xi(\lambda)\|_G \\ &\qquad + \big\|\big((T^*)^h - T^*\big)\Big([N'(u(\lambda))]^*\xi - \mathcal{F}'(u(\lambda))\Big)\big\|_{Y^*}\Big\}.\end{aligned} \tag{3.24}$$

Proof. Assumptions (H16) and (H17) ensure that $\mathcal{G} \in C^2(\mathcal{X}, \mathcal{Y})$ and $D^2\mathcal{G}$ maps bounded sets of $\Lambda \times \mathcal{X}$ into bounded sets of $\mathcal{Y}$. By Lemma 3.3, assumptions (H18) and (H19) imply that (3.5) holds. By Lemma 3.2, assumptions (H13)-(H15) imply that (3.7) holds. Then, since $\mathcal{Z}$ is compactly imbedded into $\mathcal{Y}$, (3.7) implies that (3.8) holds. Thus, all the hypotheses of Theorem 3.1 are verified. Then, a direct application of Theorem 3.1 yields (3.23) and (3.24) follows from (3.9). □

It is easily seen that (3.23) and (3.24) are equivalent to

$$\lim_{h\to 0}\Big\{\|u(\lambda) - u^h(\lambda)\|_X + \|g(\lambda) - g^h(\lambda)\|_G + \|\xi(\lambda) - \xi^h(\lambda)\|_{Y^*}\Big\} = 0$$

uniformly in $\lambda \in \Lambda$ and that there exists a constant C, independent of h and λ, such that

$$\begin{aligned}\|u(\lambda) - u^h(\lambda)\|_X &+ \|g(\lambda) - g^h(\lambda)\|_G + \|\xi(\lambda) - \xi^h(\lambda)\|_{Y^*}\\ &\le C\lambda\Big\{\|(T^h - T)\big(N(u(\lambda)) + Kg(\lambda)\big)\|_X + \|(E^h - E)K^*\xi(\lambda)\|_G\\ &\qquad + \|\big((T^*)^h - T^*\big)\big([N'(u(\lambda))]^*\xi(\lambda) - \mathcal{F}'(u(\lambda))\big)\|_{Y^*}\Big\}.\end{aligned}$$

If, in (3.9), the operator $\mathcal{T}$ is invertible, we have, using (3.4), that

$$\|\psi^h(\lambda) - \psi(\lambda)\|_{\mathcal{X}} \le C\|(\mathcal{T}^h\mathcal{T}^{-1} - I)\psi(\lambda)\|_{\mathcal{X}} \quad \forall\, \lambda \in \Lambda\,.$$

Thus, if the operator T from Y to X is invertible, we have that (3.24) is equivalent to

$$\begin{aligned}\|u(\lambda) - u^h(\lambda)\|_X &+ \|g(\lambda) - g^h(\lambda)\|_G + \|\xi(\lambda) - \xi^h(\lambda)\|_{Y^*}\\ &\le C\Big\{\|(T^hT^{-1} - I)u(\lambda)\|_X + \|(E^hE^{-1} - I)g(\lambda)\|_G\\ &\qquad + \|\big((T^*)^h(T^*)^{-1} - I\big)\xi(\lambda)\|_{Y^*}\Big\}.\end{aligned} \tag{3.25}$$

4. Applications. We now apply the framework and analyses developed in §§2 and 3 to some concrete problems, all of which feature constraints on admissible states and controls that take the form of a system of nonlinear partial differential equations. In each application, we use a different control mechanism, so the discussion provided in this section illustrates the treatment of a variety of such mechanisms. However, one could use any of the control mechanisms discussed in any of the applications in any other application or, in fact, use any combination of such mechanisms.

Before examining any specific application, we establish some notation. Further notation will be established as needed when the individual applications are considered.

Throughout, C will denote a positive constant whose meaning and value changes with context. Also, $H^s(\mathcal{D})$ for $s \in \mathbb{R}$ denotes the standard real Sobolev space of order s with respect to the set $\mathcal{D}$, where $\mathcal{D}$ could either be a bounded domain $\Omega \in \mathbb{R}^d$, $d = 2, 3$, or part of the boundary Γ of such a domain. Of particular interest are the spaces $H^0(\mathcal{D}) = L^2(\mathcal{D})$,

$$H^1(\mathcal{D}) = \left\{\phi \in L^2(\mathcal{D}) \;\middle|\; \frac{\partial\phi}{\partial x_j} \in L^2(\mathcal{D}) \quad \text{for } j = 1, \ldots, d\right\},$$

and

$$H^2(\mathcal{D}) = \left\{\phi \in L^2(\mathcal{D}) \;\middle|\; \frac{\partial\phi}{\partial x_j}\,,\ \frac{\partial^2\phi}{\partial x_j\partial x_k} \in L^2(\mathcal{D}) \quad \text{for } j, k = 1, \ldots, d\right\}.$$

Also of interest is the subspace

$$H_0^1(\mathcal{D}) = \left\{ \phi \in H^1(\mathcal{D}) \;\middle|\; \phi = 0 \quad \text{on } \partial\mathcal{D} \right\},$$

where $\partial\mathcal{D}$ denotes the boundary of $\mathcal{D}$.

Dual spaces will be denoted by $(\cdot)^*$. Duality pairings between spaces and their duals will be denoted by $\langle \cdot, \cdot \rangle$. Norms of functions belonging to $H^s(\Omega)$ and $H^s(\Gamma)$ are denoted by $\| \cdot \|_s$ and $\| \cdot \|_{s,\Gamma}$, respectively. Of particular interest are the $L^2(\Omega)$-norm $\| \cdot \|_0$, the $H^1(\Omega)$-norm

$$\|\phi\|_1^2 = \sum_{j=1}^{d} \left\| \frac{\partial \phi}{\partial x_j} \right\|_0^2 + \|\phi\|_0^2,$$

and the $H^2(\Omega)$-norm

$$\|\phi\|_2^2 = \sum_{j,k=1}^{d} \left\| \frac{\partial^2 \phi}{\partial x_j \partial x_k} \right\|_0^2 + \|\phi\|_1^2.$$

Corresponding Sobolev spaces of real, vector-valued functions having r components will be denoted by $\mathbf{H}^s(\mathcal{D})$, e.g., $\mathbf{H}^1(\mathcal{D}) = [H^1(\mathcal{D})]^r$. Of particular interest will be the spaces $\mathbf{L}^2(\mathcal{D}) = \mathbf{H}^0(\mathcal{D}) = [L^2(\mathcal{D})]^r$,

$$\mathbf{H}^1(\mathcal{D}) = \left\{ v_j \in L^2(\mathcal{D}) \;\middle|\; \frac{\partial v_j}{\partial x_k} \in L^2(\mathcal{D}) \quad \text{for } j = 1, \ldots, r \text{ and } k = 1, \ldots, d \right\},$$

and

$$\mathbf{H}^2(\mathcal{D}) = \left\{ v_j \in L^2(\mathcal{D}) \;\middle|\; \frac{\partial v_j}{\partial x_k} \in L^2(\mathcal{D}),\ \frac{\partial^2 v_j}{\partial x_k \partial x_\ell} \in L^2(\Omega) \right.$$
$$\left. \text{for } j = 1, \ldots, r \text{ and } k, \ell = 1, \ldots, d \right\},$$

where v_j, $j = 1, \ldots, r$, denote the components of $\mathbf{v}$. Also of interest is the subspace

$$\mathbf{H}_0^1(\mathcal{D}) = \left\{ \mathbf{v} \in \mathbf{H}^1(\mathcal{D}) \;\middle|\; v_j = 0 \text{ on } \partial\mathcal{D},\ j = 1, \ldots, r \right\}.$$

Norms for spaces of vector-valued functions will be denoted by the same notation as that used for their scalar counterparts. For example,

$$\|\mathbf{v}\|_s^2 = \sum_{j=1}^{r} \|v_j\|_s^2 \quad \text{and} \quad \|\mathbf{v}\|_{s,\Gamma}^2 = \sum_{j=1}^{r} \|v_j\|_{s,\Gamma}^2.$$

We denote the $L^2(\Omega)$ and $\mathbf{L}^2(\Omega)$ inner products by $(\cdot, \cdot)$; i.e., for $p, q \in L^2(\Omega)$ and $\mathbf{u}, \mathbf{v} \in \mathbf{L}^2(\Omega)$

$$(p, q) = \int_\Omega pq \, d\Omega \quad \text{and} \quad (\mathbf{u}, \mathbf{v}) = \int_\Omega \mathbf{u} \cdot \mathbf{v} \, d\Omega.$$

Similarly, we denote by $(\cdot, \cdot)_\Gamma$ the $L^2(\Gamma)$ and $\mathbf{L}^2(\Gamma)$ inner products; i.e., for $p, q \in L^2(\Gamma)$ and $\mathbf{u}, \mathbf{v} \in \mathbf{L}^2(\Gamma)$

$$(p, q)_\Gamma = \int_\Gamma pq \, d\Gamma \quad \text{and} \quad (\mathbf{u}, \mathbf{v})_\Gamma = \int_\Gamma \mathbf{u} \cdot \mathbf{v} \, d\Gamma.$$

Since in all cases L^2-spaces will be used as pivot spaces, the above inner product notation can also be used to denote duality pairings between functions defined on H^s-spaces and their dual spaces.

For details concerning the notation employed, one may consult, e.g., [1].

4.1. Distributed controls for the von Kármán plate equations. For this application we will use distributed controls; i.e., control is effected through a source term in the governing partial differential equations. Let Ω be a bounded, convex polygonal domain in $\mathbb{R}^2$, and let Γ denote the boundary of Ω. The von Kármán equations for a clamped plate are given by (see, e.g., [9] or [18])

$$\Delta^2\psi_1 + \frac{1}{2}[\psi_2, \psi_2] = 0 \quad \text{in } \Omega,$$

$$\Delta^2\psi_2 - [\psi_1, \psi_2] = \lambda g \quad \text{in } \Omega,$$

and

$$\psi_1 = \frac{\partial\psi_1}{\partial n} = \psi_2 = \frac{\partial\psi_2}{\partial n} = 0 \quad \text{on } \Gamma,$$

where

$$[\psi, \phi] = \frac{\partial^2\psi}{\partial x_1^2}\frac{\partial^2\phi}{\partial x_2^2} + \frac{\partial^2\psi}{\partial x_2^2}\frac{\partial^2\phi}{\partial x_1^2} - 2\frac{\partial^2\psi}{\partial x_1 x_2}\frac{\partial^2\phi}{\partial x_1 x_2}.$$

Here, ψ_1 denotes the Airy stress function, ψ_2 the deflection of the plate in the direction normal to the plate, λg is an external load normal to the plate that depends on the loading parameter λ, and $\partial(\cdot)/\partial n$ is the normal derivative in the direction of the outer normal to Γ.

By introducing appropriate rescalings, i.e., by replacing ψ_1 by $\lambda\psi_1$, ψ_2 by $\lambda\psi_2$, and g by λg, we can rewrite the von Kármán equations as follows:

$$\Delta^2\psi_1 + \frac{\lambda}{2}[\psi_2, \psi_2] = 0 \quad \text{in } \Omega, \tag{4.1}$$

$$\Delta^2\psi_2 - \lambda\,[\psi_1, \psi_2] = \lambda\, g \quad \text{in } \Omega, \tag{4.2}$$

and

$$\psi_1 = \frac{\partial\psi_1}{\partial n} = \psi_2 = \frac{\partial\psi_2}{\partial n} = 0 \quad \text{on } \Gamma. \tag{4.3}$$

We introduce the spaces

$$H_0^2(\Omega) = \left\{ \psi \in H^2(\Omega) \;\middle|\; \psi = 0,\ \frac{\partial\psi}{\partial n} = 0 \ \text{ on } \Gamma \right\},$$

$$\mathbf{H}_0^2(\Omega) = [H_0^2(\Omega)]^2, \quad H^{-2}(\Omega) = \left(H_0^2(\Omega)\right)^*, \quad \text{and} \quad \mathbf{H}^{-2}(\Omega) = \left(\mathbf{H}_0^2(\Omega)\right)^*$$

and the bilinear form

$$a(\psi, \phi) = \int_\Omega \Delta\psi\, \Delta\phi\, d\Omega \quad \forall\, \psi, \phi \in H^2(\Omega)$$

in order to define the following weak formulation of the von Kármán equations (4.1)-(4.3). Find $\boldsymbol{\psi} = (\psi_1, \psi_2) \in \mathbf{H}_0^2(\Omega)$ such that

$$a(\psi_1, \phi_1) + \frac{\lambda}{2}\big([\psi_2, \psi_2], \phi_1\big) = 0 \quad \forall\, \phi_1 \in H_0^2(\Omega) \tag{4.4}$$

and

$$a(\psi_2, \phi_2) - \lambda\big([\psi_1, \psi_2], \phi_2\big) = \lambda\langle g, \phi_2\rangle \quad \forall\, \phi_2 \in H_0^2(\Omega). \tag{4.5}$$

Using the identity

$$([\psi, \phi], \zeta) = ([\psi, \zeta], \phi) \quad \forall\, \psi, \phi, \zeta \in H_0^2(\Omega)\,, \tag{4.6}$$

one can show that for each $g \in H^{-2}(\Omega)$, (4.4)-(4.5) possesses at least one solution $\boldsymbol{\psi} = (\psi_1, \psi_2) \in \mathbf{H}_0^2(\Omega)$ and that all solutions of (4.4)-(4.5) satisfy the a priori estimate

$$\|\psi_1\|_2 + \|\psi_2\|_2 \leq C\|g\|_{-2}\,; \tag{4.7}$$

see, e.g., [18], for details. In the sequel a solution to (4.1)-(4.3) will be understood in the sense of (4.4)-(4.5).

Given a desired state $\boldsymbol{\psi}_0 = (\psi_{10}, \psi_{20}) \in \mathbf{L}^2(\Omega)$, we define for any $\boldsymbol{\psi} = (\psi_1, \psi_2) \in \mathbf{H}_0^2(\Omega)$ and $g \in L^2(\Omega)$ the functional

$$\begin{aligned} \mathcal{J}(\boldsymbol{\psi}, g) &= \mathcal{J}(\psi_1, \psi_2, g) \\ &= \frac{\lambda}{2}\int_\Omega \left((\psi_1 - \psi_{10})^2 + (\psi_2 - \psi_{20})^2\right) d\Omega + \frac{\lambda}{2}\int_\Omega g^2\, d\Omega\,. \end{aligned} \tag{4.8}$$

We then consider the following optimal control problem associated with the von Kármán plate equations:

$$\min\left\{\, \mathcal{J}(\boldsymbol{\psi}, g) \mid \boldsymbol{\psi} \in \mathbf{H}_0^2(\Omega)\,,\ g \in \Theta \,\right\} \quad \text{subject to} \quad (4.4)\text{-}(4.5)\,, \tag{4.9}$$

where Θ is a subset of $L^2(\Omega)$.

We define the spaces $X = \mathbf{H}_0^2(\Omega)$, $Y = \mathbf{H}^{-2}(\Omega)$, $G = L^2(\Omega)$, and $Z = \mathbf{L}^1(\Omega)$. By compact imbedding results, $Z \hookrightarrow\hookrightarrow Y$. For the time being, we assume that the admissible set Θ for the control g is a closed, convex subset of $G = L^2(\Omega)$.

Let the continuous linear operator $T \in \mathcal{L}(Y; X)$ be defined as follows. For $\mathbf{f} = (f_1, f_2) \in Y = \mathbf{H}^{-2}(\Omega)$, $T\mathbf{f} = \boldsymbol{\psi} \in X = \mathbf{H}_0^2(\Omega)$ is the unique solution of

$$a(\psi_1, \phi_1) = \langle f_1, \phi_1\rangle \quad \forall\, \phi_1 \in H_0^2(\Omega)$$

and

$$a(\psi_2, \phi_2) = \langle f_2, \phi_2\rangle \quad \forall\, \phi_2 \in H_0^2(\Omega)\,.$$

It can be easily verified that T is self-adjoint.

We define the (differentiable) nonlinear mapping $N : X \to Y$ by

$$N(\boldsymbol{\psi}) = \begin{pmatrix} \frac{1}{2}[\psi_2, \psi_2] \\ -[\psi_1, \psi_2] \end{pmatrix} \quad \forall\, \boldsymbol{\psi} \in X$$

or, equivalently,

$$\langle N(\boldsymbol{\psi}), \boldsymbol{\phi}\rangle = \frac{1}{2}([\psi_2, \psi_2], \phi_1) - ([\psi_1, \psi_2], \phi_2) \quad \forall\, \boldsymbol{\phi} = (\phi_1, \phi_2) \in X$$

and define $K : g \in L^2(\Omega) \to Y$ by

$$Kg = -\begin{pmatrix} 0 \\ g \end{pmatrix}$$

or, equivalently,

$$\langle Kg, \boldsymbol{\phi}\rangle = -\langle g, \phi_2\rangle \quad \forall\, \boldsymbol{\phi} = (\phi_1, \phi_2) \in X\,.$$

Clearly, the constraint equations (4.4)-(4.5) can be expressed as

$$\boldsymbol{\psi} + \lambda\, T N(\boldsymbol{\psi}) + \lambda\, T K g = 0\,,$$

i.e., in the form (2.2). With the obvious definitions for $\mathcal{F}(\cdot)$ and $\mathcal{E}(\cdot)$, i.e.,

$$\mathcal{F}(\boldsymbol{\psi}) = \frac{1}{2}\int_\Omega \Big((\psi_1 - \psi_{10})^2 + (\psi_2 - \psi_{20})^2\Big)\, d\Omega \quad \forall\, \boldsymbol{\psi} \in X$$

and

$$\mathcal{E}(g) = \frac{1}{2}\int_\Omega g^2\, d\Omega \quad \forall\, g \in G\,,$$

the functional (4.8) can be expressed as

$$\mathcal{J}(\boldsymbol{\psi}, g) = \lambda\, \mathcal{F}(\boldsymbol{\psi}) + \lambda\, \mathcal{E}(g)\,,$$

i.e., in the form (2.1). Thus, the minimization problem (4.9) is in the form of the minimization problem (2.3).

We are now in a position to verify, for the minimization problem (4.9), all the hypotheses of §§2 and 3.

4.1.1. Verification of the hypotheses for the existence of optimal solutions. We first verify that the hypotheses (H1)-(H6) hold in the current setting.

(H1) is obviously satisfied with a lower bound 0.

(H2) holds with $\alpha = 1$ and $\beta = 2$.

(H3) is verified with the choice $(\boldsymbol{\psi}^{(0)}, g^{(0)}) \in X \times \Theta$, where $g^{(0)}$ is an arbitrarily chosen element in Θ and $\boldsymbol{\psi}^{(0)} = \big(\psi_1^{(0)}, \psi_2^{(0)}\big)$ is a solution of

$$\Delta^2 \psi_1^{(0)} + \frac{\lambda}{2}[\psi_2^{(0)}, \psi_2^{(0)}] = 0 \quad \text{in } \Omega\,,$$

$$\Delta^2 \psi_2^{(0)} - \lambda\, [\psi_1^{(0)}, \psi_2^{(0)}] = \lambda\, g^{(0)} \quad \text{in } \Omega\,,$$

and

$$\psi_1^{(0)} = \frac{\partial \psi_1^{(0)}}{\partial n} = \psi_2^{(0)} = \frac{\partial \psi_2^{(0)}}{\partial n} = 0 \quad \text{on } \Gamma\,.$$

In order to verify (H4), we assume $\{g^{(n)}\} \subset \Theta$ is a sequence satisfying $g^{(n)} \rightharpoonup g$ in $L^2(\Omega)$; then, we have $g^{(n)} \rightharpoonup g$ in $H^{-2}(\Omega)$, so $\lim_{n\to\infty}\langle g^{(n)}, z\rangle = \langle g, z\rangle$ for all $z \in H^2(\Omega)$, i.e., $Kg^{(n)} \rightharpoonup Kg$ in Y. Assume that the sequence $\{\boldsymbol{\psi}^{(n)}\} \subset \mathbf{H}_0^2(\Omega)$ satisfies $\boldsymbol{\psi}^{(n)} \rightharpoonup \boldsymbol{\psi}$ in $\mathbf{H}_0^2(\Omega)$; then, $(\partial^2 \boldsymbol{\psi}^{(n)}/\partial x_i \partial x_j) \rightharpoonup (\partial^2 \boldsymbol{\psi}/\partial x_i \partial x_j)$ in $\mathbf{L}^2(\Omega)$ and, by using a compact imbedding result, $\boldsymbol{\psi}^{(n)} \to \boldsymbol{\psi}$ in $\mathbf{L}^2(\Omega)$. Now, using the identity (4.6),

$$\begin{aligned}
\langle N(\boldsymbol{\psi}^{(n)}), \boldsymbol{\phi}\rangle &= \frac{1}{2}\Big([\psi_2^{(n)}, \psi_2^{(n)}], \phi_1\Big) - \Big([\psi_1^{(n)}, \psi_2^{(n)}], \phi_2\Big) \\
&= \frac{1}{2}\Big([\psi_2^{(n)}, \phi_1], \psi_2^{(n)}\Big) - \Big([\psi_1^{(n)}, \phi_2], \psi_2^{(n)}\Big) \\
&\to \frac{1}{2}\,([\psi_2, \phi_1], \psi_2) - ([\psi_1, \phi_2], \psi_2) \\
&= \frac{1}{2}\,([\psi_2, \psi_2], \phi_1) - ([\psi_1, \psi_2], \phi_2) = \langle N(\boldsymbol{\psi}), \boldsymbol{\phi}\rangle\,.
\end{aligned}$$

Hence, (H4) is verified.

The verification of (H5) follows directly from the observation that the mappings $\boldsymbol{\phi} \mapsto \mathcal{F}(\boldsymbol{\phi}) = (1/2)\|\boldsymbol{\phi} - \boldsymbol{\psi}_0\|_0^2$ and $g \mapsto \mathcal{E}(g) = (1/2)\|g\|_0^2$ are convex.

The verification of (H6) is a trivial consequence of the a priori estimate (4.7).

It is now just a matter of citing Theorem 2.1 to prove the existence of an optimal solution that minimizes (4.8) subject to (4.4)-(4.5).

THEOREM 4.1. *There exists a* $(\boldsymbol{\phi}, g) \in \mathbf{H}_0^2(\Omega) \times \Theta$ *such that* (4.8) *is minimized subject to* (4.4)-(4.5). □

4.1.2. Verification of the hypotheses for the existence of Lagrange multipliers. We now assume $(\boldsymbol{\psi}, g)$ is an optimal solution and turn to the verification of hypotheses (H7)-(H9).

The validity of (H7) is obvious.

(H8) holds since the mapping $g \mapsto \mathcal{E}(g) = (1/2)\|g\|_0^2$ is convex.

(H9) can be verified as follows. For any $\boldsymbol{\psi} \in X$, the operator $N'(\boldsymbol{\psi}) : X \to Y$ is given by

$$N'(\boldsymbol{\psi}) \cdot \boldsymbol{\phi} = \begin{pmatrix} [\psi_2, \phi_2] \\ -[\psi_1, \phi_2] - [\psi_2, \phi_1] \end{pmatrix} \quad \forall\, \boldsymbol{\phi} = (\phi_1, \phi_2) \in X\,.$$

Thus, using the definition of $[\cdot, \cdot]$, we obtain that $N'(\boldsymbol{\psi}) \cdot \boldsymbol{\phi} \in \mathbf{L}^1(\Omega) = Z$.

The Lagrangian is given by

$$\mathcal{L}(\boldsymbol{\psi}, g, \boldsymbol{\eta}, k) = k\,\mathcal{J}(\boldsymbol{\psi}, g) - \Big\{ a(\psi_1, \eta_1) + \frac{\lambda}{2}\big([\psi_2, \psi_2], \eta_1\big) + a(\psi_2, \eta_2) - \lambda\big([\psi_1, \psi_2], \eta_2\big) - \lambda(g, \eta_2) \Big\}$$

for all $(\boldsymbol{\psi}, g, \boldsymbol{\eta}, k) \in X \times G \times X \times \mathbb{R} = \mathbf{H}_0^2(\Omega) \times L^2(\Omega) \times \mathbf{H}_0^2(\Omega) \times \mathbb{R}$. Note that in this form of the Lagrangian, the Lagrange multiplier $\boldsymbol{\eta} \in X = Y^*$, so we have already introduced the change of variables indicated between (2.17)-(2.18) and (2.19)-(2.21).

Having verified the hypotheses (H7)-(H9), we may apply Theorem 2.4 to conclude that there exist a Lagrange multiplier $\boldsymbol{\eta} \in X = \mathbf{H}_0^2(\Omega)$ and a real number k such that

$$\boldsymbol{\eta} + \lambda\, T^*\big([N'(\boldsymbol{\psi})]^* \cdot \boldsymbol{\eta} - k\,\mathcal{F}'(\boldsymbol{\psi})\big) = \mathbf{0} \tag{4.10}$$

and

$$\mathcal{L}(\boldsymbol{\psi}, g, \boldsymbol{\eta}, k) \le \mathcal{L}(\boldsymbol{\psi}, z, \boldsymbol{\eta}, k) \quad \forall\, z \in \Theta \tag{4.11}$$

and that for almost all values of λ, we may choose $k = 1$.

Recall that T is self-adjoint. Also, note that for any $\boldsymbol{\psi} \in X = \mathbf{H}_0^2(\Omega)$,

$$[N'(\boldsymbol{\psi})]^* \cdot \boldsymbol{\eta} = \begin{pmatrix} -[\psi_2, \eta_2] \\ [\psi_2, \eta_1] - [\psi_1, \eta_2] \end{pmatrix} \quad \forall\, \boldsymbol{\eta} = (\eta_1, \eta_2) \in X\,.$$

Thus, (4.10), with $k = 1$, can be rewritten as

$$a(\zeta_1, \eta_1) - \lambda\big([\psi_2, \eta_2], \zeta_1\big) = \lambda(\psi_1 - \psi_{10}, \zeta_1) \quad \forall\, \zeta_1 \in H_0^2(\Omega) \tag{4.12}$$

and

$$a(\zeta_2, \eta_2) + \lambda\big([\psi_2, \eta_1], \zeta_2\big) - \lambda\big([\psi_1, \eta_2], \zeta_2\big) = \lambda(\psi_2 - \psi_{20}, \zeta_2) \quad \forall\, \zeta_2 \in H_0^2(\Omega)\,. \tag{4.13}$$

Using the definition of the Lagrangian functional, (4.11), with $k = 1$, can be rewritten as

$$\frac{\lambda}{2}(z, z) + \lambda(z, \eta_2) - \frac{\lambda}{2}(g, g) - \lambda(g, \eta_2) \geq 0 \quad \forall\, z \in \Theta\,.$$

Note that, in the above expression, we have already employed hypothesis (H12), which in the current context is trivially satisfied with E the identity operator on $G^* = G = L^2(\Omega)$. For each $\epsilon \in (0, 1)$ and each $t \in \Theta$, set $z = \epsilon t + (1 - \epsilon)g \in \Theta$ in the last equation to obtain

$$\frac{\epsilon^2}{2}(t - g, t - g) + \epsilon\,(t - g, g) + \epsilon\,(t - g, \eta_2) \geq 0 \quad \forall\, t \in \Theta$$

so that, after dividing by $\epsilon > 0$ and then letting $\epsilon \to 0^+$, we obtain

$$(t - g, g + \eta_2) \geq 0 \quad \forall\, t \in \Theta\,. \tag{4.14}$$

We see that for almost all values of λ, necessary conditions for an optimum are that (4.4)-(4.5) and (4.12)-(4.14) are satisfied. The system formed by these equations will be called an *optimality system*.

We now specialize to the case $\Theta = L^2(\Omega)$. Note that the hypothesis (H10) is satisfied. Then, using Theorem 2.5, we see that the inequality (4.14) becomes an equality and, by letting $z = t - g$ vary arbitrarily in $L^2(\Omega)$, we now have, instead of (4.14),

$$(z, g + \eta_2) = 0 \quad \forall\, z \in L^2(\Omega)\,. \tag{4.15}$$

Thus, according to that theorem, we have that for almost all λ, an optimality system of equations is now given by (4.4)-(4.5), (4.12)-(4.13), and (4.15). However, we can go further and verify that the hypothesis (H11)′ is valid, which in turn will justify the existence of a Lagrange multiplier satisfying the optimality system for *all* $\lambda \in \Lambda$. We now assume the domain Ω is a convex polygon with no angles greater than 126°.

Let λ be given such that $1/\lambda$ is an eigenvalue of $-TN'(\boldsymbol{\psi})$, where $(\boldsymbol{\psi}, g) \in \mathbf{H}_0^2(\Omega) \times L^2(\Omega)$ is an optimal pair that minimizes (4.8) subject to (4.4)-(4.5). We wish to show that for each $\tilde{\mathbf{f}} \in \mathbf{H}^{-2}(\Omega)$, there exists a $\tilde{g} \in L^2(\Omega)$ and a $\tilde{\boldsymbol{\psi}} \in \mathbf{H}_0^2(\Omega)$ such that

$$\tilde{\boldsymbol{\psi}} + \lambda\, TN'(\boldsymbol{\psi}) \cdot \tilde{\boldsymbol{\psi}} + \lambda\, TK\tilde{g} = \tilde{\mathbf{f}}\,;$$

i.e.,

$$a(\tilde{\psi}_1, \phi_1) + \lambda\big([\psi_2, \tilde{\psi}_2], \phi_1\big) = \langle \tilde{f}_1, \phi_1 \rangle \quad \forall\, \phi_1 \in H_0^2(\Omega) \tag{4.16}$$

and

$$\begin{aligned} a(\tilde{\psi}_2, \phi_2) - \lambda\big([\tilde{\psi}_1, \psi_2], \phi_2\big) - \lambda\big([\psi_1, \tilde{\psi}_2], \phi_2\big) - \lambda(\tilde{g}, \phi_2) \\ = \langle \tilde{f}_2, \phi_2 \rangle \quad \forall\, \phi_2 \in H_0^2(\Omega)\,. \end{aligned} \tag{4.17}$$

To show this, we first let $\tilde{\boldsymbol{\psi}} \in \mathbf{H}_0^2(\Omega)$ be a solution of

$$a(\tilde{\psi}_1, \phi_1) + \lambda\big([\psi_2, \tilde{\psi}_2], \phi_1\big) = \langle \tilde{f}_1, \phi_1 \rangle \quad \forall\, \phi_1 \in H_0^2(\Omega)$$

and

$$a(\tilde{\psi}_2, \phi_2) - \lambda\big([\tilde{\psi}_1, \psi_2], \phi_2\big) = \langle \tilde{f}_2, \phi_2 \rangle \quad \forall\, \phi_2 \in H_0^2(\Omega)\,.$$

The existence of such a $\tilde{\boldsymbol{\psi}}$ can be shown in a manner similar to that for showing the existence of a solution to the von Kármán equation; the key step is that by adding the two equations with the test function $\boldsymbol{\phi}$ replaced by $\tilde{\boldsymbol{\psi}}$, we have the a priori estimate

$$a(\tilde{\psi}_1, \tilde{\psi}_1) + a(\tilde{\psi}_2, \tilde{\psi}_2) = \langle \tilde{f}_1, \tilde{\psi}_1 \rangle + \langle \tilde{f}_2, \tilde{\psi}_2 \rangle .$$

Then, we choose $\tilde{g} = -[\psi_1, \tilde{\psi}_2]$. Note that regularity results for the biharmonic equation applied to (4.4)-(4.5) yield $\boldsymbol{\psi} \in \mathbf{H}^4(\Omega)$ (see [3]). Hence, using imbedding theorems we deduce that $\tilde{g} \in L^2(\Omega)$. It is obvious that $\tilde{g}$ and $\tilde{\boldsymbol{\psi}}$ satisfy (4.16)-(4.17); i.e., we have verified (H11)′. Hence we conclude that for *all* λ, the optimality system (4.4)-(4.5), (4.12)-(4.13), and (4.15) has a solution. Thus, we have Theorem 2.6, which, in the present context, is given as follows.

THEOREM 4.2. *Let* $(\boldsymbol{\psi}, g) \in \mathbf{H}_0^2(\Omega) \times L^2(\Omega)$ *denote an optimal solution that minimizes* (4.8) *subject to* (4.5)-(4.6). *Then, for all* $\lambda \in \Lambda$, *there exists a nonzero Lagrange multiplier* $\boldsymbol{\eta} \in \mathbf{H}_0^2(\Omega)$ *satisfying the Euler equations* (4.12)-(4.13) *and* (4.15). □

4.1.3. Verification of the hypotheses for approximations and error estimates. We finally verify the hypotheses (H13)-(H19) that are used in connection with approximations and error estimates.

A finite-element discretization of the optimality system (4.4)-(4.5), (4.12)-(4.13), and (4.15) is defined in the usual manner. We first choose families of finite-dimensional subspaces $X^h \subset \mathbf{H}_0^2(\Omega)$ and $G^h \subset L^2(\Omega)$ parameterized by a parameter h that tends to zero and satisfying the following approximation properties. There exist a constant C and an integer r such that

$$\inf_{\boldsymbol{\phi}^h \in X^h} \|\boldsymbol{\phi} - \boldsymbol{\phi}^h\|_2 \le Ch^m \|\boldsymbol{\phi}\|_{m+2} \quad \forall\, \boldsymbol{\phi} \in \mathbf{H}^{m+2}(\Omega),\ 1 \le m \le r , \tag{4.18}$$

and

$$\inf_{z^h \in G^h} \|z - z^h\|_0 \le Ch^m \|z\|_m \quad \forall\, z \in H^m(\Omega),\ 1 \le m \le r . \tag{4.19}$$

One may consult, e.g., [8] for some finite-element spaces satisfying (4.18) and (4.19). For example, one may choose $X^h = V^h \times V^h$ where V^h is the piecewise quintic-$C^1(\overline{\Omega})$ finite-element space constrained to satisfy the given boundary conditions and defined with respect to a family of triangulations of Ω. In this case, h is a measure of the grid size. For simplicity, one may choose $G^h = V^h$.

Once the approximating spaces have been chosen, we may formulate the approximate problem for the optimality system (4.4)-(4.5), (4.12)-(4.13), and (4.15). Seek $\boldsymbol{\psi}^h \in X^h$, $g^h \in G^h$, and $\boldsymbol{\eta}^h \in X^h$ such that

$$a(\psi_1^h, \phi_1^h) + \frac{\lambda}{2}\big([\psi_2^h, \psi_2^h], \phi_1^h\big) = 0 \quad \forall\, \phi_1^h \in V^h , \tag{4.20}$$

$$a(\psi_2^h, \phi_2^h) - \lambda\big([\psi_1^h, \psi_2^h], \phi_2^h\big) = (g^h, \phi_2^h) \quad \forall\, \phi_2^h \in V^h , \tag{4.21}$$

$$a(\zeta_1^h, \eta_1^h) - \lambda\big([\psi_2^h, \eta_2^h], \zeta_1^h\big) = \lambda(\psi_1^h - \psi_{10}, \zeta_1^h) \quad \forall\, \zeta_1^h \in V^h , \tag{4.22}$$

$$a(\zeta_2^h, \eta_2^h) + \lambda\big([\psi_2^h, \eta_1^h], \zeta_2^h\big) - \lambda\big([\psi_1^h, \eta_2^h], \zeta_2^h\big) = \lambda(\psi_2^h - \psi_{20}, \zeta_2^h) \quad \forall\, \zeta_2^h \in V^h , \tag{4.23}$$

and

$$(z^h, g^h + \eta_2^h) = 0 \quad \forall\, z^h \in G^h . \tag{4.24}$$

The operator $T^h \in \mathcal{L}(Y; X^h)$ is defined as follows. For $\mathbf{f} \in Y$, $T^h\mathbf{f} = \boldsymbol{\psi}^h \in X^h$ is the unique solution of

$$a(\psi_1^h, \phi_1^h) = \langle f_1, \phi_1^h \rangle \quad \forall\, \phi_1^h \in V^h$$

and

$$a(\psi_2^h, \phi_2^h) = \langle f_2, \phi_2^h \rangle \quad \forall\, \phi_2^h \in V^h\,.$$

Since $T = T^*$, we define $(T^*)^h = T^h$.

We define the operator $E^h : L^2(\Omega) \to G^h$ as the $L^2(\Omega)$-projection on G^h; i.e., for each $g \in L^2(\Omega)$,

$$(E^h g, \phi^h) = (g, \phi^h) \quad \forall\, \phi^h \in G^h\,.$$

Since $G = L^2(\Omega)$ is reflexive, E^h is in fact an operator from $G^* \to G^h$.

By the well-known results concerning the approximation of biharmonic equations (see, e.g., [2] or [8]), we obtain

$$\|(T - T^h)\mathbf{f}\|_X \to 0$$

as $h \to 0$, for all $\mathbf{f} \in Y$. This is simply a restatement of (H13).

(H14) follows trivially from (H13) and the fact that T is self-adjoint, and we have chosen $(T^*)^h = T^h$.

(H15) follows from the best approximation property of $L^2(\Omega)$-projections and (4.19).

(H16) and (H17) follow from the fact that N and $\mathcal{F}$ are polynomials. Here we also use imbedding theorems and Cauchy inequalities.

We set $\hat{Z} = Z = \mathbf{L}^1(\Omega)$. For each $\boldsymbol{\eta} \in \mathbf{H}_0^2(\Omega)$ and $\boldsymbol{\zeta} \in \mathbf{H}_0^2(\Omega)$, Sobolev imbedding theorems imply that

$$[N'(\boldsymbol{\psi})]^* \cdot \boldsymbol{\eta} = \begin{pmatrix} -[\psi_2, \eta_2] \\ [\psi_2, \eta_1] - [\psi_1, \eta_2] \end{pmatrix} \in \hat{Z}\,,$$

$$\big([N''(\boldsymbol{\psi})]^* \cdot \boldsymbol{\zeta}\big) \cdot \boldsymbol{\eta} = \begin{pmatrix} -[\zeta_2, \eta_2] \\ [\zeta_2, \eta_1] - [\zeta_1, \eta_2] \end{pmatrix} \in \hat{Z}\,,$$

and

$$\big(\mathcal{F}''(\boldsymbol{\psi}) \cdot \boldsymbol{\zeta}\big) \cdot \boldsymbol{\eta} = \begin{pmatrix} \eta_1\zeta_1 \\ \eta_2\zeta_2 \end{pmatrix} \in \hat{Z}\,.$$

These relations verify (H18).

From the definition of the operator K we see that K maps $L^2(\Omega)$ into $\mathbf{L}^1(\Omega)$, i.e., K maps G into Z. Thus (H19) is verified.

Hence, we are now in a position to apply Theorem 3.5 to derive error estimates for the approximate solutions of the optimality system (4.4)–(4.5), (4.12)–(4.13) and (4.15). It should be noted that Lemma 3.4 implies that for almost all values of λ, the solutions of the optimality system are regular.

THEOREM 4.3. *Assume that Λ is a compact interval of $\mathbb{R}_+$ and that there exists a branch $\{(\lambda, \boldsymbol{\psi}(\lambda), g(\lambda), \boldsymbol{\eta}(\lambda)) : \lambda \in \Lambda\}$ of regular solutions of the optimality system* (4.4)–(4.5), (4.12)–(4.13), *and* (4.15). *Assume that the finite-element spaces X^h and G^h satisfy the hypotheses* (4.18)–(4.19). *Then, there exist a $\delta > 0$ and an $h_0 > 0$ such that for $h \leq h_0$, the discrete optimality system* (4.20)–(4.24) *has a unique branch of solutions $\{(\lambda, \boldsymbol{\psi}^h(\lambda), g^h(\lambda), \boldsymbol{\eta}^h(\lambda)) : \lambda \in \Lambda\}$ satisfying*

$$\{\|\boldsymbol{\psi}^h(\lambda) - \boldsymbol{\psi}(\lambda)\|_2 + \|g^h(\lambda) - g(\lambda)\|_0 + \|\boldsymbol{\eta}^h(\lambda) - \boldsymbol{\eta}(\lambda)\|_2\} < \delta \quad \forall\, \lambda \in \Lambda.$$

Moreover,

$$\lim_{h\to 0}\{\|\boldsymbol{\psi}^h(\lambda)-\boldsymbol{\psi}(\lambda)\|_2+\|g^h(\lambda)-g(\lambda)\|_0+\|\boldsymbol{\eta}^h(\lambda)-\boldsymbol{\eta}(\lambda)\|_2\}=0\,,$$

uniformly in $\lambda\in\Lambda$.

If, in addition, the solution of the optimality system satisfies $\big(\boldsymbol{\psi}(\lambda),g(\lambda),\boldsymbol{\eta}(\lambda)\big)\in \mathbf{H}^{m+2}(\Omega)\times H^m(\Omega)\times\mathbf{H}^{m+2}(\Omega)$ *for* $\lambda\in\Lambda$, *then there exists a constant* C, *independent of* h, *such that*

$$\begin{aligned}\|\boldsymbol{\psi}(\lambda)-\boldsymbol{\psi}^h(\lambda)\|_2+\|g(\lambda)-g^h(\lambda)\|_0+\|\boldsymbol{\eta}(\lambda)-\boldsymbol{\eta}^h(\lambda)\|_2\\ \le Ch^m\big(\|\boldsymbol{\psi}(\lambda)\|_{m+2}+\|g(\lambda)\|_m+\|\boldsymbol{\eta}(\lambda)\|_{m+2}\big)\,,\end{aligned}$$

uniformly in $\lambda\in\Lambda$.

Proof. All results follow from Theorem 3.5. For the last result, we also use (3.25) and the estimates (see, e.g., [2] or [8])

$$\|(T^hT^{-1}-I)\boldsymbol{\psi}\|_2\le Ch^m\|\boldsymbol{\psi}\|_{m+2}\qquad\text{for }\boldsymbol{\psi}\in\mathbf{H}^{m+2}(\Omega)\,,$$

$$\|\big((T^*)^h(T^*)^{-1}-I\big)\boldsymbol{\eta}\|_2=\|(T^hT^{-1}-I)\boldsymbol{\eta}\|_2\le Ch^m\|\boldsymbol{\eta}\|_{m+2}\qquad\text{for }\boldsymbol{\eta}\in\mathbf{H}^{m+2}(\Omega)\,,$$

and

$$\|(E^hE^{-1}-I)g\|_0\le Ch^m\|g\|_m\qquad\text{for } g\in H^m(\Omega)\,.$$

In these estimates, the constant C is independent of h, $\boldsymbol{\psi}$, g, $\boldsymbol{\eta}$, and λ. □

Remark. In fact, we obtain from (4.15) that $g=-\eta_2$, so the term $\|g(\lambda)\|_m$ in the right-hand side of the error estimate is redundant. □

Remark. By using (4.15) again, along with (4.24) and the error estimate in Theorem 4.3, we have the following improved error estimate for the approximation of the control g:

$$\|g(\lambda)-g^h(\lambda)\|_2=\|\eta_2(\lambda)-\eta_2^h(\lambda)\|_2\le Ch^m\big(\|\boldsymbol{\psi}(\lambda)\|_{m+2}+\|\boldsymbol{\eta}(\lambda)\|_{m+2}\big)\,.$$

Of course, we also use the fact that we have chosen $G^h=V^h\subset H^2(\Omega)$. □

4.2. Neumann boundary controls for the Ginzburg–Landau superconductivity equations. For this application we will use Neumann boundary controls; i.e., control is effected through the data in a Neumann boundary condition. Let Ω be a bounded open domain in $\mathbb{R}^d$, $d=2$ or 3, and let Γ be its boundary. A simplified Ginzburg–Landau model for superconductivity is given by

$$\begin{aligned}-\Delta\psi_1+\big(\psi_1^2+\psi_2^2+|\mathbf{A}|^2-1\big)\psi_1-\nabla\cdot(\mathbf{A}\psi_2)-\mathbf{A}\cdot\nabla\psi_2=0\quad&\text{in }\Omega\,,\\ -\Delta\psi_2+\big(\psi_1^2+\psi_2^2+|\mathbf{A}|^2-1\big)\psi_2+\nabla\cdot(\mathbf{A}\psi_1)+\mathbf{A}\cdot\nabla\psi_1=0\quad&\text{in }\Omega\,,\\ \mathbf{n}\cdot(\nabla\psi_1+\mathbf{A}\psi_2)=\lambda g_1\quad&\text{on }\Gamma\,,\end{aligned}$$

and

$$\mathbf{n}\cdot(\nabla\psi_2-\mathbf{A}\psi_1)=\lambda g_2\quad\text{on }\Gamma\,.$$

Here, ψ_1 and ψ_2 denote the real and imaginary parts, respectively, of the complex-valued order parameter, $\mathbf{A}$ is a given real magnetic potential, g_1 and g_2 are related to the normal component of the current at the boundary, and $\lambda>0$ is a "current loading" parameter. These equations are a special case of a more general model for superconductivity wherein $\mathbf{A}$ is also unknown;

see, e.g., [22] for a derivation of the general model. It can be shown that in certain limits, e.g., high values of the applied field, the above simpler model is valid; see [7].

By introducing appropriate rescalings, i.e., by replacing ψ_j by $\sqrt{\lambda}\psi_j$ and g_j by $\sqrt{\lambda}g_j$, $j = 1, 2$, we can rewrite the above Ginzburg-Landau equations as follows:

$$-\Delta\psi_1 + (|\mathbf{A}|^2 - 1)\psi_1 - \nabla\cdot(\mathbf{A}\psi_2) - \mathbf{A}\cdot\nabla\psi_2 + \lambda\left(\psi_1^2 + \psi_2^2\right)\psi_1 = 0 \quad \text{in } \Omega\,, \tag{4.25}$$

$$-\Delta\psi_2 + (|\mathbf{A}|^2 - 1)\psi_2 + \nabla\cdot(\mathbf{A}\psi_1) + \mathbf{A}\cdot\nabla\psi_1 + \lambda\left(\psi_1^2 + \psi_2^2\right)\psi_2 = 0 \quad \text{in } \Omega\,, \tag{4.26}$$

$$\mathbf{n}\cdot(\nabla\psi_1 + \mathbf{A}\psi_2) = \lambda g_1 \quad \text{on } \Gamma\,, \tag{4.27}$$

and

$$\mathbf{n}\cdot(\nabla\psi_2 - \mathbf{A}\psi_1) = \lambda g_2 \quad \text{on } \Gamma\,. \tag{4.28}$$

We introduce the bilinear forms

$$a(\psi, \phi) = \int_\Omega \left(\nabla\psi\cdot\nabla\phi + (|\mathbf{A}|^2 - 1)\psi\phi\right) d\Omega \quad \forall\, \psi, \phi \in H^1(\Omega)$$

and

$$b(\psi, \phi) = \int_\Omega \mathbf{A}\cdot(\psi\nabla\phi - \phi\nabla\psi)\, d\Omega \quad \forall\, \psi, \phi \in H^1(\Omega)\,.$$

We assume that $\mathbf{A} \in \mathbf{H}^1(\Omega)$. Note that

$$a(\psi, \phi) = a(\phi, \psi) \quad \text{and} \quad b(\psi, \phi) = -b(\phi, \psi)\,.$$

Then, a weak formulation of the Ginzburg-Landau equations (4.25)-(4.28) is defined as follows. Seek $\boldsymbol{\psi} = (\psi_1, \psi_2) \in \mathbf{H}^1(\Omega)$ such that

$$a(\psi_1, \phi_1) + b(\psi_2, \phi_1) + \lambda\big((\psi_1^2 + \psi_2^2)\psi_1, \phi_1\big) = \lambda\langle g_1, \phi_1\rangle_\Gamma \quad \forall\, \phi_1 \in H^1(\Omega) \tag{4.29}$$

and

$$a(\psi_2, \phi_2) - b(\psi_1, \phi_2) + \lambda\big((\psi_1^2 + \psi_2^2)\psi_2, \phi_2\big) = \lambda\langle g_2, \phi_2\rangle_\Gamma \quad \forall\, \phi_2 \in H^1(\Omega)\,. \tag{4.30}$$

It can be shown that, for each $\mathbf{g} = (g_1, g_2) \in \mathbf{H}^{-1/2}(\Gamma)$, (4.29) and (4.30) possess at least one solution $\boldsymbol{\psi} \in \mathbf{H}^1(\Omega)$ and that all solutions of (4.29) and (4.30) satisfy the a priori estimate

$$\|\psi_1\|_1 + \|\psi_2\|_1 \le C\left(\|g_1\|_{-1/2,\Gamma} + \|g_2\|_{-1/2,\Gamma}\right)\,; \tag{4.31}$$

see, e.g., [11], for details. In the sequel, a solution of (4.25)-(4.28) will be understood in the sense of (4.29)-(4.30).

Given a desired state $\boldsymbol{\psi}_0 = (\psi_{10}, \psi_{20}) \in \mathbf{L}^2(\Omega)$, we define for any $\boldsymbol{\psi} = (\psi_1, \psi_2) \in \mathbf{H}^1(\Omega)$ and $\mathbf{g} = (g_1, g_2) \in \mathbf{L}^2(\Gamma)$ the functional

$$\mathcal{J}(\boldsymbol{\psi}, \mathbf{g}) = \frac{\lambda}{2}\int_\Omega \left((\psi_1 - \psi_{10})^2 + (\psi_2 - \psi_{20})^2\right) d\Omega + \frac{\lambda}{2}\int_\Gamma (g_1^2 + g_2^2)\, d\Gamma\,. \tag{4.32}$$

We then consider the following optimal control problem associated with the Ginzburg-Landau equations for superconductivity:

$$\min\left\{\, \mathcal{J}(\boldsymbol{\psi}, \mathbf{g}) \mid \boldsymbol{\psi} \in \mathbf{H}^1(\Omega)\,,\ \mathbf{g} \in \Theta \,\right\} \quad \text{subject to} \quad (4.29) \text{ and } (4.30)\,, \tag{4.33}$$

where Θ is a subset of $\mathbf{L}^2(\Gamma)$.

We define the spaces $X = \mathbf{H}^1(\Omega)$, $Y = (\mathbf{H}^1(\Omega))^*$, $G = \mathbf{L}^2(\Gamma)$, and $Z = [\mathbf{H}^{1/2+\epsilon}(\Omega)]^*$ where $\epsilon \in (0, 1/2)$ is chosen such that $\mathbf{H}^1(\Omega) \hookrightarrow\hookrightarrow \mathbf{H}^{1/2+\epsilon}(\Omega) \hookrightarrow \mathbf{L}^4(\Omega)$. By compact imbedding results, $\mathbf{L}^{4/3}(\Omega) \hookrightarrow Z \hookrightarrow\hookrightarrow Y$. For the time being, we assume that the admissible set Θ for the control $\mathbf{g}$ is a closed convex subset of $G = \mathbf{L}^2(\Gamma)$.

Let the continuous linear operator $T \in \mathcal{L}(Y; X)$ be defined as follows. For each $\mathbf{f} = (f_1, f_2) \in Y = (\mathbf{H}^1(\Omega))^*$, $T\mathbf{f} = \boldsymbol{\psi} \in X = \mathbf{H}^1(\Omega)$ is the unique solution of

$$a(\psi_1, \phi_1) + b(\psi_2, \phi_1) = \langle f_1, \phi_1 \rangle \quad \forall\, \phi_1 \in H^1(\Omega)$$

and

$$a(\psi_2, \phi_2) - b(\psi_1, \phi_2) = \langle f_2, \phi_2 \rangle \quad \forall\, \phi_2 \in H^1(\Omega)\,.$$

It can be easily verified that T is self-adjoint. Also, it can be shown that for most choices of $\mathbf{A}$, the operator T is well defined; see [11].

We define the (differentiable) nonlinear mapping $N : X \to Y$ by

$$N(\boldsymbol{\psi}) = \begin{pmatrix} (\psi_1^2 + \psi_2^2)\psi_1 \\ (\psi_1^2 + \psi_2^2)\psi_2 \end{pmatrix} \quad \forall\, \boldsymbol{\psi} \in X$$

or, equivalently,

$$\langle N(\boldsymbol{\psi}), \boldsymbol{\phi} \rangle = \big((\psi_1^2 + \psi_2^2)\psi_1, \phi_1\big) + \big((\psi_1^2 + \psi_2^2)\psi_2, \phi_2\big) \quad \forall\, \boldsymbol{\phi} = (\phi_1, \phi_2) \in X$$

and define $K : \mathbf{H}^{-1/2}(\Gamma) \to Y$ as the injection mapping

$$\langle K\mathbf{z}, \mathbf{v} \rangle = -\langle \mathbf{z}, \mathbf{v} \rangle_\Gamma \quad \forall\, \mathbf{z} \in \mathbf{H}^{-1/2}(\Gamma)\,, \forall\, \mathbf{v} \in \mathbf{H}^1(\Omega)\,.$$

Clearly, the constraint equations (4.29)-(4.30) can be expressed as

$$\boldsymbol{\psi} + \lambda\, TN(\boldsymbol{\psi}) + \lambda\, TK\mathbf{g} = 0\,,$$

i.e., in the form (2.2). With the obvious definitions for $\mathcal{F}(\cdot)$ and $\mathcal{E}(\cdot)$, i.e.,

$$\mathcal{F}(\boldsymbol{\psi}) = \frac{1}{2}\int_\Omega \Big((\psi_1 - \psi_{10})^2 + (\psi_2 - \psi_{20})^2\Big)\, d\Omega \quad \forall\, \boldsymbol{\psi} \in X$$

and

$$\mathcal{E}(\mathbf{g}) = \frac{1}{2}\int_\Gamma (g_1^2 + g_2^2)\, d\Gamma \quad \forall\, \mathbf{g} \in G\,,$$

the functional (4.32) can be expressed as

$$\mathcal{J}(\boldsymbol{\psi}, g) = \lambda\, \mathcal{F}(\boldsymbol{\psi}) + \lambda\, \mathcal{E}(\mathbf{g})\,,$$

i.e., in the form (2.1). Thus, the minimization problem (4.33) is in the form of the minimization problem (2.3).

We are now in a position to verify, for the minimization problem (4.33), all the hypotheses of §§2 and 3.

4.2.1. Verification of the hypotheses for the existence of optimal solutions. We first verify that the hypotheses (H1)-(H6) hold in the current setting.

(H1) is obviously satisfied with a lower bound 0.

(H2) holds with $\alpha = 1$ and $\beta = 2$.

(H3) is verified since $\boldsymbol{\psi} = \mathbf{0}$ and $\mathbf{g} = \mathbf{0}$ is obviously a solution of (4.29)-(4.30).

In order to verify (H4), we assume $\{\mathbf{g}^{(n)}\} \subset \Theta \subset \mathbf{L}^2(\Gamma)$ is a sequence satisfying $\mathbf{g}^{(n)} \rightharpoonup \mathbf{g}$ in $\mathbf{L}^2(\Gamma)$; then, we have $\mathbf{g}^{(n)} \rightharpoonup \mathbf{g}$ in $\mathbf{H}^{-1/2}(\Gamma)$, so $\lim_{n\to\infty}\langle \mathbf{g}^{(n)}, \mathbf{v}\rangle_\Gamma = \langle \mathbf{g}, \mathbf{v}\rangle_\Gamma$ for all $\mathbf{v} \in \mathbf{H}^1(\Omega)$, i.e., $K\mathbf{g}^{(n)} \rightharpoonup K\mathbf{g}$ in Y. Assume that the sequence $\{\boldsymbol{\psi}^{(n)}\} \subset \mathbf{H}^1(\Omega)$ satisfies $\boldsymbol{\psi}^{(n)} \rightharpoonup \boldsymbol{\psi}$ in $\mathbf{H}^1(\Omega)$; then, by using the compact imbedding $\mathbf{H}^1(\Omega) \hookrightarrow\hookrightarrow \mathbf{L}^4(\Omega)$, $\boldsymbol{\psi}^{(n)} \to \boldsymbol{\psi}$ in $\mathbf{L}^4(\Omega)$. Now,

$$\begin{aligned}\langle N(\boldsymbol{\psi}^{(n)}), \boldsymbol{\phi}\rangle &= \Big(\big((\psi_1^{(n)})^2 + (\psi_2^{(n)})^2\big)\psi_1^{(n)}, \phi_1\Big) + \Big(\big((\psi_1^{(n)})^2 + (\psi_2^{(n)})^2\big)\psi_2^{(n)}, \phi_2\Big)\\ &\to \Big((\psi_1^2 + \psi_2^2)\psi_1, \phi_1\Big) + \Big((\psi_1^2 + \psi_2^2)\psi_2, \phi_2\Big) = \langle N(\boldsymbol{\psi}), \boldsymbol{\phi}\rangle\,.\end{aligned}$$

Hence, (H4) is verified.

The verification of (H5) follows directly from the observation that the mappings $\boldsymbol{\phi} \mapsto \mathcal{F}(\boldsymbol{\phi}) = (1/2)\|\boldsymbol{\phi} - \boldsymbol{\psi}_0\|_0^2$ and $\mathbf{g} \mapsto \mathcal{E}(\mathbf{g}) = (1/2)\|\mathbf{g}\|_{0,\Gamma}^2$ are convex.

The verification of (H6) is a trivial consequence of the a priori estimate (4.31).

It is now just a matter of citing Theorem 2.1 to prove the existence of an optimal solution that minimizes (4.32) subject to (4.29)-(4.30).

THEOREM 4.4. *There exists a* $(\boldsymbol{\phi}, \mathbf{g}) \in \mathbf{H}^1(\Omega) \times \Theta$ *such that* (4.32) *is minimized subject to* (4.29)-(4.30). □

4.2.2. Verification of the hypotheses for the existence of Lagrange multipliers. We now assume $(\boldsymbol{\psi}, \mathbf{g})$ is an optimal solution and turn to the verification of hypotheses (H7)-(H9).

The validity of (H7) is obvious.

(H8) holds since the mapping $\mathbf{g} \mapsto \mathcal{E}(\mathbf{g}) = (\frac{1}{2})\int_\Gamma |\mathbf{g}|^2\, d\Gamma$ is convex.

(H9) can be verified as follows. For any $\boldsymbol{\psi} \in X$, the operator $N'(\boldsymbol{\psi}) : X \to Y$ is given by

$$N'(\boldsymbol{\psi})\cdot\boldsymbol{\phi} = \begin{pmatrix}(3\psi_1^2 + \psi_2^2)\phi_1 + (2\psi_1\psi_2)\phi_2\\ (3\psi_2^2 + \psi_1^2)\phi_2 + (2\psi_1\psi_2)\phi_1\end{pmatrix} \quad \forall\, \boldsymbol{\phi} = (\phi_1, \phi_2) \in X\,.$$

Thus, we obtain that $N'(\boldsymbol{\psi})\cdot\boldsymbol{\phi} \in \mathbf{L}^{4/3}(\Omega) \hookrightarrow [\mathbf{H}^{1/2+\epsilon}(\Omega)]^* = Z$.

The Lagrangian is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\psi}, \mathbf{g}, \boldsymbol{\eta}, k) &= k\,\mathcal{J}(\boldsymbol{\psi}, \mathbf{g})\\ &\quad - \Big\{a(\psi_1, \eta_1) + b(\psi_2, \phi_1) + \lambda\big((\psi_1^2 + \psi_2^2)\psi_1, \eta_1\big) - \lambda(g_1, \eta_1)_\Gamma\\ &\qquad + a(\psi_2, \eta_2) - b(\psi_1, \phi_2) - \lambda\big((\psi_1^2 + \psi_2^2)\psi_2, \eta_2\big) - \lambda(g_2, \eta_2)_\Gamma\Big\}\end{aligned}$$

for all $(\boldsymbol{\psi}, \mathbf{g}, \boldsymbol{\eta}, k) \in X \times G \times X \times \mathbb{R} = \mathbf{H}^1(\Omega) \times \mathbf{L}^2(\Gamma) \times \mathbf{H}^1(\Omega) \times \mathbb{R}$. Note that in this form of the Lagrangian, the Lagrange multiplier $\boldsymbol{\eta} \in X = Y^*$, so we have already introduced the change of variables indicated between (2.17)-(2.18) and (2.19)-(2.21).

Having verified the hypotheses (H7)-(H9), we may apply Theorem 2.4 to conclude that there exist a Lagrange multiplier $\boldsymbol{\eta} \in X = \mathbf{H}^1(\Omega)$ and a real number k such that

$$\boldsymbol{\eta} + \lambda\, T^*\big([N'(\boldsymbol{\psi})]^*\cdot\boldsymbol{\eta} - k\,\mathcal{F}'(\boldsymbol{\psi})\big) = \mathbf{0} \tag{4.34}$$

and

$$\mathcal{L}(\boldsymbol{\psi}, \mathbf{g}, \boldsymbol{\eta}, k) \leq \mathcal{L}(\boldsymbol{\psi}, \mathbf{z}, \boldsymbol{\eta}, k) \quad \forall\, z \in \Theta \tag{4.35}$$

and that for almost all values of λ, we may choose $k = 1$.

Recall that T is self-adjoint. Also, note that for any $\boldsymbol{\psi} \in X = \mathbf{H}^1(\Omega)$,

$$[N'(\boldsymbol{\psi})]^* \cdot \boldsymbol{\eta} = \begin{pmatrix} (3\psi_1^2 + \psi_2^2)\eta_1 + (2\psi_1\psi_2)\eta_2 \\ (3\psi_2^2 + \psi_1^2)\eta_2 + (2\psi_1\psi_2)\eta_1 \end{pmatrix} \quad \forall\, \boldsymbol{\eta} = (\eta_1, \eta_2) \in X\,.$$

Thus, $N'(\boldsymbol{\psi})$ is self-adjoint as well and (4.34), with $k = 1$, can be rewritten as

$$\begin{aligned} a(\zeta_1, \eta_1) - b(\zeta_1, \eta_2) &+ \lambda\big((3\psi_1^2 + \psi_2^2)\eta_1, \zeta_1\big) \\ &+ \lambda\big((2\psi_1\psi_2)\eta_2, \zeta_1\big) = \lambda(\psi_1 - \psi_{10}, \zeta_1) \quad \forall\, \zeta_1 \in H^1(\Omega) \end{aligned} \tag{4.36}$$

and

$$\begin{aligned} a(\zeta_2, \eta_2) + b(\zeta_2, \eta_1) &+ \lambda\big((3\psi_2^2 + \psi_1^2)\eta_2, \zeta_2\big) \\ &+ \lambda\big((2\psi_1\psi_2)\eta_1, \zeta_2\big) = \lambda(\psi_2 - \psi_{20}, \zeta_2) \quad \forall\, \zeta_2 \in H^1(\Omega)\,. \end{aligned} \tag{4.37}$$

Using the definition of the Lagrangian functional, (4.35), with $k = 1$, can be rewritten as

$$\frac{\lambda}{2}(\mathbf{z}, \mathbf{z})_\Gamma + \lambda(\mathbf{z}, \boldsymbol{\eta})_\Gamma - \frac{\lambda}{2}(\mathbf{g}, \mathbf{g})_\Gamma - \lambda(\mathbf{g}, \boldsymbol{\eta})_\Gamma \geq 0 \quad \forall\, \mathbf{z} \in \Theta\,.$$

Note that, in the above expression, we have already employed hypothesis (H12), which in the current context is trivially satisfied with E, the identity operator on $G^* = G = \mathbf{L}^2(\Gamma)$. For each $\epsilon \in (0, 1)$ and each $\mathbf{t} \in \Theta$, set $\mathbf{z} = \epsilon\mathbf{t} + (1 - \epsilon)\mathbf{g} \in \Theta$ in the last equation to obtain

$$\frac{\epsilon^2}{2}(\mathbf{t} - \mathbf{g}, \mathbf{t} - \mathbf{g})_\Gamma + \epsilon\,(\mathbf{t} - \mathbf{g}, \mathbf{g})_\Gamma + \epsilon\,(\mathbf{t} - \mathbf{g}, \boldsymbol{\eta})_\Gamma \geq 0 \quad \forall\, \mathbf{t} \in \Theta\,,$$

so, after dividing by $\epsilon > 0$ and then letting $\epsilon \to 0^+$, we obtain

$$(\mathbf{t} - \mathbf{g}, \mathbf{g} + \boldsymbol{\eta})_\Gamma \geq 0 \quad \forall\, \mathbf{t} \in \Theta\,. \tag{4.38}$$

We see that for almost all values of λ, necessary conditions for an optimum are that (4.29)-(4.30) and (4.36)-(4.38) are satisfied. Again, the system formed by these equations will be called an *optimality system*.

We now specialize to the case $\Theta = \mathbf{L}^2(\Gamma)$. Note that the hypothesis (H10) is satisfied. Then, using Theorem 2.5, we see that the inequality (4.38) becomes an equality and, by letting $\mathbf{z} = \mathbf{t} - \mathbf{g}$ vary arbitrarily in $\mathbf{L}^2(\Gamma)$, we now have, instead of (4.38),

$$(\mathbf{z}, \mathbf{g} + \boldsymbol{\eta})_\Gamma = 0 \quad \forall\, \mathbf{z} \in \mathbf{L}^2(\Gamma)\,. \tag{4.39}$$

Thus, according to that theorem, we have that for almost all λ, an optimality system of equations is now given by (4.29)-(4.30), (4.36)-(4.37), and (4.39). However, we can go further and verify that the hypothesis (H11) is valid, which in turn will justify the existence of a Lagrange multiplier satisfying the optimality system for *all* $\lambda \in \Lambda$.

To verify (H11), we first note that, via the change of variable $\xi = T^*v$, that assumption can be equivalently stated as

$$\text{if } \xi \in Y^* \text{ satisfies } \big(I + \lambda\, T^*[N'(u)]^*\big)\xi = 0 \text{ and } K^*\xi = 0, \text{ then } \xi = 0\,.$$

To verify this version of (H11), we assume that $\boldsymbol{\xi} \in Y^* = \mathbf{H}^1(\Omega)$ satisfies $\left(I + \lambda T^*[N'(\boldsymbol{\psi})]^*\right)\boldsymbol{\xi} = \mathbf{0}$ and $K^*\boldsymbol{\xi} = 0$, i.e.,

$$a(\zeta_1, \xi_1) - b(\zeta_1, \xi_2) + \lambda\left((3\psi_1^2 + \psi_2^2)\xi_1, \zeta_1\right) + \lambda\left((2\psi_1\psi_2)\xi_2, \zeta_1\right) = 0 \quad \forall\, \zeta_1 \in H^1(\Omega),$$

$$a(\zeta_2, \xi_2) + b(\zeta_2, \xi_1) + \lambda\left((3\psi_2^2 + \psi_1^2)\xi_2, \zeta_2\right) + \lambda\left((2\psi_1\psi_2)\xi_1, \zeta_2\right) = 0 \quad \forall\, \zeta_2 \in H^1(\Omega),$$

and

$$\boldsymbol{\xi} = \mathbf{0} \quad \text{on } \Gamma.$$

(Note that $K^*\boldsymbol{\xi} = \boldsymbol{\xi}|_\Gamma$.) Let Ω' be a smooth extension of Ω such that $\overline{\Omega}$ is a compact subset of Ω'. We then define $\boldsymbol{\xi}'$, $\boldsymbol{\psi}'$, and $\mathbf{A}'$ to be the extensions, by zero outside Ω, of $\boldsymbol{\xi}$, $\boldsymbol{\psi}$ and $\mathbf{A}$, respectively. Let the forms $a'(\cdot,\cdot)$, $b'(\cdot,\cdot)$, and $(\cdot,\cdot)'$ defined over Ω' be the analogues of corresponding forms defined over Ω. We may show from the last three equations that

$$\boldsymbol{\xi}' \in \mathbf{H}^1(\Omega'), \qquad \boldsymbol{\psi}' \in \mathbf{L}^6(\Omega'),$$

$$a'(\zeta_1, \xi_1') - b'(\zeta_1, \xi_2') + \lambda\left((3\psi_1'^2 + \psi_2'^2)\xi_1', \zeta_1\right)' + \lambda\left((2\psi_1'\psi_2')\xi_2', \zeta_1\right)' = 0 \quad \forall\, \zeta_1 \in \mathbf{H}_0^1(\Omega'),$$

and

$$a'(\zeta_2, \xi_2') + b'(\zeta_2, \xi_1') + \lambda\left((3\psi_2'^2 + \psi_1'^2)\xi_2', \zeta_2\right)' + \lambda\left((2\psi_1'\psi_2')\xi_1', \zeta_2\right)' = 0 \quad \forall\, \zeta_2 \in \mathbf{H}_0^1(\Omega').$$

In the sense of distribution, $\boldsymbol{\xi}'$ satisfies

$$-\Delta\xi_1' - 2\mathbf{A}' \cdot \nabla\xi_2' + \left(|\mathbf{A}'|^2 + \lambda(3\psi_1'^2 + \psi_2'^2) - 1\right)\xi_1' - (\nabla \cdot \mathbf{A}' - 2\lambda\psi_1'\psi_2')\xi_2' = 0 \quad \text{in } \Omega' \tag{4.40}$$

and

$$-\Delta\xi_2' + 2\mathbf{A}' \cdot \nabla\xi_1' + (\nabla \cdot \mathbf{A}' + 2\lambda\psi_1'\psi_2')\xi_1' + \left(|\mathbf{A}'|^2 + \lambda(3\psi_2'^2 + \psi_1'^2) - 1\right)\xi_2' = 0 \quad \text{in } \Omega'. \tag{4.41}$$

We now quote the following unique continuation result whose proof can be found in [17]. See also [12] and [19].

LEMMA 4.5. *Let Ω' be an open and connected subset of $\mathbb{R}^d$, $d = 2$ or 3. Let the functions $\mathbf{V} \in [L^q_{\text{loc}}(\Omega')]^{d\times d}$ for some $q \geq 2$ and $\mathbf{W} \in [L^{2d-1}_{\text{loc}}(\Omega')]^{d\times d\times d}$ be given. If $\boldsymbol{\xi} \in \mathbf{H}^1_{\text{loc}}(\Omega')$, $-\Delta\xi_i + \sum_{j=1}^d \sum_{k=1}^d W_{ijk}(\partial\xi_k/\partial x_j) + \sum_{j=1}^d V_{ij}\xi_j = 0$ (in the sense of distributions), $i = 1, \ldots, d$, and $\boldsymbol{\xi} = \mathbf{0}$ on an open, nonempty subset of Ω', then $\boldsymbol{\xi} = \mathbf{0}$ on Ω'.* □

Since $\mathbf{A} \in \mathbf{H}^1(\Omega)$ and $\boldsymbol{\psi} \in \mathbf{H}^1(\Omega)$, it is easy to see that the coefficients in (4.40)-(4.41) satisfy the regularity requirements of Lemma 4.5. Also note that $\boldsymbol{\xi}' = \mathbf{0}$ on $(\Omega' \setminus \Omega)$, which contains an open set. Thus we obtain that $\boldsymbol{\xi}' = \mathbf{0}$ in Ω', or $\boldsymbol{\xi} = \mathbf{0}$ in Ω; i.e., (H11) is verified.

Hence we conclude that for *all* λ, the optimality system (4.29)-(4.30), (4.36)-(4.37), and (4.39) has a solution. Thus, we have Theorem 2.6, which, in the present context, is given as follows.

THEOREM 4.6. *Let $(\boldsymbol{\psi}, \mathbf{g}) \in \mathbf{H}^1(\Omega) \times \mathbf{L}^2(\Gamma)$ denote an optimal solution that minimizes* (4.32) *subject to* (4.29)-(4.30). *Then, for all $\lambda \in \Lambda$, there exists a nonzero Lagrange multiplier $\boldsymbol{\eta} \in \mathbf{H}^1(\Omega)$ satisfying the Euler equations* (4.36)-(4.37) *and* (4.39). □

4.2.3. Verification of the hypotheses for approximations and error estimates. We finally verify the hypotheses (H13)-(H19) that are used in connection with approximations and to derive error estimates.

A finite-element discretization of the optimality system (4.29)-(4.30), (4.36)-(4.37), and (4.39) is defined in the usual manner. We first choose families of finite-dimensional subspaces $X^h \subset \mathbf{H}^1(\Omega)$ and $G^h \subset \mathbf{L}^2(\Gamma)$ parameterized by a parameter h that tends to zero and satisfying the following approximation properties. There exists a constant C and an integer r such that

$$\inf_{\boldsymbol{\phi}^h \in X^h} \|\boldsymbol{\phi} - \boldsymbol{\phi}^h\|_1 \le Ch^m \|\boldsymbol{\phi}\|_{m+1} \quad \forall\, \boldsymbol{\phi} \in \mathbf{H}^{m+1}(\Omega)\,,\ 1 \le m \le r\,, \tag{4.42}$$

and

$$\inf_{z^h \in G^h} \|\mathbf{z} - \mathbf{z}^h\|_{0,\Gamma} \le Ch^m \inf_{\mathbf{v} \in \mathbf{H}^{m+1/2}(\Omega),\, \mathbf{v}|_\Gamma = \mathbf{z}} \|\mathbf{v}\|_{m+1/2} \quad \forall\, \mathbf{z} \in \mathbf{H}^{m+1/2}(\Omega)|_\Gamma\,,\ 1 \le m \le r\,. \tag{4.43}$$

One may consult, e.g., [8] and [15], for some finite-element spaces satisfying (4.42) and (4.43). For example, one may choose $X^h = V^h \times V^h$ where V^h is the piecewise linear or quadratic finite-element space defined with respect to a family of triangulations of Ω. In this case, h is a measure of the grid size. For simplicity we may choose $G^h = (X^h)|_\Gamma$, i.e., the functions in G^h are the restrictions to the boundary Γ of functions belonging to X^h.

Once the approximating spaces have been chosen, we may formulate the approximate problem for the optimality system (4.29)-(4.30), (4.36)-(4.37), and (4.39). Seek $\boldsymbol{\psi}^h \in X^h$, $\mathbf{g}^h \in G^h$, and $\boldsymbol{\eta}^h \in X^h$ such that

$$a(\psi_1^h, \phi_1^h) + b(\psi_2^h, \phi_1^h) + \lambda\big\{\big((\psi_1^h)^2 + (\psi_2^h)^2\big)\psi_1^h, \phi_1^h\big\} = \lambda\langle g_1^h, \phi_1^h\rangle_\Gamma \quad \forall\, \phi_1^h \in V^h\,, \tag{4.44}$$

$$a(\psi_2^h, \phi_2^h) - b(\psi_1^h, \phi_2^h) + \lambda\big\{\big((\psi_1^h)^2 + (\psi_2^h)^2\big)\psi_2^h, \phi_2^h\big\} = \lambda\langle g_2^h, \phi_2^h\rangle_\Gamma \quad \forall\, \phi_2^h \in V^h\,, \tag{4.45}$$

$$\begin{aligned} a(\zeta_1^h, \eta_1^h) - b(\zeta_1^h, \eta_2^h) &+ \lambda\big((3(\psi_1^h)^2 + (\psi_2^h)^2)\eta_1^h, \zeta_1^h\big) \\ &+ \lambda\big((2\psi_1^h\psi_2^h)\eta_2^h, \zeta_1^h\big) = \lambda(\psi_1^h - \psi_{10}, \zeta_1^h) \quad \forall\, \zeta_1^h \in V^h\,, \end{aligned} \tag{4.46}$$

$$\begin{aligned} a(\zeta_2^h, \eta_2^h) + b(\zeta_2^h, \eta_1^h) &+ \lambda\big((3(\psi_2^h)^2 + (\psi_1^h)^2)\eta_2^h, \zeta_2^h\big) \\ &+ \lambda\big((2\psi_1^h\psi_2^h)\eta_1^h, \zeta_2^h\big) = \lambda(\psi_2^h - \psi_{20}, \zeta_2^h) \quad \forall\, \zeta_2^h \in V^h\,, \end{aligned} \tag{4.47}$$

and

$$(\mathbf{z}^h, \mathbf{g}^h + \boldsymbol{\eta}^h)_\Gamma = 0 \quad \forall\, \mathbf{z}^h \in G^h\,. \tag{4.48}$$

The operator $T^h \in \mathcal{L}(Y; X^h)$ is defined as follows. For $\mathbf{f} \in Y$, $T^h\mathbf{f} = \boldsymbol{\psi}^h \in X^h$ is the solution for

$$a(\psi_1^h, \phi_1^h) + b(\psi_2^h, \phi_1^h) = \langle f_1, \phi_1^h\rangle \quad \forall\, \phi_1^h \in V^h$$

and

$$a(\psi_2^h, \phi_2^h) - b(\psi_1^h, \phi_2^h) = \langle f_2, \phi_2^h\rangle \quad \forall\, \phi_2^h \in V^h\,.$$

Since $T = T^*$, we define $(T^*)^h = T^h$.

We define the operator $E^h : \mathbf{L}^2(\Gamma) \to G^h$ as the $\mathbf{L}^2(\Gamma)$-projection on G^h; i.e., for each $\mathbf{g} \in \mathbf{L}^2(\Gamma)$,

$$(E^h\mathbf{g}, \mathbf{z}^h)_\Gamma = (\mathbf{g}, \mathbf{z}^h)_\Gamma \quad \forall\, \mathbf{z}^h \in G^h\,.$$

Since $G = \mathbf{L}^2(\Gamma)$ is reflexive, E^h is in fact an operator from $G^* \to G^h$.

By results concerning the approximation of the Ginzburg-Landau equations (see, e.g., [11]), we obtain

$$\|(T - T^h)\mathbf{f}\|_X \to 0$$

as $h \to 0$, for all $\mathbf{f} \in Y$. This is simply a restatement of (H13).

(H14) follows trivially from (H13) and the fact that T is self-adjoint and we have chosen $(T^*)^h = T^h$.

(H15) follows from the best approximation property of $\mathbf{L}^2(\Gamma)$-projections and (4.43).

(H16) and (H17) follow from the fact that N and $\mathcal{F}$ are polynomials. Here we also use imbedding theorems and Cauchy inequalities.

Setting $\hat{Z} = Z = \mathbf{H}^{1/2+\epsilon}(\Omega)$, we have that $\hat{Z} \hookrightarrow\hookrightarrow [\mathbf{H}^1(\Omega)]^* = X^*$. For each $\boldsymbol{\eta} \in \mathbf{H}^1(\Omega)$ and $\boldsymbol{\zeta} \in \mathbf{H}^1(\Omega)$, Sobolev imbedding theorems imply that

$$[N'(\boldsymbol{\psi})]^* \cdot \boldsymbol{\eta} = \begin{pmatrix} (3\psi_1^2 + \psi_2^2)\eta_1 + (2\psi_1\psi_2)\eta_2 \\ (3\psi_2^2 + \psi_1^2)\eta_2 + (2\psi_1\psi_2)\eta_1 \end{pmatrix} \in \mathbf{L}^{4/3}(\Omega) \subset \hat{Z}\,,$$

$$\big([N''(\boldsymbol{\psi})]^* \cdot \boldsymbol{\zeta}\big) \cdot \boldsymbol{\eta} = \begin{pmatrix} (6\psi_1\zeta_1 + 2\psi_2\zeta_2)\eta_1 + (2\psi_1\zeta_2)\eta_2 + (2\zeta_1\psi_2)\eta_2 \\ (6\psi_2\zeta_2 + 2\psi_1\zeta_1)\eta_2 + (2\psi_1\zeta_2)\eta_1 + (2\zeta_1\psi_2)\eta_1 \end{pmatrix} \in \mathbf{L}^{4/3}(\Omega) \subset \hat{Z}\,,$$

and

$$\big(\mathcal{F}''(\boldsymbol{\psi}) \cdot \boldsymbol{\zeta}\big) \cdot \boldsymbol{\eta} = \begin{pmatrix} \eta_1\zeta_1 \\ \eta_2\zeta_2 \end{pmatrix} \in \mathbf{L}^{4/3}(\Omega) \subset \hat{Z}\,.$$

These relations verify (H18).

From the definition of the operator K we see that K maps $\mathbf{L}^2(\Gamma)$ into $[\mathbf{H}^{1/2+\epsilon}(\Omega)]^*$, i.e., K maps G into Z. Thus (H19) is verified.

Hence, we are now in a position to apply Theorem 3.5 to derive error estimates for the approximate solutions of the optimality system (4.29)-(4.30), (4.36)-(4.37), and (4.39). It should be noted that Lemma 3.4 implies that for almost all values of λ, the solutions of the optimality system are regular.

THEOREM 4.7. *Assume that Λ is a compact interval of $\mathbb{R}_+$ and that there exists a branch $\{(\lambda, \boldsymbol{\psi}(\lambda), \mathbf{g}(\lambda), \boldsymbol{\eta}(\lambda)) : \lambda \in \Lambda\}$ of regular solutions of the optimality system* (4.29)-(4.30), (4.36)-(4.37), *and* (4.39). *Assume that the finite-element spaces X^h and G^h satisfy the hypotheses* (4.42)-(4.43). *Then there exist a $\delta > 0$ and an $h_0 > 0$ such that for $h \le h_0$, the discrete optimality system* (4.44)-(4.48) *has a unique branch of solutions $\{\big(\lambda, \boldsymbol{\psi}^h(\lambda), g^h(\lambda), \boldsymbol{\eta}^h(\lambda)\big) : \lambda \in \Lambda\}$ satisfying*

$$\{\|\boldsymbol{\psi}^h(\lambda) - \boldsymbol{\psi}(\lambda)\|_1 + \|\mathbf{g}^h(\lambda) - \mathbf{g}(\lambda)\|_{0,\Gamma} + \|\boldsymbol{\eta}^h(\lambda) - \boldsymbol{\eta}(\lambda)\|_1\} < \delta \quad \forall\, \lambda \in \Lambda.$$

Moreover,

$$\lim_{h\to 0}\{\|\boldsymbol{\psi}^h(\lambda) - \boldsymbol{\psi}(\lambda)\|_1 + \|\mathbf{g}^h(\lambda) - \mathbf{g}(\lambda)\|_{0,\Gamma} + \|\boldsymbol{\eta}^h(\lambda) - \boldsymbol{\eta}(\lambda)\|_1\} = 0\,,$$

uniformly in $\lambda \in \Lambda$.

If, in addition, the solution of the optimality system satisfies $\big(\boldsymbol{\psi}(\lambda), \mathbf{g}(\lambda), \boldsymbol{\eta}(\lambda)\big) \in \mathbf{H}^{m+1}(\Omega) \times \mathbf{H}^{m+1/2}(\Omega)|_\Gamma \times \mathbf{H}^{m+1}(\Omega)$ for $\lambda \in \Lambda$, then there exists a constant C, independent of h, such that

$$\begin{aligned} &\|\boldsymbol{\psi}(\lambda) - \boldsymbol{\psi}^h(\lambda)\|_1 + \|\mathbf{g}(\lambda) - \mathbf{g}^h(\lambda)\|_{0,\Gamma} + \|\boldsymbol{\eta}(\lambda) - \boldsymbol{\eta}^h(\lambda)\|_1 \\ &\quad \le Ch^m\big(\|\boldsymbol{\psi}(\lambda)\|_{m+1} + \inf_{\mathbf{v}\in\mathbf{H}^{m+1/2}(\Omega),\mathbf{v}|_\Gamma=\mathbf{g}} \|\mathbf{v}\|_{m+1/2} + \|\boldsymbol{\eta}(\lambda)\|_{m+1}\big)\,, \end{aligned}$$

uniformly in $\lambda \in \Lambda$.

Proof. All results follow from Theorem 3.5. For the last result, we also use (3.25) and the estimates (see [11])

$$\|(T^h T^{-1} - I)\boldsymbol{\psi}\|_1 \le Ch^m \|\boldsymbol{\psi}\|_{m+1} \qquad \text{for } \boldsymbol{\psi} \in \mathbf{H}^{m+1}(\Omega),$$
$$\|\big((T^*)^h (T^*)^{-1} - I\big)\boldsymbol{\eta}\|_1 = \|(T^h T^{-1} - I)\boldsymbol{\eta}\|_1 \le Ch^m \|\boldsymbol{\eta}\|_{m+1} \qquad \text{for } \boldsymbol{\eta} \in \mathbf{H}^{m+1}(\Omega),$$

and (see, e.g., [2], [8], and [15])

$$\|(E^h E^{-1} - I)\mathbf{g}\|_{0,\Gamma} \le Ch^m \inf_{\mathbf{v} \in \mathbf{H}^{m+1/2}(\Omega), \mathbf{v}|_\Gamma = \mathbf{g}} \|\mathbf{v}\|_{m+1/2} \qquad \text{for } \mathbf{g} \in \mathbf{H}^{m+1/2}(\Omega)|_\Gamma .$$

In these estimates, the constant C is independent of h, $\boldsymbol{\psi}$, $\mathbf{g}$, $\boldsymbol{\eta}$, and λ. □

Remark. In fact, we obtain from (4.39) that $\mathbf{g} = -\boldsymbol{\eta}|_\Gamma$, which implies

$$\inf_{\mathbf{v} \in \mathbf{H}^{m+1/2}(\Omega), \mathbf{v}|_\Gamma = \mathbf{g}} \|\mathbf{v}\|_{m+1/2} \le \|\boldsymbol{\eta}\|_{m+1/2} \le \|\boldsymbol{\eta}\|_{m+1},$$

so the term $\big(\inf_{\mathbf{v} \in \mathbf{H}^{m+1/2}(\Omega), \mathbf{v}|_\Gamma = \mathbf{g}} \|\mathbf{v}\|_{m+1/2}\big)$ in the right-hand side of the error estimate is redundant. □

Remark. By using (4.39) again, along with (4.48) and the error estimate in Theorem 4.7, we have the following improved error estimate for the approximation of the control $\mathbf{g}$:

$$\|\mathbf{g}(\lambda) - \mathbf{g}^h(\lambda)\|_{1/2,\Gamma} \le C\|\boldsymbol{\eta}(\lambda) - \boldsymbol{\eta}^h(\lambda)\|_1 \le Ch^m \big(\|\boldsymbol{\psi}(\lambda)\|_{m+1} + \|\boldsymbol{\eta}(\lambda)\|_{m+1}\big).$$

Of course, we also use the fact that we have chosen $G^h = (X^h)|_\Gamma \subset H^{1/2}(\Gamma)$. □

4.3. Dirichlet boundary control for the Navier–Stokes equations of incompressible, viscous flow. For this application we will use Dirichlet boundary controls, i.e., control is effected through the data in a Dirichlet boundary condition. Let Ω denote a bounded domain in $\mathbb{R}^d$, $d = 2$ or 3, with a boundary denoted by Γ. Let $\mathbf{u}$ and p denote the velocity and pressure fields in Ω. The Navier–Stokes equations for a viscous, incompressible flow are given by (see, e.g., [13], [14], or [20])

$$-\nu \nabla \cdot \big((\nabla \mathbf{u}) + (\nabla \mathbf{u})^T\big) + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega,$$
$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega,$$

and

$$\mathbf{u} = \mathbf{b} + \mathbf{g} \quad \text{on } \Gamma,$$

where $\mathbf{f}$ is a given body force, $\mathbf{b}$ and $\mathbf{g}$ are boundary velocity data with $\int_\Gamma \mathbf{b} \cdot \mathbf{n}\, d\Gamma = 0$ and $\int_\Gamma \mathbf{g} \cdot \mathbf{n}\, d\Gamma = 0$, and ν denotes the (constant) kinematic viscosity. We have absorbed the constant density into the pressure and the body force. If the variables in these equations are nondimensionalized, then ν is simply the inverse of the Reynolds number Re.

Setting $\lambda = 1/\nu = Re$ and replacing p with p/λ, $\mathbf{b}$ with $\lambda\mathbf{b}$, and $\mathbf{g}$ with $\lambda\mathbf{g}$, we may write the Navier–Stokes equations in the form

$$-\nabla \cdot \big((\nabla \mathbf{u}) + (\nabla \mathbf{u})^T\big) + \nabla p + \lambda \mathbf{u} \cdot \nabla \mathbf{u} = \lambda \mathbf{f} \quad \text{in } \Omega, \tag{4.49}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \tag{4.50}$$

and

$$\mathbf{u} = \lambda(\mathbf{b} + \mathbf{g}) \quad \text{on } \Gamma. \tag{4.51}$$

We introduce the subspaces

$$L_0^2(\Omega) = \left\{ p \in L^2(\Omega) \;\middle|\; \int_\Omega p \, d\Omega = 0 \right\}$$

and

$$\mathbf{H}_n^1(\Gamma) = \left\{ \mathbf{g} \in \mathbf{H}^1(\Gamma) \;\middle|\; \int_\Gamma \mathbf{g} \cdot \mathbf{n} \, d\Gamma = 0 \right\}.$$

We also introduce the bilinear forms

$$a(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \int_\Omega \left((\nabla \mathbf{u}) + (\nabla \mathbf{u})^T \right) : \left((\nabla \mathbf{v}) + (\nabla \mathbf{v})^T \right) d\Omega \quad \forall\, \mathbf{u}, \mathbf{v} \in \mathbf{H}^1(\Omega)$$

and

$$b(\mathbf{v}, q) = -\int_\Omega q \, \nabla \cdot \mathbf{v} \, d\Omega \quad \forall\, \mathbf{v} \in \mathbf{H}^1(\Omega) \text{ and } \forall\, p \in L^2(\Omega)$$

and the trilinear form

$$c(\mathbf{u}, \mathbf{v}, \mathbf{w}) = \int_\Omega (\mathbf{u} \cdot \nabla) \mathbf{v} \cdot \mathbf{w} \, d\Omega \quad \forall\, \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbf{H}^1(\Omega).$$

These forms are continuous over the spaces of definition indicated above. Moreover, we have the coercivity properties

$$a(\mathbf{v}, \mathbf{v}) \geq C_a \|\mathbf{v}\|_1^2 \quad \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega) \tag{4.52}$$

and

$$\sup_{\mathbf{0} \neq \mathbf{v} \in \mathbf{H}_0^1(\Omega)} \frac{b(\mathbf{v}, q)}{\|\mathbf{v}\|_1} \geq C_b \|q\|_0 \quad \forall\, q \in L_0^2(\Omega) \tag{4.53}$$

for some constants C_a and $C_b > 0$. For details concerning the notation employed and/or for (4.52)-(4.53), one may consult [13], [14], and [20].

We recast the Navier-Stokes equations (4.49)-(4.51) into the following particular weak form (see, e.g., [15]). Seek $(\mathbf{u}, p, \mathbf{t}) \in \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma)$ such that

$$a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) - \langle \mathbf{t}, \mathbf{v} \rangle_\Gamma + \lambda\, c(\mathbf{u}, \mathbf{u}, \mathbf{v}) = \lambda \langle \mathbf{f}, \mathbf{v} \rangle \quad \forall\, \mathbf{v} \in \mathbf{H}^1(\Omega), \tag{4.54}$$

$$b(\mathbf{u}, q) = 0 \quad \forall\, q \in L_0^2(\Omega), \tag{4.55}$$

and

$$\langle \mathbf{s}, \mathbf{u} \rangle_\Gamma - \lambda \langle \mathbf{s}, \mathbf{g} \rangle_\Gamma = \lambda \langle \mathbf{s}, \mathbf{b} \rangle_\Gamma \quad \forall\, \mathbf{s} \in \mathbf{H}^{-1/2}(\Gamma). \tag{4.56}$$

Formally we have

$$\mathbf{t} = \left[-p\mathbf{n} + \left(\nabla \mathbf{u} + (\nabla \mathbf{u})^T \right) \cdot \mathbf{n} \right]_\Gamma,$$

i.e., $\mathbf{t}$ is the stress force on the boundary. The existence of a solution $(\mathbf{u}, p, \mathbf{t})$ for the system (4.54)-(4.56) was established in [15].

Given a desired velocity field $\mathbf{u}_0$, we define for any $(\mathbf{u}, p, \mathbf{t}) \in \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma)$ and $\mathbf{g} \in \mathbf{H}_n^1(\Gamma)$ the functional

$$\mathcal{J}(\mathbf{u}, p, \mathbf{t}, \mathbf{g}) = \frac{\lambda}{4} \int_\Omega |\mathbf{u} - \mathbf{u}_0|^4 \, d\Omega + \frac{\lambda}{2} \int_\Gamma \left(|\nabla_s \mathbf{g}|^2 + |\mathbf{g}|^2 \right) d\Gamma, \tag{4.57}$$

where ∇_s denotes the surface gradient.

We define the spaces $X = \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma)$, $Y = [\mathbf{H}^1(\Omega)]^* \times L_0^2(\Omega) \times \mathbf{H}^{1/2}(\Gamma)$, $G = \mathbf{H}_n^1(\Gamma)$, and $Z = \mathbf{L}^{3/2}(\Omega) \times \{0\} \times \mathbf{H}^1(\Gamma)$. By compact imbedding results, Z is compactly imbedded into Y. For the time being, we assume that the admissible set Θ for the control $\mathbf{g}$ is a closed, convex subset of $G = \mathbf{H}_n^1(\Gamma)$.

We then consider the following optimal control problem associated with the Navier-Stokes equations:

$$\min\{\mathcal{J}(\mathbf{u}, p, \mathbf{t}, \mathbf{g}) : (\mathbf{u}, p, \mathbf{t}) \in X, \mathbf{g} \in \Theta\} \quad \text{subject to} \quad (4.54)\text{-}(4.56). \tag{4.58}$$

We define the continuous linear operator $T \in \mathcal{L}(Y; X)$ as follows. For each $(\boldsymbol{\zeta}, \eta, \boldsymbol{\kappa}) \in Y$, $T(\boldsymbol{\zeta}, \eta, \boldsymbol{\kappa}) = (\tilde{\mathbf{u}}, \tilde{p}, \tilde{\mathbf{t}}) \in X$ is the unique solution of

$$a(\tilde{\mathbf{u}}, \mathbf{v}) + b(\mathbf{v}, \tilde{p}) - \langle \tilde{\mathbf{t}}, \mathbf{v} \rangle_\Gamma = \langle \boldsymbol{\zeta}, \mathbf{v} \rangle \quad \forall\, \mathbf{v} \in \mathbf{H}^1(\Omega),$$

$$b(\tilde{\mathbf{u}}, q) = (\eta, q) \quad \forall\, q \in L_0^2(\Omega),$$

and

$$\langle \mathbf{s}, \tilde{\mathbf{u}} \rangle_\Gamma = \langle \mathbf{s}, \boldsymbol{\kappa} \rangle_\Gamma \quad \forall\, \mathbf{s} \in \mathbf{H}^{-1/2}(\Gamma).$$

It can be easily verified that T is self-adjoint.

We define the (differentiable) nonlinear mapping $N : X \to Y$ by

$$N(\mathbf{u}, p, \mathbf{t}) = -\begin{pmatrix} \mathbf{f} - \mathbf{u} \cdot \nabla \mathbf{u} \\ 0 \\ \mathbf{b} \end{pmatrix}$$

or, equivalently,

$$\langle N(\mathbf{u}, p, \mathbf{t}), (\mathbf{v}, q, \mathbf{s}) \rangle = -(\mathbf{f}, \mathbf{v}) + c(\mathbf{u}, \mathbf{u}, \mathbf{v}) - \langle \mathbf{s}, \mathbf{b} \rangle_\Gamma \quad \forall\, (\mathbf{v}, q, \mathbf{s}) \in X$$

and define $K : \mathbf{H}^{1/2}(\Gamma) \to Y$ by

$$K\mathbf{g} = -\begin{pmatrix} 0 \\ 0 \\ \mathbf{g} \end{pmatrix}$$

or, equivalently,

$$\langle K\mathbf{g}, (\mathbf{v}, q, \mathbf{s}) \rangle = -\langle \mathbf{s}, \mathbf{g} \rangle_\Gamma \quad \forall\, \mathbf{g} \in \mathbf{H}^{1/2}(\Gamma), \forall\, (\mathbf{v}, q, \mathbf{s}) \in X.$$

Clearly, the constraint equations (4.54)-(4.56) can be expressed as

$$(\mathbf{u}, p, \mathbf{t}) + \lambda T N(\mathbf{u}, p, \mathbf{t}) + \lambda T K \mathbf{g} = 0,$$

i.e., in the form (2.2). With the obvious definitions for $\mathcal{F}(\cdot)$ and $\mathcal{E}(\cdot)$, i.e.,

$$\mathcal{F}(\mathbf{u}, p, \mathbf{t}) = \frac{1}{4} \int_\Omega |\mathbf{u} - \mathbf{u}_0|^4 \, d\Omega \quad \forall\, (\mathbf{u}, p, \mathbf{t}) \in X$$

and

$$\mathcal{E}(\mathbf{g}) = \frac{1}{2} \int_\Gamma \left(|\nabla_s \mathbf{g}|^2 + |\mathbf{g}|^2 \right) d\Gamma,$$

the functional (4.57) can be expressed as

$$\mathcal{J}(\mathbf{u}, p, \mathbf{t}, \mathbf{g}) = \lambda \mathcal{F}(\mathbf{u}, p, \mathbf{t}) + \lambda \mathcal{E}(\mathbf{g}),$$

i.e., in the form (2.3).

We are now in a position to verify, for the minimization problem (4.58), all the hypotheses of §§2 and 3.

4.3.1. Verification of the hypotheses for the existence of optimal solutions. We first verify the that the hypotheses (H1)-(H6) hold in the current setting.

(H1) is obviously satisfied with a lower bound 0.

(H2) holds with $\alpha = 1$ and $\beta = 2$.

(H3) is verified with the choice $(\mathbf{u}^{(0)}, p^{(0)}, \mathbf{t}^{(0)}, \mathbf{0}) \in X \times \Theta$ where $(\mathbf{u}^{(0)}, p^{(0)})$ is a solution to the Navier-Stokes equations with Dirichlet boundary conditions, and $\mathbf{t}^{(0)} = [-p^{(0)}\mathbf{n} + (\nabla \mathbf{u}^{(0)} + (\nabla \mathbf{u}^{(0)})^T) \cdot \mathbf{n}]_\Gamma$; see, e.g., [13] or [20].

In order to verify (H4), we assume $\{\mathbf{g}^{(n)}\} \subset \Theta \subset \mathbf{H}_n^1(\Gamma)$ is a sequence satisfying $\mathbf{g}^{(n)} \rightharpoonup \mathbf{g}$ in $\mathbf{H}^1(\Gamma)$; then we have $\mathbf{g}^{(n)} \rightharpoonup \mathbf{g}$ in $\mathbf{H}^{1/2}(\Gamma)$, so $\lim_{n\to\infty} \langle \mathbf{g}^{(n)}, \mathbf{v}\rangle_\Gamma = \langle \mathbf{g}, \mathbf{v}\rangle_\Gamma$ for all $\mathbf{v} \in \mathbf{H}^1(\Omega)$, i.e., $K\mathbf{g}^{(n)} \rightharpoonup K\mathbf{g}$ in Y. Assume that the sequence $\{\mathbf{u}^{(n)}\} \subset \mathbf{H}^1(\Omega)$ satisfies $\mathbf{u}^{(n)} \rightharpoonup \mathbf{u}$ in $\mathbf{H}^1(\Omega)$; then $\mathbf{u}^{(n)} \to \mathbf{u}$ in $\mathbf{L}^4(\Omega)$ by the compactness of the imbedding $\mathbf{H}^1(\Omega) \hookrightarrow\hookrightarrow \mathbf{L}^4(\Omega)$. Now,

$$\begin{aligned}\langle N(\mathbf{u}^{(n)}), \mathbf{v}\rangle &= c(\mathbf{u}^{(n)}, \mathbf{u}^{(n)}, \mathbf{v}) = c(\mathbf{u}, \mathbf{u}^{(n)}, \mathbf{v}) + c(\mathbf{u}^{(n)} - \mathbf{u}, \mathbf{u}^{(n)}, \mathbf{v}) \\ &\to c(\mathbf{u}, \mathbf{u}, \mathbf{v}) + 0 = \langle N(\mathbf{u}), \mathbf{v}\rangle \qquad \text{as } n \to \infty .\end{aligned}$$

Hence, (H4) is verified.

The verification of (H5) follows directly from the observation that the mappings $(\mathbf{u}, p, \mathbf{t}) \mapsto \mathcal{F}(\mathbf{u}, p, \mathbf{t}) = (1/4)\,\|\mathbf{u} - \mathbf{u}\|^4_{\mathbf{L}^4(\Omega)}$ and $\mathbf{g} \mapsto \mathcal{E}(\mathbf{g}) = (1/2)\,\|\mathbf{g}\|^2_{1,\Gamma}$ are convex.

To verify (H6), we combine a priori estimates obtained from the constraint equations and the functional. Let $\{\mathbf{u}^{(k)}, p^{(k)}, \mathbf{t}^{(k)}, \mathbf{g}^{(k)}\} \subset \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma) \times \mathbf{H}_n^1(\Gamma)$ be a sequence such that

$$\mathcal{J}(\mathbf{u}^{(k)}, \mathbf{g}^{(k)}) \le C , \tag{4.59}$$

$$a(\mathbf{u}^{(k)}, \mathbf{v}) + b(\mathbf{v}, p^{(k)}) - \langle \mathbf{t}^{(k)}, \mathbf{v}\rangle_\Gamma + \lambda\, c(\mathbf{u}^{(k)}, \mathbf{u}^{(k)}, \mathbf{v}) = \lambda\, \langle \mathbf{f}, \mathbf{v}\rangle \quad \forall\, \mathbf{v} \in \mathbf{H}^1(\Omega) , \tag{4.60}$$

$$b(\mathbf{u}^{(k)}, q) = 0 \quad \forall\, q \in L_0^2(\Omega) , \tag{4.61}$$

and

$$\langle \mathbf{s}, \mathbf{u}^{(k)}\rangle_\Gamma - \lambda \langle \mathbf{s}, \mathbf{g}^{(k)}\rangle_\Gamma = \lambda \langle \mathbf{s}, \mathbf{b}\rangle_\Gamma \quad \forall\, \mathbf{s} \in \mathbf{H}^{-1/2}(\Gamma) . \tag{4.62}$$

First, (4.59) implies that $(\mathbf{u}^{(k)}, \mathbf{g}^{(k)})$ is uniformly bounded in $\mathbf{L}^4(\Omega) \times \mathbf{H}^1(\Gamma)$. For each $\mathbf{g}^{(k)}$, we may choose a $(\mathbf{w}^{(k)}, r^{(k)}) \in \mathbf{H}^1(\Omega) \times L_0^2(\Omega)$ that satisfies the Stokes problem

$$a(\mathbf{w}^{(k)}, \mathbf{v}) + b(\mathbf{v}, r^{(k)}) = \langle \mathbf{f}, \mathbf{v}\rangle \quad \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega) , \tag{4.63}$$

$$b(\mathbf{w}^{(k)}, q) = 0 \quad \forall\, q \in L_0^2(\Omega), \tag{4.64}$$

and

$$\mathbf{w}^{(k)} = \lambda(\mathbf{g}^{(k)} + \mathbf{b}) \quad \text{on } \Gamma . \tag{4.65}$$

Furthermore, the estimate

$$\|\mathbf{w}^{(k)}\|_1 \le C(\|\mathbf{f}\|_0 + \|\mathbf{b}\|_{1/2,\Gamma} + \|\mathbf{g}^{(k)}\|_{1,\Gamma}) \tag{4.66}$$

holds. By subtracting (4.63) from (4.60) with $\mathbf{v} = \mathbf{u}^{(k)} - \mathbf{w}^{(k)}$, also using (4.61) and (4.64), we obtain

$$\begin{aligned}a(\mathbf{u}^{(k)} - \mathbf{w}^{(k)}, \mathbf{u}^{(k)} - \mathbf{w}^{(k)}) &= -\lambda\, c(\mathbf{u}^{(k)}, \mathbf{u}^{(k)}, \mathbf{u}^{(k)} - \mathbf{w}^{(k)}) \\ &= \lambda\, c(\mathbf{u}^{(k)}, \mathbf{u}^{(k)} - \mathbf{w}^{(k)}, \mathbf{u}^{(k)}) .\end{aligned} \tag{4.67}$$

Note that

$$
\begin{aligned}
&|c(\mathbf{u}^{(k)}, \mathbf{u}^{(k)} - \mathbf{w}^{(k)}, \mathbf{u}^{(k)})| \\
&= \frac{1}{2}\left|\int_\Omega \mathbf{u}^{(k)} \cdot \left((\nabla(\mathbf{u}^{(k)} - \mathbf{w}^{(k)})) + (\nabla(\mathbf{u}^{(k)} - \mathbf{w}^{(k)}))^T\right) \cdot \mathbf{u}^{(k)}\, d\Omega\right| \\
&\le C \left\|(\nabla(\mathbf{u}^{(k)} - \mathbf{w}^{(k)})) + (\nabla(\mathbf{u}^{(k)} - \mathbf{w}^{(k)}))^T\right\|_0 \left\|\mathbf{u}^{(k)}\right\|_{\mathbf{L}^4(\Omega)} \\
&\le \frac{1}{4\lambda} \left\|(\nabla(\mathbf{u}^{(k)} - \mathbf{w}^{(k)})) + (\nabla(\mathbf{u}^{(k)} - \mathbf{w}^{(k)}))^T\right\|_0^2 + C_\lambda \left\|\mathbf{u}^{(k)}\right\|_{\mathbf{L}^4(\Omega)}^4
\end{aligned}
$$

so that, using (4.67), we have that

$$
\frac{1}{4}\left\|\left(\nabla(\mathbf{u}^{(k)} - \mathbf{w}^{(k)})\right) + \left(\nabla(\mathbf{u}^{(k)} - \mathbf{w}^{(k)})\right)^T\right\|_0^2 \le C_\lambda \left\|\mathbf{u}^{(k)}\right\|_{\mathbf{L}^4(\Omega)}^4 .
$$

Then, by (4.66) and the triangle inequality, we have that

$$
\left\|(\nabla \mathbf{u}^{(k)}) + (\nabla \mathbf{u}^{(k)})^T\right\|_0 \le C\{\|\mathbf{f}\|_0 + \|\mathbf{b}\|_{1/2,\Gamma} + \|\mathbf{g}^{(k)}\|_{1,\Gamma} + \|\mathbf{u}^{(k)}\|_{\mathbf{L}^4(\Omega)}^2\} .
$$

Thus,

$$
\begin{aligned}
&\|(\nabla \mathbf{u}^{(k)}) + (\nabla \mathbf{u}^{(k)})^T\|_0 + \|\mathbf{u}^{(k)}\|_{0,\Gamma} \\
&\le \|(\nabla \mathbf{u}^{(k)}) + (\nabla \mathbf{u}^{(k)})^T\|_0 + \|\mathbf{b}\|_{0,\Gamma} + \|\mathbf{g}^{(k)}\|_{0,\Gamma} \\
&\le C(\|\mathbf{f}\|_0 + \|\mathbf{b}\|_{1/2,\Gamma} + \|\mathbf{g}^{(k)}\|_{1,\Gamma} + \|\mathbf{u}^{(k)}\|_{\mathbf{L}^4(\Omega)}^2) .
\end{aligned}
$$

Since the mapping $\mathbf{u} \mapsto \|\nabla \mathbf{u} + (\nabla \mathbf{u})^T\|_0 + \|\mathbf{u}\|_{0,\Gamma}$ defines a norm on $\mathbf{H}^1(\Omega)$ equivalent to the standard $\mathbf{H}^1(\Omega)$-norm, we have that

$$
\|\mathbf{u}^{(k)}\|_1 \le C\{\|\mathbf{f}\|_0 + \|\mathbf{b}\|_{1/2,\Gamma} + \|\mathbf{g}^{(k)}\|_{1,\Gamma} + \|\mathbf{u}^{(k)}\|_{\mathbf{L}^4(\Omega)}^2\} ;
$$

since $\|\mathbf{u}^{(k)}\|_{\mathbf{L}^4(\Omega)}$ and $\|\mathbf{g}^{(k)}\|_{1,\Gamma}$ are uniformly bounded, we conclude that $\|\mathbf{u}^{(k)}\|_1$ is uniformly bounded as well. One easily concludes from (4.60) that $\|\mathbf{t}^{(k)}\|_{-1/2,\Gamma}$ is uniformly bounded. Thus (H6) is verified.

It is now just a matter of citing Theorem 2.1 to conclude the existence of an optimal solution that minimizes (4.57) subject to (4.54)–(4.56).

THEOREM 4.8. *There exists a* $(\mathbf{u}, p, \mathbf{t}, \mathbf{g}) \in \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Omega) \times \Theta$ *such that* (4.57) *is minimized subject to* (4.54)–(4.56). □

4.3.2. Verification of the hypotheses for the existence of Lagrange multipliers. We now assume $(\mathbf{u}, p, \mathbf{t}, \mathbf{g})$ is an optimal solution and turn to the verification of hypotheses (H7)–(H9).

The validity of (H7) is obvious.

(H8) holds since the mapping $\mathbf{z} \mapsto \mathcal{E}(\mathbf{g}) = (1/2)\int_\Gamma \left(|\nabla_s \mathbf{g}|^2 + |\mathbf{g}|^2\right) d\Gamma$ is convex.

(H9) can be verified as follows. For any $(\mathbf{u}, p, \mathbf{t}) \in X$, the operator $N'(\mathbf{u}, p, \mathbf{t}) : X \to Y$ is given by

$$
N'(\mathbf{u}, p, \mathbf{t}) \cdot (\mathbf{v}, q, \mathbf{s}) = -\begin{pmatrix} \mathbf{u} \cdot \nabla \mathbf{v} + \mathbf{v} \cdot \nabla \mathbf{u} \\ 0 \\ \mathbf{0} \end{pmatrix}
$$

for all $(\mathbf{v}, q, \mathbf{s}) \in \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma)$. Thus we obtain $N'(\mathbf{u}, p, \mathbf{t}) \cdot (\mathbf{v}, q, \mathbf{s}) \in \mathbf{L}^{3/2}(\Omega) \times \{0\} \times \mathbf{H}^1(\Gamma) = Z$.

The Lagrangian is given by

$$\begin{aligned}\mathcal{L}(\mathbf{u}, p, \mathbf{t}, \mathbf{g}, \boldsymbol{\nu}, \phi, \boldsymbol{\tau}, k) &= k\,\mathcal{J}(\mathbf{u}, \mathbf{g}) - \{a(\mathbf{u}, \boldsymbol{\nu}) + \lambda\, c(\mathbf{u}, \mathbf{u}, \boldsymbol{\nu}) + b(\boldsymbol{\nu}, p) + b(\mathbf{u}, \phi) - \langle \boldsymbol{\tau}, \mathbf{u}\rangle_\Gamma \\ &\qquad - \langle \mathbf{t}, \boldsymbol{\nu}\rangle_\Gamma - \lambda\,\langle \mathbf{f}, \boldsymbol{\nu}\rangle_\Gamma + \lambda\,\langle \boldsymbol{\tau}, \mathbf{b}\rangle_\Gamma + \lambda\,\langle \boldsymbol{\tau}, \mathbf{g}\rangle_\Gamma\}\end{aligned}$$

for all $(\mathbf{u}, p, \mathbf{t}, \mathbf{g}, \boldsymbol{\nu}, \phi, \boldsymbol{\tau}, k) \in X \times G \times X \times \mathbb{R} = \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma) \times \mathbf{H}_n^1(\Gamma) \times \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma) \times \mathbb{R}$. Note that in this form of the Lagrangian, the Lagrange multiplier $(\boldsymbol{\nu}, \phi, \boldsymbol{\tau}) \in X = Y^*$, so we have already introduced the change of variables indicated between (2.17)-(2.18) and (2.19)-(2.21).

Having verified the hypotheses (H7)-(H9), we may apply Theorem 2.4 to conclude that there exist a Lagrange multiplier $(\boldsymbol{\nu}, \phi, \boldsymbol{\tau}) \in X = \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma)$ and a real number k such that

$$(\boldsymbol{\nu}, \phi, \boldsymbol{\tau}) + \lambda\, T^*\Big([N'(\mathbf{u}, p, \mathbf{t})]^* \cdot (\boldsymbol{\nu}, \phi, \boldsymbol{\tau}) - k\,\mathcal{F}'(\mathbf{u}, p, \mathbf{t})\Big) = 0 \tag{4.68}$$

and

$$\mathcal{L}(\mathbf{u}, p, \mathbf{t}, \mathbf{z}, \boldsymbol{\nu}, \phi, \boldsymbol{\tau}, k) \le \mathcal{L}(\mathbf{u}, p, \mathbf{t}, \mathbf{g}, \boldsymbol{\nu}, \phi, \boldsymbol{\tau}, k) \quad \forall\, \mathbf{z} \in \Theta \tag{4.69}$$

and that for almost all values of λ, we may choose $k = 1$.

Recall that $T^* = T$. Also, note that for $(\mathbf{u}, p, \mathbf{t}) \in X = \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma)$, the operator $[N'(\mathbf{u}, p, \mathbf{t})]^* : X \to Y$ is given by

$$[N'(\mathbf{u}, p, \mathbf{t})]^* \cdot (\mathbf{v}, q, \mathbf{s}) = \begin{pmatrix} -\mathbf{u} \cdot \nabla \mathbf{v} + \mathbf{v} \cdot (\nabla \mathbf{u})^T \\ 0 \\ \mathbf{0} \end{pmatrix} \quad \forall\, (\mathbf{v}, q, \mathbf{s}) \in X.$$

Thus, (4.68), with $k = 1$, can be rewritten as

$$\begin{aligned} &a(\mathbf{w}, \boldsymbol{\nu}) + \lambda\, c(\mathbf{w}, \mathbf{u}, \boldsymbol{\nu}) + \lambda\, c(\mathbf{u}, \mathbf{w}, \boldsymbol{\nu}) + b(\mathbf{w}, \phi) - \langle \boldsymbol{\tau}, \mathbf{w}\rangle_\Gamma \\ &\qquad = \lambda\,\big((\mathbf{u} - \mathbf{u}_0)^3, \mathbf{w}\big) \quad \forall\, \mathbf{w} \in \mathbf{H}^1(\Omega)\,, \end{aligned} \tag{4.70}$$

$$b(\boldsymbol{\nu}, r) = 0 \quad \forall\, r \in L_0^2(\Omega)\,, \tag{4.71}$$

and

$$\langle \mathbf{y}, \boldsymbol{\nu}\rangle_\Gamma = 0 \quad \forall\, \mathbf{y} \in \mathbf{H}^{-1/2}(\Gamma)\,. \tag{4.72}$$

In the right-hand side of (4.70), we use the notation $(\mathbf{v}^3, \mathbf{w}) = \sum_{j=1}^d (v_j^3, w_j)$.

Using the definition of the Lagrangian functional, (4.69), with $k = 1$, can be rewritten as

$$\begin{aligned} &\frac{\lambda}{2}\,(\nabla_{\mathrm{s}}\mathbf{z}, \nabla_{\mathrm{s}}\mathbf{z})_\Gamma + \frac{\lambda}{2}\,(\mathbf{z}, \mathbf{z})_\Gamma - \frac{\lambda}{2}\,(\nabla_{\mathrm{s}}\mathbf{g}, \nabla_{\mathrm{s}}\mathbf{g})_\Gamma \\ &\qquad - \frac{\lambda}{2}\,(\mathbf{g}, \mathbf{g})_\Gamma - \lambda\langle \boldsymbol{\tau}, \mathbf{z}\rangle_\Gamma + \lambda\langle \boldsymbol{\tau}, \mathbf{g}\rangle_\Gamma \ge 0 \quad \forall\, \mathbf{z} \in \Theta\,. \end{aligned}$$

For each $\epsilon \in (0, 1)$ and each $\mathbf{z} \in \Theta$, by plugging $\epsilon\mathbf{z} + (1 - \epsilon)\mathbf{g} \in \Theta$ into the last inequality we obtain

$$\begin{aligned} &\epsilon\,\big(\nabla_{\mathrm{s}}\mathbf{g}, \nabla_{\mathrm{s}}(\mathbf{z} - \mathbf{g})\big)_\Gamma + \epsilon\,(\mathbf{g}, \mathbf{z} - \mathbf{g})_\Gamma + \frac{\epsilon^2}{2}\,\big(\nabla_{\mathrm{s}}(\mathbf{z} - \mathbf{g}), \nabla_{\mathrm{s}}(\mathbf{z} - \mathbf{g})\big)_\Gamma \\ &\qquad + \frac{\epsilon^2}{2}\,(\mathbf{z} - \mathbf{g}, \mathbf{z} - \mathbf{g})_\Gamma - \epsilon\langle \boldsymbol{\tau}, \mathbf{z} - \mathbf{g}\rangle_\Gamma \ge 0 \quad \forall\, \mathbf{z} \in \Theta\,; \end{aligned}$$

hence, after dividing by $\epsilon > 0$ and then letting $\epsilon \to 0^+$, we obtain

$$(4.73) \qquad \big(\nabla_s \mathbf{g}, \nabla_s(\mathbf{z}-\mathbf{g})\big)_\Gamma + (\mathbf{g}, \mathbf{z}-\mathbf{g})_\Gamma - \langle \boldsymbol{\tau}, \mathbf{z}\rangle_\Gamma \geq 0 \quad \forall\, \mathbf{z} \in \Theta\,.$$

We see that for almost all values of λ, necessary conditions for an optimum are that (4.54)-(4.56), (4.70)-(4.72) and (4.73) are satisfied. Again, the system formed by these equations will be called an *optimality system*.

We now specialize to the case $\Theta = \mathbf{H}_n^1(\Gamma)$. Note that hypothesis (H10) is satisfied. Then using Theorem 2.5, we see that inequality (4.73) becomes an equality, and by letting $\mathbf{z} = \mathbf{k} - \mathbf{g}$ vary arbitrarily in $\mathbf{H}_n^1(\Gamma)$, we now have, instead of (4.73),

$$(4.74) \qquad (\nabla_s \mathbf{g}, \nabla_s \mathbf{z})_\Gamma + (\mathbf{g}, \mathbf{z})_\Gamma - \langle \boldsymbol{\tau}, \mathbf{z}\rangle_\Gamma = 0 \quad \forall\, \mathbf{z} \in \Theta = \mathbf{H}_n^1(\Gamma)\,.$$

Thus, according to that theorem, we have that for almost all λ, an optimality system of equations is now given by (4.54)-(4.56), (4.70)-(4.72) and (4.74). However, we can go further and verify that hypothesis (H11) is valid, which in turn will justify the existence of a Lagrange multiplier satisfying the optimality system for *all* $\lambda \in \Lambda$.

We now verify (H11), which we again note can be equivalently stated as follows.

$$\text{If } \xi \in Y^* \text{ satisfies } \big(I + \lambda\, T^*[N'(u)]^*\big)\xi = 0 \text{ and } K^*\xi = 0, \text{ then } \xi = 0\,.$$

To verify this hypothesis, we assume that $(\boldsymbol{\xi}, \sigma, \boldsymbol{\theta}) \in Y^* = \mathbf{H}^1(\Omega) \times L^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma)$ satisfies $\big(I + \lambda\, T^*[N'(\mathbf{u}, p, \mathbf{t})]^*\big)(\boldsymbol{\xi}, \sigma, \boldsymbol{\theta}) = (\mathbf{0}, 0, \mathbf{0})$ and $K^*(\boldsymbol{\xi}, \sigma, \boldsymbol{\theta}) = \mathbf{0}$, i.e.,

$$a(\mathbf{w}, \boldsymbol{\xi}) + \lambda\, c(\mathbf{w}, \mathbf{u}, \boldsymbol{\xi}) + \lambda\, c(\mathbf{u}, \mathbf{w}, \boldsymbol{\xi}) + b(\mathbf{w}, \sigma) - \langle \boldsymbol{\theta}, \mathbf{w}\rangle_\Gamma = 0 \quad \forall\, \mathbf{w} \in \mathbf{H}^1(\Omega)\,,$$

$$b(\boldsymbol{\xi}, r) = 0 \quad \forall\, r \in L_0^2(\Omega)\,,$$

$$\langle \mathbf{y}, \boldsymbol{\xi}\rangle_\Gamma = 0 \quad \forall\, \mathbf{y} \in \mathbf{H}^{-1/2}(\Gamma)\,,$$

and

$$\boldsymbol{\theta} = \mathbf{0} \quad \text{on } \Gamma\,.$$

(Note that $K^*(\boldsymbol{\xi}, \sigma, \boldsymbol{\theta}) = \boldsymbol{\theta}$.) Let Ω' be a smooth extension of Ω such that $\overline{\Omega}$ is a compact subset of Ω'. We then set $\boldsymbol{\xi}'$, σ', and $\mathbf{u}'$ to be the extension, by zero outside Ω, of $\boldsymbol{\xi}$, σ, and $\mathbf{u}$, respectively. We may show from the last four equations that

$$\boldsymbol{\xi}' \in \mathbf{H}^1(\Omega')\,, \qquad \sigma' \in L_0^2(\Omega')\,,$$

$$(4.75) \qquad a'(\mathbf{w}, \boldsymbol{\xi}') + \lambda\, c'(\mathbf{w}, \mathbf{u}', \boldsymbol{\xi}') + \lambda\, c'(\mathbf{u}', \mathbf{w}, \boldsymbol{\xi}') + b'(\mathbf{w}, \sigma') = 0 \quad \forall\, \mathbf{w} \in \mathbf{H}_0^1(\Omega')\,,$$

and

$$(4.76) \qquad b'(\boldsymbol{\xi}', r) = 0 \quad \forall\, r \in L_0^2(\Omega')\,,$$

where the forms $a'(\cdot,\cdot)$, $b'(\cdot,\cdot)$, and $c'(\cdot,\cdot,\cdot)$ defined over Ω' are the analogues of corresponding forms defined over Ω. Using a unique continuation result for the system (4.75)-(4.76) that was established in [16] or [17], we obtain $\boldsymbol{\xi}' = \mathbf{0}$ and $\sigma' = 0$ in Ω', or $\boldsymbol{\xi} = \mathbf{0}$ and $\sigma = 0$ in Ω. Thus (H11) is verified.

Hence we conclude that for *all* λ, the optimality system (4.54)-(4.56), (4.70)-(4.72), and (4.74) has a solution. Thus, we have Theorem 2.6, which, in the present context, is given as follows.

THEOREM 4.9. *Let* $(\mathbf{u}, p, \mathbf{t}, \mathbf{g}) \in \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma) \times \mathbf{H}_n^1(\Gamma)$ *denote an optimal solution that minimizes* (4.57) *subject to* (4.54)-(4.56). *Then, for all* $\lambda \in \Lambda$, *there exists a*

nonzero Lagrange multiplier $(\boldsymbol{\nu}, \phi, \boldsymbol{\tau}) \in \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma)$ *satisfying the Euler equations* (4.70)–(4.72) *and* (4.74). □

Note that, in the above expression, we have already employed hypothesis (H12), which in the current context is easily seen to be satisfied with $E : G \to G^*$ defined by

$$\langle E\mathbf{g}, \mathbf{z}\rangle = \int_\Gamma (\nabla_s \mathbf{g} \cdot \nabla_s \mathbf{z} + \mathbf{g} \cdot \mathbf{z})\, d\Gamma \quad \forall\, \mathbf{z} \in \mathbf{H}_n^1(\Gamma) = G\,.$$

We also note that for each fixed $\boldsymbol{\tau}$, (4.74), with $\mathbf{g} \in \mathbf{H}_n^1(\Gamma)$, is equivalent to

$$(\nabla_s \mathbf{g}, \nabla_s \mathbf{k})_\Gamma + (\mathbf{g}, \mathbf{k})_\Gamma + \gamma \int_\Gamma \mathbf{k} \cdot \mathbf{n}\, d\Gamma = \langle \boldsymbol{\tau}, \mathbf{k}\rangle_\Gamma \quad \forall\, \mathbf{k} \in \mathbf{H}^1(\Gamma) \tag{4.77}$$

and

$$\int_\Gamma \mathbf{g} \cdot \mathbf{n}\, d\Gamma = 0\,, \tag{4.78}$$

where $\gamma \in \mathbb{R}$ is an additional unknown constant that accounts for the single integral constraint of (4.78). The equivalence can be shown as follows. First, an application of the Lax–Milgram Lemma to (4.74) on the space $\mathbf{H}_n^1(\Gamma)$ guarantees the existence and uniqueness of a solution $\mathbf{g} \in \mathbf{H}_n^1(\Gamma)$ to (4.74); this solution $\mathbf{g}$ clearly satisfies (4.77)–(4.78) with $\gamma = \int_\Gamma \big(\boldsymbol{\tau} \cdot \mathbf{n} - \nabla_s \mathbf{g} : \nabla_s \mathbf{n} - \mathbf{g} \cdot \mathbf{n}\big)\, d\Gamma$. Conversely, any solution $(\mathbf{g}, \gamma)$ of (4.77)–(4.78) trivially satisfies (4.74). Although (4.74) and (4.77)–(4.78) are equivalent, the latter is more easily discretized.

4.3.3. Verification of the hypotheses for approximations and error estimates. We finally verify hypotheses (H13)–(H19) that are used in connection with approximations and error estimates.

A finite-element discretization of the optimality system (4.54)–(4.56), (4.70)–(4.72), and (4.74) is defined as follows. First, one chooses families of finite-dimensional subspaces $\mathbf{V}^h \subset \mathbf{H}^1(\Omega)$ and $S^h \subset L^2(\Omega)$. These families are parameterized by the parameter h that tends to zero; commonly, this parameter is chosen to be some measure of the grid size in a subdivision of Ω into finite elements. We let $S_0^h = S^h \cap L_0^2(\Omega)$ and $\mathbf{V}_0^h = \mathbf{V}^h \cap \mathbf{H}_0^1(\Omega)$.

One may choose any pair of subspaces $\mathbf{V}^h$ and S^h that can be used for finding finite-element approximations of solutions of the Navier–Stokes equations. Thus, concerning these subspaces, we make the following standard assumptions, which are exactly those employed in well-known finite-element methods for the Navier–Stokes equations. First, we have the approximation properties: there exist an integer k and a constant C, independent of h, $\mathbf{v}$, and q, such that

$$\inf_{\mathbf{v}^h \in \mathbf{V}^h} \|\mathbf{v} - \mathbf{v}^h\|_1 \le Ch^m \|\mathbf{v}\|_{m+1} \quad \forall\, \mathbf{v} \in \mathbf{H}^{m+1}(\Omega)\,,\ 1 \le m \le k\,, \tag{4.79}$$

and

$$\inf_{q^h \in S_0^h} \|q - q^h\|_0 \le Ch^m \|q\|_m \quad \forall\, q \in H^m(\Omega) \cap L_0^2(\Omega)\,,\ 1 \le m \le k\,. \tag{4.80}$$

Next, we assume the *inf-sup condition*, or *Ladyzhenskaya–Babuska–Brezzi condition*: there exists a constant C, independent of h, such that

$$\inf_{0 \ne q^h \in S_0^h} \sup_{0 \ne \mathbf{v}^h \in \mathbf{V}^h} \frac{b(\mathbf{v}^h, q^h)}{\|\mathbf{v}^h\|_1\, \|q^h\|_0} \ge C\,. \tag{4.81}$$

This condition ensures the stability of finite-element discretizations of the Navier-Stokes equations. For thorough discussions of the approximation properties (4.79)-(4.80), see, e.g., [2] or [8]; for like discussions of the stability condition (4.81), see, e.g., [13] or [14]. The latter references may also be consulted for a catalogue of finite-element subspaces that meet the requirements of (4.79)-(4.81).

Next, let $\mathbf{P}^h = \mathbf{V}^h|_\Gamma$; i.e., $\mathbf{P}^h$ consists of the restriction, to the boundary Γ, of functions belonging to $\mathbf{V}^h$. For all choices of conforming finite-element spaces $\mathbf{V}^h$, e.g., Lagrange-type finite-element spaces, we have that $\mathbf{P}^h \subset \mathbf{H}^{-1/2}(\Gamma)$. For the subspaces $\mathbf{P}^h = \mathbf{V}^h|_\Gamma$, we can show the following approximation property: there exist an integer k and a constant C, independent of h and $\mathbf{s}$, such that

$$\begin{aligned} &\inf_{\mathbf{s}^h \in \mathbf{P}^h} \|\mathbf{s} - \mathbf{s}^h\|_{-1/2,\Gamma} \\ &\qquad \le C h^m \inf_{\mathbf{v} \in \mathbf{H}^m(\Omega), \mathbf{v}|_\Gamma = \mathbf{s}} \|\mathbf{v}\|_m \qquad \forall\, \mathbf{s} \in \mathbf{H}^m(\Omega)|_\Gamma\,,\ 1 \le m \le k\,. \end{aligned} \tag{4.82}$$

We also use the following inverse assumption: there exists a constant C, independent of h and $\mathbf{s}^h$, such that

$$\|\mathbf{s}^h\|_{s,\Gamma} \le C h^{s-q} \|\mathbf{s}^h\|_{q,\Gamma} \quad \forall\, \mathbf{s}^h \in \mathbf{P}^h\,,\ -1/2 \le q \le s \le 1/2\,. \tag{4.83}$$

See [2] or [8] for details concerning (4.82) and (4.83). See also [15] for (4.82).

Now, let $\mathbf{Q}^h = \mathbf{V}^h|_\Gamma$; i.e., $\mathbf{Q}^h$ consists of the restriction, to the boundary Γ, of functions belonging to $\mathbf{V}^h$. Again, for all choices of conforming finite-element spaces $\mathbf{V}^h$ we then have that $\mathbf{Q}^h \subset \mathbf{H}^1(\Gamma)$. We can show the approximation property: there exist an integer k and a constant C, independent of h and $\mathbf{k}$, such that for $1 \le m \le k$, $0 \le s \le 1$, and $\mathbf{k} \in \mathbf{H}^{m+1}(\Omega)|_\Gamma$,

$$\inf_{\mathbf{k}^h \in \mathbf{Q}^h} \|\mathbf{k} - \mathbf{k}^h\|_{s,\Gamma} \le C h^{m-s+\frac{1}{2}} \inf_{\mathbf{v} \in \mathbf{H}^{m+1}(\Omega), \mathbf{v}|_\Gamma = \mathbf{k}} \|\mathbf{v}\|_{m+1}\,. \tag{4.84}$$

This property follows from (4.79), once one notes that the same type of polynomials are used in $\mathbf{Q}^h$ as are used in $\mathbf{V}^h$. We set $G^h = \mathbf{Q}^h \cap \mathbf{H}^1_n(\Gamma)$.

Once the approximating subspaces have been chosen we seek $\mathbf{u}^h \in \mathbf{V}^h$, $p^h \in S^h_0$, $\mathbf{t}^h \in \mathbf{P}^h$, $\mathbf{g}^h \in \mathbf{Q}^h$, $\boldsymbol{\nu}^h \in \mathbf{V}^h$, $\phi^h \in S^h_0$, $\boldsymbol{\tau}^h \in \mathbf{P}^h$, and $\gamma^h \in \mathbb{R}$ such that

$$a(\mathbf{u}^h, \mathbf{v}^h) + \lambda\, c(\mathbf{u}^h, \mathbf{u}^h, \mathbf{v}^h) + b(\mathbf{v}^h, p^h) - \langle \mathbf{v}^h, \mathbf{t}^h \rangle_\Gamma = \lambda\, \langle \mathbf{f}, \mathbf{v}^h \rangle \quad \forall\, \mathbf{v}^h \in \mathbf{V}^h\,, \tag{4.85}$$

$$b(\mathbf{u}^h, q^h) = 0 \quad \forall\, q^h \in S^h_0\,, \tag{4.86}$$

$$\langle \mathbf{u}^h, \mathbf{s}^h \rangle_\Gamma - \lambda \langle \mathbf{g}^h, \mathbf{s}^h \rangle_\Gamma = \lambda \langle \mathbf{b}, \mathbf{s}^h \rangle_\Gamma \quad \forall\, \mathbf{s}^h \in \mathbf{P}^h\,, \tag{4.87}$$

$$(\nabla_s \mathbf{g}^h, \nabla_s \mathbf{k}^h)_\Gamma + \langle \mathbf{g}^h, \mathbf{k}^h \rangle_\Gamma + \gamma^h \int_\Gamma \mathbf{k}^h \cdot \mathbf{n}\, d\Gamma = \langle \boldsymbol{\tau}^h, \mathbf{k}^h \rangle_\Gamma \quad \forall\, \mathbf{k}^h \in \mathbf{Q}^h\,, \tag{4.88}$$

$$\int_\Gamma \mathbf{g}^h \cdot \mathbf{n}\, d\Gamma = 0\,, \tag{4.89}$$

$$\begin{aligned} &a(\mathbf{w}^h, \boldsymbol{\nu}^h) + \lambda\, c(\mathbf{w}^h, \mathbf{u}^h, \boldsymbol{\nu}^h) + \lambda\, c(\mathbf{u}^h, \mathbf{w}^h, \boldsymbol{\nu}^h) + b(\mathbf{w}^h, \phi^h) - \langle \mathbf{w}^h, \boldsymbol{\tau}^h \rangle_\Gamma \\ &\qquad = \lambda \left((\mathbf{u}^h - \mathbf{u}_0)^3, \mathbf{w}^h \right) \quad \forall\, \mathbf{w}^h \in \mathbf{V}^h\,, \end{aligned} \tag{4.90}$$

$$b(\boldsymbol{\nu}^h, r^h) = 0 \quad \forall\, r^h \in S^h_0\,, \tag{4.91}$$

and

$$\langle \boldsymbol{\nu}^h, \mathbf{y}^h \rangle = 0 \quad \forall\, \mathbf{y}^h \in \mathbf{P}^h\,. \tag{4.92}$$

Note that if (4.85)-(4.92) are satisfied, then necessarily $\mathbf{g}^h \in G^h$. Also, in the right-hand side of (4.90), we use a notation similar to that used in the right-hand side of (4.70).

The operator $T^h \in \mathcal{L}(Y, X^h)$ is defined as the solution operator for

$$a(\mathbf{u}^h, \mathbf{v}^h) + b(\mathbf{v}^h, p^h) - \langle \mathbf{v}^h, \mathbf{t}^h \rangle_\Gamma = \langle \mathbf{f}, \mathbf{v}^h \rangle \quad \forall\, \mathbf{v}^h \in \mathbf{V}^h\,,$$
$$b(\mathbf{u}^h, q^h) = 0 \quad \forall\, q^h \in S_0^h\,,$$

and

$$\langle \mathbf{u}^h, \mathbf{s}^h \rangle_\Gamma = \langle \mathbf{b}, \mathbf{s}^h \rangle_\Gamma \quad \forall\, \mathbf{s}^h \in \mathbf{P}^h\,;$$

i.e, for each $\mathbf{f} \in Y$, $T^h\mathbf{f} = \boldsymbol{\psi}^h \in X^h$ is the solution of the above system of equations.

Since $T = T^*$, we define $(T^*)^h = T^h$.

We define the operator $E^h : G^* \to G^h$ as follows. For each $\boldsymbol{\tau} \in G^*$, $\mathbf{g}^h = E^h\boldsymbol{\tau}$ if and only if

$$(\nabla_s \mathbf{g}^h, \nabla_s \mathbf{z}^h)_\Gamma + \langle \mathbf{g}^h, \mathbf{z}^h \rangle_\Gamma + \gamma^h \int_\Gamma \mathbf{z}^h \cdot \mathbf{n}\, d\Gamma = \langle \boldsymbol{\tau}^h, \mathbf{z}^h \rangle_\Gamma \quad \forall\, \mathbf{z}^h \in \mathbf{Q}^h$$

and

$$\int_\Gamma \mathbf{g}^h \cdot \mathbf{n}\, d\Gamma = 0\,.$$

The existence and uniqueness of a solution $(\mathbf{g}^h, \gamma^h) \in \mathbf{Q}^h \times \mathbb{R}$ are guaranteed by the Brezzi theory for mixed finite-element methods (see [4] or [5]) and the inequalities

$$(\nabla_s \mathbf{k}^h, \nabla_s \mathbf{k}^h)_\Gamma + (\mathbf{k}^h, \mathbf{k}^h)_\Gamma \geq C\|\mathbf{k}^h\|_{1,\Gamma}^2 \quad \forall\, \mathbf{k}^h \in \mathbf{Q}^h \subset \mathbf{H}^1(\Gamma) \tag{4.93}$$

and

$$\sup_{\mathbf{0} \neq \mathbf{k}^h \in \mathbf{Q}^h} \frac{\gamma^h \int_\Gamma \mathbf{k}^h \cdot \mathbf{n}\, d\Gamma}{\|\mathbf{k}^h\|_{1,\Gamma}} \geq C|\gamma^h| \quad \forall\, \gamma^h \in \mathbb{R}\,. \tag{4.94}$$

The solution necessarily satisfies $\mathbf{g}^h \in G^h$. Thus the operator E^h is well defined.

With these definitions we see that (4.85)-(4.92) can be written in the form (3.1)-(3.3).

By results concerning the approximation of the Navier-Stokes equations with inhomogeneous boundary conditions (see [15]), we obtain

$$\|(T - T^h)f\|_X \to 0$$

as $h \to 0$, for all $f = (\boldsymbol{\zeta}, \eta, \boldsymbol{\kappa}) \in Y$. This is simply a restatement of (H13).

(H14) follows trivially from (H13), the fact that T is selfadjoint, and the choice $(T^*)^h = T^h$.

To verify (H15), we note that the nondiscretized version of (4.93)-(4.94) certainly also holds; i.e.,

$$(\nabla_s \mathbf{k}, \nabla_s \mathbf{k})_\Gamma + (\mathbf{k}, \mathbf{k})_\Gamma \geq C\|\mathbf{k}\|_{1,\Gamma}^2 \quad \forall\, \mathbf{k} \in \mathbf{H}^1(\Gamma)$$

and

$$\sup_{\mathbf{0} \neq \mathbf{k} \in \mathbf{H}^1(\Gamma)} \frac{\gamma \int_\Gamma \mathbf{k} \cdot \mathbf{n}\, d\Gamma}{\|\mathbf{k}\|_{1,\Gamma}} \geq C|\gamma| \quad \forall\, \gamma \in \mathbb{R}\,.$$

Using the Brezzi theory for the mixed finite-element method (see [4] or [5]), we obtain that

$$\|(E - E^h)\boldsymbol{\tau}\|_{1,\Gamma} \to 0 \quad \text{as } h \to 0\,,$$

which verifies (H15).

(H16) and (H17) follow from the fact that N and $\mathcal{F}$ are polynomials. Here we also use imbedding theorems and Cauchy inequalities.

We set $\hat{Z} = \mathbf{L}^{3/2}(\Omega) \times \{0\} \times \{\mathbf{0}\}$. For each $(\mathbf{v}, q, \mathbf{s}) \in X = \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma)$ and $(\mathbf{w}, r, \mathbf{k}) \in X = \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma)$, Sobolev imbedding theorems imply that

$$[N'(\mathbf{u}, p, \mathbf{t})]^* \cdot (\mathbf{v}, q, \mathbf{s}) = -\begin{pmatrix} -(\mathbf{u} \cdot \nabla)\mathbf{v} + \mathbf{v} \cdot (\nabla \mathbf{u})^T \\ 0 \\ \mathbf{0} \end{pmatrix} \in \hat{Z},$$

$$\big([N''(\mathbf{u}, p, \mathbf{t})]^* \cdot (\mathbf{v}, q, \mathbf{s})\big) \cdot (\mathbf{w}, r, \mathbf{k}) = -\begin{pmatrix} -(\mathbf{w} \cdot \nabla)\mathbf{v} + \mathbf{v} \cdot (\nabla \mathbf{w})^T \\ 0 \\ \mathbf{0} \end{pmatrix} \in \hat{Z},$$

and

$$\big(\mathcal{F}''(\mathbf{u}, p, \mathbf{t}) \cdot (\mathbf{v}, q, \mathbf{s})\big) \cdot (\mathbf{w}, r, \mathbf{k}) = \begin{pmatrix} \begin{pmatrix} 3(u_1 - u_{01})^2 w_1 v_1 \\ \vdots \\ 3(u_d - u_{0d})^2 w_d v_d \end{pmatrix} \\ 0 \\ \mathbf{0} \end{pmatrix} \in \hat{Z},$$

where d (= 2 or 3) is the space dimension. These relations verify (H18).

From the definition of the operator K we see that K maps $\mathbf{H}_n^1(\Gamma)$ into $\mathbf{L}^{3/2}(\Omega) \times \{0\} \times \mathbf{H}^1(\Gamma)$, i.e., K maps G into Z. Thus (H19) is verified.

Hence, we are now in a position to apply Theorem 3.5 to derive error estimates for the approximate solutions of the optimality system (4.54)-(4.56), (4.70)-(4.72) and (4.74). It should be noted that Lemma 3.4 implies that for almost all values of λ, the solutions of the optimality system are regular.

THEOREM 4.10. *Assume that Λ is a compact interval of $\mathbb{R}_+$ and that there exists a branch $\{(\lambda, \mathbf{u}(\lambda), p(\lambda), \mathbf{t}(\lambda), \mathbf{g}(\lambda), \boldsymbol{\nu}(\lambda), \phi(\lambda), \boldsymbol{\tau}(\lambda)) : \lambda \in \Lambda\}$ of regular solutions of the optimality system* (4.54)-(4.56), (4.70)-(4.72), *and* (4.74). *Assume that the finite-element spaces X^h and G^h satisfy the hypotheses* (4.79)-(4.84). *Then there exist a $\delta > 0$ and an $h_0 > 0$ such that for $h \le h_0$, the discrete optimality system* (4.85)-(4.92) *has a unique branch of solutions $\{(\lambda, \mathbf{u}^h(\lambda), p^h(\lambda), \mathbf{t}^h(\lambda), \mathbf{g}^h(\lambda), \boldsymbol{\nu}^h(\lambda), \phi^h(\lambda), \boldsymbol{\tau}^h(\lambda)) : \lambda \in \Lambda\}$ satisfying*

$$\Big(\|\mathbf{u}(\lambda) - \mathbf{u}^h(\lambda)\|_1 + \|p(\lambda) - p^h(\lambda)\|_0 + \|\mathbf{t}(\lambda) - \mathbf{t}^h(\lambda)\|_{-1/2,\Gamma} + \|\mathbf{g}(\lambda) - \mathbf{g}^h(\lambda)\|_{1,\Gamma} + \|\boldsymbol{\nu}(\lambda) - \boldsymbol{\nu}^h(\lambda)\|_1 + \|\phi(\lambda) - \phi^h(\lambda)\|_0 + \|\boldsymbol{\tau}(\lambda) - \boldsymbol{\tau}^h(\lambda)\|_{-1,2,\Gamma}\Big) < \delta \qquad \text{for all } \lambda \in \Lambda.$$

Moreover,

$$\lim_{h \to 0} \Big(\|\mathbf{u}(\lambda) - \mathbf{u}^h(\lambda)\|_1 + \|p(\lambda) - p^h(\lambda)\|_0 + \|\mathbf{t}(\lambda) - \mathbf{t}^h(\lambda)\|_{-1/2,\Gamma} + \|\mathbf{g}(\lambda) - \mathbf{g}^h(\lambda)\|_{1,\Gamma} + \|\boldsymbol{\nu}(\lambda) - \boldsymbol{\nu}^h(\lambda)\|_1 + \|\phi(\lambda) - \phi^h(\lambda)\|_0 + \|\boldsymbol{\tau}(\lambda) - \boldsymbol{\tau}^h(\lambda)\|_{-1,2,\Gamma}\Big) = 0$$

uniformly in $\lambda \in \Lambda$.

If, in addition, the solution satisfies $\big(\mathbf{u}(\lambda), p(\lambda), \mathbf{t}(\lambda), \mathbf{g}(\lambda), \boldsymbol{\nu}(\lambda), \phi(\lambda), \boldsymbol{\tau}(\lambda)\big) \in \mathbf{H}^{m+1}(\Omega) \times H^m(\Omega) \times \mathbf{H}^m(\Omega)|_\Gamma \times \mathbf{H}^{m+1}(\Omega)|_\Gamma \times \mathbf{H}^{m+1}(\Omega) \times H^m(\Omega) \times \mathbf{H}^m(\Omega)|_\Gamma$ for $\lambda \in \Lambda$, then there

exists a constant C, independent of h, such that

$$\Big(\|\mathbf{u}(\lambda)-\mathbf{u}^h(\lambda)\|_1+\|p(\lambda)-p^h(\lambda)\|_0+\|\mathbf{t}(\lambda)-\mathbf{t}^h(\lambda)\|_{-1/2,\Gamma}+\|\mathbf{g}(\lambda)-\mathbf{g}^h(\lambda)\|_{1,\Gamma}$$
$$+\|\boldsymbol{\nu}(\lambda)-\boldsymbol{\nu}^h(\lambda)\|_1+\|\phi(\lambda)-\phi^h(\lambda)\|_0+\|\boldsymbol{\tau}(\lambda)-\boldsymbol{\tau}^h(\lambda)\|_{-1,2,\Gamma}\Big)$$
$$\le Ch^{m-1/2}\Big(\|\mathbf{u}(\lambda)\|_{m+1}+\|p(\lambda)\|_m+\inf_{\mathbf{v}\in\mathbf{H}^m(\Omega),\mathbf{v}|_\Gamma=\mathbf{t}}\|\mathbf{v}\|_m$$
$$+\inf_{\mathbf{v}\in\mathbf{H}^{m+1}(\Omega),\mathbf{v}|_\Gamma=\mathbf{g}}\|\mathbf{v}\|_{m+1}+\|\boldsymbol{\nu}(\lambda)\|_{m+1}+\|\phi(\lambda)\|_m+\inf_{\mathbf{w}\in\mathbf{H}^m(\Omega),\mathbf{w}|_\Gamma=\boldsymbol{\tau}}\|\mathbf{w}\|_m\Big),$$

uniformly in $\lambda\in\Lambda$.

Proof. All results follow from Theorem 3.5. For the last result, we also use (3.25) and the estimates (see, e.g., [16] or [17])

$$\|(T^hT^{-1}-I)(\mathbf{u},p,\mathbf{t})\|_X\le Ch^m(\|\mathbf{u}\|_{m+1}+\|p\|_m+\inf_{\mathbf{v}\in\mathbf{H}^m(\Omega),\mathbf{v}|_\Gamma=\mathbf{t}}\|\mathbf{v}\|_m)$$
$$\text{for } \mathbf{u}\in\mathbf{H}^{m+1}(\Omega),\ p\in H^m(\Omega),\ \text{and } \mathbf{t}\in\mathbf{H}^m(\Omega)|_\Gamma,$$

$$\|\big((T^*)^h(T^*)^{-1}-I\big)(\boldsymbol{\nu},\phi,\boldsymbol{\tau})\|_{Y^*}=\|(T^hT^{-1}-I)(\boldsymbol{\nu},\phi,\boldsymbol{\tau})\|_X$$
$$\le Ch^m(\|\boldsymbol{\nu}\|_{m+1}+\|\phi\|_m+\inf_{\mathbf{w}\in\mathbf{H}^m(\Omega),\mathbf{w}|_\Gamma=\boldsymbol{\tau}}\|\mathbf{w}\|_m)$$
$$\text{for } \boldsymbol{\nu}\in\mathbf{H}^{m+1}(\Omega),\phi\in H^m(\Omega),\ \text{and } \boldsymbol{\tau}\in\mathbf{H}^m(\Omega)|_\Gamma,$$

and

$$\|(E^hE^{-1}-I)\mathbf{g}\|_{1,\Gamma}\le Ch^{m-1/2}\inf_{\mathbf{v}\in\mathbf{H}^{m+1}(\Omega),\mathbf{v}|_\Gamma=\mathbf{g}}\|\mathbf{v}\|_{m+1}\qquad\text{for } \mathbf{g}\in\mathbf{H}^{m+1}(\Omega)|_\Gamma.$$

In these estimates, the constant C is independent of h, $\mathbf{u}$, p, $\mathbf{t}$, $\mathbf{g}$, $\boldsymbol{\nu}$, ϕ, $\boldsymbol{\tau}$, and λ. □

Remark. If the control $\mathbf{g}\in\mathbf{H}^{m+3/2}(\Omega)|_\Gamma$, then the exponent of h in the error estimate of Theorem 4.10 can be increased from $(m-1/2)$ to m. □

REFERENCES

[1] R. ADAMS, *Sobolev Spaces*, Academic, New York, 1975.

[2] I. BABUSKA AND A. AZIZ, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. Aziz, ed., Academic Press, New York, 1972, pp. 3-359.

[3] H. BLUM AND R. RANNACHER, *On the boundary value problem of the biharmonic operator on domains with angular corners*, Math. Methods Appl. Sci., 2 (1980), pp. 556-581.

[4] F. BREZZI, *On the existence, uniqueness, and approximation of saddle-point problems arising from Lagrange multipliers*, RAIRO Modél. Math. Anal. Numér., 8 (1974), pp. 129-151.

[5] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.

[6] F. BREZZI, J. RAPPAZ, AND P. RAVIART, *Finite-dimensional approximation of nonlinear problems. Part* I: *Branches of nonsingular solutions*, Numer. Math., 36 (1980), pp. 1-25.

[7] J. CHAPMAN, Q. DU, M. GUNZBURGER, AND J. PETERSON, *Simplified Ginzburg-Landau models for superconductivity valid for high kappa and high fields*, Adv. Math. Sci. Appl., 5 (1995), pp. 193-218.

[8] P. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[9] P. CIARLET, *Les Equations de von Karman*, Springer, Berlin, 1980.

[10] M. CROUZEIX AND J. RAPPAZ, *On Numerical Approximation in Bifurcation Theory*, Masson, Paris, 1990.

[11] Q. DU, M. GUNZBURGER, AND J. PETERSON, *Analysis and approximation of the Ginzburg-Landau model of superconductivity*, SIAM Rev., 34 (1992), pp. 54-81.

[12] V. GEORGESCU, *On the unique continuation property for Schrödinger Hamiltonians*, Helv. Phys. Acta, 52 (1979), pp. 655-670.

[13] V. GIRAULT AND P. RAVIART, *Finite Element Methods for Navier-Stokes Equations*, Springer, Berlin, 1986.
[14] M. GUNZBURGER, *Finite Element Methods for Viscous Incompressible Flows: A Guide to Theory, Practice, and Algorithms*, Academic Press, Boston, MA, 1989.
[15] M. GUNZBURGER AND L. HOU, *Treating inhomogeneous essential boundary conditions in finite element methods and the calculation of boundary stresses*, SIAM J. Numer. Anal., 29 (1992), pp. 390–424.
[16] M. GUNZBURGER, L. HOU, AND T. SVOBODNY, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with Dirichlet controls*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 711–748.
[17] L. HOU AND T. SVOBODNY, *Optimization problems for the Navier-Stokes equations with regular boundary controls*, J. Math. Anal. Appl., 177 (1993), pp. 342–367.
[18] J. LIONS, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Gauthier-Villars, Paris, 1969.
[19] J. SAUT AND B. SCHEURER, *Unique continuation and uniqueness of the Cauchy problem for elliptic operators with unbounded coefficients*, in Applications of Nonlinear Partial Differential Equations, H. Brezis and J. Lions, eds., Longman, New York, 1982, pp. 260–275.
[20] R. TEMAM, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1979.
[21] V. TIKHOMIROV, *Fundamental Principles of the Theory of Extremal Problems*, Wiley, Chichester, 1982.
[22] M. TINKHAM, *Introduction to Superconductivity*, McGraw-Hill, New York, 1975.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 1044-1070, May 1996

013

TOPOLOGICAL ASPECTS OF UNIVERSAL ADAPTIVE STABILIZATION*

STUART TOWNLEY†

Abstract. In this paper we consider two problems in "non-identifier-based," universal adaptive control within the framework of Mårtensson [*Adaptive Stabilization*, Ph.D. thesis, Lund Institute of Technology, 1986]. In this framework, any linear system stabilizable by constant linear output feedback is adaptively stabilized by an adaptive piecewise-linear output feedback control law. The essential feature we exploit is that of a piecewise-linear output feedback which switches through a set of feedback matrices, with switching controlled by an output-driven differential equation. For each initial condition the state of the system converges to zero and the time-varying gain matrix converges to a "limit gain." In this setting we consider two related problems. The first concerns the sensitivity of closed-loop solutions under small perturbations of the initial data. The second concerns generic properties, with respect to the set of initial conditions, of stabilization by the limit gain. We adopt a topological approach, based on a decomposition of the dynamics of the resultant nonlinear, closed-loop system into a sequence of homeo/diffeomorphisms derived from the switching nature of the dynamics. Using this decomposition we show that the set of initial conditions for which solutions are stable under small perturbations and the limiting gain is stabilizing has full Lebesgue measure and dense interior. This latter result has been conjectured in the literature. The results are illustrated by examples of planar control systems where the sets of initial conditions yielding nonstabilizing limit gains are computed.

Key words. adaptive control systems, adaptive stabilization, sequential switching, feedback control, convergence analysis

AMS subject classifications. 93C40, 93D15, 93D21, 34D04

1. Introduction.

For a known universum, Σ, of linear systems

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) \in \mathbb{R}^n, \tag{1.1a}$$

$$y(t) = Cx(t), \tag{1.1b}$$

with $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$, a universal adaptive stabilizer is an output feedback control law

$$u(t) = f(y(t), k(t)) \tag{1.2}$$

with a "gain" parameter, $k(t)$, adapted according to the law

$$\dot{k}(t) = g(y(t), k(t), u(t)) \tag{1.3}$$

such that for each (unknown) element of Σ, each $x(0) = x_0 \in \mathbb{R}^n$, and $k(0) = k_0 \in \mathbb{R}$, the state $x(t)$ and gain parameter $k(t)$ given by (1.1)-(1.3) satisfy

$$x(t) \to 0 \text{ and } k(t) \to k(\infty, x_0) < \infty \text{ as } t \to \infty.$$

The pioneering work in this area, by Mårtensson (1986), Morse (1983), Nussbaum (1983), and Willems and Byrnes (1984), opened up the area of universal or "non-identifier-based" adaptive control.

For a detailed survey and comprehensive bibliography, see Ilchmann (1991). The control/adaptation laws are divided into three types according to whether f is smooth, piecewise-smooth, or piecewise-constant in k. Each type of controller is then a universal adaptive stabilizer for certain universa of systems. For example, the Willems-Byrnes *smooth control* law (Willems and Byrnes (1984)),

$$u(t) = k^2(t) \cos k(t)\, y(t), \quad \dot{k}(t) = |y(t)|^2,$$

*Received by the editors June 1, 1992; accepted for publication (in revised form) January 27, 1995.

†Department of Mathematics and Centre for Systems and Control Engineering, University of Exeter, Exeter EX4 4QE, UK.

is a universal adaptive stabilizer for the universum of all *minimum-phase* single-input, single-output systems with $CB \neq 0$.

A second example is the *piecewise-linear* (piecewise-constant in k) dense searching control law

$$u(t) = K_{S(k(t))}y(t), \quad \dot{k}(t) = \|y(t)\|^2 + \|u(t)\|^2,$$

with switching function

$$S(k) = i \text{ if } k \in [\tau_i, \tau_{i+1}),$$

and $K_j = M_i$ if $2j = q(q+1) + 2i$ for some q and i, where $\{M_1, M_2, \ldots\}$ is a dense subset of $\mathbb{R}^{m\times p}$ and $\tau_{i+1} = \tau_i^2$, $\tau_0 > 1$, due to Mårtensson. This controller is a universal adaptive stabilizer for the universum of all systems stabilizable by constant output feedback.

Despite the considerable interest shown in the problem of universal adaptive stabilization, with a few exceptions, the only guaranteed stability properties of the closed-loop system are convergence of the internal state to zero and convergence of the gain parameter. There is no sensitivity analysis for the solutions under small perturbations in the initial conditions nor, for example, guaranteed exponential decay of the solutions. More important, the issue of stabilization of (1.1) by the limiting feedback control

$$u(t) = f(y(t), k(\infty, x_0))$$

has received only minor attention.

There have been attempts to improve asymptotic behaviour by making minor modifications to the control law. In works by Ilchmann and Owens, exponential decay of each solution is guaranteed for the class of minimum-phase, relative-degree-one systems by adding exponential weighting to the adaptation law (1990) or using nondifferential gain adaptation (1991). In Miller and Davison (1991) the adaptation mechanisms are modified to improve transient responses and guarantee arbitrary fast decay of the output to a prespecified neighbourhood of the origin.

As we shall see in §2, if f is linear in y and piecewise-constant right continuous in k so that

$$f(y(t), k(t)) = K(t, x_0)y(t)$$

with $K(t, x_0) \in \mathbb{R}^{m\times p}$, piecewise-constant in t, then exponential decay of each solution is guaranteed. In Ilchmann (1994) universal feedback control laws of this type are constructed for the universum of minimum-phase, relative-degree-one systems, with the property that for each given x_0 the sequence of gain matrices yielding $\sigma(A + BK(\infty, x_0)C) \subset \mathbb{C}_-$ is dense. This result constitutes a preliminary step toward understanding the role of the switching sequence in the adaptation scheme. However, nothing is said either about stability of each solution under small perturbation of the initial data or about exponential stability of the limiting system matrix, $A + BK(\infty, x_0)C$, for a given gain sequence.

In this paper we show, for any universal adaptive stabilization scheme of this piecewise-constant type, that both of these properties are generic (that is, they hold in an open and dense set) for initial conditions $x_0 \in \mathbb{R}^n$. This generic stability property has been conjectured in the literature; see Ilchmann (1991). The significance of this generic property is that a stabilizing gain can be derived from a single closed-loop response (experiment) if the initial condition lies in a certain set of full Lebesgue measure. In addition, the generic properties imply regularity of the apparently erratic dynamics. Our approach is topological, replacing the piecewise-smooth

differential equation with a nonlinear discrete-time dynamical system, described in terms of a sequence of homeomorphisms/diffeomorphisms. Using this approach we can analyse the mapping $x_0 \mapsto k(\infty, x_0)$ from which we deduce our main results.

The paper is organized as follows. Section 2 contains preliminary material which establishes the framework for the remainder of the paper. We motivate the problem by considering the first-order case in detail where we characterize the limiting gain explicitly. The characterization appears in the form of inequalities on the partial sums of series generated by the switching and gain sequences. We close the section with an illustration of the difficulties arising in the second-order case.

In §3 we present our main results on stability of solutions and stabilization by the limiting gain matrix. We show that for all initial conditions in an open and dense set with full Lebesgue measure, the limiting gain is exponentially stabilizing and the solutions are insensitive to small perturbations. In §4 the results are applied to second-order, minimum-phase, relative-degree-one systems.

2. Preliminaries.

2.1. System formulation. We are interested in qualitative properties of closed-loop systems arising from universal adaptive stabilization of m-input, p-output linear systems

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0 \in \mathbb{R}^n, \tag{2.1}$$

$$y(t) = Cx(t), \tag{2.2}$$

by *piecewise-linear* controllers, that is, where f in (1.2) is piecewise-linear in the output y. Perhaps the most striking example of a universal adaptive stabilizer of this type is the "dense searching" controller in the context of Mårtensson's "existence of a stabilizing regulator is sufficient for universal adaptive stabilization."

Example 2.1 (Mårtensson (1986)). If (2.1), (2.2) is exponentially stabilizable by a proportional feedback $K \in \mathbb{R}^{m\times p}$, then a piecewise-linear universal adaptive stabilizer is given by

$$u(t) = K_{S(k(t))}y(t), \tag{2.3}$$

$$\dot{k}(t) = \|y(t)\|^2 + \|u(t)\|^2, \qquad k(0) = \tau_0, \tag{2.4}$$

where

$S(k) = i$ if $k \in [\tau_i, \tau_{i+1})$,
$(K_0, K_1, K_2, \ldots) = (M_0, M_0, M_1, M_0, M_1, M_2, \ldots)$,
$\tau_{i+1} = \tau_i^2$, $\tau_0 > 1$, and
$\{M_0, M_1, M_2, \ldots\}$ is a dense subset of $\mathbb{R}^{m\times p}$ (so that at least one of the M_i will stabilize).

In this case $f(y, k) = K_{S(k)}y$ and $g(y, k, u) = \|y\|^2 + \|u\|^2$.

A second important example of piecewise-linear universal adaptive stabilization arises when a piecewise-constant *Nussbaum function* is used in the Willems–Byrnes adaptive stabilization of minimum-phase systems.

Example 2.2 (Willems–Byrnes; Ilchmann and Logemann (1992)). Assume that in (2.1), (2.2) $m = p$ and the system is minimum phase so that

$$\text{Rank}\begin{bmatrix} sI - A & B \\ C & 0 \end{bmatrix} = n + m \text{ for all } s \in \mathbb{C}_+. \tag{2.5}$$

We have three cases.

(a) If $\det(CB) \neq 0$, then a piecewise-linear universal adaptive stabilizer is given by

$$u(t) = N(k(t))K_{S(k(t))}y(t), \tag{2.6}$$

$$\dot{k}(t) = \|y(t)\|^2, \qquad k(0) = \tau_0, \tag{2.7}$$

where

$(N(k), S(k)) = (\tau_i, i)$ if $k \in [\tau_i, \tau_{i+1})$;
$(K_0, K_1, K_2, \ldots) = (M_0, M_1, M_2, \ldots, M_L, M_0, M_1, M_2, \ldots)$, preserving the ordering;
$\{M_0, M_1, M_2, \ldots, M_L\}$ is a spectrum unmixing set for $m \times m$ invertible matrices (i.e., for each invertible $m \times m$ matrix M there exists $i \in \{0, 1, 2, \ldots, L\}$ such that $\sigma(M_i M) \subset \mathbb{C}_-$); and
$\{\tau_i\}$ is a sequence of positive real numbers with $\lim_{i\to\infty} \tau_i/\tau_{i+1} = 0$.

In this case $f(y, k) = N(k)K_{S(k)}y$ and $g(y, k, u) = \|y\|^2$.

(b) If $\sigma(CB) \subset \mathbb{C}_-$ or $\mathbb{C}_+$, then we can take $L = 1$, H_0 equal to the $m \times m$ identity matrix, $H_1 = -H_0$, and choose $\tau_i > 0$ so that

$$\inf_{l\to\infty} \frac{\sum_{i=0}^{l}(-1)^i \tau_i(\tau_{i+1} - \tau_i)}{\tau_l - \tau_0} = -\infty, \quad \sup_{l\to\infty} \frac{\sum_{i=0}^{l}(-1)^i \tau_i(\tau_{i+1} - \tau_i)}{\tau_l - \tau_0} = +\infty. \tag{2.8}$$

In this case

$$u(t) = (-1)^j \tau_j y(t) \text{ when } k(t) \in [\tau_j, \tau_{j+1}) \tag{2.9}$$

is a universal adaptive stabilizer.

(c) If it is known a priori that $\sigma(CB) \subset \mathbb{C}_+$, then a universal adaptive stabilizer is given by (2.7) with the simple high-gain feedback

$$u(t) = -\tau_j y(t), \qquad k(t) \in [\tau_j, \tau_{j+1}). \tag{2.10}$$

In this case $\{\tau_i\}$ is any increasing sequence with $\lim_{i\to\infty} \tau_i = \infty$.

Remark 2.3. We will not recall the proofs of universal adaptive stabilization for the systems in Examples 2.1 and 2.2. However, we will have cause to borrow some ideas from these proofs. For those readers not familiar with universal adaptive control, a flavour of this area can be obtained by looking at Proposition 3.20 and Lemma 4.2.

Examples 2.1 and 2.2 have common properties. The dynamics are switched according to the nonlinear adaptation of k, and in between switches in k, the dynamics of the x component are linear. Indeed both examples fit into a single framework of a linear system (2.1), (2.2) and a piecewise-linear (in y) nonlinear adaptive feedback control

$$u(t) = K_i y(t), \tag{2.11}$$

with adaptation

$$\dot{k}(t) = \|C_i x(t)\|^2, \qquad k(0) = \tau_0 \in \mathbb{R}, \tag{2.12}$$

where the right-hand sides of (2.11) and (2.12) are determined by $k(t) \in [\tau_i, \tau_{i+1})$ with $\{\tau_i\}$ an increasing sequence with $\lim_{i\to\infty} \tau_i = \infty$. The particular features of importance are as follows.

- The feedback gain matrix $K(t, x_0)$, defined by

$$K(t, x_0) = K_i, \quad k(t) \in [\tau_i, \tau_{i+1}),$$

is switched according to a nonlinear differential equation-based adaptation law. Note the dependence on x_0.

- $\mathcal{K} = \{K_i\}_{i=1}^{\infty}$ is an ordered set of potential gain matrices. In Example 2.1, $\mathcal{K}$ is determined by recursive searching through a dense subset of $\mathbb{R}^{m\times p}$, whilst in Example 2.2, $\mathcal{K}$ is determined by cycling through the spectrum unmixing set and multiplying by the Nussbaum gain.
- The right-hand side of (2.12) is allowed to depend on k, as is the case in Example 2.1, where

$$C_i = \begin{bmatrix} C \\ K_i C \end{bmatrix} \text{ so that } \|y\|^2 + \|u\|^2 = \|C_i y\|^2.$$

In Example 2.2, $C_i = C$ for all i.

Remark 2.4. Whilst we will state most of the results for the general class of closed-loop systems given by (2.1), (2.2), (2.11), and (2.12), in §4 we will focus on the class of systems given in Example 2.2 in the single-input, single-output case.

A simple consequence of the piecewise-smoothness of (2.1), (2.2), (2.11), and (2.12) is the existence and uniqueness of solutions on maximal intervals of existence. This is summarised in the following lemma.

LEMMA 2.5 (existence and uniqueness of solutions on maximal intervals of existence). *For each $x(0) = x_0 \in \mathbb{R}^n$ and $k(0) = \tau_0 \in \mathbb{R}$ there exists $\omega > 0$ and a unique continuous and piecewise-differentiable solution $(x(t), k(t))$ on the maximal interval $[0, \omega)$, satisfying* (2.1), (2.2), (2.11), *and* (2.12) *almost everywhere.*

Proof. If $k(t)$ is bounded, then there exists $t_j < \infty$ such that $\tau_j \le k(t) < \tau_{j+1}$ for all $t \ge t_j$,

$$\dot{x}(t) = (A + BK_jC)x(t) \text{ for all } t \ge t_j$$

and $\omega = \infty$. If $k(t)$ is unbounded, then by monotonicity of $k(\cdot)$ there exists a sequence of times $t_0 = 0, t_1, t_2, \ldots$ such that $k(t_j) = \tau_j$. Since

$$\dot{x}(t) = (A + BK_jC)x(t) \text{ for all } t \in [t_j, t_{j+1}),$$

$x(\cdot)$ is smooth on $[t_j, t_{j+1})$. If $t_j \to t^* < \infty$, then $\omega = t^* < \infty$; otherwise $\omega = \infty$. □

Remark 2.6. We consider (2.1), (2.2), (2.11), and (2.12) as a class of systems containing the essential features of nonlinear systems arising in universal adaptive stabilization. Recall, in particular, that this means that for each $x(0)$ and $k(0) = \tau_0$, $x(\cdot)$ and $k(\cdot)$ exist for all $t \ge 0$ and

$$\lim_{t\to\infty} x(t) = 0 \text{ and } \lim_{t\to\infty} k(t) = k(\infty, x(0)) < \infty. \tag{2.13}$$

We now assume throughout that (2.11) and (2.12) are a universal adaptive stabilizer for (2.1), (2.2) so that (2.13) holds. We denote by

$$\Phi_t : (x(0), k(0)) \mapsto (x(t), k(t))$$

the flow of the nonlinear system (2.1), (2.2), (2.11), and (2.12), which is defined for all $t \ge 0$.

2.2. Problem motivation. In the context of adaptive stabilization the notion of stability is relaxed so that a priori the only guaranteed properties of the closed-loop system are those specified by (2.13). The aim of this paper is to characterise information obtained as a necessary consequence of universal adaptive stabilization. In particular, we are interested in whether the limiting gain, $K(\infty, x_0)$, given by

$$K(\infty, x_0) = K_i \text{ if } k(\infty, x_0) \le \tau_{i+1} \text{ and } k(t, x_0) \ge \tau_i \text{ for some } t \ge 0,$$

results in an exponentially stable limit system

$$\dot{x}(t) = (A + BK(\infty, x_0)C)x(t). \tag{2.14}$$

Anticipating our results, we are also interested in whether exponential stability of the limit system is a generic (open and dense) property in the parameter space of initial conditions.

To focus ideas, consider the simplest case of a first-order controllable system:

$$\dot{y}(t) = ay(t) + bu(t), \quad \text{where } y = x \text{ and } b \neq 0. \tag{2.15}$$

This system can be stabilized by (2.7) and (2.9). In this simple case the limiting gain can be characterised explicitly and is stabilizing unless $y = 0$.

PROPOSITION 2.7. *If $y_0 \neq 0$, then $u(t) = K(\infty, y_0)y(t)$ is a constant linear stabilizing feedback control for* (2.15). *Moreover,*

$$K(\infty, y_0) = (-1)^{j^*}\tau_{j^*},$$

where j^* is determined by

$$j^* = \inf\left\{ j \mid y_0^2 + 2\sum_{i=0}^{j}(a + (-1)^i b\tau_i)(\tau_{i+1} - \tau_i) \leq 0 \right\}.$$

If $y_0 = 0$, then $K(\infty, 0) = \tau_0$.

Proof. Let y_0 be fixed, and suppose $k(\infty, y_0) \in (\tau_{j^*}, \tau_{j^*+1}]$. We have

$$\frac{d}{dt}(y(t)^2) = 2(a + (-1)^i b\tau_i)\frac{dk}{dt} \quad \text{provided that} \quad k(t) \in [\tau_i, \tau_{i+1}). \tag{2.16}$$

Let t^* be such that $k(t^*) = \tau_{j^*}$. Integrating (2.16) from 0 to t^* we have

$$0 < y^2(t^*) = y_0^2 + 2\sum_{i=0}^{j^*-1}(a + (-1)^i b\tau_i)(\tau_{i+1} - \tau_i). \tag{2.17}$$

Integrating (2.16) from $t = t^*$ to $t = \infty$, we have

$$-y^2(t^*) = 2(a + (-1)^{j^*} b\tau_{j^*})(k(\infty, y_0) - \tau_{j^*}). \tag{2.18}$$

Equation (2.17) implies that $a + (-1)^{j^*} b\tau_{j^*} < 0$. It follows that

$$(a + (-1)^{j^*} b\tau_{j^*})(\tau_{j^*+1} - \tau_{j^*}) \leq -y^2(t^*). \tag{2.19}$$

Adding (2.17) and (2.19) gives

$$y_0^2 + 2\sum_{i=0}^{j^*}(a + (-1)^i b\tau_i(t_{i+1} - \tau_i)) \leq 0. \qquad \square$$

Even in this simplest situation, if τ_0 is not stabilizing and $y_0 = 0$, then $K(\infty, y_0)$ is not stabilizing.

DEFINITION 2.8. *We denote by $\mathcal{U} \subset \mathbb{R}^n$ the set of initial conditions for which the limit system* (2.14) *is not exponentially stable, that is,*

$$\mathcal{U} = \{x_0 \in \mathbb{R}^n \mid \sigma(A + BK(\infty, x_0)C) \cap \overline{\mathbb{C}_+} \neq \emptyset\}. \tag{2.20}$$

The complement of $\mathcal{U}$ is partitioned into two parts, $\mathcal{S}$ and $\mathcal{T}$:

$$\mathcal{S} = \{x_0 \in \mathbb{R}^n \mid x_0 \mapsto K(\infty, x_0) \text{ is continuous}\}, \tag{2.21}$$

$$\mathcal{T} = \{x_0 \in \mathbb{R}^n \mid x_0 \mapsto K(\infty, x_0) \text{ is discontinuous}\}. \tag{2.22}$$

To see that in general $\mathcal{U}$ is not empty nor equal to $\{0\}$, consider the second-order, minimum-phase, relative-degree-one system

$$\dot{x}(t) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} x(t) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t), \quad y(t) = [1 \;\; 0]\, x(t),$$

which is stabilized by (2.7) and (2.9). Observe that if $x_0 \in \{0\} \times \mathbb{R}$, then $y(t) = 0, \dot{k}(t) = 0$, there are no switches in gain, $K(\infty, x_0) = \tau_0$, and yet $x(t) \to 0$. If τ_0 is not stabilizing, then $K(\infty, x_0)$ is not stabilizing and

$$\{0\} \times \mathbb{R} \subset \mathcal{U}.$$

Of course this observation is hardly surprising since $\operatorname{Ker}[1 \;\; 0] = \{0\} \times \mathbb{R}$ is a stable A-invariant subspace. A more striking example, which we will consider in detail in §4, is obtained if (2.7), (2.10) is applied to a relative-degree-one, minimum-phase, *controllable* and *observable* system.

Example 2.9. Consider the following minimum-phase, relative-degree-one system:

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ 6 & 1 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t), \quad y(t) = [1 \;\; 1]\, x(t), \tag{2.23}$$

expressed in control canonical form.[1] If we assume that knowledge of the sign of the high-frequency gain is available information, then we can use (2.10) as a universal adaptive stabilizer. Let $\tau_0 = 0$ and $\tau_1 = 3$. If

$$x_0 \in \left\{ \delta \begin{pmatrix} 1 \\ -2 \end{pmatrix} \mid \delta^2 \leq 2\sqrt{3} \right\},$$

then $x(t) = e^{-2t} x_0$, $k(\infty, x_0) \leq 3$, and the first switch is never reached. Hence

$$A + BK(\infty, x_0)C = A + \tau_0 BC = A, \quad \sigma(A + BK(\infty, x_0)C) = \sigma(A) \not\subset \mathbb{C}_-,$$

and

$$\left\{ \delta \begin{pmatrix} 1 \\ -2 \end{pmatrix} \mid \delta^2 \leq 2\sqrt{3} \right\} \subset \mathcal{U}. \tag{2.24}$$

In §4 we will amplify this example, using the control canonical form structure to show that the structure of this bad set $\mathcal{U}$ can be very complicated. The main result of §3 is that despite the possible complexity of $\mathcal{U}$, in measure theoretic and topological terms this bad set is very small.

MAIN RESULT. *$\mathcal{U}$ is a nowhere dense and Lebesgue-measure-zero set. Moreover, $\mathcal{S}$ is open and dense and has full Lebesgue measure.*

Our approach is to exploit the sequential nature of the closed-loop dynamics to essentially replace the flow Φ_t with a sequence of diffeomorphisms. This section is concluded with notation related to this sequence of diffeomorphisms.

[1] The control canonical form is not necessary at this point but is useful for the analysis in §4.

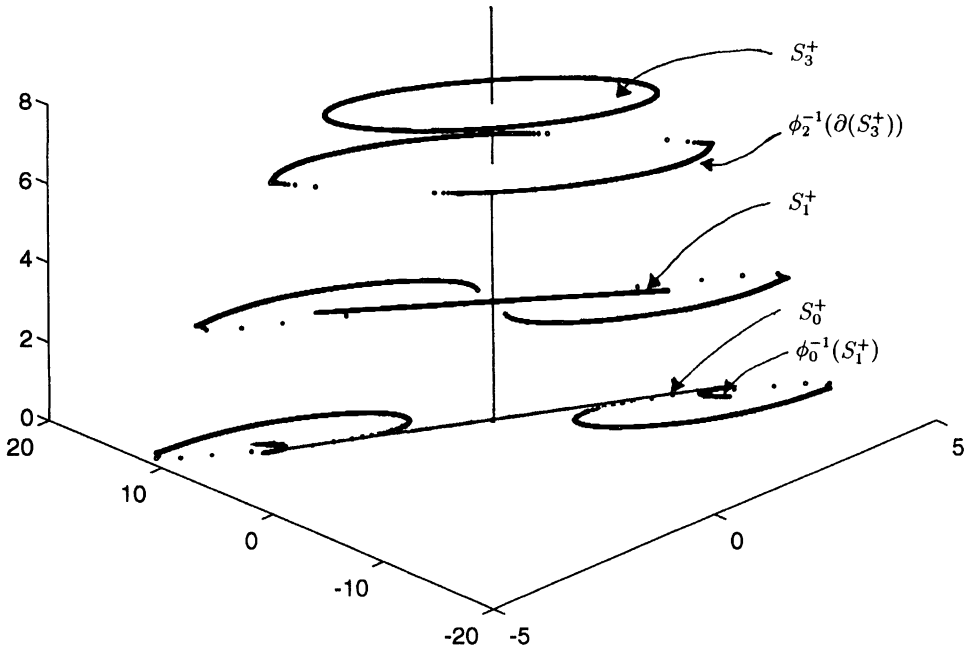


FIG. 1. S_3^+ *mapped to* H_0, *together with* S_0^+, S_1^+, *and* $\phi_0^{-1}(S_1^+)$.

2.3. Notation. Within the natural framework of $\mathbb{R}^{n+1}$ for the dynamics of (2.1), (2.2), (2.11), and (2.12) we adopt the following notation. For each $i \in \mathbb{N}$

(a) $H_i \subset \mathbb{R}^{n+1}$ is the hyperplane

$$H_i = \{(x, k) \mid k = \tau_i\}.$$

(b) $A_i = (A + BK_iC)$ with stable subspace V_i^s and unstable subspace V_i^u.

(c) $W_i(t) = \int_0^t e^{A_i^T \tau} C_i^T C_i e^{A_i \tau} d\tau$ is the observability Gramian at time t for the pair (A_i, C_i).

(d) It is clear from Example 2.9 that certain subsets of $V_i^s \times \{\tau_i\}$ and $V_i^u \times \{\tau_{i+1}\}$ of H_i and H_{i+1}, respectively, defined by the switching condition, will play important roles. These subsets, denoted by S_i^+ and S_{i+1}^-, are defined by

$$S_i^+ = \{(x, \tau_i) \mid \langle x, W_i(\infty)x\rangle \leq \tau_{i+1} - \tau_i\} \subset H_i,$$
$$S_{i+1}^- = \{(x, \tau_{i+1}) \mid \langle x, W_i(-\infty)x\rangle \leq \tau_{i+1} - \tau_i\} \subset H_{i+1}.$$

The subscripts in S_i^+ and S_{i+1}^- refer to the hyperplane in which they are contained, whilst the superscripts refer to the direction of flow defining them.

(e) The complements of S_i^+, S_{i+1}^- are denoted by R_i^+, R_{i+1}^-.

Finally,

(f) $J = \{i \in \mathbb{N}_0 \mid \sigma(A + BK_iC) \cap \overline{\mathbb{C}_+} \neq \emptyset\}$ is the indexing set for nonstabilizing gains.

These are illustrated in Fig. 1.

Remark 2.10. i) Between the hyperplanes H_i and H_{i+1} the dynamics of the state component $x(t)$ evolve according to the linear time invariant system

$$\dot{x}(t) = A_i x(t).$$

ii) S_i^+ is a *closed* subset of the affine space $V_i^s \times \{\tau_i\}$. In Example 2.9

$$S_0^+ = \left\{ \delta \begin{pmatrix} 1 \\ -2 \end{pmatrix} | \delta^2 \leq 2\sqrt{3} \right\} \times \{0\}.$$

S_{i+1}^- is a *closed* subset of the affine space $V_i^u \times \{\tau_{i+1}\}$. S_i^+ and S_{i+1}^- play fundamental roles in our approach. If $(x(t), k(t)) \in S_i^+$ for some $t \geq 0$, then H_{i+1} is not reached, whereas S_{i+1}^- cannot be reached from H_i.

iii) The sets $R_i^+ \subset H_i$ and $R_{i+1}^- \subset H_{i+1}$ are open. They are, respectively, the points in H_i which reach H_{i+1} and the points in H_{i+1} which are reached from H_i under the flow Φ_t. Note that they are naturally homeomorphic. In the case when (A_i, C_i) is an observable pair, they are both punctured n-dimensional space. If $\operatorname{Ker} C_i$ contains a stable A_i-invariant subspace V, then both R_i^+ and R_{i+1}^- are homeomorphic to $\mathbb{R}^n \setminus V$.

3. Generic stability properties of universal adaptive stabilization. In this section we present and prove the main results. We assume throughout that (2.11) and (2.12) result in a universal adaptive stabilizer for (2.1) and (2.2) as defined by (2.13). We show that in the partition

$$\mathbb{R}^n = \mathcal{S} \cup \mathcal{T} \cup \mathcal{U} \tag{3.1}$$

of the set of initial conditions introduced in Definition 2.8,

- $\mathcal{U}$ is nowhere dense and has zero Lebesgue measure,
- $\mathcal{T}$ is nowhere dense and has zero Lebesgue measure,
- S is open and dense and has full Lebesgue measure.

We characterise $\mathcal{S}$, $\mathcal{T}$, and $\mathcal{U}$ explicitly via smooth functions which arise from a decomposition of the flow Φ_t into a sequence of *homeomorphisms* $\{\phi_i : R_i^+ \to R_{i+1}^-\}$.

The first thing to note is the importance of the sets S_i. Indeed, for each $x_0 \in \mathbb{R}^n$, boundedness of $k(t, x_0)$ implies that there are at most finitely many switches in $K(t, x_0)$. It follows that there must exist $t \geq 0$ such that $(x(t), k(t)) \in S_q$ for some $q \geq 0$ and that this, moreover, determines $K(\infty, x_0)$.

LEMMA 3.1. *$K(\infty, x_0) = K_q$ if and only if $(x(t), k(t)) \in S_q^+$ for some $t < \infty$.*

Remark 3.2. As an immediate consequence of Lemma 3.1 and the definition of S_q^+ it follows that for each $x_0 \in \mathbb{R}^n$ there exists $M, \lambda > 0$ (depending on x_0) such that

$$\|x(t)\| \leq Me^{-\lambda t}.$$

(See also Ilchmann (1994).)

If $K(\infty, x_0) = K_q$, then there exists a sequence of times $0 = t_0 < t_1 < \cdots < t_q < \infty$ such that $(x(t_j), k(t_j)) \in R_j^+$, $j = 0, 1, \ldots, q-1$, and $(x(t_q), k(t_q)) \in S_q^+$.

This simple observation is formalised as a decomposition of Φ_t into a sequence of maps ϕ_i.

DEFINITION 3.3. *For each $i \in \mathbb{N}$, $\phi_i : R_i^+ \to R_{i+1}^-$ is defined implicitly as follows: Let $(x, \tau_i) \in R_i^+$, then there exists $t_i(x)$ such that*

$$\tau_{i+1} - \tau_i = \langle x, W_i(t_i(x))x \rangle.$$

We set

$$\phi_i(x, \tau_i) = (e^{A_i(t_i(x))}x, \tau_{i+1}).$$

Remark 3.4. The explicit reference to τ_i and τ_{i+1} is not really important but simply serves to emphasize the domain ($\subset H_i$) and range ($\subset H_{i+1}$) of ϕ_i. Indeed we often think of ϕ_i as a map defined on a subset of $\mathbb{R}^n$ rather than on R_i^+ and drop the explicit reference to τ_i and τ_{i+1}.

Example 3.5. (a) For the first-order system (2.15) and (2.9) with $a = b = 1$,

$$R_i^+ = \{x \mid x^2 + 2(\tau_{i+1} + \tau_i)((-1)^i\tau_i + 1) > 0\} \times \{0\}$$

and

$$\phi_i(x) = \sqrt{x^2 + 2(\tau_{i+1} - \tau_i)((-1)^i\tau_i + 1)}$$

with

$$\phi_i^{-1}(y) = \sqrt{y^2 - 2(\tau_{i+1} - \tau_i)((-1)^i\tau_i + 1)}.$$

(b) Figures 1 and 2(a)-(c) show the typical qualitative effect of ϕ_i^{-1} for second-order systems. Figure 1 shows an ellipse mapped down through the H_i under ϕ_i^{-1} for $i = 2, 1, 0$. Figure 2(a) shows the image of the unit circle under ϕ_i^{-1} when A_i has two positive real eigenvalues, and Fig. 2(b) shows the image when A_i has two negative real eigenvalues. Figure 2(c) shows the hyperbolic case when A_i has one negative and one positive real eigenvalue. Note that if A_i has hyperbolic eigenvalues and $\tau_{i+1} - \tau_i$ is large, then the image of the unit circle is split into two connected components and the image of the unit circle has cusp points (at which, as we will see, ϕ_i has a singular derivative).

The development of the topological approach pivots on smoothness properties of the ϕ_i. A word of caution: for each x, and $i \in \mathbb{N}$ there is a different time $t_i(x)$ required to reach H_{i+1} from H_i, and so ϕ_i is truly nonlinear and not simply the linear flow $x \mapsto e^{A_i t}x$.

PROPOSITION 3.6. *For each $i \in \mathbb{N}$, $\phi_i : R_i^+ \to R_{i+1}^-$ is a homeomorphism.*

Proof. If $(x, \tau_i) \in R_i^+$, then there exists $t(x) < \infty$ such that $\langle x, W_i(t(x))x\rangle = \tau_{i+1} - \tau_i$. Let $\epsilon > 0$ be arbitrary. For each $\hat{x} \in \mathbb{R}^n$ and $\delta > 0$

$$\sqrt{\langle (x + \delta\hat{x}), W_i(t(x) + \epsilon)(x + \delta\hat{x})\rangle} \geq \sqrt{\langle x, W_i(t(x) + \epsilon)x\rangle} - \delta\sqrt{\langle \hat{x}, W_i(t(x) + \epsilon)\hat{x}\rangle}$$

and

$$\sqrt{\langle (x + \delta\hat{x}), W_i(t(x) - \epsilon)(x + \delta\hat{x})\rangle} \leq \sqrt{\langle x, W_i(t(x) - \epsilon)x\rangle} + \delta\sqrt{\langle \hat{x}, W_i(t(x) - \epsilon)\hat{x}\rangle},$$

where $\sqrt{\langle x, W_i(t(x) - \epsilon)x\rangle} < \tau_{i+1} - \tau_i$ and $\sqrt{\langle x, W_i(t(x) + \epsilon)x\rangle} > \tau_{i+1} - \tau_i$. If $\|\hat{x}\| = 1$ and δ is sufficiently small, then

$$\langle (x + \delta\hat{x}), W_i(t(x) + \epsilon)(x + \delta\hat{x})\rangle > \tau_{i+1} - \tau_i$$

and

$$\langle (x + \delta\hat{x}), W_i(t(x) - \epsilon)(x + \delta\hat{x})\rangle < \tau_{i+1} - \tau_i.$$

It follows by continuity of $W_i(\cdot)$ (in the operator topology) that $t(y) \in (t(x) - \epsilon, t(x) + \epsilon)$ for each $y \in B_\delta(x)$. Hence $t(\cdot)$ is continuous at x, and therefore ϕ_i (as a composition and product of continuous functions) is continuous at (x, τ_i). By definition of R_{i+1}^- it follows that ϕ_i is onto. Injectivity of ϕ_i follows from analyticity of $t \mapsto C_i e^{A_i t}x$ together with the fact that ϕ_i is defined only on R_i^+ (those points in H_i reaching H_{i+1}). Hence $\phi_i^{-1} : R_{i+1}^- \to R_i^+$ exists. (In R_{i+1}^- we are interested only in points coming from below H_{i+1}. Those points on H_{i+1} which are on flows of Φ_t lying entirely in H_{i+1} are not in R_{i+1}^-.) A similar argument to the one above, reversing time, shows that ϕ_i^{-1} is continuous and therefore ϕ_i is a homeomorphism. □

Remark 3.7. If $K(\infty, x_0) = K_q$, then a trajectory flows from $(x_0, k_0) \in H_0$ (under Φ_t) via the homeomorphisms ϕ_i through each of the hyperplanes $H_i, i = 1, 2, \ldots, q - 1$ and finally hits S_q^+.

For each $i \in \mathbb{N}$ let $\mathcal{S}_i \subset \mathbb{R}^n$ be defined by

$$\mathcal{S}_i = \{x_0 \in \mathbb{R}^n \mid \Phi_t(x_0, \tau_0) \in S_i^+ \text{ for some } t < \infty\}.$$

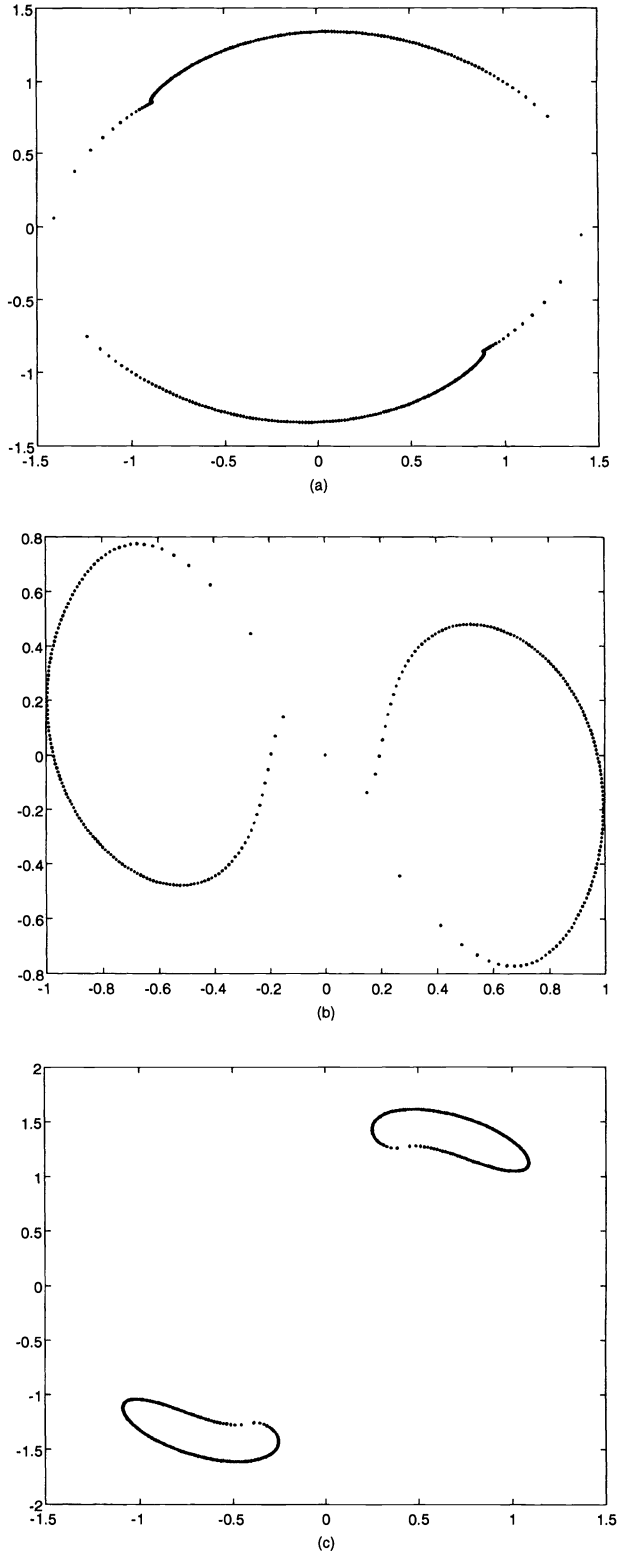


FIG. 2. *The image of the unit circle under* ϕ_i *when* A_i *has* (a) *unstable real,* (b) *stable real, and* (c) *hyperbolic eigenvalues.*

Using the decomposition of the flow Φ_t into the sequence of homeomorphisms ϕ_i, $\mathcal{S}_i$ is defined equivalently by

$$\mathcal{S}_i = \{x_0 \in \mathbb{R}^n \mid \phi_{i-1}(\phi_{i-2}(\ldots(\phi_0(x_0, \tau_0))\ldots)) \in S_i^+\}.$$

LEMMA 3.8. (a) *For each* $i \notin J$

$$\operatorname{int}(\mathcal{S}_i) = \{x_0 \in \mathbb{R}^n \mid \Phi_t(x_0, \tau_0) \in \operatorname{int}(S_i^+) \text{ for some } t \geq 0\}$$

and

(b) $\mathcal{S} = \cup_{i \notin J} (\operatorname{int}(\mathcal{S}_i))$.

Proof. (a) This follows immediately from the homeomorphic structure of the ϕ_i.

(b) It is clear that $\bigcup (\operatorname{int}(\mathcal{S}_i)) \subset \mathcal{S}$. To show the opposite inclusion we must verify that if $x_0 \notin \bigcup (\mathcal{S}_i)$, then $x \mapsto K(\infty, x)$ is discontinuous at x_0, so $x_0 \notin \mathcal{S}$. Now $K(\infty, x_0) = K_j$ for some j, and since $x_0 \notin \operatorname{int}(\mathcal{S}_j)$ there must exist x_j with $(x_j, \tau_j) \in \partial S_j^+$ such that $x(t) = x_j$ for some t. (Here ∂S_j^+ is the boundary of S_j^+.) Since S_j^- cannot be reached from H_{j-1} it is clear that $(x_j, \tau_j) \notin S_j^-$. Hence, we can use the continuity of the ϕ_i^{-1} and the fact that S_j^- is closed to find, given $\epsilon > 0$, a small enough neighbourhood, $\mathcal{N}_j$, of (x_j, τ_j) (in H_j) such that $\|\phi_0^{-1}(\ldots(\phi_{j-1}^{-1}(x, \tau_j))) - (x_0, \tau_0)\| < \epsilon$ for all $(x, \tau_j) \in \mathcal{N}_j$. Finally, using the closedness of S_j^+ we can choose $z_j \in \mathcal{N}_j$ so that with x defined by $(x, \tau_0) = \phi_0^{-1}(\ldots(\phi_{j-1}^{-1}(z_j, \tau_j)))$, we have $\|(x, \tau_0) - (x_0, \tau_0)\| < \epsilon$ and $K(\infty, x) \neq K_j$. □

COROLLARY 3.9. $\mathcal{S}$ *is open.*

$\mathcal{S}$ has very nice properties. By definition, for each $x_0 \in \mathcal{S}$, the limiting gain $K(\infty, x_0)$ is stabilizing. Moreover, as a direct consequence of the homeomorphic structure of the ϕ_i, the continuity of each $x \mapsto t_i(x)$, and the openness of $\mathcal{S}$, we can prove stability of solutions on $[0, \infty)$ under small perturbations of initial data x_0 in $\mathcal{S}$.

THEOREM 3.10. *If* $x_0 \in \mathcal{S}$, *then given* $\epsilon > 0$ *there exists* $\delta > 0$ *so that if* $\|x - x_0\| < \delta$, *then* $\|\Phi_t(x, \tau_0) - \Phi_t(x_0, \tau_0)\| < \epsilon$ *for all* $t \geq 0$.

Remark 3.11. Whilst this result is intuitively clear, the reappearance of time means that a detailed proof is required.

Proof. It is sufficient to prove the result in the case when $1 \notin J$, $\sigma(A_1) \subset \mathbb{C}_-$, and $x_0 \in \operatorname{int}(\mathcal{S}_1)$. The general case follows immediately because only a finite number of ϕ_i are involved.

Let $x_0 \in (\operatorname{int} \mathcal{S}_1)$ and $\epsilon > 0$ be given. Choose $\delta > 0$ small enough so that $B_\delta(\phi_0(x_0)) \subset \operatorname{int} S_1^+$. Here $B_\delta(x) = \{z \in \mathbb{R}^n \mid \|z - x\| < \delta\}$. Now choose $\eta > 0$ so that $\phi_0(B_\eta(x_0)) \subset B_\delta(\phi_0(x_0))$. Given $\mu > 0$, reduce $\eta > 0$ if required so that

$$\left| \max_{x \in B_\eta(x_0)} t(x) - \min_{x \in B_\eta(x_0)} t(x) \right| < \mu.$$

If $x \in B_\eta(x_0)$, then $t_{\min} \leq t_0(x), t_0(x_0) \leq t_{\max}$. Without loss of generality, suppose that $t_0(x_0) \leq t_0(x)$. We now check that $\|x(t, x) - x(t, x_0)\| < \epsilon$ on each interval $[0, t_0(x_0)]$, $(t_0(x_0), t_0(x))$, and $[t_0(x), \infty)$. If $t \geq t_0(x)$, then

$$\|x(t, x) - x(t, x_0)\| = \|e^{A_1(t - t_0(x))}\phi_0(x) - e^{A_1(t - t_0(x_0))}\phi_0(x_0)\|$$

$$\leq \|e^{A_1(t - t_0(x))}(\phi_0(x) - \phi_0(x_0))\|$$

$$+ \|(e^{A_1(t - t_0(x))} - e^{A_1(t - t_0(x_0))})\phi_0(x_0)\| < \epsilon$$

if μ and hence η are small enough. (Here $\phi_0(\cdot)$ means only the x-component.)

If $t \leq t_0(x_0)$ and μ (hence η) is small enough, then

$$\|x(t,x) - x(t,x_0)\| = \|e^{A_0 t}(x - x_0)\| < \epsilon.$$

Finally if $t \in (t_0(x_0), t_0(x))$ and μ is small enough, then

$$\|x(t,x) - x(t,x_0)\| = \|e^{A_0 t}x - e^{A_1(t-t_0(x_0))}e^{A_0 t_0(x_0)}x_0\|$$

$$\leq \|e^{A_0 t}x - e^{A_1(t-t_0(x_0))}e^{A_0(t_0(x_0))}x\| + \|e^{A_1(t-t_0(x_0))}\|\|e^{A_0(t_0(x_0))}\|\|x - x_0\|$$

$$\leq \|[e^{A_0(t-t_0(x_0))} - e^{A_1(t-t_0(x_0))}]e^{A_0 t_0(x_0)}x\| + \frac{\epsilon}{2} < \epsilon.$$

Similar estimates hold for $|k(t,x) - k(t,x_0)|$ (and hence $\|\Phi_t(x,\tau_0) - \Phi_t(x_0,\tau_0)\|$) using continuity of $t_0(\cdot)$ and $W_0(\cdot)$. □

Remark 3.12. Simulations of closed-loop systems derived from universal adaptive control algorithms can exhibit erratic transient behaviour. Theorem 3.10 gives *robustness* of stability of solutions with respect to perturbations of initial data for solutions starting in the open set $\mathcal{S}$. This shows that there is inherent regularity underlying the erratic behaviour, even in the case of the Mårtensson dense search controller of Example 2.1.

Implicit in the analysis above is an algorithm by which we can compute $\mathcal{U}$. Whilst complicated by the restricted domains and codomains of the ϕ_i we have

$$\mathcal{U} = \bigcup_{j \in J} \mathcal{U}_j,$$

where for each $j \in J$, $\mathcal{U}_j$ is defined recursively by

$$\begin{cases} \mathcal{U}_j = \phi_0^{-1}(R_1^+ \cap Z_1), \\ Z_k = \phi_k^{-1}(R_{k+1}^+ \cap Z_{k+1}), \qquad k = 1, \ldots, j-1, \\ Z_j = S_j. \end{cases}$$

The partition of $\mathbb{R}^n$ is completed by setting

$$\mathcal{T} = \mathbb{R}^n \setminus (\mathcal{S} \cup \mathcal{U}) = \bigcup_{j \notin J} \mathcal{T}_j,$$

where $\mathcal{T}_j = \phi_0^{-1}(\ldots(\phi_{j-1}^{-1}(\partial S_j^+)))$.

Both of the sets $\mathcal{U}$ and $\mathcal{T}$ are constructed via *countable* unions of sets derived (by preimage) from sets $\{S_j^+\}_{j \in J}$ and $\{\partial S_j^+\}_{j \notin J}$, which are nowhere dense and have zero Lebesgue measure. It is therefore reasonable to expect that $\mathcal{U}$ and $\mathcal{T}$ are themselves nowhere dense and have zero Lebesgue measure. However, these sets become very complicated as they map under the homeomorphisms ϕ_i^{-1} to H_0. (See §4.) Moreover, homeomorphisms do not, in general, preserve zero Lebesgue measure. To overcome these potential difficulties we establish that if the domains of ϕ_i and ϕ_i^{-1} are restricted still further, then they are diffeomorphisms.

PROPOSITION 3.13. *For each* $i = 0, 1, \ldots,$

i) $\phi_i : R_i^+ \to R_{i+1}^-$ *is continuously differentiable at every* $x \notin \phi_i^{-1}(\operatorname{Ker} C_i)$,

ii) $\phi_i^{-1} = \phi_i^{-1} : R_{i+1}^- \to R_i^+$ *is continuously differentiable at every* $y \notin \phi_i(\operatorname{Ker} C_i)$, *and*

iii) $\phi_i : R_i^+ \setminus ((\operatorname{Ker} C_i) \cup \phi_i^{-1}(\operatorname{Ker} C_i)) \to R_{i+1}^- \setminus ((\operatorname{Ker} C_i) \cup \phi_i(\operatorname{Ker} C_i))$ *is a diffeomorphism.*

Proof. For $x \in R_i^+$, $\phi_i(x) = (e^{A_i t(x)}x, \tau_i)$, where $t(x)$ is defined implicitly by

$$\langle x, W_i(t(x))x\rangle = \tau_{i+1} - \tau_i.$$

Clearly the map $F : \mathbb{R}^{n+1} \to \mathbb{R}$, $(t, x) \mapsto \langle x, W_i(t)x\rangle$ is continuously differentiable with

$$\frac{\partial F}{\partial t}|_{(t,x)} = \|C_i e^{A_i t}x\|^2 \text{ and } \frac{\partial F}{\partial x}|_{(t,x)} = 2x^T W_i(t).$$

If $\phi_i(x) \notin \operatorname{Ker} C_i$, then

$$\frac{\partial F}{\partial t}|_{(t(x),x)} = \|C_i\phi_i(x)\|^2 > 0.$$

It now follows from the implicit function theorem that $t(\cdot)$ is *continuously differentiable* in a neighbourhood of x. Hence ϕ_i is continuously differentiable on $(\phi_i^{-1}(\operatorname{Ker} C_i))^c$ with the derivative of ϕ_i, $D\phi_i$, given by

$$D\phi_i|_x = e^{A_i t(x)} - \frac{2A_i\phi_i(x)x^T W_i(t(x))}{\|C_i\phi_i(x)\|^2}. \tag{3.2}$$

Arguing similarly, reversing time, we have that ϕ_i^{-1} is continuously differentiable on $(\phi_i(\operatorname{Ker} C_i))^c$.

Finally, $y \in \operatorname{Ker} C_i \cup \phi_i(\operatorname{Ker} C_i)$ if and only if $\phi_i^{-1}(y) \in \operatorname{Ker} C_i \cup \phi_i^{-1}(\operatorname{Ker} C_i)$ and

$$\phi_i : R_i^+ \setminus ((\operatorname{Ker} C_i) \cup \phi_i^{-1}(\operatorname{Ker} C_i)) \to R_{i+1}^- \setminus ((\operatorname{Ker} C_i) \cup \phi_i(\operatorname{Ker} C_i))$$

is a diffeomorphism. □

Remark 3.14. We observe that $D\phi_i^{-1}|_x \to \infty$ as either $\phi_i^{-1}(x) \to \operatorname{Ker} C_i$ or $x \to S_{i+1}^-$. Similarly $D\phi_i|_x \to \infty$ as $\phi(x) \to \operatorname{Ker} C_i$ or as $x \to S_i^+$. It follows that $D\phi_i^{-1}|_x$ and $D\phi_i|_x$ become singular as $x \to \operatorname{Ker} C_i$. At first sight this does not seem obvious from (3.2). However, we can use the alternative formula

$$D\phi_i|_x = e^{A_i t(x)} - \frac{2A_i e^{A_i t(x)}xx^T W_i(t(x))}{\|C_i x\|^2 + 2x^T W_i(t(x))A_i x},$$

and now taking determinants we have

$$\operatorname{Det}(D\phi_i|_x) = \operatorname{Det}(e^{A_i t(x)})\operatorname{Det}\left(I - \frac{2A_i xx^T W_i(t(x))}{\|C_i x\|^2 + 2x^T W_i(t(x))A_i x}\right),$$

which is zero since

$$\left(1 - \frac{2x^T W_i(t(x))A_i x}{\|C_i x\|^2 + 2x^T W_i(t(x))A_i x}\right) = 0$$

when $C_i x = 0$.

Example 3.15. (a) In the one-dimensional case

$$\frac{d\phi_i}{dx} = \frac{x}{\sqrt{x^2 + 2(\tau_{i+1} - \tau_i)((-1)^i\tau_i + 1)}}.$$

If $((-1)^i\tau_i + 1) > 0$, then $\frac{d\phi_i}{dx} > 1$ and Φ_t expands the area between R_i^+ and R_{i+1}^-.

(b) If we look again at Fig. 2, we clearly see cusp points in the image of the ellipse. These are due precisely to the singularity of $D\phi_1^{-1}$ on $\operatorname{Ker} C_i$ since functions with singular derivatives do not necessarily map smooth curves to smooth curves. We also see that divergence

of $D\phi_1^{-1}|_x$, as $\phi_1^{-1}(x)$ approaches Ker C_i, is reflected in the sparsity of points in Figs. 2(a)-(c) near Ker C_i.

To establish denseness and, more importantly, full Lebesgue measure of $\mathcal{S}$, we use the differentiability of ϕ_i and a simple result from measure theory, modified slightly to suit our purposes.

LEMMA 3.16. *If $X \subset H_{i+1}$ has zero Lebesgue measure, then $\phi_i^{-1}(X) \subset H_i$ has zero Lebesgue measure.*

Proof. It is well known that diffeomorphisms preserve Lebesgue measure. We have to take care of places where ϕ_i^{-1} is not differentiable. Since S_i^- and KerC_i are *closed sets*, we can decompose $\phi_i^{-1}(X)$ as

$$\phi_i^{-1}(X) = \left(\bigcup_{n=1}^{\infty} \phi_i^{-1}(X_n)\right) \cup (\text{Ker } C_i \cap \phi_i^{-1}(X)),$$

where $X_n = X \setminus B_{\frac{1}{n}}(S_{i+1}^-) \cup \phi(B_{\frac{1}{n}}(\text{Ker } C_i))$. Note continuity of ϕ_i implies that X_n is measurable for each n.

Since ϕ_i^{-1} is differentiable on X_n, it follows that $\phi_i^{-1}(X_n)$ has zero Lebesgue measure for each n and therefore $\phi_i^{-1}(X)$ has zero Lebesgue measure. □

We can now state the main result of this section.

THEOREM 3.17. *If* (2.11), (2.12) *satisfy* (2.13), *then $\mathcal{S}$ is open and dense and has full Lebesgue measure*; *that is, the property that* (2.14) *is exponentially stable is generic.*

Proof. For each $i \notin J$ and $i \in J$, both ∂S_i^+ and S_i^+, respectively, have zero Lebesgue measure. Under each "diffeomorphism" ϕ_k^{-1}, $k = i - 1, \ldots, 0$, it follows from Lemma 3.16 that this zero measure is preserved. Hence, $\mathcal{T}$ and $\mathcal{U}$ have zero Lebesgue measure as countable unions of sets with Lebesgue measure zero. Now

$$\mathbb{R}^n = \mathcal{S} \cup \mathcal{T} \cup \mathcal{U},$$

so $\mathcal{S}$ has full Lebesgue measure. Moreover, since $\mathcal{S}$ is open, $\mathcal{T} \cup \mathcal{U}$ is closed. Therefore, $\mathcal{T} \cup \mathcal{U}$ cannot have interior, and therefore $\mathcal{S}$ is dense. □

COROLLARY 3.18. *Under the conditions of Theorem* 3.17, $\mathbb{R}^n$ *is decomposed as*

$$\mathbb{R}^n = \mathcal{S} \cup \mathcal{T} \cup \mathcal{U},$$

where

- *$\mathcal{U}$ is nowhere dense and has zero Lebesgue measure,*
- *$\mathcal{T}$ is nowhere dense and has zero Lebesgue measure, and*
- *$\mathcal{S}$ is open and dense with full Lebesgue measure.*

Remark 3.19. We see that the set of initial conditions x_0, producing non-stabilizing limit gains $K(\infty, x_0)$, is restricted to a closed set with Lebesgue measure zero. Moreover, Theorem 3.10, concerning stability of solutions under small perturbations, applies to all initial conditions in an open, dense, and full-measure set of initial conditions.

If an *experiment* is defined as the solution $\{\Phi_t(x_0, \tau_0) | \ t \geq 0\}$ of (2.11), (2.12) corresponding to a single initial condition $x_0 \in \mathbb{R}^n$, then except for initial conditions in a Lebesgue measure zero set, a *single experiment* will guarantee the identification of a stabilizing output feedback matrix. Note that $\mathcal{U}$ is made up of a countable union and that any bounded subset of $\mathbb{R}^n$ may intersect infinitely many components of $\mathcal{U}$ corresponding to different limit gains. This would lead to highly sensitive closed-loop dynamics. However, in the case of Examples 2.1 and 2.2 we can rule out this possibility. This adds to the regularity of these control algorithms.

PROPOSITION 3.20. *For each closed-loop system in either Example* 2.1 *or* 2.2 *we have bounded switching number on compact subsets*; *i.e., for each $M > 0$ there exists $\kappa \geq 0$ such that if $\|x_0\| \leq M$, then $k(\infty, x_0) \leq \kappa$.*

Proof. We prove this result for the closed-loop system in Example 2.1. The same ideas work for the closed-loop system in Example 2.2, although the technical details are more involved. The details are omitted.

For each $x_0 \in \mathbb{R}^n$ an upper bound on the number of switches and hence on $k(\infty, x_0)$ is determined as follows. Let M_q be a feedback gain which is stabilizing for the given system. We define sequences of times t_n, t'_n (depending on x_0) such that $k(t_n, x_0) = \tau_{i_n}$ and $k(t'_n, x_0) = \tau_{i_n+1}$ with $K_{i_n} = M_q$. (If no such t_n exists, then $k(\infty, x_0) < \tau_{q(q+1)/2}$.) Now

$$\int_{t_n}^{t'_n} \|y(s)\|^2 + \|u(s)\|^2 ds = \tau_{i_n+1} - \tau_{i_n}.$$

But the feedback gain on the interval $[t_n, t_{n+1})$ is stabilizing so that there exists M (not depending on $\{\tau_i\}$) such that

$$\int_{t_n}^{t'_n} \|y(s)\|^2 + \|u(s)\|^2 ds \leq M\|x(t_n)\|^2$$

and

$$x(t) \leq M\left(\|x_0\| + \sqrt{\int_0^t \|y(s)\|^2 + \|u(t)\|^2 ds}\right).$$

Putting these two inequalities together we obtain

$$\tau_{i_{n+1}} - \tau_{i_n} \leq c\|x_0\| + c(\tau_{i_n} - \tau_0) \tag{3.3}$$

for some c. Switching is, of course, terminated no later than when (3.3) is violated, and since $\|x_0\|$ is bounded, this inequality is violated for some τ_{i_n} independent of x_0. (The subsequence $\{\tau_{i_n}\}$ is determined by the condition that $K_{i_n} = M_q$, and the length of the sequence depends only on x_0. □

So far we have concentrated on qualitative results for the dynamics of the closed-loop system which are necessary consequences of the assumed existence of a universal adaptive stabilizer. We can use the same approach to analyse those properties which are necessarily required of the underlying system.

PROPOSITION 3.21 (a necessary condition for universal adaptive stabilization).
Let

$$\left.\begin{aligned} \dot{x}(t) &= A_i x(t) \\ \dot{k}(t) &= \|C_i x(t)\|^2 \end{aligned}\right\} \text{ for } k \in [\tau_i, \tau_{i+1}).$$

Suppose for each $x_0 \in \mathbb{R}^n$, *with* $k(0) = \tau_0$, *that* $x(t) \to 0$ *and* $k(t) \to k(\infty, x_0) < \infty$. *Then there exists* $j \in \mathbb{N}$ *such that* $\sigma(A_j) \subset \mathbb{C}_-$.

Proof. If no A_i is stabilizing, then by Corollary 3.18 we would have

$$\mathbb{R}^n = \mathcal{U},$$

which is not possible because $\text{int}(\mathcal{U}) = \phi$. □

Remark 3.22. This is the piecewise-constant analogue of the necessary conditions for universal adaptive stabilization established in Byrnes, Helmke, and Morse (1986).

4. Qualitative and topological properties for minimum-phase, relative-degree-one systems. In §3 we saw that in a sequence of "experiments" (i.e., applying the same control to the same system but with differing initial conditions), generically the limiting linear system (2.14) will be exponentially stable. As we have clearly seen in Example 2.9, we cannot really expect better than this, and in most cases $\mathcal{U}$ will be nonempty. Indeed, it is easy to see that if a piecewise-linear universal adaptive stabilizer is designed for a large enough class of systems (containing, of course, some open-loop unstable ones), then $\mathcal{U}$ will be nontrivial for some realization of the system to be controlled.

We now consider the case when only one experiment is performed so that the initial condition is fixed. We focus on the effect that the controller, given by (2.11) and (2.12), has on the stability properties of the limit system (2.14) and in particular whether the limiting system is exponentially stable. We will restrict attention to Example 2.2 in the single-input, single-output case. We have a system

$$\dot{x}(t) = Ax(t) + bu(t), \qquad y(t) = c^T x(t), \tag{4.1}$$

which can be stabilized by (2.9) and (2.7), which we recall is given by

$$u(t) = (-1)^i \tau_i y(t), \ \dot{k}(t) = y^2(t) \ \text{ if } \ k(t) \in [\tau_i, \tau_{i+1}).$$

Note that the controller is parametrized by the sequence $\{\tau_i\}$, which must satisfy (2.8). We refer to such a sequence as a *Nussbaum sequence*.

In this problem x_0 is fixed and the limiting gain k_∞ is now a function of $\{\tau_i\}$ only, and so $k_\infty = k(\infty, \{\tau_i\})$. Using the relative-degree-one structure we can rewrite (4.1) as

$$\begin{aligned} \dot{y}(t) &= ay(t) + c^T bu(t) + a_{12} z(t), \\ \dot{z}(t) &= a_{21} y(t) + A_{22} z(t), \end{aligned}$$

where $a \in \mathbb{R}$, $a_{12} \in \mathbb{R}^{1\times(n-1)}$, $a_{21} \in \mathbb{R}^{(n-1)\times 1}$, $A_{22} \in \mathbb{R}^{(n-1)\times(n-1)}$, and (by minimum phase) $\sigma(A_{22}) \subset \mathbb{C}_-$.

LEMMA 4.1 (Ilchmann and Owens (1991)). *For each x_0 there exists $K > 0$ such that*

$$0 \le y^2(t) \le K + K \int_0^t y^2(s) + c^T b \int_0^t y(s)u(s)ds. \tag{4.2}$$

LEMMA 4.2. *Let $\{\tau_i\}$ be fixed. For each $R > 0$ there exists q such that $k(\infty, \{\hat{\tau}_i\}) \le \tau_q$ for all $\{\hat{\tau}_i\}$ with $\|\{\tau_i\} - \{\hat{\tau}_i\}\|_\infty \le R$.*

Proof. Using Lemma 4.1 and (2.7) we have for each $\{\hat{\tau}_i\}$ that switching terminates before the inequality

$$0 \le y^2(t) + K(\hat{\tau}_i - \hat{\tau}_0) + c^T b \sum (-1)^i \hat{\tau}_i (\hat{\tau}_i - \hat{\tau}_{i-1})$$

is violated. This will be uniformly bounded on $\|\{\tau_i\} - \{\hat{\tau}_i\}\|_\infty \le R$ for $\{\hat{\tau}_i\}$ satisfying (2.8). □

Remark 4.3. If $\{\tau_i\}$ is fixed, then it follows from Lemma 4.2 that any l_∞-neighbourhood of $\{\tau_i\}$ is, with regard to the possible values of the limit gain, essentially an $\mathbb{R}^{q+1}$-neighbourhood of $(\tau_0, \tau_1, \ldots, \tau_q)$ for some q.

In Ilchmann (1994) it is shown that for each fixed x_0 the set of Nussbaum sequences which results in an exponentially stable linear limiting system is dense with full measure. To be precise, let $G(x_0)$, given by

$$G(x_0) = \{\{\tau_i\} \mid (2.14) \text{ is exponentially stable}\},$$

be the totality of all Nussbaum sequences which produce exponentially stable limit systems when (2.9) is applied to (4.1).

PROPOSITION 4.4 (Ilchmann (1994)). *If (A, b, c^T) is controllable and observable and x_0 is fixed, then $G(x_0)$ is dense and has full Lebesgue measure in the following sense: for each $\{\tau_i\}$ and every $\epsilon > 0$, $G(x_0)$ is dense in the l_∞ ϵ-neighbourhood around $\{\tau_i\}$. If q is the uniform bound on the number of switches given in Lemma 4.2 for this ϵ-neighbourhood, then $\Pi_q(G(x_0))$ has full Lebesgue measure in the $\mathbb{R}^{q+1}$ ϵ-neighbourhood around $(\tau_0, \tau_1, \ldots, \tau_q)$.*

Here Π_q is the projection $\{\tau_0, \tau_1, \ldots\} \mapsto (\tau_0, \tau_1, \ldots, \tau_q)$.

We can make a minor improvement to this result by exploiting continuity properties of the ϕ_i with respect to the controller parameters $\{\tau_i\}$.

LEMMA 4.5. *$(\tau_i, \tau_{i+1}) \mapsto e^{A(t_i(x))}x$ is continuous with respect to (τ_i, τ_{i+1}) for every x such that*

$$\langle x, W_i(t)x\rangle > \tau_{i+1} - \tau_i.$$

PROPOSITION 4.6. *$\operatorname{int}(G(x_0))$ is dense in any l_∞-neighbourhood of a Nussbaum sequence.*

Proof. Let

$$IG(x_0) = \{\{\tau_i\} \mid x_0 \in \mathcal{S}(\{\tau_i\})\}.$$

We claim that $IG(x_0)$ is dense in $G(x_0)$ and open.

If $\epsilon > 0$ and $\{\tau_i\}$ are given, choose $\{\hat{\tau}_i\} \in G(x_0)$, using Proposition 4.4, such that $\|\{\hat{\tau}_i\} - \{\tau_i\}\|_\infty < \epsilon$. Let q be such that $x_0 \in \mathcal{S}_q(\{\hat{\tau}_i\})$ for some q.[2] By increasing $\hat{\tau}_{q+1}$ by an arbitrary small amount, we can arrange that $x_0 \in \operatorname{int}(\mathcal{S}_q(\{\hat{\tau}_i\})) \subset \mathcal{S}$ and therefore $IG(x_0)$ is dense.

Almost by definition $IG(x_0)$ is open. To make this precise, let $\{\tau_i\} \in IG(x_0)$. We must show that

$$\{\{\hat{\tau}_i\} \mid \|\{\hat{\tau}_i\} - \{\tau_i\}\|_\infty < \epsilon\} \subset IG(x_0)$$

for some $\epsilon > 0$. By Lemma 4.2 we know that there exists q such that small changes to $\tau_{q+1}, \tau_{q+2}, \ldots$ are unimportant. We therefore have to consider only small perturbations to finitely many of the τ_i, and so the result follows if we can verify the following two claims:

(1) if $(x, \tau_q) \in \operatorname{int}(S_q^+)$, then there exists $\epsilon > 0$ such that $(\hat{x}, \hat{\tau}_q) \in \operatorname{int}(S_q^+(\{\hat{\tau}_q\}))$ for all $\hat{x}$, $\hat{\tau}_q$, and $\hat{\tau}_{q+1}$ with $\|x - \hat{x}\|, |\hat{\tau}_j - \tau_j| < \epsilon$, $j = q, q+1$;

(2) for all $\epsilon > 0$ there exists $\delta > 0$ such that $\|\{\tau_q\} - \{\hat{\tau}_q\}\|_\infty < \delta$ and $\|x - \hat{x}\| < \delta$ implies $\|\phi_q(x) - \hat{\phi}_q(\hat{x})\| < \epsilon$,

since by applying (2) repeatedly we can move from $\hat{H}_q$ to $\hat{H}_0$. (We use the obvious notation of $\hat{\phi}_i$ denoting ϕ_i for $\{\hat{\tau}_i\}$.)

Claim (1) follows from the continuity of $\hat{x}^T \hat{W}_q(\infty)\hat{x} = \tau_{i+1} - \tau_i$ with respect to $\hat{x}$ and $\{\hat{\tau}_q\}$. Claim (2) follows from the continuity of ϕ_q and Lemma 4.5 since

$$\|\phi_q(x) - \hat{\phi}_q(\hat{x})\| \le \|\phi_q(x) - \phi_q(\hat{x})\| + \|\phi_q(\hat{x}) - \hat{\phi}_q(\hat{x})\|. \qquad \square$$

Remark 4.7. Putting these results specific to minimum-phase, relative-degree-one, single-input single-output systems together with the general results of §3 gives a complete picture in that stability of the limit system is a generic and full-Lebesgue-measure property with respect to both plant initial conditions and controller parameters.

[2] Recall $\mathcal{S}_q = \{x_0 \in \mathbb{R}^n \mid \phi_t(x_o, \tau_i) \in S_i^+ \text{ some } t < \infty\}$.

The remainder of this paper concerns a detailed quantitative and qualitative analysis of the second-order case. We assume that $c^T b > 0$ and (A, b, c^T) is minimal. By Proposition 3.20 we have on each compact subset of $\mathbb{R}^2$ a finite partition of $\mathcal{U}$, $\mathcal{T}$, and $\mathcal{S}$ according to a finite number of potential limit gains. The qualitative nature of this partition depends strongly on whether information about the high-frequency gain $c^T b$ is available and on the root locus of the system to be controlled.

i) If the information $c^T b > 0$ is available, then we can simply use (2.10) and (2.7).

ii) If the high-frequency gain information is not available, then we must use (2.9) and (2.7).

It is easy to see that there are five types of second-order, minimum-phase, relative-degree-one systems characterized according to pole-zero locations of the system transfer function

$$c^T(sI - A)^{-1}b = \frac{p(s)}{q(s)} = \frac{c^T b(s - \gamma)}{(s - \lambda)(s - \mu)}.$$

N.B. γ, the zero, is negative.

Case 1 (types I, II, and III) (real poles).

I. $\lambda \leq \mu < \gamma$.

II. $\lambda < \gamma < \mu$ (pole-zero interlacing).

III. $\gamma > \lambda \geq \mu$.

Case 2 (types IV and V) (complex poles).

$\mu = \bar{\lambda}$ with $\mathrm{Im}\lambda \neq 0$.

IV. $\mathrm{Re}\lambda < 0$.

V. $\mathrm{Re}\lambda \geq 0$.

Type I: The root locus lies completely in the left half plane. If $c^T b > 0$ is known, then $\mathcal{U} = \phi$.

Type II: The root locus lies on the real axis for all positive values of gain.

Type III: The root locus has a departure from the real axis at some positive gain.

Types IV and V: The root locus is a subset of the root locus in III. If, for some positive gain τ^*, the root locus intersects the real axis, then the intersection is in the left half plane and all gains greater than τ^* are stabilizing.

If $\mathrm{sign}(c^T b)$ is not known, then the complete $\pm$ gain root locus is relevant. In this case I, II, IV, and V are equivalent up to translation of the complex plane parallel to the real axis.

We are mainly interested in properties of $\mathcal{U}$. Because we have a second-order system, the only contribution to $\mathcal{U}$ can come from those $j \in J$ for which the x-dynamics are hyperbolic (A_j has one positive and one negative eigenvalue). Let

$$J^H = \{j \in J \mid x_j \to e^{A_j t}x \text{ is hyperbolic}\}$$

and, for each $j \in J^H$, λ_j and μ_j be the negative and positive eigenvalues of A_j. How the S_j^+, $j \in J_H$ ultimately contribute to $\mathcal{S}$ is dependent on the overall dynamics, which in turn are determined by the possible transitions between linear x-dynamics as we map down to H_0.

Qualitative and quantitative results when knowledge of $c^T b > 0$ is available. If $c^T b > 0$ is known and we use (2.10), then either $J^H = \emptyset$ or there exists $q \in \mathbb{N}$ such that

$$J^H = \{0, 1, \ldots, q\}.$$

It is important to analyse the transition from one hyperbolic system to another, as this causes the distortion and twisting of S_i^+.

LEMMA 4.8. *If* $i, i + 1 \in J^H$, *then*

i) $\phi_i^{-1}(S_{i+1}^+)$ *lies between* $V_i^s \times \{\tau_i\}$ *and* $V_i^u \times \{\tau_i\}$,

ii) $\overline{\phi_i^{-1}(S_{i+1}^+)} \cap V_i^s = \partial S_i^+ \setminus \{0\} := \{(s_i^+, \tau_i), (-s_i^+, \tau_i)\}$, *where* $\pm s_i^+ = \delta_i \binom{1}{\lambda_i}$ *and* δ_i *is determined by*

$$\delta_i^2 \int_0^\infty \left| c^T e^{\lambda_i t} \begin{pmatrix} 1 \\ \lambda_i \end{pmatrix} \right|^2 dt = \tau_{i+1} - \tau_i,$$

iii) $\phi_i^{-1}(S_{i+1}^+)$ *is smooth and* $\phi_i^{-1}(S_{i+1}^+) \cup S_i^+$ *is continuous.*

Proof. i) Let $V_i^s = \text{span}\{v_i^s\}$, $V_i^u = \text{span}\{v_i^u\}$ with eigenvalues λ_i and μ_i. If $(x, \tau_i) \in S_{i+1}^+$ with $x = \alpha v_i^S + \beta v_i^u$, $\alpha, \beta \in \mathbb{R}$, then

$$\phi_i^{-1}(x) = \alpha \rho^{-\lambda_i} v_i^S + \beta \rho^{-\mu_i} v$$

where $\rho = e^{t(x)} > 1$.[3] The result follows.

ii) Because of the continuity of ϕ_i^{-1} we need only to consider points $(x_\epsilon, \tau_i) \in S_{i+1}^+$, with

$$x_\epsilon = \epsilon \begin{pmatrix} 1 \\ \lambda_{i+1} \end{pmatrix}$$

and $\epsilon \to 0$. Then

$$x_\epsilon = \frac{\epsilon}{(\lambda_i - \mu_i)} \left\{ (\lambda_i - \lambda_{i+1}) \begin{pmatrix} 1 \\ \lambda_i \end{pmatrix} - (\mu_i - \lambda_{i+1}) \begin{pmatrix} 1 \\ \mu_i \end{pmatrix} \right\}$$

and

$$\phi_i^{-1}(x_\epsilon) = \frac{\epsilon}{(\lambda_i - \mu_i)} \left\{ (\lambda_i - \lambda_{i+1}) \rho^{-\lambda_i} \begin{pmatrix} 1 \\ \lambda_i \end{pmatrix} - (\mu_i - \lambda_{i+1}) \rho^{-\mu_i} \begin{pmatrix} 1 \\ \mu_i \end{pmatrix} \right\},$$

where $\rho \to \infty$, $\epsilon \rho^{-\lambda} \to \delta_i$ as $\epsilon \to 0$, and

$$\delta_i^2 \int_o^\infty \left| c^T e^{\lambda t} \begin{pmatrix} 1 \\ \lambda \end{pmatrix} \right|^2 dt = \tau_{i+1} - \tau_i. \tag{4.3}$$

It follows that $\lim_{\epsilon \to 0} \phi_i^{-1}(x_\epsilon) \in \partial S_i^+ = \{(s_i^+, \tau_i), (-s_i^+, \tau_i)\}$,

iii) Note that $\text{Ker}\, c^T = \text{span}\binom{1}{\gamma}$ is positioned anticlockwise between V_i^s and V_i^u. Hence, by i) $\text{Ker}\, c^T \cap \phi_i^{-1}(S_{i+1}^+) = \emptyset$ and $\text{Ker}\,(c^T) \cap S_{i+1}^+ = \emptyset$. Using Proposition 3.13, it follows that $\phi_i^{-1}(S_{i+1}^+)$ is smooth. Finally by ii), $\phi_i^{-1}(S_{i+1}^+) \cup S_i^+$ is a continuous union connected at $\{(s_i^+, \tau_i), (-s_i^+, \tau_i)\}$. □

PROPOSITION 4.9. *If* $(A, b, c^T) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 1} \times \mathbb{R}^{1\times 2}$ *is minimal and minimum phase,* $c^T b > 0$ *is known,* (2.10) *is used, and* $J_H \neq \emptyset$, *then* $\mathcal{U}$ *is a closed continuous curve differentiable at all but* $2(|J^H| - 1)$ *points and*

$$\inf\{\|x - y\| \mid x \in \text{Ker} c^T,\ \|x\| = 1,\ y \in \mathcal{U},\ \|y\| = 1\} > 0.$$

Before illustrating some features of $\mathcal{U}$ in the case $c^T b > 0$ known, we have a lemma which further aids the qualitative analysis. This is to cope with transitions caused by switching from stable linear x-dynamics to hyperbolic x-dynamics.

[3] We are treating ϕ_i as a function defined on $\mathbb{R}^2$.

LEMMA 4.10. *If* $i \in J^H$, $(i+1) \notin J$, *and* $S_{i+1}^- \not\subset S_{i+1}^+$, *then* $\phi_i^{-1}(S_{i+1}^+)$ *has exactly two connected components.*

Proof. ϕ_i^{-1} is continuous but not defined on S_{i+1}^-. Hence ϕ_i^{-1} is continuous at each point $x \in S_{i+1}^+ \setminus S_{i+1}^-$. However, since $S_{i+1}^- \not\subset S_{i+1}^+$, S_{i+1}^- cuts S_{i+1}^+ into two halves. Let $(x_\epsilon, \tau_{i+1}) \in S_{i+1}^+$ approach S_{i+1}^- so that

$$x_\epsilon = \epsilon u_i + v_i$$

with $u_i \in V_i^s$, $v_i \in V_i^u$, and $\epsilon \to 0$.

Arguing as in Lemma 4.8 we have that

$$\phi_i^{-1}(x_\epsilon) = \epsilon\rho^{-\lambda_i} u_i + \rho^{-\mu_i} v_i$$

with $\lim_{\epsilon\to 0} \rho = 0$, $\lim_{\epsilon\to 0} \epsilon\rho^{-\lambda} = c^*$, where

$$c^* = \sqrt{\frac{2\mu(\tau_{i+1} - \tau_i) - |c^T v_i|^2}{2\mu |c^T u_i|^2}}.$$

(N.B. $S_{i+1}^- \not\subset S_{i+1}^+$ guarantees that $2\mu\delta\tau_i > |c^T v_i|^2$ for each x_ϵ approaching S_{i+1}^- in S_{i+1}^+.) □

Example 4.11 (Example 2.9 revisited–known high-frequency gain). For the system given in Example 2.9 the open-loop poles and zeros are interlaced so that the root locus, i.e., the zeros of $(s^2 - s - 6) + k(s+1)$, lies entirely on the real axis. We take $\tau_0 = 0$, $\tau_1 = 3$, and $\tau_2 = 6.5$. Simple calculations yield

$$S_0^+ = \left\{ \delta \begin{pmatrix} 1 \\ -2 \end{pmatrix} \mid |\delta| \le 2\sqrt{3} \right\},$$

$$S_1^- = \left\{ \delta \begin{pmatrix} 1 \\ 3 \end{pmatrix} \mid |\delta| \le \frac{\sqrt{3}}{2} \right\},$$

$$S_1^+ = \left\{ \delta \begin{pmatrix} 1 \\ -3 \end{pmatrix} \mid |\delta| \le \sqrt{\frac{10.5}{2}} \right\},$$

$$S_2^- = \left\{ \delta \begin{pmatrix} 1 \\ 1 \end{pmatrix} \mid |\delta| \le 2\sqrt{3.5} \right\}.$$

If τ_3 is not too large, then the line S_2^- is not contained in the ellipse S_2^+. $\tau_3 = 8$ is a suitable choice. Hence by Lemma 4.10, $\phi_1^{-1}(S_2^+)$ has two connected components.

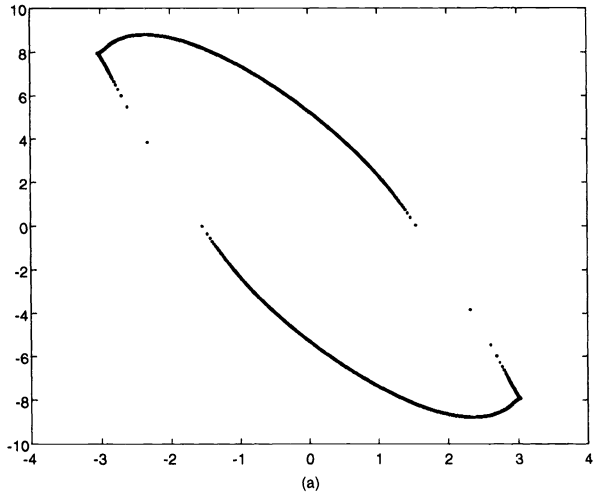
The remainder of the mapping depends on the relative sizes of $\tau_4, \tau_5, \ldots$ and information given by Lemmas 4.8 and 4.10. If $\tau_4 = 10$, then $S_3^- \subset S_3^+$, and therefore by Lemma 4.10, $\phi_2^{-1}(S_3^+)$ is not split into two components. However, $S_2^- \not\subset \phi_2^{-1}(S_3^+)$, and therefore $\phi_1^{-1}(\phi_2^{-1}(\psi_3^+))$ is in two components. These features are illustrated in Fig. 3, where we show S_3^+ mapping under ϕ_2^{-1}, ϕ_1^{-1}, and ϕ_0^{-1} to H_0. Note the cusp features in $\phi_2^{-1}(\partial S_3^+)$ caused by intersection with Ker (c^T).

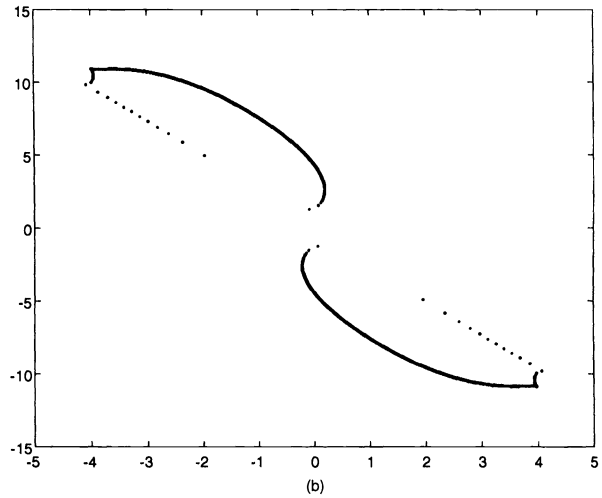
Finally we consider the structure of $\mathcal{U}$. In this case $J = \{0, 1\}$ so that

$$\mathcal{U} = \mathcal{U}_0 \bigcup \mathcal{U}_1.$$

We can calculate $\mathcal{U}_0$ and $\mathcal{U}_1$ explicitly. Indeed,

$$\mathcal{U}_0 = S_0^+$$





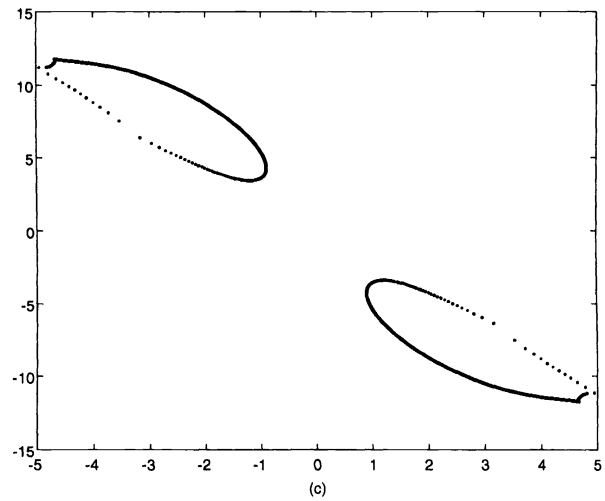


FIG. 3. S_3^+ *mapped to* H_0 *for Example* 4.11.

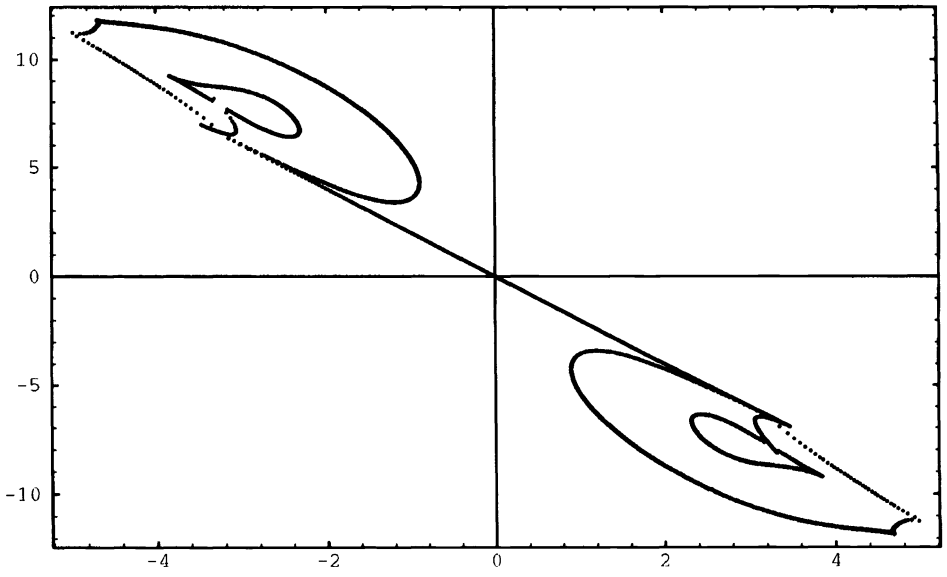


FIG. 4. $\mathcal{U}_0 \cup \mathcal{U}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$ *for Example* 4.11.

and

$$\mathcal{U}_1 = \phi_0^{-1}(S_1^+) = \left\{ \phi_0^{-1}\left(\delta \begin{pmatrix} 1 \\ -3 \end{pmatrix} \right) \mid |\delta| \leq \sqrt{5.25} \right\} = \frac{\delta(\rho)}{5} \left\{ 6\rho^2 \begin{pmatrix} 1 \\ -2 \end{pmatrix} - \rho^{-3} \begin{pmatrix} 1 \\ 3 \end{pmatrix} \right\},$$

where

$$\delta(\rho) = \frac{\pm 15\rho^3}{\sqrt{27\rho^{10} + 125\rho^6 - 144\rho^5 - 8}}$$

and $\rho \in [1.41, \infty)$. Figure 4 shows the complete picture for $\mathcal{U}$ together with $\mathcal{T}_2 = \partial\mathcal{S}_2$ and $\mathcal{T}_3 = \partial\mathcal{S}_3$. Note that $\mathcal{U}$ is a closed, continuous curve with two points of non-differentiability as predicted by Proposition 4.9.

Qualitative and quantitative results when knowledge of the sign of the high-frequency gain is not available. If $c^T b > 0$ is unknown, then the structure of $\mathcal{U}$ is more complicated because as k increases, the hyperbolic dynamics are interlaced with source and sink dynamics.

The most interesting feature, which did not occur in the case of known sign of the high-frequency gain, is that caused by transitions from hyperbolic dynamics to complex source dynamics and back to hyperbolic dynamics. Let

$$J^{cs} = \{j \in J \mid \mathrm{Im}(\sigma(A_j)) \neq 0\};$$

i.e., for each $j \in J^{cs}$, $x \mapsto e^{A_j t}x$ has complex source dynamics.

LEMMA 4.12. *If* $j \in J^{cs}$, $(j+1) \in J^H$, *and* $S_{j+1}^+ \not\subset S_{j+1}^-$, *then* $\phi_j^{-1}(S_{j+1}^+) \subset H_j$ *is a doubly infinite spiral about* $\{0\}$.

If, in addition, $(j-1) \in J^H$, *then* $\phi_{j-1}^{-1}(\phi_j^{-1}(S_{j+1}^+))$ *is a pair of infinite spirals centred at* $\partial(S_{j-1}^+) = \{(\pm s_{j-1}^+, \tau_{j-1})\}$.

Proof. Let $(x, \tau_{j+1}) \in S_{j+1}^+ \setminus S_{j+1}^-$, with x parameterized by ϵ, approach S_{j+1}^- as $\epsilon \to 0$. It follows that $\lim_{\epsilon \to 0} t_j(x_\epsilon) = \infty$, where $t_j(\cdot)$ is the time taken to flow from H_{j+1} to H_j. But

$$\phi_j^{-1}(x_\epsilon) = e^{-\sigma t_j(x_\epsilon)}(\alpha_\epsilon \cos w t_j(x_\epsilon)\, u + \beta_\epsilon \sin w t_j(x_\epsilon) v)$$

for some fixed $\sigma > 0$, $w > 0$, $u, v \in \mathbb{R}^2$. Hence $\phi_j^{-1}(S_{j+1}^+)$ is a spiral winding infinitely many times about $(0, \tau_j)$. Arguing as in Lemma 4.10, part of $\phi_j^{-1}(S_{j+1}^+)$ is cut by S_j^-. It follows that all points in $\phi_j^{-1}(\partial S_{j+1}^+)$ which are on one side of S_j^- become an infinite nest of loops, $\phi_{j-1}^{-1}(\phi_j^{-1}(S_{j+1}^+))$. Since the spiral $\phi_j^{-1}(S_{j+1}^+)$ is centred at $(0, \tau_j)$ and $\partial(S_{j-1}^+)$ "maps" to $(0, \tau_j)$ under ϕ_{j-1}, it follows that $\{(\pm s_{j-1}^+, \tau_{j-1})\} = \partial(S_{j-1}^+)$ are the centres of the nested loops. □

If, in Lemma 4.12, $j+2 \notin J$, then S_{j+2}^+ is an ellipse and $\phi_{j+1}^{-1}(S_{j+2}^+)$ contains part of S_{j+1}^+. Hence the infinite spirals $\phi_j(S_{j+1}^+)$ contributing ultimately to $\mathcal{U}$ carry with them contributions to $\mathcal{S}$, with the ellipse S_{j+2}^+ in H_{j+2} becoming an infinitely spiralling band in H_j. These features are illustrated in Example 4.13. However, it is difficult to make general statements as to whether these features will always reach H_0 when greater numbers of switches are considered.

Example 4.13 (unknown sign of the high-frequency gain). Let

$$A = \begin{bmatrix} 0 & 1 \\ 0.2 & 0.8 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad c^T = [1\ 1].$$

If we do not know the sign of $c^T b$, then we use (2.7) and (2.9). If $\tau_0 = 0$, $\tau_1 = 0.6$, $\tau_2 = 0.8$, and $\tau_3 = 1$, then $0, 2 \in J^H$, $1 \in J^{cs}$, and $3 \notin J$. If $\tau_4 = 3.22$, then

$$S_3^+ = \{(x, y) \mid 29x^2 + 10xy + 45y^2 \leq 160/9\},$$

$$S_2^+ = \left\{ \delta \begin{pmatrix} 1 \\ -0.48 \end{pmatrix} \mid |\delta| \in [0.2, 0.84] \right\},$$

$$S_2^- = \{(x, y) \mid 2.65x^2 - 2.5xy + 3.5y^2 \leq 0.2\},$$

$$S_1^+ = \{0\}.$$

In Figs. 5(a)-(c) we show S_3^+ mapping to H_0 (shown in a dotted line) together with S_2^+ in Fig. 5(b) (shown in boldface). Figures 6(a) and 6(b) show part of $\mathcal{U}$ coming from $\mathcal{U}_0$ and $\mathcal{U}_2$ (shown in boldface) together with $\mathcal{T}_3$ (shown dotted). Figure 6(b) is a blow-up of part of Fig. 6(a). In order to emphasize the detail, in Figs. 6(a) and 6(b) we have rotated coordinates, putting S_0^+ on the vertical axis.

In this example we see that even for minimum-phase, relative-degree-one systems, the possibility of unknown sign of the high-frequency gain induces a significant degree of complexity in the partition of $\mathbb{R}^2 = \mathcal{S} \cup \mathcal{T} \cup \mathcal{U}$. In Example 4.13, $J^H = \{0, 2, 4, 6, \ldots\}$ and $J = J^H \cup \{1\}$.

The points exterior to

$$\mathcal{U}_0 \cup \mathcal{U}_2 \cup \mathcal{S}_3,$$

represent those initial conditions whose gain evolution is not terminated until four or more switches in gain have taken place.

Observation. Dropping explicit reference to τ_0, we have in Example 4.13 that $\pm s_0^+$ are the centres of the infinite spirals $\phi_0^{-1}(\phi_1^{-1}(S_2^+)) = \mathcal{S}_2$. Hence, for all $\epsilon > 0$ there exist x_1, x_2, and x_3 within ϵ of s_0 such that

$$K(\infty, s_0^+) = \tau_0, \quad K(\infty, x_1) = \tau_2, \quad K(\infty, x_2) = -\tau_3, \quad K(\infty, x_3) = (-1)^j \tau_j, \quad j \geq 4.$$

Hence, starting arbitrarily close to s_0^+ we have four qualitatively distinct types of closed-loop-systems behaviour, according to none, two, three, and more than four switches switches in gain. However, from Proposition 3.20 we know that on each bounded set we have only finitely many switches so that this erratic nature of the closed-loop system in a neighbourhood of s_0^+ is confined to solutions switching at most a finite number of different times.

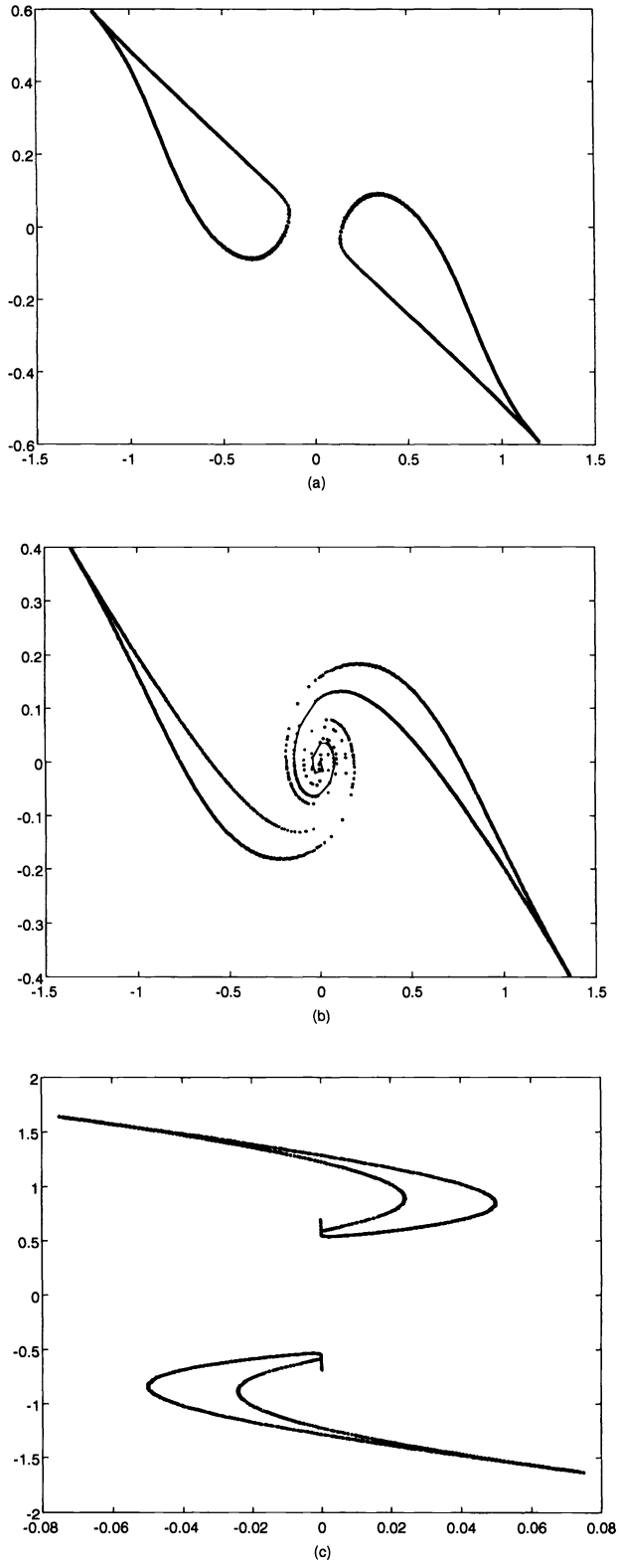


FIG. 5. S_3^+ *mapping to* H_0 *(dotted) and* S_2 *mapping from* H_2 *to* H_1 *(boldface) for Example* 4.13.

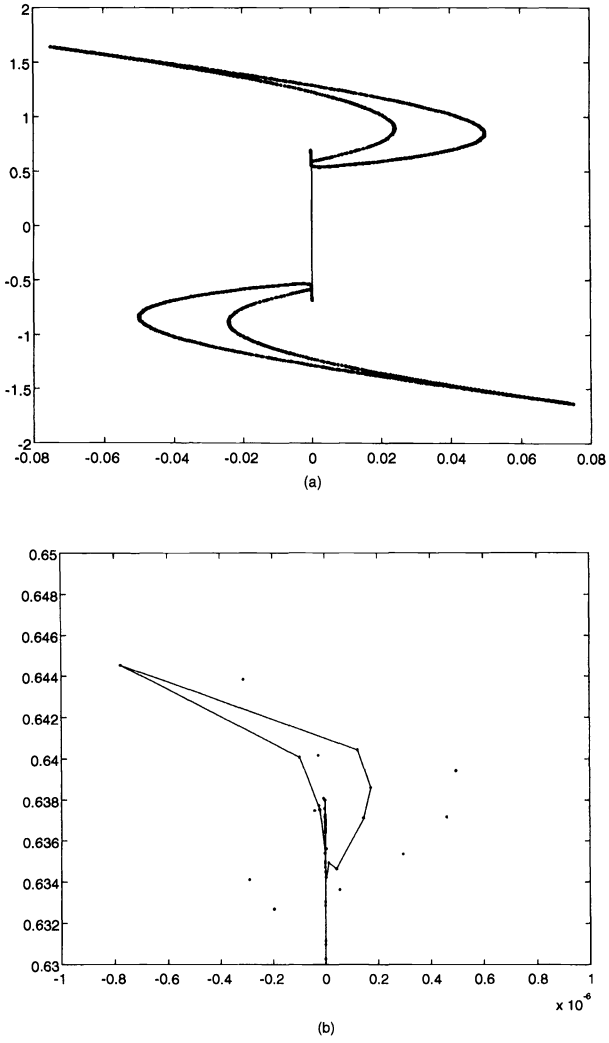


FIG. 6. $\mathcal{U}_0 \cup \mathcal{U}_2 \cup \mathcal{S}_3 \cup \mathcal{T}_3$ *for Example* 4.13. (b) *is a blow-up of part of* (a).

5. Concluding remarks. In this paper we have considered a new approach to an analysis of those nonlinear systems arising from adaptive control of linear time-invariant systems. The particular problems considered were stability of solutions under small perturbation of the initial data and the important question of generic stabilization by the limit gain matrix.

Instead of considering the closed-loop system as a differential equation we reduced the analytical problem to one of the topological properties of a sequence of *flow induced* homeo/diffeomorphisms. The approach is applicable to all universal adaptive stabilization schemes in which the feedback gain matrix is piecewise-constant and right continuous and where the first gain and subsequent ordering of the gains is independent of initial data. This encompasses every known sequential universal adaptive stabilization scheme.

We have established that the totality of initial conditions, $\{x_0 \in \mathbb{R}^n\}$, is partitioned according to

- an open and dense full-Lebesgue-measure set $\mathcal{S}$ on which stability of solutions is preserved under small perturbations and the limit gain is stabilizing,
- a nowhere dense and Lebesgue-measure-zero set $\mathcal{T}$ on which on which stability of solutions is not preserved but the limiting gain is still stabilizing,

• a nowhere dense and Lebesgue-measure-zero set $\mathcal{U}$ on which the limiting gain is not stabilizing.

For the class of minimum-phase, relative-degree-one, second-order systems with known sign of the high-frequency gain we showed the latter to be a closed, connected curve with at most finitely many points of nonsmoothness. We also obtained necessary conditions under which universal adaptive stabilization by switching sequence controllers is possible.

This new approach suggests several interesting problems. For example, how does the geometric structure of the sets $\mathcal{S}$, $\mathcal{T}$, and $\mathcal{U}$ break up as we perturb either the system (A, B, C) or the gain and switching sequences? In particular, is the structure stable under small nonlinear perturbations? Can we exploit knowledge of this structure to specify a finite number of experiments by which a stabilizing gain can be "identified"? Moreover, what happens if the system is *nonlinear* between switches? Can we still conclude these same generic properties?

Finally, in Logemann and Mårtensson (1990) piecewise-constant universal adaptive stabilization is considered for a large class of distributed parameter systems. It would be interesting to investigate the topological structure of $\mathcal{U}$, $\mathcal{S}$, and $\mathcal{T}$ in this case.

Beyond the piecewise-constant case we could also consider generic properties for adaptation laws in which f in (1.1)-(1.3) is smooth or piecewise-smooth. Similar results are anticipated in this case. However, the analytical techniques will be different.

Acknowledgment. I would like to thank B. D. Mestel of the University of Exeter for several illuminating discussions.

REFERENCES

C. I. Byrnes, U. Helmke, and A. S. Morse (1986), *Necessary conditions in adaptive control*, in Modelling Identification and Robust Control, C. I. Byrnes and A. Linquist, eds., North-Holland, Amsterdam, pp. 3-14.

A. Ilchmann (1991), *Non-identifier based adaptive control of dynamical systems: A survey*, IMA J. Math. Control Inform., 8, pp. 321-366.

——— (1994), *Adaptive controllers and root loci of multivariable minimum phase systems*, Dynamics Control, 4, pp. 123-146.

A. Ilchmann and H. Logemann (1993), *High-gain adaptive stabilization of multivariable linear systems—revisited*, Systems Control Lett., 18, pp. 35-364.

A. Ilchmann and D. H. Owens (1990), *Adaptive stabilization with exponential decay*, Systems Control Lett., 14, pp. 437-443.

——— (1991), *Exponential stabilization using non-differential gain adaptation*, IMA J. Math. Control Inform., 7, pp. 339-349.

H. Logemann and B. Mårtensson (1992), *Adaptive stabilization of infinite dimensional systems*, IEEE Trans. Automat. Control, 37, pp. 1869-1883.

B. Mårtensson (1986), *Adaptive Stabilization*, Ph.D. thesis, Lund Institute of Technology, Lund, Sweden.

——— (1991), *The unmixing problem*, IMA J. Math. Control Inform., 8, pp. 367-377.

D. E. Miller and E. J. Davison (1991), *An adaptive controller which provides an arbitrarily good transient and steady state response*, IEEE Trans. Automat. Control, AC-36, pp. 68-81.

A. S. Morse (1983), *Recent problems in parameter adaptive control*, in I. D. Landau, ed., Outils et Modèles Mathématiques pour l'Automatique, l'Analyse de Système et le Traitment du Signal, vol. 3, Editions du CNRS, Paris, pp. 733-740.

R. D. Nussbaum (1983), *Some remarks on a conjecture in parameter adaptive control*, Systems Control Lett., 3, pp. 243-246.

J. C. Willems and C. I. Byrnes (1984), *Global Adaptive Stabilization in the Absence of Information on the Sign of the High Frequency Gain*, Lecture Notes in Control and Inform. Sci. 62, Springer-Verlag, New York, pp. 49-57.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 3, pp. 1071-1097, May 1996

014

$\mathcal{H}_\infty$ CONTROL OF NONLINEAR SYSTEMS: DIFFERENTIAL GAMES AND VISCOSITY SOLUTIONS*

PIERPAOLO SORAVIA†

Abstract. Dealing with disturbances is one of the most important questions for controlled systems. $\mathcal{H}_\infty$ optimal control theory is a deterministic way to tackle the problem in the presence of unfavorable disturbances. The theory of differential games and the study of the associated Hamilton-Jacobi-Isaacs equation appear to be basic tools of the theory. We consider a general, nonlinear system and prove that the existence of a continuous, local viscosity supersolution of the Isaacs equation corresponding to the $\mathcal{H}_\infty$ control problem is sufficient for its solvability. We also show that the existence of a lower semicontinuous viscosity supersolution is necessary.

Key words. $\mathcal{H}_\infty$ control, differential games, viscosity solutions, Isaacs equation, nonlinear systems

AMS subject classifications. viiipt 93B36, 49L25, 90D25

1. Introduction.

In this paper we consider a general, nonlinear, controlled, dynamical system subject to unknown disturbances

$$\dot{y} = f(y, a, b), \tag{1.1}$$

with output or response $h(y, a, b)$, where a is the control and b is the disturbance. The disturbances are modelled deterministically as functions of time, and we want to optimize the performance of the system using the worst case criterion.

We are given a closed set $\mathcal{T}$ with respect to which the undisturbed system ($b = 0$) is (expected to be) stable (or asymptotically stable), and for some prescribed $\gamma > 0$ we look for control functionals (or strategies) $\alpha = \alpha[b]$ of the controller that achieve the stability and such that, for all possible disturbances $b(\cdot)$, the trajectories solutions of (1.1) starting at a point of $\mathcal{T}$ satisfy

$$\int_0^t |h(y, \alpha, b)|^2 ds \leq \gamma^2 \int_0^t |b|^2 ds \quad \text{for all } t > 0. \tag{1.2}$$

If we can find such an α (ideally in feedback form), we say that the $\mathcal{H}_\infty$ suboptimal control problem is solvable with disturbance attenuation level γ. The definition we use, which we basically take from Van der Shaft [30], is here given in an informal way and we refer the reader to the next section, where (1.2) is also generalized in particular to points off $\mathcal{T}$, and to Remarks 2.2 and 2.3, where we discuss the connections with previous literature.

As formulated, the problem appears to be a differential game for the system (1.1), as first observed by Basar and Bernhard [10], and indeed the $\mathcal{H}_\infty$ problem is solvable if and only if a suitably defined, nonnegative value function is zero on $\mathcal{T}$ and admits optimal strategies (for the definition of value function we follow Elliott and Kalton [13]; for a description of the relationship between differential games and viscosity solutions and more recent references, see also Evans and Souganidis [14] and the author [25]). This fact makes the problem similar to that of proving stability and asymptotic stability of dynamical systems with competitive controls, which we studied in [26], [27]. We approach the $\mathcal{H}_\infty$ problem using the same method of the mentioned papers with relevant additional difficulties, namely, the unboundedness of the sets of controls and disturbances, the unboundedness of the dynamics with respect to the parameters, and the fact that the running cost of the trajectories, that is,

$$|h(y, a, b)|^2 - \gamma^2 |b|^2,$$

*Received by the editors April 21, 1994; accepted for publication (in revised form) February 14, 1995.

†Dipartimento di Matematica Pura e Applicata, via Belzoni, 7, 35131 Padova, Italy (soravia@pdmat1.math.unipd.it).

does not have a prescribed sign. To tackle the first two we use a change of variables and adapt to differential games a general, classical method to study unbounded control problems, which consists of the reparametrization of the trajectories. This idea leads to a more regular Hamiltonian while leaving unchanged the value function and was used to this purpose for control problems by Barles [6]. The problem shows also the unusual fact that the payoff functional of the differential game, which we recover from (1.2), is a maximum cost type functional, as previously studied by Lions [24], Barles and Perthame [7], and Barron and Ishii [9] in optimal stopping time control problems and by Barron [8], Krassowski and Subbotin [23], and the author [27] for differential games. There is also the subtle question of the existence of the value of differential games that we have to take into account, and this fact will be discussed in Remarks 2.4 and 3.5.

Our approach is based on the study of the Hamilton-Jacobi-Isaacs equation corresponding to the differential game and the use of the theory of viscosity solutions for fully nonlinear first- (and second-) order partial differential equations initiated by Crandall and Lions [12]. This is due to the fact that, in the case of nonlinear systems even in a special form as nonlinear in the state and affine in the controls, the Isaacs equation does not have in general classical solutions. The goal is to show a rigorous relationship between the $\mathcal{H}_\infty$ problem and the Isaacs equation and to prove necessary and sufficient conditions for its solvability. We prove that if there is a continuous, nonnegative, viscosity supersolution of the Isaacs equation, null on $\mathcal{T}$, then the $\mathcal{H}_\infty$ suboptimal control problem is solvable. Moreover, if the value function is finite, then it is a discontinuous solution of the equation; when it is continuous, it is the minimal nonnegative, continuous supersolution. Our first result can be stated also locally in a neighborhood of $\mathcal{T}$, with a suitable definition of the local $\mathcal{H}_\infty$ suboptimal problem. In our case, the Isaacs equation has no unique solution in general, and this creates a further difficulty for the problem since the usual comparison theorems for viscosity solutions do not apply and we need to look for new optimality principles for supersolutions. The presentation of the results is almost self-contained, and we remark that they can be easily generalized to a wider class of nonlinear differential games.

Our results are the parallel of those of James [20] for dissipative systems, first studied by Willems [32] (see also Hill and Moylan [16]), where, however, the stability requirement is not an issue. In that case, the control $a(\cdot)$ ranges in a set that is a singleton, or equivalently a smooth feedback control $a(x)$ is fixed in the dynamics (1.1); this makes the proof much easier to obtain. However in that special case, the results of James [20] are stronger in the sense that, to prove the sufficiency part, the supersolution is required only to be lower semicontinuous. We could do the same in our situation if the control set A is compact, allowing relaxed strategies or requiring convexity of the sets $(f(x, A, b), h(x, A, b))$. Therefore our results can be extended to contain those of James [20], but the details will be presented elsewhere. One of the referees pointed out to us the recent paper by Ball and Helton [2], where the results in [20] are applied to get necessary conditions for the solution of the nonlinear $\mathcal{H}_\infty$ control problem in the case of nonlinear affine systems using viscosity solutions.

In this paper we consider a system with complete information. However, the $\mathcal{H}_\infty$ control problem for partially observed systems has great relevance for the applications and has been studied by Isidori and Astolfi [18] in the case of nonlinear systems affine in the controls (but see also Basar and Bernhard [10]). Some results in this direction are contained in the paper by the author [28].

It is well known that disturbances can be also modelled as stochastic processes. There is an interesting link between the two theories, by means of the risk-sensitive control approach, where a small parameter is introduced to measure the noise intensity in the stochastic system and to affect the payoff functional, which is in a certain log-exp form. In the limit as the

parameter goes to zero, it can be proven that the related value functions converge to the value function of a differential game that is contained in the class we consider. This result, originally obtained by Jacobson [19] for the linear exponential quadratic gaussian problem and formally by Whittle [31] for nonlinear systems, has been proved by Fleming and McEneaney [15] and independently by James [21], for finite horizon problems.

We finally recall that the $\mathcal{H}_\infty$ control problem was originally formulated for linear systems by Zames [33] in the frequency domain, where the name $\mathcal{H}_\infty$ has a clear meaning. For detailed descriptions of the problem as well as important recent developments of the theory and long lists of references, we also refer to Basar and Bernhard [10], Van der Shaft [29], [30], and Ball, Helton, and Walker [3].

The paper is organized as follows. In §2 we discuss assumptions and definitions. In §3 we state the main results of the paper with some comments. In §4 we prove some results concerning an auxiliary problem. In §5 we give the proofs of the main results. In §6 we briefly discuss the solution of the problem and the existence of optimal strategies in feedback form.

2. Preliminaries and differential games. We consider the following controlled dynamical system:

$$\begin{cases} \dot{y} = f(y,a,b), & y(0) = x \in \mathbb{R}^N, \\ z = h(y,a,b). \end{cases} \tag{2.1}$$

We assume that $A \subset \mathbb{R}^M$ and $0 \in B \subset \mathbb{R}^M$ are closed, $f : \mathbb{R}^N \times A \times B \to \mathbb{R}^N$ and $h : \mathbb{R}^N \times A \times B \to \mathbb{R}$ are continuous, and the following are satisfied:

$$\begin{cases} |f(x,a,b) - f(y,a,b)| \le L(1 + |a|^q + |b|^q)|x - y|, \\ |f(x,a,b)| \le L(1 + |x| + |a|^q + |b|^q) \quad \text{for all } x, y, a, b \text{ and some } 0 \le q < 2, \\ |h(x,a,b) - h(y,a,b)| \le L_R(1 + |a|^2 + |b|^2)|x - y|, \\ C_R^0|a|^2 - L_R(1 + |a|^q + |b|^2) \le h(x,a,b) \le L_R(1 + |a|^2 + |b|^q), \quad (C_R^0 > 0), \\ \text{for all } |x|, |y| \le R, a \in A, b \in B, R > 0, \text{ and some } 0 \le q < 2. \end{cases} \tag{2.2}$$

In our main results, we will also require that

$$h(x,a,0) \ge 0. \tag{2.3}$$

No big changes have to be made if in (2.2) we substitute 2 with $p \ge 1$. In our notation, y is the state space variable, z is the running cost, a is the controlled input, and b is the disturbance on the system. Usually in the applications $h(x,a,b) = |\bar{h}(x,a)|^2$, where $\bar{h} : \mathbb{R}^N \times A \to \mathbb{R}^P$ is the output to be controlled; however, since some results of this paper have independent interest and in order to extend the known results in the case of dissipative systems, we do not restrict ourselves to nonnegative costs, allowing h to depend also on b. The assumptions (2.2), (2.3) are satisfied by the usual linear quadratic model. The same is true for nonlinear systems affine in the controls with quadratic cost.

The admissible controls for our problem are given by the two sets

$$\begin{cases} \mathcal{A} = L^2_{loc}(\mathbb{R}_+, A), \\ \mathcal{B} = L^2_{loc}(\mathbb{R}_+, B); \end{cases}$$

note in particular that they are not required to be small for large times, as usually done in the literature. We will denote by $y_x(\cdot; a, b)$ or simply by $y_x(\cdot)$ or $y(\cdot)$ the unique solution

of (2.1) corresponding to the choice of the controls $a \in \mathcal{A}$, $b \in \mathcal{B}$. We also define the admissible strategies for the controller as *nonanticipating* (or causal) functionals $\alpha : \mathcal{B} \to \mathcal{A}$, i.e., satisfying the condition

$$\text{for all } t > 0,\ b = \bar{b} \text{ a.e. in } [0,t] \text{ implies } \alpha[b] = \alpha[\bar{b}] \text{ a.e. in } [0,t], \tag{2.4}$$

and indicate by Δ the set of such functionals.

Remark 2.1. Definition (2.4), which corresponds to the full-information situation, means that, when acting at time t, the controller a knows completely the control $b(\cdot)$ in the past $[0,t]$ but has no preview of the future behavior of the disturbance. There are at least two typical examples of nonanticipating strategies. The first one is a constant mapping $\alpha[b] \equiv a \in \mathcal{A}$, for all $b \in \mathcal{B}$; the second one is provided by (static state) feedback controls. Let $a : \mathbb{R}^N \to \mathcal{A}$ be a function such that for all $x \in \mathbb{R}^N$ and $b \in \mathcal{B}$, the system $\dot{y} = f(y, a(y), b)$, $y(0) = x$, has a unique, absolutely continuous, global solution. Then for any trajectory $y(\cdot)$ the position $\alpha[b](t) = a(y(t))$ defines a strategy, if $a(\cdot)$ is sufficiently smooth. A little variant of this class of strategies is defined by means of the so-called (see Van der Shaft [30] and the references therein) nonlinear compensators. To the system (2.1) we add the set of equations

$$\dot{\xi} = k(\xi, y), \qquad \xi(0) \in \mathbb{R}^P,$$

and consider a feedback control for this extended system or dynamic state feedback control, $a : \mathbb{R}^{N+P} \to A$. We then define the functional $\alpha[b](t) = a(\xi(t), y(t))$, that, with suitable assumptions on k and a, is a strategy.

We are given a closed set $\mathcal{T} \subset \mathbb{R}^N$, with respect to which we want to study the stability of (2.1). We will not require it as an assumption in the statements, but the following is a necessary condition for fulfilling the assumptions of our main Theorem 3.2, namely: for every $x \in \mathcal{T}$ there is $a \in A$ such that $h(x, a, 0) = 0$ and $f(0, a, 0) = 0$.

Given an open set Ω, we introduce the following functional, or exit time of the trajectories from Ω, as

$$t_x = t_x(a,b) = \inf\{t \geq 0 : y_x(t) \notin \Omega\} \leq +\infty, \tag{2.5}$$

where $t_x = +\infty$ if $y(t) \in \Omega$ for all $t \geq 0$. Observe that trivially if $x \notin \Omega$, then $t_x(a,b) = 0$ for all a, b.

Let $\mathcal{T} \subset \Omega \subset \mathbb{R}^N$ be an open set. We say that the system has *finite gain* in Ω measured by γ if the two following conditions are satisfied:

(i) (Ω is viable for the undisturbed system) For all $x \in \Omega$ there is $a \in \mathcal{A}$ such that the solution $y_x(t; a, 0) \in \Omega$ for all $t \geq 0$ or equivalently $t_x(a, 0) = +\infty$.

(ii) The function

$$V_\gamma^\Omega(x) = \inf_{\alpha \in \Delta_\Omega} \sup_{b \in \mathcal{B}_{\Omega,\alpha}} \sup_{t \in \mathbb{R}_+} \int_0^t (h(y_x, \alpha[b], b) - \gamma^2 |b|^2) ds \tag{2.6}$$

is finite for all $x \in \Omega$ and such that $V_\gamma^\Omega \equiv 0$ and continuous at the points of $\mathcal{T}$.

In (2.6) we denoted

$$\mathcal{B}_{\Omega,\alpha} = \{b \in \mathcal{B} : y_x(t; \alpha[b], b) \in \Omega \text{ for all } t \in \mathbb{R}_+\} \tag{2.7}$$

and set that $\alpha \in \Delta_\Omega$ if and only if $0 \in \mathcal{B}_{\Omega,\alpha}$. We observe that condition (i) implies $\Delta_\Omega \neq \emptyset$ and then $V_\gamma^\Omega \geq 0$ (to prove this fact it is sufficient to select $0 = t \in \mathbb{R}_+$ and $b \equiv 0$ as special values in the right-hand side of (2.6)), but in general V_γ^Ω may assume the value $+\infty$ at some points.

We say that the *local* $\mathcal{H}_\infty$ *suboptimal control problem* with attenuation level index $\gamma > 0$ is solvable if for any open set $\mathcal{U} \supset \mathcal{T}$, there is an open $\mathcal{T} \subset \Omega \subset \mathcal{U}$ such that the system has finite gain in Ω measured by γ.

If instead the undisturbed system is open-loop Lyapunov stable to $\mathcal{T}$ and the system has finite gain in R^N, namely, the (lower) value function

$$V_\gamma(x) = \inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \sup_{t\in\mathbb{R}_+} \int_0^t (h(y, \alpha[b], b) - \gamma^2|b|^2)ds, \tag{2.8}$$

is finite, equal to zero, and continuous on $\mathcal{T}$, that is, the above condition (ii) is satisfied for $\Omega = \mathbb{R}^N$ ((i) is automatically true in this case), we say that the $\mathcal{H}_\infty$ *suboptimal control problem* is solvable.

The solvability of the local $\mathcal{H}_\infty$ suboptimal control problem clearly implies that the undisturbed system is Lyapunov stable with respect to $\mathcal{T}$ by means of open-loop controls even if this is not explicitly required as for the solvability of the $\mathcal{H}_\infty$ suboptimal problem. We recall in fact that open-loop Lyapunov stability to $\mathcal{T}$ of the undisturbed system means that for any neighborhood $\mathcal{U} \supset \mathcal{T}$ there is $\Omega \supset \mathcal{T}$ such that if $x \in \Omega$ we can find a control $a \in \mathcal{A}$ so that the trajectory solution of

$$\dot{y} = f(y, a, 0), \qquad y(0) = x,$$

satisfies $y(t) \in \mathcal{U}$ for all $t \in \mathbb{R}_+$. If we also want our system to be asymptotically stable, then some condition on the sign of h stronger than (2.3) is to be assumed. We will discuss this point in Remark 3.3. We prefer to prove the stability of the system rather than assuming it in advance as in some previous literature, since as we will see in Remark 3.3, as a consequence of the proof of Theorem 3.2, it is almost a consequence of the finite gain condition.

Remark 2.2 (on previous literature). When applied to linear systems, our definition of $\mathcal{H}_\infty$ suboptimal control problem is less restrictive than the standard one (see, e.g., Basar and Bernhard [10] and the references therein), requiring the existence of a feedback control $a : \mathbb{R}^N \to A$ (or at least a dynamic state feedback control) such that the closed-loop system

$$\dot{y} = f(y, a(y), b), \qquad y(0) = x,$$

has a unique solution for all initial conditions, is asymptotically stable when $b \equiv 0$, and has L^2 gain less than or equal to γ; i.e., for all $y(0) = x \in \mathcal{T}$ the solution satisfies

$$\int_0^\infty h(y, a(y), b)ds \le \gamma^2 \int_0^\infty |b|^2 ds \quad \text{for all } b \in \mathcal{B}. \tag{2.9}$$

In fact, let us assume that the $\mathcal{H}_\infty$ suboptimal control problem is solvable in the sense above and h is nonnegative. If there is an optimal strategy for the value function (2.8), i.e., there is $\alpha \in \Delta$ such that

$$V_\gamma(x) = \sup_{b\in\mathcal{B}} \sup_{\mathbb{R}_+} \int_0^t (h(y, \alpha[b], b) - \gamma^2|b|^2)ds,$$

then we immediately obtain

$$\int_0^\infty h(y, \alpha[b], b)ds \le \gamma^2 \int_0^\infty |b|^2 ds + V_\gamma(x) \quad \text{for all } b \in \mathcal{B}. \tag{2.10}$$

If moreover the optimal strategy can be chosen in feedback form and the initial point of the dynamics is $x \in \mathcal{T}$, then (2.9) is satisfied. It is also clear that feedback stabilizability implies

open-loop Lyapunov asymptotic stability. It is important to observe however that the inequality (2.10) gives information about any point of the space (or at least of a neighborhood of $\mathcal{T}$ in the local problem) and not only about those of the equilibrium set. This is a natural additional request for nonlinear systems and was first proposed by Van der Shaft [30]. As a matter of fact, by known results (see, e.g., Van der Shaft [29] and our Theorem 3.2), when the system is linear and asymptotically stabilizable by linear feedback, (2.9) implies (2.10). For general nonlinear systems such an estimate for trajectories starting at points in a neighborhood of $\mathcal{T}$ cannot in general be deduced from (2.9). Note that if (2.10) holds for some strategy $\alpha \in \Delta$, then choosing $b \in \mathcal{B}$ with support in $[0, t]$, we easily obtain that, when h is nonnegative, α is optimal for $V_\gamma(x)$. The extra positive term in the right-hand side of (2.10) is needed when starting the system at $x \notin \mathcal{T}$, as one realizes choosing $b \equiv 0$, if $h(x, a, 0) > 0$ for all $a \in A$.

For nonlinear systems, the choice of the wider class of strategies that we consider is more appropriate than just feedback controls for developing the dynamic programming approach and the connections with the Hamilton-Jacobi equation. One can later look for optimal strategies in smaller classes. The class of so-called feedback strategies, i.e., causal functionals of the state, can also be used in our problem, with the same results that we prove, provided that the Isaacs condition (see Remark 3.3) is satisfied by the system. We will not discuss this point in detail; see, however, [28]. As much as the definition in the nonlinear case is concerned, according to Van der Shaft [30] (but relaxing the set of strategies), we solve the nonlinear $\mathcal{H}_\infty$ suboptimal control problem if we find a nonnegative function $U : \mathbb{R}^N \to \mathbb{R}$, which is null on $\mathcal{T}$, such that for all $x \in \mathbb{R}^N$ there is a strategy $\alpha_x \in \Delta$ satisfying

$$\int_0^t h(y, \alpha_x[b], b)ds \leq \gamma^2 \int_0^t |b|^2 ds + U(x) \quad \text{for all } b \in \mathcal{B} \text{ and } t \geq 0, \tag{2.11}$$

and the controls $\alpha_x[0]$ provide the open-loop local Lyapunov stability. Arguing as above, we can check that this is equivalent to solving the two following steps. First prove the solvability of the problem as previously defined and then prove the existence of optimal strategies at points of $\mathcal{T}$ (ideally showing that they can be chosen in feedback form) and that they satisfy the stability requirement. The main goal of this paper is to find necessary and sufficient conditions for the solution of the first step. Moreover we will show in the proof of Theorem 3.2 that the stability is guaranteed even by almost optimal strategies of the value function V_γ. We refer to the last section for a discussion and some results about the existence of optimal strategies. This part can be studied by fairly classical methods, as far as a nonconstructive proof is concerned, while it is almost open in the general case if we seek explicit formulas for optimal strategies, except for very special systems.

We need to mention that the stronger gain condition (2.10) or (2.11) involving points off $\mathcal{T}$ is not required in some other previous papers (see, e.g., Van der Shaft [29] and Ball, Helton, and Walker [3]); ours is probably the first to require the continuity of V_γ on $\mathcal{T}$. We feel that both requests provide desirable information about the nonlinear system, and as a matter of fact we will get them as consequences of the proof of Theorem 3.2, without need of additional assumptions, so we decided to include them in the definition. Moreover the definition we use allows us to characterize the solvability of the $\mathcal{H}_\infty$ suboptimal control problem in terms of the existence of suitable continuous solutions of the Hamilton-Jacobi-Isaacs equation, extending the results of the linear case.

We remark that the $\mathcal{H}_\infty$ problem is studied in the previous literature only with respect to a single equilibrium point of the system and not to a general closed set. We finally recall that in the literature, the $\mathcal{H}_\infty$ control problem is to find the smallest $\gamma^* \geq 0$ such that the $\mathcal{H}_\infty$ suboptimal control problem is solvable for all $\gamma > \gamma^*$.

Remark 2.3. For the definition of the local $\mathcal{H}_\infty$ suboptimal control problem, again we follow Van der Shaft [30]. However, as an alternative to our finite gain condition in Ω, one

may consider the function

$$V(x) = \inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \sup_{[0,t_x[} \int_0^t (h(y, \alpha[b], b) - \gamma^2|b|^2)ds, \tag{2.12}$$

rather than the one in (2.6). The statement of our main Theorem 3.2 with a similar proof will hold true even if we change (ii) and ask that V in (2.12) is finite in Ω, null and continuous on $\mathcal{T}$.

Remark 2.4. The function in (2.8) is called the lower value function of the differential game (2.1) with payoff functional given by

$$P(x, a, b) = \sup_{\mathbb{R}_+} \int_0^t (h(y, a, b) - \gamma^2|b|^2)ds,$$

which is a maximum cost type functional. The function in (2.6) is a localized version of (2.8) in Ω that we introduce for our problem. In this paper we will mostly study properties of functions like (2.8), and this will lead to prove properties also for the local functions (2.6) or (2.12). Our definition of the lower value function follows Elliott and Kalton [13] and the references therein. The definition is not symmetric since b maximizes using controls while a minimizes using strategies. As usual in the theory of differential games it is not possible, in general, to define a value function in any reasonable way. We decided here to develop the lower value approach, but using the upper value given by

$$\sup_{\beta\in\Gamma} \inf_{a\in\mathcal{A}} P(x, a, \beta[a]),$$

where Γ is the corresponding set of strategies for "player" b, we would obtain completely analogous results with obvious changes in the statements. We will come back to this and to the question of existence of value in Remark 3.5.

3. Viscosity solutions and main results. We start this section by recalling the definition of viscosity solution of a nonlinear, first-order, partial differential equation, in the discontinuous case (we refer to Crandall and Lions [12], Ishii [17], and the more recent Crandall, Ishii, and Lions [11] for more details).

Let $w : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^N$ open, be a locally bounded function. We define its lower and upper semicontinuous envelopes as, respectively,

$$w_*(x) = \lim_{r\to 0^+} \inf\{w(y) : |x - y| \le r\}, \qquad w^*(x) = \lim_{r\to 0^+} \sup\{w(y) : |x - y| \le r\}.$$

DEFINITION 3.1. *Let $F : \Omega \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ be a continuous function. The lower (resp., upper) semicontinuous function $u : \Omega \to \mathbb{R}$ is a viscosity supersolution (resp., subsolution) of*

$$F(x, u, Du) = 0 \quad \text{in } \Omega,$$

if for all $\varphi \in C^1(\Omega)$ and $x \in \operatorname{argmin}_{x\in\Omega}(u - \varphi)$, (resp., $x \in \operatorname{argmax}_{x\in\Omega}(u - \varphi)$), we have

$$F(x, u(x), D\varphi(x)) \ge 0 \qquad (\text{resp., } F(x, u(x), D\varphi(x)) \le 0).$$

We also say that $D\varphi(x) \in D^- u(x)$, the subdifferential of u at x (resp., $D\varphi(x) \in D^+ u(x)$, the superdifferential). A locally bounded function u is a viscosity solution of $F(x, u, Du) = 0$ if u_* is a supersolution and u^* is a subsolution.

We now state the main results of this paper, whose proof can be found at the end of §5.

THEOREM 3.2. *Consider the system* (2.1) *and assume* (2.2), (2.3). *Let* $\Omega \supset \mathcal{T}$ *be open, and suppose there is a continuous and nonnegative function* $U : \overline{\Omega} \to \mathbb{R}$ *such that* $U(x) \equiv U_0$ *if* $x \in \partial\Omega$, $U(x) < U_0$ *for* $x \in \Omega$, $U \equiv 0$ *on* $\mathcal{T}$, *and* U *is a viscosity supersolution of*

$$(3.1)\quad \mathcal{H}(x, DU(x)) = \inf_{b\in B}\sup_{a\in A}\{-f(x,a,b)\cdot DU(x) - h(x,a,b) + \gamma^2|b|^2\} \geq 0 \qquad \text{in } \Omega.$$

Then the system has finite gain in Ω *measured by* γ. *If the family of open sets* $\{x \in \Omega : U(x) < \varepsilon\}_{\varepsilon>0}$ *is a local base for* $\mathcal{T}$, *then the local* $\mathcal{H}_\infty$ *suboptimal control problem is solvable. If, moreover,* $\Omega = \mathbb{R}^N$, *then the* $\mathcal{H}_\infty$ *control problem is solvable.*

Remark 3.3. Observe that by Definition (3.1) of $\mathcal{H}$, the assumptions of Theorem 3.2 imply that the function U is also a supersolution of

$$\max_{a\in A}\{-f(x,a,0)\cdot DU - h(x,a,0)\} \geq 0 \quad \text{in } \Omega;$$

therefore by the sign condition (2.3) it can be seen as a Liapunov function for the undisturbed control system in the sense we defined in [27]. The stability of the undisturbed system that comes as a consequence of the proof of Theorem 3.2 is then not a surprise. If, moreover, the output h satisfies the condition

$$\text{for all } \varepsilon > 0, \qquad h(x,a,0) \geq C_\varepsilon > 0 \qquad \text{in } (\mathbb{R}^N \backslash B(\mathcal{T},\varepsilon)) \times A,$$

where $B(\mathcal{T},\varepsilon) = \{x : \operatorname{dist}(x,\mathcal{T}) < \varepsilon\}$ and A is compact, then the undisturbed system is also asymptotically stable, as proved in the paper by the author [26]. To achieve asymptotic stability, the previous condition can be relaxed by asking the so-called zero-state detectability, which we do not state here in detail. If the set $\mathcal{T}$ is compact, as for example in the standard case $\mathcal{T} = \{0\}$, then the result follows without the assumption on the sublevel sets of the continuous supersolution U of (3.1) by assuming instead that it is positive outside $\mathcal{T}$. To see this, we can apply the first part of the statement to the family $\{\Omega_\varepsilon\}_{\varepsilon>0}$ of open neighborhoods of $\mathcal{T}$ constructed as follows. Given a family $\{\mathcal{U}_\varepsilon\}_{\varepsilon>0}$ of compact neighborhoods of $\mathcal{T}$, define $\Omega_\varepsilon = \{x \in \mathcal{U}_\varepsilon : U(x) < \min_{\partial\mathcal{U}_\varepsilon} U\}$. Therefore when $\mathcal{T}$ is compact, the finite gain condition implies the open-loop stability if the value function V_γ is continuous and positive outside $\mathcal{T}$.

We observe that if the assumptions of the theorem are satisfied for some $\gamma^* > 0$, then they are obviously satisfied for all $\gamma > \gamma^*$ by the same function U. We also note that, since U of Theorem 3.2 is nonnegative, $U(x) = 0$ implies $0 \in D^-U(x)$; therefore $\mathcal{H}(x,0) \geq 0$, and then, by the assumption (2.3), there is $a \in A$ such that $h(x,a,0) = 0$. This in particular holds on $\mathcal{T}$ and is a necessary condition on the data for the assumptions of Theorem 3.2 to hold.

When studying the local $\mathcal{H}_\infty$ suboptimal control problem, it is interesting to have a priori information on the size of the neighborhood of $\mathcal{T}$ where the gain condition is satisfied. Theorem 3.2 provides an indirect answer to the question, saying that after computing, even locally around $\mathcal{T}$, a nonnegative, continuous viscosity supersolution of the Isaacs equation, which is null on $\mathcal{T}$, the system has a finite gain in any of its sublevel sets.

The next result completes the parallel with the theory of dissipative systems; see for example Willems [32], Hill and Moylan [16], and James [20], where U plays the role of storage function and V_γ defined in (2.8) plays the role of available storage. It also characterizes the lower value function V_γ when it is continuous, and this is useful to prove the existence of the value of the game for our problem; see Remarks 2.3 and 3.2.

THEOREM 3.4. *Assume* (2.2); *let* V_γ *be defined as in* (2.8) *and* $\Omega \subset \mathbb{R}^N$ *be open*; *and suppose that* V_γ *is locally bounded in* Ω.

(i) *Then* V_γ *is a viscosity solution of*

$$(3.2)\qquad \min\{\mathcal{H}(x, DV_\gamma(x)), V_\gamma\} = 0 \quad \text{in } \Omega$$

and, if it is continuous in $\Omega = \mathbb{R}^N$, *is the minimal continuous supersolution of* (3.2).

(ii) *If moreover* (2.3) *holds, then* V_γ *is a viscosity solution of*

$$\mathcal{H}(x, DV_\gamma(x)) = 0 \quad \text{in } \Omega \tag{3.3}$$

and, when continuous, is the minimal, nonnegative, continuous supersolution of (3.3).

Remark 3.5. Theorem 3.4 shows that a necessary condition for the solvability of the $\mathcal{H}_\infty$ suboptimal problem is the existence of a lower semicontinuous, nonnegative viscosity supersolution, null on $\mathcal{T}$ ($V_{\gamma *}$ to be specific), of

$$\mathcal{H}(x, DU(x)) \geq 0 \quad \text{in } \mathbb{R}^N.$$

Theorem 3.2 shows that a sufficient condition for the solvability is the existence of a continuus supersolution with the same sign properties. The gap between necessary and sufficient conditions can be filled up, when A is compact, by further extending the class of strategies, namely, allowing relaxed strategies or instead assuming the convexity of the sets $(f(x, A, b), h(x, A, b))$, for all x, b. The details will be presented elsewhere; see [34].

Our last remark concerns the upper value approach. When considering the upper value, the associated Hamiltonian would be $\overline{\mathcal{H}}(x, p)$ defined as in (3.1), interchanging the roles of inf and sup. We can get analogous results in this case, but of course this requires obvious modifications in the statements of the theorems. When the so-called Isaacs condition holds, namely if $\mathcal{H}(x, p) = \overline{\mathcal{H}}(x, p)$, for all $x, p \in \mathbb{R}^N$, then the two approaches are equivalent even if Theorem 3.4 does not prove, in general, that the lower (2.8) and the upper (see Remark 2.4) values coincide. The value does exist for our problem, as a consequence of the characterization of Theorem 3.4, when both upper and lower values are real valued and continuous in $\mathbb{R}^N$ (and the Isaacs condition holds).

Remark 3.6. Recently many efforts have been made to compute numerically the value function of differential games. We just refer here to the paper of Bardi, Falcone, and the author [1] and the references therein, which are more related to the techniques used in this paper, based on the study of the Isaacs equation. As pointed out in Remark 3.3 and as a result of Theorem 3.2, it looks interesting to study numerically the equations (3.2) or (3.3) to compute a solution (or at least a supersolution) or have a description of its sublevel sets. See also the concluding remark of the last section, where the importance of numerical approximations is pointed out for the computation of optimal strategies and feedbacks. We leave this as a future plan of research. Some results are already available, however, to compute the $\mathcal{H}_\infty$ norm in the case of dissipative systems (when the set A is a singleton) and are contained in a paper by James and Yuliar [22].

4. Some preliminary results. In this section we consider a slightly different problem from that of §2. The connection will become clear in the next section.

For the rest of this section we assume that A, B are closed, but we do not specify the classes of admissible controls $\mathcal{A}$, $\mathcal{B}$. We only require that if $a \in \mathcal{A}$, then $a : \mathbb{R}_+ \to A$ is measurable and that for all $s > 0$ we have $a(\cdot + s) \in \mathcal{A}$. Moreover if $a_1, a_2 \in \mathcal{A}$ and $t_1 > 0$, then also the measurable function defined by

$$a(t) = \begin{cases} a_1(t), & t \in [0, t_1[, \\ a_2(t - t_1), & t \in [t_1, +\infty[, \end{cases}$$

belongs to $\mathcal{A}$. The same properties will hold for the set of controls $\mathcal{B}$.

We also consider continuous functions $\phi : A \times B \to [1, +\infty)$ and $g : \mathbb{R}^P \times A \times B \to \mathbb{R}^P$ such that $g(\cdot, a, b)$ is continuous uniformly in a and b and $g(\cdot, \cdot, b)$ is bounded on $K \times A$ for any $K \subset \mathbb{R}^P$ compact. We assume that for all admissible controls $a \in \mathcal{A}, b \in \mathcal{B}$, there is a unique absolutely continuous, global solution $z(\cdot) = z_x(\cdot) = z_z(\cdot; a, b)$ of the Cauchy

problem

$$\dot{z}(\tau) = g(z(\tau), a(t(\tau)), b(t(\tau))), \quad z(0) = x, \tag{4.1}$$

where $t_{a,b}(\cdot) = t(\cdot) = \tau^{-1}(\cdot)$ and $\tau(\cdot) = \tau_{a,b}(\cdot)$ is the increasing change of parameter

$$\tau(t) = \int_0^t \phi(a(s), b(s))ds,$$

and suppose that it satisfies the condition

$$|z_x(s) - x| \leq \omega_R(\tau) \qquad \text{for all } |x| \leq R, s \in [0, \tau], a \in \mathcal{A}, b \in \mathcal{B}, \tag{4.2}$$

where $\omega_R : \mathbb{R}_+ \to \mathbb{R}_+$ is nondecreasing, continuous at zero and $\omega_R(0) = 0$ (ω_R is a modulus). In particular, from (4.2) we can conclude that $|z_x(\tau) - x| \leq o(1)$, as $\tau \to 0^+$, uniformly in a, b, and for x varying in a compact set. We will briefely denote in the following the reparametrized controls as $a(\tau) = a(t(\tau))$, $b(\tau) = b(t(\tau))$, for all $a \in \mathcal{A}$, $b \in \mathcal{B}$.

Remark 4.1. If, for example, the system (4.1) has unique global solution and g satisfies the growth condition $|g(x, a, b)| \leq L(1 + |x|)$, then for all $a \in \mathcal{A}$ and $b \in \mathcal{B}$, we have that

$$|z_x(s; a, b) - x| \leq \int_0^s |g(z, a, b)|dr \leq L[(1 + |x|)s + \int_0^s |z_x - x|dr] \qquad \text{for all } s \geq 0,$$

and then, by the Gronwall lemma,

$$|z_x(s) - x| \leq L[(1 + |x|)\tau] \exp(L\tau) \quad \text{for } s \in [0, \tau],$$

and (4.2) holds.

Assume also that $h = h(x, a, b)$ is continuous and satisfies the growth condition $|h(x, a, b)| \leq C_R$, for all $|x| \leq R, a \in A, b \in B$. Then we can also derive the estimate

$$\int_0^\tau |h(z_x, a, b)|ds \leq C_{R_\tau}\tau \qquad \text{for all } |x| \leq R, a \in \mathcal{A}, b \in \mathcal{B},$$

where we used the fact that $|z_x(s)| \leq |z_x(s) - x| + |x| \leq L(1 + R)\tau \exp(L\tau) + R =: R_\tau$, for all $|x| \leq R, s \in [0, \tau]$. Therefore also the trajectories $z = (y, r)$ of the dynamical system

$$\begin{cases} y' = g(y, a, b), & y(0) = x, \\ r' = h(y, a, b), & (0) = r_0, \end{cases}$$

have the property (4.2).

We continue considering the following class of functions. Let $\Omega \subset \mathbb{R}^P$ open, $\lambda \geq 0$, and $u : \overline{\Omega} \to \mathbb{R}$ be a continuous function. We define a (lower) value function for the differential game (4.1) by setting

$$V^\lambda(x) = \inf_{\alpha \in \Delta} \sup_{b \in \mathcal{B}} \sup_{[0,\tau_x]} \left\{ \int_0^\tau \exp(-\lambda s) l(z_x(s), \alpha[b](t(s)), b(t(s)))ds \right. \tag{4.3}$$
$$\left. + \exp(-\lambda\tau)\, u(z_x(\tau)) \right\},$$

where $l : \mathbb{R}^P \times A \times B \to \mathbb{R}$ is continuous, $l(\cdot, a, b)$ is continuous uniformly in a, b and satisfies $|l(x, a, b)| \leq C_R$ for all $|x| \leq R, a, b$, $R > 0$, z_x is here the solution of (4.1), and τ_x

is the exit time of $z_x(\cdot)$ from Ω. It is clear by definition that $V^\lambda \geq u$. Throughout this section we will assume that V^λ is locally bounded.

We also denote the Hamiltonian function associated with the differential game as

$$H(x, p) = \inf_{b\in B} \sup_{a\in A}\{-g(x, a, b) \cdot p - l(x, a, b)\}$$

and assume it is continuous in $\mathbb{R}^{2P}$.

We will prove that the value function (4.3), under some assumptions on the functions involved, namely u, l, g, H, and on the superdifferential of V^λ itself, satisfies a suitable variational inequality in the viscosity sense. This proof, besides a technical lemma, combines modifications of several known arguments, in particular of Evans and Souganidis [14], Ishii [17], and the author [27], but we will give it for the sake of completeness. The main new difficulties of which we need to take care are the unboundedness of the control sets and the presence of reparametrized controls in the dynamics (4.1).

Remark 4.2. By the definition of the reparametrization, we observe that for all $a \in \mathcal{A}$, $b \in \mathcal{B}$, and $t_1, t_2 > 0$, we have

$$\tau_{a,b}(t_1 + t_2) = \tau_{a,b}(t_1) + \tau_{a(\cdot+t_1),b(\cdot+t_1)}(t_2),$$

and then for all $\tau_1, \tau_2 > 0$,

$$t_{a,b}(\tau_1 + \tau_2) = t_{a,b}(\tau_1) + t_{a(\cdot+t_{a,b}(\tau_1)),b(\cdot+t_{a,b}(\tau_1))}(\tau_2).$$

As a consequence of this equality, for all $a \in \mathcal{A}$, $b \in \mathcal{B}$, and $\tau_1 > 0$, we get

$$a(t_{a,b}(\tau_1 + \cdot)) = a(t_{a,b}(\tau_1) + t_{a(\cdot+t_{a,b}(\tau_1)),b(\cdot+t_{a,b}(\tau_1))}(\cdot)).$$

Moreover for all $a_1, a_2 \in \mathcal{A}$, $b_1, b_2 \in \mathcal{B}$, $\tau_1 > 0$, if we define

$$a(t) = \begin{cases} a_1(t), & [0, t_{a_1,b_1}(\tau_1)), \\ a_2(t - t_{a_1,b_1}(\tau_1)), & [t_{a_1,b_1}(\tau_1), +\infty), \end{cases}$$

and $b(\cdot)$ correspondingly, then we have

$$\begin{aligned} a(t_{a,b}(\tau)) &= \begin{cases} a(t_{a_1,b_1}(\tau)), & [0, \tau_1), \\ a(t_{a_1,b_1}(\tau_1) + t_{a_2,b_2}(\tau - \tau_1)), & [\tau_1, +\infty), \end{cases} \\ &= \begin{cases} a_1(t_{a_1,b_1}(\tau)), & [0, \tau_1), \\ a_2(t_{a_2,b_2}(\tau - \tau_1)), & [\tau_1, +\infty). \end{cases} \end{aligned}$$

The same properties hold true for the elements of the set of controls $\mathcal{B}$.

The reparametrization functional also has the following nonanticipating property. For any fixed $\overline{\tau} > 0$ and $a_1, a_2 \in \mathcal{A}$, $b_1, b_2 \in \mathcal{B}$, if $a_1 = a_2$ and $b_1 = b_2$ a.e. in $[0, t_{a_1,b_1}(\overline{\tau})]$, then $t_{a_1,b_1}(\overline{\tau}) = t_{a_2,b_2}(\overline{\tau})$. In fact we have by definition

$$\tau_{a_2,b_2}(t_{a_1,b_1}(\overline{\tau})) = \int_0^{t_{a_1,b_1}(\overline{\tau})} \phi(a_2(s), b_2(s))ds = \int_0^{t_{a_1,b_1}(\overline{\tau})} \phi(a_1(s), b_1(s))ds = \overline{\tau}.$$

The following lemma is an important step of the proof and can be viewed as the dynamic programming principle for (4.3) (see also [27]). Main ingredients of the proof are contained in Remark 4.2.

LEMMA 4.3. *For all $\tau \geq 0$ and $x \in \Omega$ we have that*

$$V^\lambda(x) \geq \inf_{\alpha\in\Delta}\sup_{b\in\mathcal{B}}\left\{\int_0^{\tau\wedge\tau_x}\exp(-\lambda s)l(z_x,\alpha[b],b)ds + \exp(-\lambda\tau\wedge\tau_x)\,V^\lambda(z_x(\tau\wedge\tau_x))\right\}. \tag{4.4}$$

Moreover if $V^{\lambda *}(x) > u(x)$, for all sequences $x_n \to x$ such that $V^\lambda(x_n) \to V^{\lambda *}(x)$ there is $\varepsilon > 0$ such that the equality holds in (4.4), *at x_n, for all $\tau \in [0,\varepsilon]$ and $|x_n - x| < \varepsilon$.*

Proof. 1. By Definition (4.3) of the value function, for every $\varepsilon > 0$ there is $\alpha \in \Delta$ such that

$$V^\lambda(x) + \varepsilon \geq \sup_{[0,\tau_x]}\left\{\int_0^\tau \exp(-\lambda s)l(z_x,\alpha[b],b)ds + \exp(-\lambda\tau)\,u(z_x(\tau))\right\} \quad \text{for all } b \in \mathcal{B}.$$

For any fixed $s > 0$, we get, for all $b \in \mathcal{B}$,

$$V^\lambda(x) + \varepsilon \geq \int_0^{s\wedge\tau_x}\exp(-\lambda r)l(z_x,\alpha[b],b)dr + \exp(-\lambda s\wedge\tau_x)\cdot\sup_{[0,\tau_{z_x}(s\wedge\tau_x)]}\left\{\int_0^\tau \exp(-\lambda r)l(z_{z_x(s\wedge\tau_x)},\alpha[b](t(r+s\wedge\tau_x)),b(t(r+s\wedge\tau_x)))dr + \exp(-\lambda\tau)\,u(z_{z_x(s\wedge\tau_x)}(\tau))\right\}.$$

Then we have, also using Remark 4.2 and the properties of the sets $\mathcal{A}$, $\mathcal{B}$,

$$V^\lambda(x) + \varepsilon \geq \int_0^{s\wedge\tau_x}\exp(-\lambda r)l(z_x,\alpha[b],b)dr + \exp(-\lambda s\wedge\tau_x)V^\lambda(z_x(s\wedge\tau_x)) \quad \text{for all } b,$$

and the inequality (4.4) follows since ε is arbitrary.

2. We now assume that $x \in \Omega$ and $V^{\lambda *}(x) > u(x)$. If the statement was false, we could find sequences $|x_n - x| < 1/n$, $0 < \tau_n \leq 1/n$ such that $V^\lambda(x_n) \to V^{\lambda *}(x)$ and

$$V^\lambda(x_n) > \inf_{\alpha\in\Delta}\sup_{b\in\mathcal{B}}\left\{\int_0^{\tau_{x_n}\wedge\tau_n}\exp(-\lambda r)l(z_{x_n},\alpha_n[b_n],b_n)dr + \exp(-\lambda\tau_n\wedge\tau_{x_n})\,V^\lambda(z_{x_n}(\tau_n\wedge\tau_{x_n}))\right\}. \tag{4.5}$$

By definition of $V^\lambda(x_n)$, using Remark 4.2 and (4.5), we can construct appropriately $\alpha_n \in \Delta$, $b_n \in \mathcal{B}$ and choose $\varepsilon_n \to 0^+$ such that

$$\sup_{[\tau_n\wedge\tau_{x_n},\tau_{x_n}]}\left\{\int_0^\tau \exp(-\lambda r)l(z_{x_n},\alpha_n[b_n],b_n)dr + \exp(-\lambda\tau)\,u(z_{x_n}(\tau))\right\} \leq V^\lambda(x_n) - \varepsilon_n < \sup_{[0,\tau_{x_n}]}\left\{\int_0^\tau \exp(-\lambda r)l(z_{x_n},\alpha_n[b_n],b_n)dr + \exp(-\lambda\tau)\,u(z_{x_n}(\tau))\right\}$$

and then

$$V^\lambda(x_n) - \varepsilon_n < \sup_{[0,\tau_n\wedge\tau_{x_n}]}\left\{\int_0^\tau \exp(-\lambda r)l(z_{x_n},\alpha_n[b_n],b_n)dr + \exp(-\lambda\tau)\,u(z_{x_n}(\tau))\right\}.$$

Therefore for some sequence $0 < s_n \le \tau_n \wedge \tau_{x_n} \le 1/n$, we have

$$\int_0^{s_n} \exp(-\lambda r) l(z_{x_n}, \alpha_n[b_n], b_n) dr + \exp(-\lambda s_n)\, u(z_{x_n}(s_n)) > V^\lambda(x_n) - \varepsilon_n. \tag{4.6}$$

By assumption (4.2) we know that

$$\sup_{[0,s_n]} |z_{x_n}(\tau) - x_n| \le \omega(s_n) = o(1) \quad \text{as } n \to +\infty.$$

If n is large, by (4.6) and the assumption on l, we then obtain for some $R > 0$ independent of n that

$$C_R s_n + \exp(-\lambda s_n) \sup_{B(x, o(1)+1/n)} u \ge V^\lambda(x_n) - \varepsilon_n.$$

As $n \to +\infty$ we finally get $u(x) = u^*(x) \ge V^{\lambda *}(x)$, a contradiction. □

We can now prove the relationship between the value function and the corresponding Hamilton-Jacobi equation.

PROPOSITION 4.4. *Assume that the following hold. For all $x_0 \in \Omega$, if $V^{\lambda *}(x_0) > u(x_0)$ and $p_0 \in D^+ V^{\lambda *}(x_0)$*

$$\sup_{a \in A, \sigma > 0} \inf \{-g(x, a, b) \cdot p - l(x, a, b) : |x - x_0| \le \sigma, |p - p_0| \le \sigma, |b| \ge 1/\sigma\} > -\lambda V^{\lambda *}(x_0). \tag{4.7}$$

Then for $\lambda \ge 0$, the value function V^λ is a viscosity solution of

$$\min\{\lambda V + H(x, DV), V - u\} = 0 \qquad \text{in } \Omega.$$

Proof. 1. We start by proving that $V (= V^\lambda)$ is a subsolution. Let $x_0 \in \arg\max(V^* - \varphi)$, where $\varphi \in C^1(\Omega)$ and $V^*(x_0) = \varphi(x_0)$. Observe that $V^*(z) \le \varphi(z)$, if $z \in \Omega$. We assume by contradiction that $V^*(x_0) > u(x_0)$ and

$$\lambda\, \varphi(x_0) + H(x_0, D\varphi(x_0)) > 0. \tag{4.8}$$

By (4.7) at $(x_0, D\varphi(x_0))$, we can find σ, $R > 0$ and $a_0 \in A$ such that

$$\lambda\varphi(z) - g(z, a_0, b) \cdot D\varphi(z) - l(z, a_0, b) \ge r_0 > 0 \quad \text{for all } z \in B(x_0, \sigma), |b| \ge R. \tag{4.9}$$

Moreover by (4.8), for all $\bar{b} \in B$, $|\bar{b}| \le R$, there are $\bar{r} > 0, \bar{a} \in A$ such that

$$\lambda\, \varphi(z) - g(z, \bar{a}, b) \cdot D\varphi(z) - l(z, \bar{a}, b) \ge \bar{r}, \qquad z \in B(x_0, \bar{r}), b \in B(\bar{b}, \bar{r}). \tag{4.10}$$

By the compactness of $\{b \in B : |b| \le R\}$, let $\{B(b_i, r_i)\}_{i=1,\dots,n}$ be a finite cover. Using (4.9) or (4.10), we proved that there are $\varepsilon > 0$ and $a_0, \dots, a_n \in A$ so that for all $b \in B$ there is $i \in \{0, \dots, n\}$ that satisfies

$$\lambda\, \varphi(z) - g(z, a_i, b) \cdot D\varphi(z) - l(z, a_i, b) \ge \varepsilon \quad \text{in } B(x_0, \varepsilon) \subset \Omega.$$

We now define the strategy $\alpha \in \Delta$ by the position

$$\alpha[b](t) = \begin{cases} a_0 & \text{if } |b(t)| > R, \\ \alpha[b](t) = a_i & \text{if } |b(t)| \le R \text{ and } b(t) \in B(b_i, r_i) \setminus \bigcap_{0<j<i} B(b_j, r_j). \end{cases}$$

Therefore by (4.2), we can select some small $\tau > 0$, such that

(4.11)
$$\lambda\,\varphi(z_x(s)) - g(z_x(s), \alpha[b](t(s)), b(t(s))) \cdot D\varphi(z_x(s)) - l(z_x(s), \alpha[b](t(s)), b(t(s))) \geq \varepsilon,$$

for all $b \in \mathcal{B}$, $s \in (0, \tau)$, $x \in B(x_0, \varepsilon/2)$, since we can assume $z_x(s) \in B(x_0, \varepsilon)$, for $s \in (0, \tau)$. If we multiply by $\exp(-\lambda s)$ and integrate on $(0, \tau)$, we obtain (also since $\tau = \tau \wedge \tau_x$ and with the agreement that the right-hand side reads $\varepsilon\tau \wedge \tau_x$ if $\lambda = 0$)

$$\begin{aligned}(4.12)\qquad \varphi(x) - \exp(-\lambda\tau \wedge \tau_x)\,\varphi(z_x(\tau \wedge \tau_x)) - \int_0^{\tau\wedge\tau_x} \exp(-\lambda r) l(z_x, \alpha[b], b)dr \\ \geq \varepsilon(1 - \exp(-\lambda\tau \wedge \tau_x))/\lambda \geq \sigma > 0 \quad \text{for all } b \in \mathcal{B},\ x \in B(x_0, \varepsilon/2).\end{aligned}$$

By Lemma 4.3, we can choose $\delta > 0$ and $x_n \to x$ such that $V(x_n) \to V^*(x_o)$, $|x_n - x_0| < \delta$, and

$$\begin{aligned}(4.13)\quad V(x_n) = \inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \left\{ \int_0^{s\wedge\tau_{x_n}} \exp(-\lambda r) l(z_{x_n}, \alpha[b], b)dr \right. \\ \left. + \exp(-\lambda s \wedge \tau_{x_n})\, V(z_{x_n}(s \wedge \tau_{x_n})) \right\}, \quad s \in (0, \delta).\end{aligned}$$

Using (4.12) at such points x_n for large values of n and the properties of φ, we obtain, for fixed τ,

$$\begin{aligned}V^*(x_0) - \exp(-\lambda\tau \wedge \tau_{x_n})\, V^*(z_{x_n}(\tau \wedge \tau_x)) + \varphi(x_n) - \varphi(x_0) \\ - \int_0^{\tau\wedge\tau_{x_n}} \exp(-\lambda r) l(z_{x_n}, \alpha[b], b)dr \geq \sigma \quad \text{for all } b \in \mathcal{B}.\end{aligned}$$

We now use (4.13) to get (we may suppose $\tau < \delta$)

$$V^*(x_0) \geq \sigma + \varphi(x_0) - \varphi(x_n) + V(x_n).$$

As $n \to +\infty$, we obtain the required contradiction.

2. We now proceed with the easier proof that V is a supersolution. We observe that of course $V_* \geq u$, since $V \geq u$ and u is continuous. Let $x_0 \in \arg\min(V_* - \varphi)$, where $\varphi \in C^1(\Omega)$ and $V_*(x_0) = \varphi(x_0)$. We assume by contradiction that

$$\lambda\,\varphi(x_0) + H(x_0, D\varphi(x_0)) < 0.$$

This implies that we can find $\bar{b} \in B$ and $\sigma > 0$ such that

$$\lambda\,\varphi(x_0) - g(x_0, a, \bar{b}) \cdot D\varphi(x_0) - l(x_0, a, \bar{b}) < -\sigma \text{ for all } a \in A.$$

By the assumptions on g and l, there exists $\varepsilon > 0$ such that

$$\lambda\,\varphi(z) - g(z, a, \bar{b}) \cdot D\varphi(z) - l(z, a, \bar{b}) \leq -\varepsilon \qquad \text{for all } a \in A,\ z \in B(x_0, \varepsilon) \subset \Omega.$$

Let $b(\cdot) \equiv \bar{b}$ be a constant control. By (4.2), we have that $|z_x(s) - x| \leq o(1)$ as $s \to 0^+$, uniformly in $x \in B(x_0, \varepsilon/2)$ and $a \in \mathcal{A}$. Therefore there exists $\tau > 0$ small enough such that $\tau < \tau_x(\alpha[b], b)$, for all $\alpha \in \Delta$, $x \in B(x_0, \varepsilon/2)$, and

$$\begin{aligned}\lambda\,\varphi(z_x(s)) - g(z_x(s), \alpha[b](s), b(s)) \cdot D\varphi(z_x(s)) - l(z_x(s), \alpha[b](s), b(s)) \leq -\varepsilon \\ \text{for all } \alpha \in \Delta,\ s \in (0, \tau).\end{aligned}$$

If we multiply by $\exp(-\lambda s)$ and integrate on $(0, \tau)$, we obtain (again the right-hand side is $-\varepsilon\tau$ if $\lambda = 0$)

$$\varphi(x) - \exp(-\lambda\tau)\,\varphi(z_x(\tau)) - \int_0^\tau \exp(-\lambda r) l(z_x, \alpha[b], b) dr \le -\varepsilon(1 - \exp(-\lambda\tau))/\lambda \le -\sigma < 0.$$

By the properties of φ we then have

$$V_*(x_0) \le -\sigma + \varphi(x_0) - \varphi(x) + \exp(-\lambda\tau)\, V(z_x(\tau)) + \int_0^\tau \exp(-\lambda r) l(z_x, \alpha[b], b) dr \quad \text{for all } \alpha \in \Delta,\ x \in B(x_0, \varepsilon/2).$$

We now use (4.4) and get

$$V_*(x_0) \le -\sigma + \varphi(x_0) - \varphi(x) + V(x) \quad \text{for all } x \in B(x_0, \varepsilon/2),$$

therefore a contradiction. □

5. The $\mathcal{H}_\infty$ suboptimal control problem. We now go back to our problem and recast it into the setting of the previous section. We first recall the Hamiltonian that one expects to be related to the value function (2.8) and precisely

$$\mathcal{H}(x, p) = \inf_{b \in B} \sup_{a \in A} \{-f(x, a, b) \cdot p - h(x, a, b) + \gamma^2 |b|^2\}. \tag{5.1}$$

PROPOSITION 5.1. *Assume* (2.2). *Then the Hamiltonian* (5.1) *satisfies the following. For all $R > 0$, we can find $C_R > 0$ such that*

$$|\mathcal{H}(x, p) - \mathcal{H}(y, p')| \le C_R[(1 + |p| \vee |p'|)^{qr} |p - p'| + (1 + |p| \vee |p'|)^{2r} |x - y|],$$

$$|\mathcal{H}(x, p)| \le C_R(1 + |p| \vee |p'|)^{2/(2-q)} \qquad \textit{for all } |x|, |y| \le R,\ p, p' \in \mathbb{R}^N,$$

where $r = 4/(2 - q)$[3]. *In particular $\mathcal{H}$ is locally Lipschitz continuous in $\mathbb{R}^{2N}$. Moreover for all $|x|, |p| \le R$,*

$$\begin{aligned} \mathcal{H}(x, p) &= \min_{b \in B_R} \sup_{a \in A} \{-f(x, a, b) \cdot p - h(x, a, b) + \gamma^2 |b|^2\} \\ &= \min_{b \in B_R} \max_{a \in A_R} \{-f(x, a, b) \cdot p - h(x, a, b) + \gamma^2 |b|^2\}, \end{aligned} \tag{5.2}$$

where $A_R = \{a \in A : |a| \le C_R\}$, $B_R = \{b \in B : |b| \le C_R\}$.

Proof. To prove the assertions, let $R > 0$ and $|x|, |y| \le R$, $p, p' \in \mathbb{R}^N$. For all $b \in B$ and $\varepsilon > 0$, we can find $a_b \in A$ such that by (2.2)

$$\begin{aligned} \mathcal{H}(x, p) - \varepsilon &\le -f(x, a_b, b) \cdot p - h(x, a_b, b) + \gamma^2 |b|^2 \\ &\le L(1 + |x| + |b|^q)|p| + L_R(1 + |b|^2) + (L|p| + L_R)|a_b|^q - C_R^0 |a_b|^2 \\ &\le C_R[(1 + |p|)^{2/(2-q)} + (1 + |b|^2)(1 + |p|)], \end{aligned} \tag{5.3}$$

as easily checked, where C_R does not depend on b and a_b. In particular for $b = 0$ and since ε is arbitrary,

$$\mathcal{H}(x, p) \le 2C_R(1 + |p|)^{2/(2-q)}.$$

We also have that for $\varepsilon > 0$ there exists $\overline{b} \in B$ such that, with the choice of $a = 0$,

$$\mathcal{H}(y, p') + \varepsilon \geq \sup_{a \in A} \{-f(y, a, \overline{b}) \cdot p' - h(y, a, \overline{b}) + \gamma^2 |\overline{b}|^2\} \tag{5.4}$$
$$\geq -L(1 + |y|)|p'| - L_R - (L|p'| + L_R)|\overline{b}|^q + \gamma^2 |\overline{b}|^2 \geq -C_R(1 + |p'|)^{2/(2-q)},$$

from which, for $(y, p') = (x, p)$, we get the second inequality. Moreover if $|\overline{b}| \geq 1$, we necessarily have by the second inequality in (5.4) that

$$\gamma^2 |\overline{b}|^{2-q} \leq L(1 + |y|)|p'| + 2L_R + L|p'| + \varepsilon + 2C_R(1 + |p'|)^{2/(2-q)},$$

and then

$$|\overline{b}|^2 \leq C_R^{\#}(1 + |p'|)^{4/(2-q)^2} \quad \text{for all } |y| \leq R,\ p' \in \mathbb{R}^N. \tag{5.5}$$

The first inequality of the statement now follows from the estimate (5.5) and the corresponding one for $a_{\overline{b}}$. The latter can be obtained from the second inequality in (5.3) at $\overline{b}$, namely, if $|a_{\overline{b}}| \geq 1$,

$$C_R^0 |a_{\overline{b}}|^{2-q} \leq C_R^{\#\#}(1 + |p| \vee |p'|)^{4/(2-q)^2}. \tag{5.6}$$

Combining the first inequalities in (5.3), (5.4) we finally get, with the choice of $\overline{b}$, $a_{\overline{b}}$ as above,

$$\mathcal{H}(x, p) - \mathcal{H}(y, p') - 2\varepsilon \leq f(y, a_{\overline{b}}, \overline{b}) \cdot p' - f(x, a_{\overline{b}}, \overline{b}) \cdot p + h(y, a_{\overline{b}}, \overline{b}) - h(x, a_{\overline{b}}, \overline{b})$$
$$\leq L(1+|a_{\overline{b}}|^q+|\overline{b}|^q)|p||x-y|+L_R(1+|a_{\overline{b}}|^2+|\overline{b}|^2)|x-y|+L(1+|x|+|a_{\overline{b}}|^q+|\overline{b}|^q)|p-p'|,$$

and then the conclusion by the two estimates (5.5) and (5.6).

The equality (5.2) again follows easily combining the two estimates (5.5), (5.6) and the inequalities (5.3), (5.4). □

The estimates of Proposition 5.1, while proving continuity of $\mathcal{H}$, are not sufficiently good to deal with this Hamiltonian directly, so we are motivated to reformulate the problem in a convenient way.

We now begin with the proof of Theorem 3.2. We will need several steps to prove the result.

LEMMA 5.2. *Assume* (2.2), *and let* $U : \Omega \to \mathbb{R}$ *be a continuous viscosity supersolution of* $\mathcal{H}(x, DU) \geq 0$, *in* Ω. *Let* $\rho : \mathbb{R} \to \mathbb{R}_+$ *be bounded, smooth such that* $M \geq \dot{\rho} > 0$ *and* $\rho(s) \to 0$, *as* $s \to -\infty$. *Define the nonnegative, bounded function* $u(x, r) = \rho(U(x) + r)$. *Then* u *is a viscosity supersolution of*

$$H(z, Du(z)) \geq 0 \quad \text{in } \Omega \times \mathbb{R}, \tag{5.7}$$

where we indicated by $z = (x, r)$, $H(z, p) = \inf_{b \in B} \sup_{a \in A} \{-\overline{g}(z, a, b) \cdot p\}$, $\overline{g}(z, a, b) = g(z, a, b)/(1+|a|^2+|b|^2) = (f(x, a, b), h(x, a, b) - \gamma^2|b|^2)/(1+|a|^2+|b|^2) = (\overline{f}(x, a, b), \overline{h}(x, a, b) - \gamma^2|b|^2/(1 + |a|^2 + |b|^2))$.

Proof. The proof is an easy consequence of the usual formulas of change of variables; see Crandall and Lions [12]. Indeed if $p_z \in D^- u(z)$, then $p_z = \dot{\rho}(\rho^{-1}(u))(p, 1)$, where $p \in D^- U(x)$. Therefore by the definition (5.1), for all $\varepsilon > 0$ and $b \in B$ there is $a_b \in A$ such that

$$-f(x, a_b, b) \cdot p - h(x, a_b, b) + \gamma|b|^2 \geq -\varepsilon;$$

hence

$$-\overline{g}(z, a_b, b) \cdot p_z \geq -\varepsilon\dot{\rho}(\rho^{-1}(u))/(1 + |a|^2 + |b|^2) \geq -M\varepsilon$$

and the conclusion follows. □

Remark 5.3. In this long remark we want to check the assumptions of §4 for the differential game associated with the Hamiltonian H. For all $a \in \mathcal{A}, b \in \mathcal{B}$, we consider the increasing change of parameter

$$\tau(t) = \tau_{a,b}(t) = \int_0^t (1 + |a|^2 + |b|^2)ds = t + ||a||^2_{L^2(0,t)} + ||b||^2_{L^2(0,t)}$$

and set $t(\cdot) = t_{a,b}(\cdot) = \tau^{-1}(\cdot)$. We will briefely denote in the following the reparametrized measurable functions as $a(\tau) = a(t(\tau))$, $b(\tau) = b(t(\tau))$ and, for such controls, solve the dynamical system

$$z'(\tau) = \overline{g}(z(\tau), a(\tau), b(\tau)), \quad z(0) \in \mathbb{R}^{N+1}. \tag{5.8}$$

For all $z(0) \in \mathbb{R}^{N+1}, a \in \mathcal{A}, b \in \mathcal{B}$, there is a unique absolutely continuous global solution to (5.8).

The system with reparametrized controls satisfies (4.2). Indeed by the assumption (2.2) it is immediate to recognize that we can apply Remark 4.1 to the vector field $\overline{g}$. Observe that we can pass from the system (2.1) to (5.8) by means of the above introduced change of parameter on the trajectories solutions. Indeed if $z(\tau)$ solves (5.8), then

$$z(\tau) = (y_x(t(\tau)), r(t(\tau))),$$

where $y_x(\cdot)$ is the solution of (2.1) corresponding to the choice of controls $(a(t), b(t))$ and $r(t(\tau)) = r_0 + \int_0^{t(\tau)} (h(y_x(t), a(t), b(t)) - \gamma^2|b|^2(t))dt$. As a notation, we will indicate by y' the derivative with respect to the new parameter τ, while $\dot{y}$ indicates the derivative with respect to the old one t, i.e., $y'(\tau) = \dot{y}(t(\tau))/(1 + |a|^2 + |b|^2)$.

This idea of reparametrizing trajectories is fairly classical in control theory and was already used in the context of the dynamic programming approach by Barles [6] to study a class of unbounded control problems. Here everything is slightly more complicated since we are not just dealing with controls in L^∞ as in his case and our problem is a differential game.

The Hamiltonian in (5.7) satisfies the following regularity condition, as easily checked:

$$|H(z, p_z) - H(\overline{z}, p_{\overline{z}})| \le L_R(|p_z||z - \overline{z}| + |p_z - p_{\overline{z}}|) \text{ for all } |z|, |\overline{z}| \le R, p_z, p_{\overline{z}} \in \mathbb{R}^{N+1}; \tag{5.9}$$

in particular, it is locally Lipschitz continuous.

Let U, u be as in Lemma 5.2. If $z(\cdot)$ is a solution of (5.8), $z(0) = (x, r_0)$, we easily compute

$$\begin{aligned} u(z(\tau)) &= \rho\left(U(y_x(t(\tau))) + r_0 + \int_0^\tau (h(y, a, b) - \gamma^2|b|^2)/(1 + |a|^2 + |b|^2)\, d\tau\right) \\ &= \rho\left(U(y_x(t(\tau))) + r_0 + \int_0^{t(\tau)} (h(y, a, b) - \gamma^2|b|^2)dt\right). \end{aligned} \tag{5.10}$$

We also observe that the exit time from $\Omega \times \mathbb{R}$ satisfies $\tau_z(a, b) = \tau_x(a, b) = \int_0^{t_x(a,b)} (1 + |a|^2 + |b|^2)dt$, where τ_x denotes the exit time of the first N components of the solution of (5.8) from Ω and, as before, $t_x(a, b)$ is the exit time of the solution of (2.1) from Ω. Therefore for all $\lambda \ge 0$, the value function in (4.3) for the system (5.8), with $l \equiv 0$, takes the form

$$V^\lambda(z) = \inf_{\alpha\in\Delta}\sup_{b\in\mathcal{B}}\sup_{[0,\tau_z)} \exp(-\lambda\tau)u(z(\tau))$$

$$= \inf_{\alpha\in\Delta}\sup_{b\in\mathcal{B}}\sup_{[0,t_x)} \left\{\exp\left(-\lambda\int_0^t (1+|a|^2+|b|^2)dt\right)\right.$$

$$\left.\cdot\rho\big(U(y_x(t))+r_0+\int_0^t (h(y,a,b)-\gamma^2|b|^2)dt\right)\Big\}.$$

Since by definition ρ is nonnegative and bounded, then also V^λ is nonnegative and bounded. In particular, since ρ is increasing, we also have that

$$V^0(x,0) = \rho\left(\inf_{\alpha\in\Delta}\sup_{b\in\mathcal{B}}\sup_{[0,t_x)}\left\{U(y_x(t))+\int_0^t (h(y,a,b)-\gamma^2|b|^2)dt\right\}\right).$$

LEMMA 5.4. *Assume* (2.2), *and let U, u be as in Lemma* 5.2, *U defined and continuous in* $\overline{\Omega}$. *Then for all* $\lambda > 0$, V^λ *is a viscosity solution of*

$$\min\{\lambda V + H(z, DV), V - u\} = 0 \qquad \text{in } \Omega\times\mathbb{R}. \tag{5.11}$$

Proof. 1. Let $z_0 \in \Omega\times\mathbb{R}$. We start proving that, if $p \in D^+V^{\lambda *}(z_0)$ and $V^{\lambda *}(z_0) > 0$, then $p_{N+1} > 0$. We recall that $p \in D^+V^{\lambda *}(z_0)$ is equivalent to

$$\limsup_{z\to z_0} (V^{\lambda *}(z) - V^{\lambda *}(z_0) - p\cdot(z-z_0))/|z-z_0| \le 0.$$

In particular for $e_{N+1} = (0,\dots,0,1)$, this implies

$$\limsup_{\sigma\to 0^+} (V^{\lambda *}(z_0+\sigma e_{N+1}) - V^{\lambda *}(z_0))/\sigma \le p_{N+1}.$$

Let $\bar\varepsilon > 0$ and $z_n \to z_0$ be such that $0 < 2\bar\varepsilon \le V^\lambda(z_n) \to V^{\lambda *}(z_0)$. For all $0 < \varepsilon \le \bar\varepsilon, \sigma > 0$ small enough, and n fixed, we can find $\alpha \in \Delta$ such that

$$V^\lambda(z_n+\sigma e_{N+1}) + \varepsilon\sigma > \sup_{b\in\mathcal{B}}\sup_{[0,\tau_{z_n+\sigma e_{N+1}})} \exp(-\lambda\tau)u(z_{z_n+\sigma e_{N+1}}(\tau)).$$

Then for such α, we can choose $b \in \mathcal{B}$, $\tau \in [0, \tau_{z_n+\sigma e_{N+1}})$ so that (observe that, by Remark 5.3, $\tau_{z_n} = \tau_{z_n+\sigma e_{N+1}}$)

$$\bar\varepsilon \le V^\lambda(z_n) - \varepsilon\sigma \le \exp(-\lambda\tau)u(z_{z_n}(\tau));$$

then $\tau \le C$, $\rho^{-1}(u(z_{z_n}(\tau))) \ge \rho^{-1}(\bar\varepsilon) \ge -C$, and by (4.2) $|z_{z_n}(\tau)| \le C$, where C is independent of ε, n, and σ. Moreover by (5.10) and the definition of u we have that

$$V^\lambda(z_n+\sigma e_{N+1}) - V^\lambda(z_n) + 2\varepsilon\sigma \ge \exp(-\lambda\tau)[u(z_{z_n+\sigma e_{N+1}}(\tau)) - u(z_{z_n}(\tau))] \tag{5.12}$$

$$\ge \exp(-\lambda C)\,[\rho(\rho^{-1}(u(z_{z_n}(\tau))) + \sigma) - \rho(\rho^{-1}(u(z_{z_n}(\tau))))].$$

We also have that by Remark 4.1

$$\int_0^{t(\tau)} h(y_{x_n}, \alpha[b], b)dt = \int_0^\tau \bar h(y_{x_n}, \alpha[b], b)d\tau \le \tau C_1 \le CC_1,$$

where C_1 is also independent of ε, σ, and n, and then, again by Remark 4.1,

$$\rho^{-1}(u(z_{z_n}(\tau))) = U(y_{x_n}(\tau)) + r_n + \int_0^{t(\tau)} (h(y_{x_n}, \alpha[b], b) - \gamma^2|b|^2)dt$$

$$\le \sup_{\overline{\Omega}\cap B(x,C)} U + r_n + CC_1.$$

We conclude by (5.12) and the definition of ρ that

$$V^\lambda(z_n + \sigma e_{N+1}) + 2\varepsilon\sigma \geq V^\lambda(z_n) + K\sigma,$$

where $K = \exp(-\lambda C)\ \min_{[-R,R]} \dot{\rho} > 0$ and R is independent of ε, n, and σ. As $n \to +\infty$

$$V^{\lambda *}(z_0 + \sigma e_{N+1}) + 2\varepsilon\sigma \geq V^{\lambda *}(z_0) + K\sigma,$$

and finally, as $\sigma \to 0^+$ and since ε is arbitrary, $p_{N+1} \geq K > 0$.

2. We now conclude with the proof of the statement. Let $z_0 \in \Omega \times \mathbb{R}$ be such that $V^{\lambda *}(z_0) > u(z_0)$, and let $p^0 \in D^+ V^{\lambda *}(z_0)$. Since u is nonnegative, it follows that $V^{\lambda *}(z_0) > 0$; then by the first part $p^0_{N+1} > 0$ and so by (2.2) and the definition of $\overline{g}$, condition (4.7) is satisfied (with $l \equiv 0$). Let us prove this last statement. Fix any $a \in A$ and $\sigma \in (0, 1)$. Let p, z, b be such that $|p - p^0| < \sigma$, $|z - z_0| < \sigma$, $|b| > 1/\sigma$; and let $R = 1 + |x_0|$. By the definition of $\overline{g}$ and (2.2) we obtain ($p = (\overline{p}, p_{N+1})$)

$$\begin{aligned} -\overline{g}(x,a,b)\cdot p &= (-f(x,a,b)\cdot\overline{p} - h(x,a,b)p_{N+1} + \gamma^2|b|^2 p_{N+1})/(1+|a|^2+|b|^2) \\ &\geq (-L(1+|x|+|a|^q+|b|^q)|\overline{p}| - L_R(1+|a|^2+|b|^q)p_{N+1} + \gamma^2|b|^2 p_{N+1})/(1+|a|^2+|b|^2) \\ &\geq -C(1+|a|^2+|b|^q)/(1+|b|^2) + \gamma^2 p_{N+1}(1+\sigma^2(1+|a|^2)) \\ &\geq -C\sigma^{2-q}(1+(1+|a|^2)\sigma^q)/(1+\sigma^2) + \gamma^2(p^0_{N+1} - \sigma)/(1+\sigma(1+|a|^2)), \end{aligned}$$

where C depends only on R and p^0 and the last inequality holds for σ sufficiently small. It is clear that the right-hand side is positive when a is kept fixed and σ is choosen sufficiently small. Therefore by Proposition 4.4 the function V^λ satisfies (5.11) as a viscosity solution. □

The following lemma is an optimality principle for viscosity supersolutions of equations of type (5.7) in Lemma 5.2.

LEMMA 5.5. *Assume* (2.2), *and let U, u be as in Lemma* 5.2. *Then*

$$u(z) = \inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \sup_{[0,\tau_z)} u(z(\tau)), \qquad z \in \Omega \times \mathbb{R}. \tag{5.13}$$

Proof. We start by proving that, for all $s > 0$, we have

$$u(z) = \inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \sup_{[0,s\wedge\tau_z)} u(z(\tau)). \tag{5.14}$$

1. We first assume that $\overline{g}$ satisfies the global, uniform Lipschitz condition

$$|\overline{g}(z,a,b) - \overline{g}(\overline{z},a,b)| \leq L|z-\overline{z}| \qquad \text{for all } z, \overline{z}, a, b$$

and prove the result for $\Omega = \mathbb{R}^N$ (in this case, for each pair of controls, $\tau_z = +\infty$). Observe that, with this assumption, the Hamiltonian H will satisfy (5.9) with $L = L_R$, independent of R. Let $s > 0$; then we denote

$$v(z) = \inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \sup_{[0,s)} u(z(\tau)).$$

It is obvious that $v(x) \geq u(x)$, so we only need to prove the opposite inequality. Therefore, let $\lambda > 0$. By Lemma 5.5, V^λ is a viscosity solution of

$$\lambda V + \min\{H(z, DV), V - (1+\lambda)u\} = 0 \quad \text{in } \mathbb{R}^{N+1}. \tag{5.15}$$

The nonnegative function u is obviously another solution of (5.15). Moreover the Hamiltonian of (5.15) satisfies the assumptions for the uniqueness of bounded solutions (see, for example, Theorem 1.1 in the paper of Bardi and the author [5]). This implies that

$$V^\lambda = u \quad \text{in } \mathbb{R}^{N+1} \text{ for all } \lambda > 0,$$

and then since u is nonnegative,

$$\exp(-\lambda s)v(z) = \exp(-\lambda s) \inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \sup_{[0,s)} u(z(\tau)) \le V^\lambda(z) = u(z) \quad \text{for all } \lambda > 0,$$

hence the conclusion as $\lambda \to 0^+$.

2. Now let Ω be an arbitrary open set, $\overline{g}$ satisfying (2.2), and $\Gamma = \Omega \times \mathbb{R}$. For $\varepsilon > 0$, let $\zeta_\varepsilon : \mathbb{R}^{N+1} \to \mathbb{R}$ be a smooth function such that $0 \le \zeta_\varepsilon(z) \le 1$, $\zeta_\varepsilon = 0$ on $\Gamma^c_{\varepsilon/2}$, $\zeta_\varepsilon = 1$ in Γ_ε, where $\Gamma_\sigma = \{z = (x, r) \in \Gamma : \operatorname{dist}(x, \partial\Omega) > \sigma, |x| < 1/\sigma\}$ and $\Gamma^c = \mathbb{R}^{N+1}\backslash\Gamma$. We extend u outside $\Gamma_{\varepsilon/2}$ as a continuous, nonnegative, and bounded function in $\mathbb{R}^{N+1}$ and call it u^ε. We therefore obtain that u^ε is a supersolution of

$$\zeta_\varepsilon(z)\, H(z, Du^\varepsilon(z)) = \inf_{b\in B} \sup_{a\in A} \{-\overline{g}_\varepsilon(z, a, b) \cdot Du^\varepsilon(z)\} \ge 0, \quad \mathbb{R}^{N+1},$$

where $\overline{g} = \overline{g}_\varepsilon$ in $\Gamma_\varepsilon \times A \times B$ and $\overline{g}_\varepsilon = 0$ in $\Gamma^c_{\varepsilon/2} \times A \times B$. We can apply the first part of our proof and deduce that for $z \in \Gamma_\varepsilon$ and $s > 0$

$$u(z) = \inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \sup_{[0,s)} u^\varepsilon(z^\varepsilon(\tau)) \ge \inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \sup_{[0,s\wedge\tau^\varepsilon_z)} u(z(\tau)), \tag{5.16}$$

where τ^ε_z is the first exit time of the trajectory $z(\cdot)$ from Γ_ε (obviously $z(\cdot) = z^\varepsilon(\cdot)$ in $[0, s\wedge\tau^\varepsilon_z)$, if $z^\varepsilon(\cdot)$ is the trajectory corresponding to the vector field $\overline{g}_\varepsilon$).

3. We now conclude by the following argument. For fixed $\varepsilon, \eta > 0$, let $\varepsilon_j = \varepsilon/2^j$. If $z_0 = z \in \Gamma_{\varepsilon_0}$, by (5.16) we can find $\alpha_0 \in \Delta$ (we can fix it as a function of $z \in \Gamma_{\varepsilon_0}$) such that

$$\sup_{[0,s\wedge\tau^{\varepsilon_0}_z]} u(z(\tau)) \le u(z) + \eta/2 \qquad \text{for all } b \in \mathcal{B}.$$

For any fixed $b_0 = \overline{b} \in \mathcal{B}$, if $\tau^{\varepsilon_0}_z \ge s$, we have nothing left to prove. Otherwise we set $\tau_1 = \tau^{\varepsilon_0}_z$, $z_1 = z(\tau_1)$, $b_1(\cdot) = b_0(\cdot + t_{\alpha_0[b_0],b_0}(\tau_1))$. Then we can find $\alpha_1 \in \Delta$ (again as a function of $z \in \Gamma_{\varepsilon_1}\backslash\Gamma_{\varepsilon_0}$) such that

$$\sup_{[0,s\wedge\tau^{\varepsilon_1}_{x_1}]} u(z_{z_1}) \le u(z_1) + \eta/2^2.$$

We proceed recursively and construct the strategy α given by the position $\alpha[\overline{b}] = \overline{a}$, where $\overline{a}$ is the control defined by setting

$$\overline{a}(s) = \begin{cases} \alpha_0[b_0](s), & s \in [0, t_{\alpha_0[b_0],b_0}(\tau_1)[, \\ \alpha_1[b_1](s - t_{\alpha_0[b_0],b_0}(\tau_1)), & s \in [t_{\alpha_0[b_0],b_0}(\tau_1), t_{\alpha_0[b_0],b_0}(\tau_1) + t_{\alpha_1[b_1],b_1}(\tau_2)[, \\ \dots. \end{cases}$$

Then, also using Remark 4.2, we easily obtain that with such an α

$$\sup_{[0,s\wedge\tau_{z_0})} u(z_{z_0}) \le u(z_0) + \eta \quad \text{for all } b \in \mathcal{B},$$

and the result follows by the arbitrarity of η, the other inequality to get (5.14) being obvious.

4. We conclude the proof of the proposition using the same technique as in point 3. Let $\eta > 0$. For $z_0 = z \in \Gamma$ and applying (5.14) with $s = 1$, we can find $\alpha_0 \in \Delta$ such that

$$\sup_{[0,\tau_{z_0}\wedge 1]} u(z_{z_0}) \le u(z) + \eta \qquad \text{for all } b \in \mathcal{B}.$$

Fix $b_0 = \overline{b} \in \mathcal{B}$. If $\tau_{z_0} \le 1$, there is nothing left to prove. Otherwise let $z_1 = z_z(1)$, $b_1(\cdot) = b_0(\cdot + t_{\alpha_0[b_0],b_0}(1))$. Then we find $\alpha_1 \in \Delta$ such that

$$\sup_{[0,\tau_{z_1}\wedge 1]} u(z_{z_1}) \le u(z_1) + \eta/2^2 \qquad \text{for all } b \in \mathcal{B}.$$

We proceed recursively and conclude the proof similarly to point 3. □

As a consequence of the previous result and Lemma 5.2, we now prove a general optimality principle for viscosity supersolutions of equations whose Hamiltonian is defined in (5.1). We remark that this result has independent interest, and its peculiarity is to hold with an equality rather than an inequality; see (5.17).

PROPOSITION 5.6. *Assume* (2.2), *and let* $U : \Omega \to \mathbb{R}$ *be a continuous viscosity supersolution of*

$$\mathcal{H}(x, DU(x)) \ge 0 \quad \text{in } \Omega.$$

Then

$$U(x) = \inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \sup_{[0,t_x)} \left\{ \int_0^t (h(y, \alpha[b], b) - \gamma^2|b|^2)ds + U(y_x(t)) \right\}, \qquad x \in \Omega. \tag{5.17}$$

Proof. We define $u(x, r) = \rho(U(x) + r)$ as in Lemma 5.2, and then by that result u is a viscosity supersolution of

$$H(z, Du(z)) \ge 0 \quad \text{in } \Omega \times \mathbb{R}.$$

By Lemma 5.5, Remark 5.3, and (5.10) we conclude that

$$\begin{aligned} u(z) &= \inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \sup_{[0,\tau_z)} u(z(\tau)) \\ &= \inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \sup_{[0,t_x)} \rho\left(\int_0^t (h(y, \alpha[b], b) - \gamma^2|b|^2)ds + U(y_x(t)) + r_0 \right); \end{aligned}$$

hence in particular

$$\rho(U(x)) = u(x, 0) = \rho\left(\inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \sup_{[0,t_x)} \int_0^t (h(y, \alpha[b], b) - \gamma^2|b|^2)ds + U(y_x(t)) \right),$$

and the conclusion follows by the choice of $\rho(\cdot)$. □

We can now give the proof of our main results.

Proof of Theorem 3.2. Let U be as in the statement of the theorem. Let $x \in \Omega$; by (5.17) of Proposition 5.6 we can choose $\varepsilon > 0$ and $\alpha \in \Delta$ so that

$$\sup_{[0,t_x)} \left\{ \int_0^t (h(y, \alpha[b], b) - \gamma^2|b|^2)dt + U(y_x(t)) \right\} \le U(x) + \varepsilon \le U_0 - \varepsilon \quad \text{for all } b \in \mathcal{B}. \tag{5.18}$$

For $b \equiv 0$, and since $h(x, a, 0)$ is nonnegative, (5.18) implies

$$\sup_{[0,t_x)} U(y_x(t)) \leq U_0 - \varepsilon;$$

therefore $t_x(\alpha[0], 0) = +\infty$ by the boundary condition and the lower semicontinuity of U on $\partial\Omega$. So part (i) of the definition is satisfied in Ω with $a = \alpha[0]$ and $\alpha \in \Delta_\Omega$. Moreover since U is nonnegative, by (5.18) we also get

$$\sup_{\mathbb{R}_+} \int_0^t (h(y, \alpha[b], b) - \gamma^2|b|^2)dt \leq U(x) + \varepsilon \qquad \text{for all } b \in \mathcal{B}_{\Omega,\alpha}.$$

Hence since ε is arbitrary and by the definition, we get $0 \leq V_\gamma^\Omega \leq U$ in Ω and also (ii) is satisfied.

Of course the argument can be repeated in each sublevel set $\{x \in \Omega : U(x) < \lambda\}$ for $\lambda \in (0, U_0)$, and if this family of neighborhoods is a local base for $\mathcal{T}$, then we proved the open-loop Lyapunov stability and the fact that the local $\mathcal{H}_\infty$ suboptimal control problem with attenuation level $\gamma > 0$ is solvable.

As for the $\mathcal{H}_\infty$ suboptimal control problem, this is clearly a special case of the above, when $\Omega = \mathbb{R}^N$, and indeed the finite gain condition is an immediate consequence of (5.17). □

Proof of Theorem 3.4. To prove the first statement concerning (3.2), for $a \in \mathcal{A}$, $b \in \mathcal{B}$, we consider again the reparametrized system, as in Remark 5.3,

$$y'(\tau) = \overline{f}(y(\tau), a(\tau), b(\tau)), \quad y(0) = x.$$

We have that, by changing variables,

$$V_\gamma(x) = \inf_{\alpha\in\Delta} \sup_{b\in\mathcal{B}} \sup_{\tau\in\mathbb{R}_+} \left\{ \int_0^\tau [\overline{h}(y, \alpha[b], b) - \gamma^2|b|^2/(1 + |a|^2 + |b|^2)]d\tau \right\}.$$

Therefore, using the results of §4, in particular Proposition 4.4, with $\lambda = 0$, $g(x, a, b) = \overline{f}(x, a, b)$, $l(x, a, b) = \overline{h}(x, a, b) - \gamma^2|b|^2/(1 + |a|^2 + |b|^2)$, and $U \equiv 0$. Since condition (4.7) is easily satisfied in this case, as one can check with the same argument as the one of the proof of point 2 in Lemma 5.4, we can conclude that V_γ, in the open sets where it is locally bounded, is a viscosity solution of the variational inequality

$$\min\{\mathcal{H}^\#(x, DV_\gamma(x)), V_\gamma\} = 0, \tag{5.19}$$

where $\mathcal{H}^\#(x, p) = \inf_{b\in B} \sup_{a\in A}\{-\overline{f}(x, a, b) \cdot p - \overline{h}(x, a, b) + \gamma^2|b|^2/(1 + |a|^2 + |b|^2)\}$. To prove that V_γ is also a solution of (3.2), we will use the fact that, by the proof of Proposition 5.1, if $B_{x,p} \subset\subset B$ satisfies

$$\mathcal{H}(x, p) = \min_{b\in B_{x,p}} \sup_{a\in A}\{-f(x, a, b) \cdot p - h(x, a, b) + \gamma^2|b|^2\},$$

then there is $A_{x,p} \subset\subset A$ such that

$$\mathcal{H}(x, p) = \min_{b\in B_{x,p}} \max_{a\in A_{x,p}} \{-f(x, a, b) \cdot p - h(x, a, b) + \gamma^2|b|^2\}.$$

Let $(x, p) \in \mathbb{R}^{2N}$ be such that $\mathcal{H}^\#(x, p) \geq 0$. Then for all $\varepsilon > 0$ and $b \in B$ we can find $a_b \in A$ such that

$$-f(x, a_b, b) \cdot p - h(x, a_b, b) + \gamma^2|b|^2 \geq -\varepsilon(1 + |a_b|^2 + |b|^2). \tag{5.20}$$

If $b \in B_{x,p}$, then for ε sufficiently small (e.g., $\varepsilon < C_R^0$, $|x|, |p| \le R$, C_R^0 defined in (2.2)), (5.20) implies that $|a_b| \le C_{x,p}$ independent of ε. So we conclude from (5.20) that $\mathcal{H}(x, p) \ge 0$.

On the other hand if $\mathcal{H}^{\#}(x, p) \le 0$, then for $\varepsilon > 0$ we can find $\bar{b} \in B$ such that for all $a \in A$

$$-f(x, a, \bar{b}) \cdot p - h(x, a, \bar{b}) + \gamma^2|\bar{b}|^2 \le \varepsilon(1 + |a|^2 + |\bar{b}|^2). \tag{5.21}$$

If in particular ε is sufficiently small (e.g., $\varepsilon < \gamma^2$) and we set $a = 0$, then (5.21) implies $|\bar{b}| \le C_{x,p}$, independent of ε, and again we conclude by computing (5.21) at the points $a \in A_{x,p}$, where $A_{x,p}$ corresponds to the choice $B_{x,p} = B(0, C_{x,p})$.

The second part of (i) is a consequence of Proposition 5.6 and the fact that supersolution of (3.2) is equivalent to the nonnegative supersolution of $\mathcal{H}(x, DV) \ge 0$.

We now prove the statements concerning (3.3). Since V_γ is a supersolution of (3.2), it is clear that it is also a supersolution of (3.3). On the other hand, by the definition of subsolution, let $p \in D^+V_\gamma^*(x)$. If $V_\gamma^*(x) > 0$, then since V_γ is a subsolution of (3.2), it follows that

$$\mathcal{H}(x, p) \le 0. \tag{5.22}$$

If instead $V_\gamma^*(x) \le 0$, then since V_γ is nonnegative, V_γ is continuous at x, which is also a minimum point of the function. Therefore $0 \in D^-V_\gamma^*(x)$. Being both semidifferentials nonempty and V_γ continuous at x, this implies $p = 0$; see Crandall and Lions [12]. By (2.3) we then have

$$\mathcal{H}(x, p) = \inf_{b \in B} \sup_{a \in A} \{\gamma^2|b|^2 - h(x, a, b)\} \le \sup_{a \in A} \{-h(x, a, 0)\} \le 0,$$

and (5.22) holds also in this case. Threfore V_γ is a subsolution of (3.3).

Again the second part of the statement (ii) is a consequence of Proposition 5.6. □

6. On the solution of the $\mathcal{H}_\infty$ control problem. This section is devoted to some results and remarks concerning the construction of optimal strategies, in particular feedback controls to solve the $\mathcal{H}_\infty$ suboptimal control problem. As a matter of fact, we will only be concerned with checking the finite gain condition, since, as seen in Remark 3.3 and in the proof of Theorem 3.2, the stability of the system is a consequence of it. We will be making considerably strong assumptions in the general case, but the problem appears still to be tough and challenging. In the following we assume that the problem is solvable as previously defined; therefore the value function V_γ is finite in $\mathbb{R}^N$, null, and continuous on $\mathcal{T}$.

We start remarking that, when A is compact, the theory of Elliott and Kalton [13] can be applied to our problem without many additional difficulties (see also the author [28] for some details concerning the full-state information and the partial-information case) and provides the existence of optimal strategies as causal mappings $\alpha : \mathcal{B} \to \mathcal{A}^r$, where $\mathcal{A}^r$ is the set of relaxed controls, namely the set of measurable functions from $\mathbb{R}_+$ to the set of probability measures on A. The optimal strategy is in the class considered in this paper if the sets $(f(x, A, b),\ h(x, A, b))$ are convex for all x, b. Of course the existence of almost optimal strategies is always guaranteed by the very definition of value function. Unfortunately the proof of existence of optimal strategies is not constructive and, therefore, not useful from the practical point of view of the applications. Moreover this class of strategies requires full information of the system; namely, the disturbance has to be known and this is thought to be too restrictive. We describe instead how one can proceed in certain special cases to construct feedback controls. The ideas we outline here contain the known results in the cases of linear and nonlinear-affine systems and show to what extent we can generalize those results. We do not address here the question of the construction of optimal dynamic state feedback controls, where however similar difficulties have to be faced.

We first reduce to the interesting case $\mathcal{T} = \{0\}$ and make the following assumption. (A1): the value function V_γ is of class C^1; therefore, there is a nonnegative, classical solution, null at the origin of the (lower) Isaacs equation

$$\mathcal{H}(x, DU(x)) = 0 \quad \text{in } \mathbb{R}^N.$$

Unfortunately existence of classical solutions (and even uniqueness) is not satisfied in general even locally around the origin, and this was the main reason we used the concept of viscosity solutions. It is clear by its mere definition that V_γ is not likely to be even continuous in general. We will outline in Remark 6.2 how to generalize when the solution is not differentiable.

For any $(x, p) \in \mathbb{R}^{2N}$ we denote

$$F_{x,p}(a, b) = -f(x, a, b) \cdot p - h(x, a, b) + \gamma^2 |b|^2,$$

so that $\mathcal{H}(x, p) = \inf_{b\in B} \sup_{a\in A} F_{x,p}(a, b)$ and $\overline{\mathcal{H}}(x, p) = \sup_{a\in A} \inf_{b\in B} F_{x,p}(a, b)$, and we look for necessary conditions. Assume that the feedback $a(x)$ solves the $\mathcal{H}_\infty$ suboptimal control problem and provides optimal strategies for V_γ; then by (2.10) we have

$$\int_0^t (h(y, a(y), b) - \gamma^2 |b|^2) ds \le V_\gamma(x) \quad \text{for all } b \in \mathcal{B},\ t \ge 0.$$

Therefore by the definition of V_γ, for any fixed $t > 0$, we get

$$\int_0^t (h(y, a(y), b) - \gamma^2 |b|^2) ds + V_\gamma(y(t)) \le V_\gamma(x) \quad \text{for all } b \in \mathcal{B}.$$

Dividing by t and letting $t \to 0^+$, we conclude

$$0 \le F_{x, DV_\gamma(x)}(a(x), b) \quad \text{for all } b \in B;$$

finally by (A1) we must have

$$0 \le \inf_{b\in B} F_{x, DV_\gamma(x)}(a(x), b) \le \overline{\mathcal{H}}(x, DV_\gamma(x)) \le \mathcal{H}(x, DV_\gamma(x)) = 0. \tag{6.1}$$

Therefore equalities hold in (6.1) and V_γ has to be also a solution of the upper Isaacs equation. Of course this is always the case when the Isaacs condition holds, namely when $\mathcal{H}(x, p) = \overline{\mathcal{H}}(x, p)$ for all x, p, and a sufficient, easy-to-check condition for it is the splitting of the system, namely,

$$f(x, a, b) = f_1(x, a) + f_2(x, b),\ \ h(x, a, b) = h_1(x, a) + h_2(x, b).$$

If in particular (6.1) does not hold, then necessarily the value of the game does not exist, and therefore it is somewhat expected that feedback controls will not solve the problem in general. The previous necessary condition of optimality will give us indications about the construction of optimal feedbacks.

In view of (6.1), we proceed and make the second assumption. (A2): There is a nonnegative, classical supersolution U of the upper Isaacs equation

$$\overline{\mathcal{H}}(x, DU(x)) = 0 \quad \text{in } \mathbb{R}^N.$$

We now define the possibly multivalued map

$$A(x, p) = \arg \max_{a\in A} \{\inf_{b\in B} F_{x,p}(a, b)\} \tag{6.2}$$

and assume (A3): there is a selection $a(x) \in A(x, DU(x))$ such that the system

$$\dot{y} = f(y, a(y), b), \qquad y(0) = x,$$

has a unique absolutely continuous solution for all $b \in \mathcal{B}$ and $x \in \mathbb{R}^N$.

PROPOSITION 6.1. *Assume* (A1)-(A3). *Then the feedback control* $a(x)$ *solves the* $\mathcal{H}_\infty$ *suboptimal control problem. If the assumptions are satisfied by* $U = V_\gamma$, *then* $a(x)$ *is optimal.*

Proof. Let us prove that the position $\alpha[b](t) = a(y(t))$ defines a strategy satisfying (2.11); then we conclude by Remark 2.2. It is clear by (A2) and the definition of $a(\cdot)$ that

$$\inf_{b \in B} F_{x, DU(x)}(a(x), b) = \overline{\mathcal{H}}(x, DU(x)) = 0$$

(if $U = V_\gamma$ the necessary condition (6.1) is satisfied by construction); therefore for all $b \in \mathcal{B}$ the solution guaranteed by (A3) satisfies

$$0 \leq -f(y, \alpha[b], b) \cdot DU(y) - h(y, \alpha[b], b) + \gamma^2 |b|^2.$$

Integrating on $[0, t]$, we obtain

$$U(y(t)) + \int_0^t (h(y, \alpha, b) - \gamma^2 |b|^2) ds \leq U(x),$$

and since U is nonnegative, we conclude

$$\sup_{b \in \mathcal{B}} \sup_{t \in \mathbb{R}_+} \int_0^t (h(y, \alpha, b) - \gamma^2 |b|^2) ds \leq U(x),$$

and the result follows.

If $U = V_\gamma$, the same inequality shows the optimality of α for $V_\gamma(x)$. □

Remark 6.2. To our knowledge, there are no general results concerning the existence of optimal feedback controls, i.e., under which condition (A3) is satisfied with $U = V_\gamma$. However, one can call the map $A(x) = A(x, DV_\gamma(x))$ a multivalued feedback synthesis in the sense that any solution of the differential inclusion

$$\dot{y} \in f(y, A(y), b), \qquad y(0) = x,$$

satisfies the gain condition (to prove this, just apply the argument of Proposition 6.1 to a selection $a(t) \in A(y(t))$ such that $\dot{y} = f(y, a, b)$). Moreover, since, as we mentioned, (A1) is too restrictive, we should really use generalized gradients (which are usually sets) and define the feedback control as a selection $a(x) \in A(x, \text{“}DV_\gamma(x)\text{”})$ satisfying (A3). This is also an open problem. Otherwise $A(x) = A(x, \text{“}DV_\gamma(x)\text{”})$ will again be a multivalued feedback synthesis.

The method we presented in Proposition 6.1, under certain nondegeneracy conditions on the cost h, really works for linear systems, as the reader can find out for example in Basar and Bernhard [10], and almost works for nonlinear-affine systems (where $f(x, a, b) = f_1(x) + f_2(x)a + f_3(x)b$ and $h(x, a, b) = |\overline{h}(x, a)|^2$, $\overline{h}(x, a) = h_1(x) + h_2(x)a$), in the sense that (A2) and (A3) are satisfied, as one easily checks (see for example Van der Shaft [30]), since the following explicit formula for (6.2) holds (A is luckily single valued in this case):

$$A(x, p) = -(h_2^T(x)h_2(x))^{-1}(f_2^T(x)p/2 + h_2^T(x)h_1(x)),$$

but again (A1) fails in general. We do not have the space here to discuss a specific counterexample, which is however not too difficult to construct and will be presented elsewhere, but

refer instead to Van der Shaft [30] for some sufficient conditions ensuring that (A1) holds at least locally around the origin.

For general systems, life is even harder. As we already observed, by (6.1) it follows that if (A2) fails, we have no hope for optimal feedback controls. However here is how one might try to proceed in this case to construct optimal strategies. We now define the set-valued map

$$A(x,b) = \arg\max_{a\in A} F_{x,DV_\gamma(x)}(a,b)$$

and assume (A4): for all $b \in \mathcal{B}$ there is an absolutely continuous solution of the differential inclusion

$$\dot{y} \in f(y, A(y,b), b), \qquad y(0) = x.$$

We now indicate by $\mathcal{Y}_{x,b}$ the set of such solutions and assume (A5): we can find a causal selection $\alpha_x[\cdot]$ satisfying: for any $b \in \mathcal{B}$, $\alpha_x[b](t) \in A(y(t), b(t))$ for some $y \in \mathcal{Y}_{x,b}$.

PROPOSITION 6.3. *Assume* (A1), (A4), *and* (A5). *Then the strategy* $\alpha_x \in \Delta$ *is optimal for* $V_\gamma(x)$.

Proof. Observe that for all $b \in \mathcal{B}$, along the solution $y \in \mathcal{Y}_{x,b}$ such that $\alpha_x[b](t) = A(y(t), b(t))$, we have

$$0 = \mathcal{H}(y, DV_\gamma(y)) \le \max_{a\in A} F_{y,DV_\gamma(y)}(a,b) = F_{y,DV_\gamma(y)}(\alpha_x[b], b);$$

then the proof proceeds as in Proposition 6.1. □

Unfortunately we do not know of reasonable sufficient conditions for (A4) and (A5). However the map $A(x,b)$ is again a multivalued optimal synthesis (not feedback in this case) in the sense of Remark 6.2.

Remark 6.4. A more appealing way to proceed would be first to use discretizations of the system (1.1) and the Isaacs equation, then to solve the corresponding discrete-time differential game by feedback controls (this is usually easier to do; see the paper of Bardi and the author [4] and also Bardi, Falcone, and the author [1] for some more details about this procedure), and in the end to prove error estimates when using these feedbacks in the original system. This direction of research looks more promising in the near future.

REFERENCES

[1] M. BARDI, M. FALCONE, AND P. SORAVIA, *Fully discrete schemes for the value function of pursuit-evasion games* in Advances in Dynamic Games and Applications, A. Haurie and T. Basar eds., Birkhauser, Cambridge, MA, 1994, pp. 89-106.

[2] J. A. BALL AND J. W. HELTON, *Viscosity solutions of Hamilton-Jacobi equations arising in nonlinear* $\mathcal{H}_\infty$ *control*, preprint, 1994.

[3] J. A. BALL, J. W. HELTON, AND M. L. WALKER, $\mathcal{H}_\infty$ *control for nonlinear systems with output feedback*, IEEE Trans. Automat. Control, 38 (1993), pp. 546-559.

[4] M. BARDI AND P. SORAVIA, *Approximation of differential games of pursuit-evasion by discrete-time games*, in Differential Games. Developing, Modeling and Computation, Lecture Notes in Control and Inform. Sci. 156, R. P. Hamalainen and H. K. Ethamo, eds., Springer-Verlag, 1991, pp. 131-143.

[5] ———, *A comparison result for Hamilton-Jacobi equations and applications to differential games lacking controllability*, Funkcial. Ekvac., 37 (1994), pp. 19-43.

[6] G. BARLES, *An approach of deterministic control problems with unbounded data*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7/8 (1990) pp. 235-258.

[7] G. BARLES and B. PERTHAME, *Discontinuous solutions of deterministic optimal stopping time problems*, RAIRO Modél. Math. Anal. Numer., 21 (1987), pp. 557-579.

[8] E. N. BARRON, *Differential games with maximum cost*, Nonlinear Anal. TMA, 14 (1990), pp. 971-989.

[9] E. N. BARRON and H. ISHII, *The Bellman equation for minimizing the maximum cost*, J. Differential Equations, 53 (1989), pp. 213-233.

[10] T. BASAR AND P. BERNHARD, *$\mathcal{H}_\infty$ Optimal Control and Related Minimax Design Problems*, Birkhauser, Cambridge, MA, 1990.

[11] M. C. CRANDALL, H. ISHII, AND P. L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1-67.

[12] M. C. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.

[13] R. J. ELLIOTT AND N. J. KALTON, *The existence of value in differential games*, Mem. Amer. Math. Soc., no. 126, Amer. Math. Soc., Providence, RI, 1972.

[14] L. C. EVANS AND P. E. SOUGANIDIS, *Differential games and representation formulas for solutions of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 773-797.

[15] W. H. FLEMING and W. M. MCENEANEY, *Risk sensitive optimal control and differential games*, in Proc. Conference on Adaptive and Stochastic Control, Univ. of Kansas, 1992, Lecture Notes and Control Inform. Sci. 184, Springer-Verlag, New York, 1992.

[16] D. HILL and P. MOYLAN, *The stability of nonlinear dissipative systems*, IEEE Trans. Automat. Control, AC-25 (1976), pp. 708-711.

[17] H. ISHII, *A boundary value problem of the Dirichlet type for Hamilton-Jacobi equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 16 (1989), pp. 105-135.

[18] A. ISIDORI and A. ASTOLFI, *Nonlinear $\mathcal{H}_\infty$ contol via measurement feedback*, J. Math. Systems Est. Control, 2 (1992), pp. 31-44.

[19] D. H. JACOBSON, *Optimal stochastic linear systems with exponential criteria and their relation to deterministic differential games*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 121-134.

[20] M. R. JAMES, *A partial differential inequality for dissipative systems*, Systems Control Lett., 21 (1993), pp. 315-320.

[21] ———, *Asymptotic analysis of nonlinear stochastic risk-sensitive control and differential games*, Math. Control Signals Systems, 5 (1992), pp. 401-417.

[22] M. R. JAMES and S. YULIAR, *Numerical approximation of the $\mathcal{H}_\infty$ norm for nonlinear systems*, Automatica, to appear.

[23] N. N. KRASSOWSKII AND A. I. SUBBOTIN, *Game-Theoretical Control Problems*, Springer, New York, 1988.

[24] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, 1982.

[25] P. SORAVIA, *The concept of value in differential games of survival and viscosity solutions of Hamilton-Jacobi equations*, Differential Integral Equations, 5 (1992), pp. 1049-1068.

[26] ———, *Pursuit-evasion problems and viscosity solutions of Isaacs equations*, SIAM J. Control Optim., 31 (1993), pp. 604-623.

[27] ———, *Stability of dynamical systems with competitive controls: the degenerate case*, J. Math. Anal. Appl., 191 (1995), pp. 428-449.

[28] ———, *Viscosity solutions to study partially observed differential games and nonlinear $\mathcal{H}_\infty$ control*, preprint.

[29] A. J. VAN DER SHAFT, *L_2 gain analysis for nonlinear systems and nonlinear $\mathcal{H}_\infty$ control*, IEEE Trans. Automat. Control, 37 (1992), pp. 770-784.

[30] ———, *Nonlinear state space $\mathcal{H}_\infty$ control theory*, in Perspective in Control, H. L. Trentelman and J. C. Willems, eds., Birkhauser, Cambridge, MA, 1993.

[31] P. WHITTLE, *Risk-Sensitive Optimal Control*, Wiley, New York, 1990.

[32] J. C. WILLEMS, *Dissipative dynamical systems. Part* I: *General theory*, Arch. Rational Mech. Anal., 45 (1972), pp. 321-351.

[33] G. ZAMES, *Feedback optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 301-320.

[34] P. SORAVIA, *Equivalence between nonlinear $\mathcal{H}_\infty$ control problems and existence of viscosity solutions of Hamilton–Jacobi equations*, preprint, 1995.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1099–1134, July 1996

INFINITE-HORIZON VARIATIONAL PROBLEMS WITH NONCONVEX INTEGRANDS*

ARIE LEIZAROWITZ† AND ALEXANDER J. ZASLAVSKI†

This paper is dedicated to the memory of Avraham Leizarowitz.

Abstract. We study *weakly optimal solutions* of infinite-horizon variational problems with first-order nonconvex integrands. This is a weakened version of the overtaking optimality criterion. These optimal solutions are closely related to the *minimal solutions* studied by Moser, Aubry, and Mather. Such solutions have definite *rotation number*, and we study the relation between the rotation number and the minimal energy growth rate. We establish the existence of a weakly optimal solution for every prescribed initial condition.

We also consider discrete-time infinite-horizon periodic control problems in R^n. Analogous to rotation numbers we consider rotation vectors and study minimization of energy growth rate for a prescribed rotation vector. This constrained problem is related to an unconstrained minimization problem that has the same class of minimizers.

Key words. infinite horizon, weakly optimal solutions, minimal solutions, overtaking optimality, rotation number

AMS subject classification. 49J15

1. Introduction.

In this paper we consider a special class of extremals, the so-called *weakly optimal solutions* of infinite horizon variational problems for real-valued functions.

We consider functionals of the form

$$I(a, b, x) = \int_a^b f(t, x(t), x'(t))\, dt, \tag{1.1}$$

where $-\infty < a < b < +\infty$ and $x \in W^{1,1}(a, b)$. By an appropriate choice of representatives, $W^{1,1}(a, b)$ can be identified with the set of absolutely continuous functions $x : [a, b] \mapsto R^1$, and we will henceforth assume that this has been done.

We will assume the integrand $f = f(t, x, p)$ satisfies the following assumption.

Assumption A. (i) $f \in C^3$, and $f(t, x, p)$ has period 1 in t, x;

(ii) $\delta \le f_{pp}(t, x, p) \le \delta^{-1}$ for every $(t, x, p) \in R^3$;

(iii) $|f_{xp}| + |f_{tp}| \le c(1 + |p|)$, $|f_{xx}| + |f_{xt}| + |f_{tt}| \le c(1 + p^2)$,

with some constants $\delta \in (0, 1)$, $c > 0$.

Clearly Assumption A implies that

$$\delta_0 p^2 - c_0 \le f(t, x, p) \le \delta_0^{-1} p^2 + c_0 \quad \text{for every } (t, x, p) \in R^3 \tag{1.2}$$

for some constants $c_0 > 0$ and $0 < \delta_0 < \delta$.

Given an $x_0 \in R^1$ we study the infinite-horizon problem of minimizing the expression

$$\int_0^T f(t, x(t), x'(t))\, dt \tag{P}$$

as T grows to infinity, where $x(\cdot) \in W^{1,1}_{\text{loc}}([0, \infty))$ satisfies the initial condition $x(0) = x_0$.

Infinite-horizon variational problems are studied mostly by researchers in economics where they are used to model, for example, economic growth. Our original motivation to study these problems came from considering questions related to continuum mechanics and

*Received by the editors September 7, 1993; accepted for publication (in revised form) January 3, 1995. This research was supported by the fund for the promotion of research at the Technion and supported in part by a grant from the Israeli Ministry of Science and the MA-AGARA special project for absorption of new immigrants.

†Department of Mathematics, Technion–Israel Institute of Technology, Haifa 32000, Israel.

to dislocations in one-dimensional crystals. Thus, this type of problems has wide-ranging applications.

Our study follows the recent papers by J. Moser [15], [16], who studied the structure of the minimal solutions of the variational problem (P). Moser's theory can be viewed as a development of the work by Aubry [2] and Mather [12] that is concerned with area-preserving mappings of an annulus or a cylinder. The works of Aubry and Mather were begun independently and with different motivations but led to similar results by different methods. While Mather studied area-preserving annulus mappings as they occur as section mappings for Hamiltonian systems of two degrees of freedom, Aubry investigated certain models of solid-state physics related to dislocations in one-dimensional crystals, which are discussed in Aubry [2], Sinai [18], and Zazlavski [19].

It is assumed in Aubry's work that the states of the model are represented by the so-called *minimal energy configurations* (*minimal solutions*). He studied the existence and the structure of such minimal solutions. From the point of view of optimal control he investigated a certain class of discrete-time infinite-horizon control systems employing the notion of minimal solutions. In the spirit of the Aubry–Mather theory, Bylayi and Polterovich [5] and Bangert [3] studied geodesic flows and geodesic rays on Riemannian two-torus. The continuous-time analog of this discrete-time problem seems to be more challenging and is of the type of problem (P) described above.

A function $x(\cdot) \in W^{1,1}_{\text{loc}}(R^1)$ will be called a *minimal solution* of the variational problem (P) if

$$\int_a^b [f(t, y(t), y'(t)) - f(t, x(t), x'(t))]\, dt \geq 0$$

for every real numbers $a < b$ and every $y \in W^{1,1}(a, b)$ satisfy $y(a) = x(a)$ and $y(b) = x(b)$ (see [2], [15], [16], [19]). It was shown in [15], [16] that (under Assumption A) minimal solutions of (P) possess numerous remarkable properties. Thus, for any minimal solution $x(\cdot)$ there exists a real number α satisfying

$$\sup\{|x(t) - \alpha t| : t \in R^1\} < \infty$$

that is called the *rotation number* of $x(\cdot)$, and given any real α, there exists a minimal solution with rotation number α.

While studying infinite-horizon optimal control problems there are several optimality notions that are considered, and the notion of minimal solution given above is the weakest among them. Clearly we can consider the notion of minimal solutions for the class of functions defined on $[0, \infty)$. A more refined optimality criterion known as the *overtaking optimality criterion* was introduced in the economic literature and was also used in studying infinite-horizon optimal control problems (see [1], [4], [6], [8]).

A function $x(\cdot) \in W^{1,1}_{\text{loc}}([0, \infty))$ will be called an *overtaking optimal solution* of the variational problem (P) if

$$\limsup_{T \to \infty} \int_0^T [f(t, x(t), x'(t)) - f(t, y(t), y'(t))]\, dt \leq 0$$

for any $y(\cdot) \in W^{1,1}_{\text{loc}}([0, \infty))$ that satisfies $y(0) = x(0)$.

Usually it is difficult to establish the existence of overtaking optimal solutions, and actually, in general they may fail to exist. Most studies that are concerned with their existence assume convex integrands f. For convex integrands the existence of overtaking optimal solutions may follow from the fact that all good trajectories converge to a unique steady state

(see Brock and Haurie [4] and Leizarowitz [10]). For nonconvex integrands the existence of overtaking optimal solutions is not guaranteed, and in this situation we look for *weakly optimal solutions*, which indeed will be established in this paper.

A function $x(\cdot) \in W^{1,1}_{\text{loc}}([0, \infty))$ will be called a *weakly optimal solution* of the variational problem (P) if

$$\liminf_{T\to\infty} \int_0^T [f(t, x(t), x'(t)) - f(t, y(t), y'(t))]\, dt \leq 0$$

for any $y(\cdot) \in W^{1,1}_{\text{loc}}([0, \infty))$ that satisfies $y(0) = x(0)$.

It will be established in this paper that under certain conditions on f there exists a weakly optimal solution $x(\cdot) \in W^{1,1}_{\text{loc}}([0, \infty))$ for any initial value $x(0) = x_0$. The analogous result also holds for the discrete-time optimal control system corresponding to problem (P) with analogous proofs. We can expect that these results will become useful when applied to Aubry's theory and in the studies of geodesic rays on Riemannian torus.

The first existence result of weakly optimal solutions without convexity assumptions was obtained by Carlson [7] for autonomous optimal control problems with vector-valued functions. Under the assumptions posed in [7], for every good trajectory $x(\cdot)$ defined on $[0, \infty)$,

$$\tau^{-1} \int_0^\tau x(t)\, dt \to \bar{x} \quad \text{as } \tau \to \infty$$

where $\bar{x}$ is a unique steady state. Using this fact, Carlson established the existence of weakly optimal solutions.

In our situation we do not have any kind of convergence property of all the good trajectories to a unique steady state. Our consideration is based on the following observation, which was described in Leizarowitz [11] for a class of variational problems:

For every initial state there exists a weakly optimal solution if all the good trajectories have the same limit point set.

Recently this was used by Zaslavski [20] to establish, for discrete-time control systems, the existence of weakly optimal solutions for generic cost functions and any initial state. Actually we will not prove that all the good trajectories of problem (P) have the same asymptotic behavior, but we will establish a result that is very close to this and is sufficient for our purpose.

We will derive the following results.

THEOREM 1.1. *Let f satisfy Assumption* A. *Then there exist a strictly convex function $E_f : R^1 \to R^1$ satisfying $E_f(\alpha) \to +\infty$ as $|\alpha| \to \infty$ and a monotonically increasing function*

$$\Gamma_f : [0, \infty) \mapsto [0, \infty)$$

such that for each real α and each minimal solution x with rotation number α the relation

$$\left| \int_S^{S+T} f(t, x(t), x'(t))\, dt - E_f(\alpha)T \right| \leq \Gamma_f(|\alpha|) \tag{1.3}$$

holds for all real numbers S and T.

It follows from Theorem 1.1 that if f satisfies Assumption A, then $\inf\{E_f(\alpha) : \alpha \in R^1\}$ is attained at a unique point, which we denote by α_f.

THEOREM 1.2. *Let f satisfy Assumption* A, *and let α_f be the unique minimizer of $\alpha \mapsto E_f(\alpha)$, namely,*

$$E_f(\alpha_f) = \inf\{E_f(\alpha) : \alpha \in R^1\}.$$

Assume that α_f is irrational. Then for any initial value $x_0 \in R^1$ there exists a weakly optimal solution $x(\cdot) \in W^{1,1}_{\text{loc}}([0, \infty))$ satisfying $x(0) = x_0$.

The proofs of these results are based on Moser's theory (see [15], [16]) and the method of reformulating the variational problem in discrete-time terms as developed in [9] and [11]. Theorem 1.1 is an analogue of Mather's theorem about the average energy function for Aubry–Mather sets generated by a diffeomorphism of the infinite cylinder, which consists of finite compositions of exact symplectic monotone twist mappings [13]. Mather's theorem is a generalization of Aubry's unpublished result, which establishes the strict convexity of the average energy function for the Frenkel–Kontorova model. A similar result appeared in Senn [17], where the strict convexity of E_f in multidimensional situations was established. In our discussion, in addition to the strict convexity of E_f, we also need the estimate (1.3), which was not considered in [17].

We also consider discrete-time infinite-horizon control problems. Let $v(\cdot, \cdot)$ be continuous and periodic on $R^n \times R^n$, namely $v(x + m, y + m) = v(x, y)$ for every $x, y \in R^n$ and every $m \in R^n$ with integer components. We consider cost expressions $D_N(\bar{x}) = \sum_{k=0}^{N-1} v(x_k, x_{k+1})$ for programs $\bar{x} = \{x_k\}_{k=0}^{\infty} \subset R^n$. Analogous to rotation numbers we define the *rotation vectors* and study the problem of minimizing the growth rate over all the programs that have a prescribed rotation vector. Let $\Phi_v(\cdot)$ be the discrete-time counterpart of the above-mentioned function $E_f(\cdot)$. It will be proved below that $\Phi_v(\cdot)$ is a convex function.

We will establish the existence of optimal programs with prescribed rotation vectors α, and we will obtain them as optimal programs for the following unconstrained related problem. Supposed that $(\alpha, \Phi_v(\alpha))$ is an exposed point of $\text{epi}\Phi_v$ (the epigraph of $\Phi_v(\cdot)$), and let λ be such that $\Phi_v(\alpha') > \Phi_v(\alpha) + \lambda \cdot (\alpha' - \alpha)$ for every $\alpha' \neq \alpha$. We define

$$v_\alpha(x, y) = v(x, y) + \lambda \cdot (x - y),$$

which like v is continuous and periodic in $R^n \times R^n$. We have the following result.

THEOREM 1.3. *The function $\Phi_v(\cdot)$ is convex. Moreover, if $(\alpha, \Phi_v(\alpha))$ is an exposed point of* $\text{epi}\Phi_v$, *then there exists a minimizer of the growth rate of programs that have a prescribed rotation vector α. This minimizer has definite limits both for the energy growth rate and the rotation vector. Actually it is a minimizer of the unconstrained problem of minimizing the cost expressions*

$$D_N^\alpha(\bar{x}) = \sum_{k=0}^{N-1} v_\alpha(x_k, x_{k+1})$$

as $N \to \infty$. In the case that $\Phi_v(\cdot)$ is strictly convex and every *point $(\alpha, \Phi_v(\alpha))$ is an exposed point of* $\text{epi}\Phi_v$, *then the assertion of this result holds for every $\alpha \in R^n$.*

2. Properties of minimal solutions. We will need the following result (see [14, Thm. 1.8.2]).

PROPOSITION 2.1. *Let $-\infty < a < b < +\infty$, $\{x_k(\cdot)\}_{k=1}^{\infty} \subset W^{1,1}(a, b)$, and $\{x_k'\}_{k=1}^{\infty}$ be a bounded sequence in $L^2[a, b]$. Furthermore, let $x(\cdot) \in W^{1,2}(a, b)$ be such that*

$$x_k(t) \to x(t) \quad \textit{uniformly in } [a, b]$$

and

$$x_k'(\cdot) \to x'(\cdot) \quad \textit{weakly in } L^2[a, b] \textit{ as } k \to \infty.$$

Then

$$I(a, b, x) \leq \liminf_{k\to\infty} I(a, b, x_k).$$

For the proof see [14, Thm. 1.8.2].

For $a, b, \alpha, \beta \in R^1$ such that $a < b$ we define

$$U(a, b, \alpha, \beta) = \inf\{I(a, b, x) : x \in W^{1,1}(a, b), x(a) = \alpha, x(b) = \beta\}. \tag{2.1}$$

Relation (1.2) and Proposition 2.1 imply the following result.

PROPOSITION 2.2. *Let $a, b, \alpha, \beta \in R^1$, $a < b$. Then there exists $x(\cdot) \in W^{1,1}(a, b)$ such that*

$$x(a) = \alpha, \qquad x(b) = \beta, \qquad I(a, b, x) = U(a, b, \alpha, \beta).$$

Assumption A, relation (1.2), and Theorem 1.10.1 in [14] imply the following result.

PROPOSITION 2.3. *Let $a, b, \alpha, \beta \in R^1$, $a < b$, and $x(\cdot) \in W^{1,1}(a, b)$ be such that*

$$x(a) = \alpha, \quad x(b) = \beta, \quad I(a, b, x) = U(a, b, \alpha, \beta).$$

Then $x(\cdot) \in C^2[a, b]$ and it satisfies the Euler–Lagrange equation.

Denote by $\mathbf{Z}$ the set of all integers. We note that for any integers j and k the translations

$$(t, x) \mapsto (t + j, x + k)$$

leave the variational problem invariant. Therefore, if $x(\cdot)$ is a minimal solution, so is $x(\cdot + j) + k$. On the torus this represents, of course, the same curve as does $x(\cdot)$. This motivates the following terminology (see [15], [16]).

We say that a function $x(\cdot) \in W^{1,1}_{\text{loc}}(R^1)$ *has no self-intersections* if for every pair of integers j, k the function $t \mapsto x(t + j) + k - x(t)$ is either always positive or always negative or identically zero. The following results, which will be needed below, were established in [16].

PROPOSITION 2.4. *If $x \in W^{1,1}_{\text{loc}}(R^1)$ is a minimal solution without self-intersections of* (P), *then there exists a unique number α such that*

$$\sup\{|x(t) - \alpha t| : t \in R^1\} < \infty$$

and there exists a constant c_1 depending only on the constants c_0 and δ_0 in (1.2) *so that*

$$|x(t + s) - x(t) - \alpha s| \leq c_1(1 + \alpha^2)^{1/2} \tag{2.2}$$

for all s and t.

Moreover, there exist constants $\epsilon \in (0, 1)$ and $\gamma_1 > 0$ independent of x but depending on $|\alpha|$ and the constants c and δ in Assumption A *such that $\|x'(\cdot)\|_{C^\epsilon} \leq \gamma_1$ (where C^ϵ denotes the space of Hölder continuous functions of order ϵ).*

The number α is uniquely determined by the minimal $x = x(t)$ and is called *the rotation number of x*.

PROPOSITION 2.5. *Given any real α there exists a non–self-intersecting minimal solution with rotation number α.*

We associate with any non–self-intersecting minimal solution $x(\cdot)$ its rotation number α and denote

$$\mathcal{M}(\alpha) = \{\, x(\cdot) : x \text{ is non–self-intersecting minimal solution with rotation number } \alpha\}. \tag{2.3}$$

Moreover, for $A > 0$ we set

$$\mathcal{M}_A = \bigcup\{\mathcal{M}(\alpha) : |\alpha| \le A\}. \tag{2.4}$$

PROPOSITION 2.6 [15, Thm. 7.4] and [16, Cor. 3.3]. *The set $\mathcal{M}_A/\mathbf{Z}$ is compact with respect to the C^1-topology on compact sets in R^1. In other words, any sequence $\{x^s\} \in \mathcal{M}_A$ possesses a subsequence, say $\{x^{s_i}\}_{i=1}^{\infty}$, and a sequence of integers $\{k_i\}_{i=1}^{\infty}$ for which $x^{s_i} - k_i$ converges with first derivatives uniformly on any compact set to a function $x^* \in \mathcal{M}_A$.*

We have the following results (see [16, Thms. 5.1, 5.2, and 5.4 and Cors. 5.5 and 5.3]).

PROPOSITION 2.7. *Let α be a rational number and q be a natural number such that $q\alpha \in \mathbf{Z}$. Set*

$$\mathcal{A}_q = \{x(\cdot) \in W^{1,1}_{\mathrm{loc}}(R^1) : x(t+q) = x(t) \text{ for every } \in R^1\}.$$

Then there exists $\hat{x}^* \in \mathcal{A}_q$ such that $x^* = \alpha t + \hat{x}^*$ minimizes $I(0, q, x)$ over all $x \in W^{(1,1)}_{\mathrm{loc}}(R^1)$ with $x - \alpha t \in \mathcal{A}_q$. Moreover, $x^* \in C^2(R^1)$, and it satisfies the Euler–Lagrange equation.

PROPOSITION 2.8. *Let α be a rational number and q be a natural number satisfying $q\alpha \in \mathbf{Z}$. Set*

$$\bar{\mathcal{A}}_q = \{x \in W^{1,1}_{\mathrm{loc}}(R^1) : x - \alpha t \in \mathcal{A}_q\},$$
$$\mathcal{M}(\alpha, q) = \{x(\cdot) \in \bar{\mathcal{A}}_q : I(0, q, x) \le I(0, q, y) \text{ for every } y \in \bar{\mathcal{A}}_q\}.$$

Then the set $\mathcal{M}(\alpha, q)$ is totally ordered; i.e., if $x, y \in \mathcal{M}(\alpha, q)$, then either $x(t) < y(t)$ for all t or $x(t) > y(t)$ for all t or $x(t) = y(t)$ identically.

COROLLARY 2.1. *If $x \in \mathcal{M}(\alpha, q)$, then $x(\cdot)$ has no self-intersections.*

PROPOSITION 2.9. *Let α be a rational number and q_1 and q_2 be natural numbers such that $q_1\alpha, q_2\alpha \in \mathbf{Z}$. Then $\mathcal{M}(\alpha, q_1) = \mathcal{M}(\alpha, q_2)$.*

For any rational number α we set

$$\mathcal{M}_{\mathrm{per}}(\alpha) = \mathcal{M}(\alpha, q),$$

where q is a natural number satisfying $\alpha \in \mathbf{Z}$.

PROPOSITION 2.10. *Let α be a rational number and $x \in \mathcal{M}_{\mathrm{per}}(\alpha)$. Then x is a non–self-intersecting minimal solution of* (P) *with the rotation number α.*

The following two results were established in [16, §4].

PROPOSITION 2.11. *Let u and v be minimal solutions of* (P). *Then the open set $\{t \in R^1 : u(t) < v(t)\}$ has no bounded components.*

PROPOSITION 2.12. *If u and v are minimal solutions, $u(t) \le v(t)$ for all $t \in R^1$, then either $u(t) = v(t)$ for all t or $u(t) < v(t)$ for all t.*

3. Proof of Theorem 1.1.

PROPOSITION 3.1. *Let $\alpha = m/n$ where m, n are integers, $n \ge 1$. Set*

$$E(\alpha) = \frac{1}{n}\int_0^n f(t, y(t), y'(t))\, dt,$$

where $y \in \mathcal{M}_{\text{per}}(\alpha)$. *Then for* α *there exists a constant* $\Gamma(\alpha)$ *that increases for* $\alpha > 0$ *and decreases for* $\alpha < 0$ *and such that for all numbers* S, T, *with* $T > 0$, *and for every* $x \in \mathcal{M}_{\text{per}}(\alpha)$,

$$\left| \int_S^{S+T} f(t, x(t), x'(t))\, dt - E(\alpha)T \right| \leq \Gamma(\alpha). \tag{3.1}$$

Proof. Let $x \in \mathcal{M}_{\text{per}}(\alpha)$. By Proposition 2.4

$$|x'(t)| \leq \gamma_1(\alpha) \quad \text{for every } t \in R^1,$$

where the constant $\gamma_1(\alpha)$ depends monotonically increasing on $|\alpha|$. We set

$$\begin{aligned} \Gamma_0(\alpha) &= 1 + 4(1 + c_0) + 48\delta_0^{-1}(c_1 + 1)^2(1 + |\alpha|)^2, \\ \beta &= \sup\{|f(t, x, y)| : |y| \leq \gamma_1(\alpha), x, t \in R^1\}, \\ \Gamma(\alpha) &= \Gamma_0(\alpha) + 3(1 + |\alpha|)\beta. \end{aligned}$$

(Recall the constants c_0 and δ_0 in (1.2) and c_1 in (2.2).)

It is easy to verify that

$$|E(\alpha)| \leq \beta|\alpha|,$$

$$\left| \int_{T_1}^{T_2} f(t, x(t), x'(t))\, dt \right| \leq \sup\{|f(t, x, y)| : t, x, y \in R^1,\ |y| \leq \gamma_1(\alpha)\}$$

for all $T_1, T_2 \in R^1$ such that $|T_1 - T_2| \leq 1$. To prove the proposition it is sufficient to show that for every *integers* T and S (with $T > 0$),

$$\left| \int_S^{S+T} f(t, x(t), x'(t))\, dt - E(\alpha)T \right| \leq \Gamma_0(\alpha). \tag{3.2}$$

The validity of (3.2) for *any integers* S and T implies that (3.1) holds for any S and T as asserted in the proposition.

Assume to the contrary that (3.2) does not hold. Then there are integers T, S such that $1 \leq T \leq n - 1, 0 \leq S \leq n - 1$, and

$$\left| \int_S^{S+T} f(t, x(t), x'(t))\, dt - E(\alpha)T \right| > \Gamma_0(\alpha).$$

If

$$\int_S^{S+T} f(t, x(t), x'(t))\, dt - E(\alpha)T > \Gamma_0(\alpha),$$

then

$$\int_{S+T}^{S+n} f(t, x(t), x'(t))\, dt - (n - T)E(\alpha) < -\Gamma_0(\alpha).$$

Therefore, we may assume without loss of generality that

$$\int_S^{S+T} f(t, x(t), x'(t))\, dt - E(\alpha)T < -\Gamma_0(\alpha). \tag{3.3}$$

Choose a large natural number k. We have

$$knE(\alpha)T = \int_S^{S+knT} f(t, x(t), x'(t))\, dt = \sum_{i=0}^{kn-1} \int_{S+iT}^{S+(i+1)T} f(t, x(t), x'(t))\, dt. \tag{3.4}$$

By (3.3) and (3.4) there is an integer j such that $1 \le j \le kn - 1$ and

$$\int_{S+jT}^{S+(j+1)T} f(t, x(t), x'(t))\, dt > E(\alpha)T + (kn)^{-1}\Gamma_0(\alpha). \tag{3.5}$$

We next construct a function $y \in W^{1,1}(S, S + n(kT + 1))$ that is equal to x on the intervals $[S, S + jT - 1]$ and $[S + (j + 1)T + 1, S + n(kT + 1)]$ and equal to the translation of x on the interval $[S + jT, S + (j + 1)T]$. It will follow from its construction and from (3.3) and (3.5) that

$$\int_S^{S+n(kT+1)} [f(t, x(t), x'(t)) - f(t, y(t), y'(t))]\, dt > 0.$$

For $t \in R^1$ we set

$$\text{Int}(t) = \sup\{i : i \in \mathbf{Z},\ i \le t\}.$$

There exists $y \in W^{1,1}(S, S + n(kT + 1))$ such that

$$y(t) = x(t) \quad \text{for every } t \in [S, S + jT - 1] \cup [S + (j + 1)T + 1, S + n(kT + 1)],$$

$$y(t) = x(t - jT) + \text{Int}(\alpha jT) \quad \text{for every } t \in [S + jT, S + (j + 1)T],$$

$$\int_\tau^{\tau+1} f(t, y(t), y'(t))\, dt = U(\tau, \tau + 1, y(\tau), y(\tau + 1)) \quad \text{for every } \tau \in [S + jT - 1, S + (j + 1)T].$$

It follows from the definition of y that

$$\int_S^{S+n(kT+1)} [f(t, x(t), x'(t)) - f(t, y(t), y'(t))]\, dt = \sigma_1 + \sigma_2 + \sigma_3 \le 0, \tag{3.6}$$

where

$$\sigma_1 = \int_{S+jT}^{S+(j+1)T} [f(t, x(t), x'(t)) - f(t, y(t), y'(t))]\, dt,$$

$$\sigma_2 = \int_{S+jT-1}^{S+jT} [f(t, x(t), x'(t)) - f(t, y(t), y'(t))]\, dt,$$

$$\sigma_3 = \int_{S+(j+1)T}^{S+(j+1)T+1} [f(t, x(t), x'(t)) - f(t, y(t), y'(t))]\, dt.$$

By (3.3) and (3.5)

$$\sigma_1 > \Gamma_0(\alpha). \tag{3.7}$$

It follows from (1.2) that

$$-c_0 - 1 - 2\delta_0|a - b| \le U(i, i + 1, a, b) \le c_0 + \delta_0^{-1}|b - a|^2 \tag{3.8}$$

for any integer $i \geq 0$ and every $a, b \in R^1$. It follows from Proposition 2.4 and (2.2) that

$$|x(\tau) - x(\tau + 1)| \leq |\alpha| + c_1(1 + |\alpha|) \quad \text{for every } \tau \in R^1.$$

Since $x \in \mathcal{M}_{\text{per}}(\alpha)$, the last inequality and (3.8) imply that for every $\tau \in [S + jT - 1, S + (j + 1)T]$

$$\left| \int_\tau^{\tau+1} f(t, x(t), x'(t))\, dt \right| \leq c_0 + 1 + 2\delta_0^{-1}[1 + (|\alpha| + c_1(1 + |\alpha|))^2]. \tag{3.9}$$

It follows from Proposition 2.4, the definition of y, and (2.2) that

$$|y(S + jT - 1) - y(S + jT)| \leq |x(S + jT - 1) - x(S) - \alpha jT| + 1 \leq (1 + |\alpha|)(1 + c_1)$$

and

$$\begin{aligned} |y(S + (j + 1)T) - y(S + 1 + (j + 1)T| &= |x(S + T) + \alpha jT - x(S + 1 + (j + 1)T)| + 1 \\ &\leq (1 + |\alpha|)(1 + c_1). \end{aligned}$$

These relations, (3.8), and the definition of y imply that

$$\left| \int_\tau^{\tau+1} f(t, y(t), y'(t))\, dt \right| \leq c_0 + 1 + 2\delta_0^{-1}[1 + (1 + |\alpha|)^2(1 + c_1)^2] \tag{3.10}$$

for $\tau \in [jT + S - 1, S + (j + 1)T]$.

In view of (3.9) and (3.10) we obtain

$$|\sigma_2|, |\sigma_3| \leq 2c_0 + 2 + 4\delta_0^{-1}(1 + (1 + |\alpha|)^2(1 + c_1)^2).$$

We now conclude from this relation, (3.6), (3.7), and the definition of $\Gamma_0(\alpha)$ that

$$0 \geq \sigma_1 + \sigma_2 + \sigma_3 > \Gamma_0(\alpha) - 4c_0 - 4 + 8\delta_0^{-1}(1 + (1 + |\alpha|^2)^2(1 + c_1)^2) > 0.$$

This contradiction concludes the proof of the proposition. □

In the paragraph that followed Proposition 2.4 we defined the rotation number for minimal solutions without self-intersections. In the next proposition we define it for *any* minimal solution.

PROPOSITION 3.2. *If x is a minimal solution of* (P), *then there exists a real number α that is called the rotation number of x such that*

$$|x(t + s) - x(t) - \alpha s| \leq c_1(1 + |\alpha|) + 6$$

for all s and t (where c_1 is as in Proposition 2.4).

Proof. Let x be a minimal solution of (P). For a natural number N set

$$r_N = \max\{i \in \mathbf{Z} : i \leq x(N) - x(-N)\}.$$

There exists $y_N \in \mathcal{M}_{\text{per}}(r_N/2N)$ such that

$$|y_N(-N) - x(-N)| \leq 1.$$

Clearly

$$|y_N(N) - x(N)| = |y_N(-N) + r_N - x(N)| \leq 2.$$

By Proposition 2.12

$$|y_N(t) - x(t)| \leq 3 \quad \text{for every } t \in [-N, N]. \tag{3.11}$$

It follows from Proposition 2.4 that

$$|y_N(t+s) - y_N(t) - (r_N/2N)s| \leq c_1(1 + |r_N|/2N)$$

for all numbers s and t.

We will next prove that the sequence $\{r_N/2N\}_{N=1}^{\infty}$ is bounded. Fix an integer $N_0 > c_1$. For $N \geq N_0$ we have

$$|y_N(N_0) - y_N(0) - (r_N/2N)N_0| \leq c_1(1 + |r_N|/2N),$$

which, in view of (3.11), implies that

$$|x(N_0) - x(0) - (r_N/2N)N_0| \leq 6 + c_1(1 + |r_N|/2N).$$

Dividing the last inequality by $|r_N|/2N$ (whenever it is nonzero) yields

$$\left| \frac{x(N_0) - x(0)}{r_N/2N} - N_0 \right| \leq c_1 + \frac{c_1 + 6}{|r_N|/2N},$$

which implies a contradiction to $N_0 > c_1$ if $\{r_N/2N\}_{N=1}^{\infty}$ is unbounded.

We have thus established that the sequence $\{r_N/2N\}_{N=N_0}^{\infty}$ is bounded, and we have that

$$y_N \in \mathcal{M}_{\text{per}}(r_N/2N) \subset \mathcal{M}(r_N/2N) \quad \text{for every } N \geq N_0$$

(recall (2.3) and (2.4)). By Proposition 2.6 there exists a subsequence $\{y_{N_k}\}_{k=1}^{\infty}$ that converges with first derivatives uniformly on any compact set to a function $z(\cdot) \in \mathcal{M}_A$ where

$$A = \sup \left\{ \frac{|r_N|}{2N} : N \geq N_0 \right\}.$$

There is a number α such that $z(\cdot) \in \mathcal{M}(\alpha)$, and it follows from (3.11) that

$$|z(t) - x(t)| \leq 3 \quad \text{for every real } t.$$

This last relation and Proposition 2.4 imply that

$$|x(t+s) - x(t) - \alpha s| \leq c_1(1 + |\alpha|) + 6$$

for all s and t. The proposition is proved. □

PROPOSITION 3.3. *Assume that x is a minimal solution of* (P), *α is the rotation number of x, and $\{\alpha_i\}$ is a sequence of rational numbers such that $\alpha_i \to \alpha$ as $i \to \infty$. Then there exists a real number $E_f(\alpha)$ such that the following limit exists*

$$E_f(\alpha) = \lim_{i \to \infty} E(\alpha_i)$$

and is independent on the sequence $\{\alpha_i\}_{i=1}^{\infty}$. Moreover, for each α there exists a real number $\Gamma_f(\alpha) > 0$ depending monotonically increasing on $|\alpha|$ such that

$$\left| \int_S^{S+T} f(t, x(t), x'(t))\, dt - E_f(\alpha)T \right| \leq \Gamma_f(\alpha)$$

for all $S, T \in R^1$.

Proof. For every natural number i there is $y_i \in \mathcal{M}_{\text{per}}(\alpha_i)$ such that

$$|y_i(0) - x(0)| \le 1. \tag{3.12}$$

Let $T_1, T_2 \in R^1$, $T_1 < T_2$. It follows from Proposition 3.2 and (3.12) that there exists an integer $i_0 \ge 1$ such that for every integer $i \ge i_0$

$$\sup\{|y_i(t) - x(t)| : t \in R^1,\ |t| \le |T_1| + |T_2| + 1\} \le 2c_1(1 + |\alpha|) + 15. \tag{3.13}$$

Let $i \ge i_0$ be an integer. We will next define a $z \in W^{1,1}_{\text{loc}}(R^1)$ that is equal to x on the intervals $(-\infty, T_1 - 1]$ and $[T_2 + 1, \infty)$ and equal to y_i on the interval $[T_1, T_2]$. We also define $u \in W^{1,1}_{\text{loc}}(R^1)$ that is equal to y_i on the intervals $(-\infty, T_1 - 1]$ and $[T_2 + 1, \infty)$ and equal to x on the interval $[T_1, T_2]$. The fact that x and y_i are minimal solutions and the definitions of z and u will imply that $|I(T_1, T_2, x) - I(T_1, T_2, y_i)|$ is bounded by a bound that depends only on $|\alpha|$.

More explicitly we define $u, z \in W^{1,1}_{\text{loc}}(R^1)$ such that

$$z(t) = \begin{cases} x(t) & \text{for } t \in (-\infty, T_1 - 1] \cup [T_2 + 1, \infty), \\ y_i(t) & \text{for } t \in [T_1, T_2]; \end{cases}$$

$$I(T_1 - 1, T_1, z) = U(T_1 - 1, T_1, x(T_1 - 1), y_i(T_1)),$$
$$I(T_2, T_2 + 1, z) = U(T_2, T_2 + 1, y_i(T_2), x(T_2 + 1));$$

and

$$u(t) = \begin{cases} y_i(t) & \text{for } t \in (-\infty, T_1 - 1] \cup [T_2 + 1, \infty), \\ x(t) & \text{for } t \in [T_1, T_2]; \end{cases}$$

$$I(T_1 - 1, T_1, u) = U(T_1 - 1, T_1, y_i(T_1 - 1), x(T_1)),$$
$$I(T_2, T_2 + 1, z) = U(T_2, T_2 + 1, x(T_2), y_i(T_2 + 1)).$$

Since x and y_i are minimal solutions, we obtain

$$\begin{aligned} I(T_1 - 1, T_2 + 1, x) &\le I(T_1 - 1, T_2 + 1, z), \\ I(T_1 - 1, T_2 + 1, y_i) &\le I(T_1 - 1, T_2 + 1, u). \end{aligned} \tag{3.14}$$

It follows from (3.13) and Proposition 3.2 that

$$|u(\tau) - u(\tau + 1)|, |z(\tau) - z(\tau + 1)| \le c_2(|\alpha| + 1)$$

with $c_2 = 3c_1 + 21$ for $\tau = T_1 - 1$ and $\tau = T_2$. These relations and the definitions of z and u and (1.2) imply that for $\tau = T_1 - 1$ and $\tau = T_2$ and for $v = z$ and $v = u$ we have

$$\begin{aligned} -c_0 &\le \int_\tau^{\tau+1} f(t, v(t), v'(t))\, dt \\ &\le \sup\{|f(t, w, p)| : (t, w, p) \in R^3,\ |p| \le |v(\tau) - v(\tau + 1)|\} \\ &\le \sup\{|f(t, w, p)| : (t, w, p) \in R^3,\ |p| \le c_2(1 + |\alpha|)\}. \end{aligned}$$

We conclude from these relations and (3.14) that

$$\begin{aligned} |I(T_1, T_2, x) - I(T_1, T_2, y_i)| &\le 2c_0 \\ &\quad + 2\sup\{|f(t, w, p)| : (t, w, p) \in R^3,\ |p| \le c_2(1 + |\alpha|)\}. \end{aligned}$$

In view of Proposition 3.1 this relation yields

$$\left|\int_{T_1}^{T_2} [f(t, x(t), x'(t)) - E(\alpha_i)]\, dt\right| \leq \Gamma(\alpha_i) + \Phi(\alpha), \tag{3.15}$$

where

$$\Phi(\alpha) = 2c_0 + 2\sup\{|f(t, w, p)| : (t, w, p) \in R^3,\ |p| \leq c_2(1 + |\alpha|)\}. \tag{3.16}$$

We have thus proved that for each $T_1, T_2 \in R^1$ satisfying $T_1 < T_2$ there exists a natural number i_0 such that for every integer $i \geq i_0$

$$\left|\int_{T_1}^{T_2} [f(t, x(t), x'(t)) - E(\alpha_i)]\, dt\right| \leq \Phi(\alpha) + \max\{\Gamma(\alpha - 1), \Gamma(\alpha + 1)\}.$$

Clearly this implies the existence of $\lim_{i\to\infty} E(\alpha_i)$, which completes the proof of the proposition. □

For any $\alpha \in R$ we set

$$E_f(\alpha) = \lim_{i\to\infty} E(\alpha_i), \tag{3.17}$$

where $\{\alpha_i\}_{i=1}^{\infty}$ is a sequence of rational numbers such that $\alpha_i \to \alpha$ as $i \to \infty$.

By Proposition 3.3 the function $E_f : R \to R$ is well defined. It follows from Propositions 3.3 and 3.1 that $E_f(\alpha) = E(\alpha)$ for every rational α. It follows easily from Proposition 3.3 that $E_f(\cdot)$ is continuous.

PROPOSITION 3.4. *The function E_f is strictly convex.*

Proof. Let $\alpha_1, \alpha_2, \beta$ be rational numbers and $\beta \in (0, 1)$, $\alpha_1 \neq \alpha_2$. We claim that it is sufficient to show that

$$E_f(\beta\alpha_1 + (1 - \beta)\alpha_2) < \beta E_f(\alpha_1) + (1 - \beta)E_f(\alpha_2).$$

This is so since this will imply that E_f is convex, hence it is strictly convex. There are natural numbers n_1, n_2, q_1, q_2 and integers m_1, m_2 satisfying

$$\beta = q_1/q_2, \quad \alpha_i = m_i/n_i, \quad i = 1, 2,$$

and $x_i \in \mathcal{M}_{\text{per}}(\alpha_i)$, $i = 1, 2$. We have

$$\beta\alpha_1 + (1 - \beta)\alpha_2 = [m_1q_1n_2 + m_2(q_2 - q_1)n_1]/(n_1n_2q_2).$$

Fix a large natural number N such that $N/(n_1n_2q_2)$ is an integer, and set

$$k_1 = Nq_1n_2, \qquad k_2 = N(q_2 - q_1)n_1.$$

We may assume without loss of generality that

$$|x_1(k_1n_1) - x_2(0)| \leq 1. \tag{3.18}$$

Let us prove that

$$E_f(\beta\alpha_1 + (1 - \beta)\alpha_2) \leq \beta E_f(\alpha_1) + (1 - \beta)E_f(\alpha_2). \tag{3.19}$$

We will construct a function $x \in W^{1,1}(0, k_1n_1 + k_2n_2)$ that is equal to x_1 on the interval $[0, k_1n_1 - 1]$, is equal to a translation of x_2 on the interval $[k_1n_1, k_1n_1 + k_2n_2 - 1]$, and satisfies

$$x(k_1n_1 + k_2n_2) = x(0) + k_1m_1 + k_2m_2.$$

Using the definition of x we will show that

$$I(0, k_1n_1 + k_2n_2, x) - I(0, k_1n_1, x_1) - I(0, k_2n_2, x_2)$$

does not exceed some constant that depends only on $|\alpha_1|$ and $|\alpha_2|$. Since $E(\beta\alpha_1+(1-\beta)\alpha_2) \le I(0, k_1n_1 + k_2n_2, x)/(q_2n_1n_2N)$, this will imply (3.19).

There exists $x \in W^{1,1}(0, k_1n_1 + k_2n_2)$ such that

$$x(t) = x_1(t)(t \in [0, k_1n_1 - 1]), \qquad x(t) = x_2(t - k_1n_1)(t \in [k_1n_1, k_1n_1 + k_2n_2 - 1]),$$
$$x(k_1n_1 + k_2n_2) = x(0) + k_1m_1 + k_2m_2,$$

$$x(\tau + t) = tx(\tau + 1) + (1 - t)x(\tau) \qquad (t \in (0, 1), \tau \in \{k_1n_1 - 1, k_1n_1 + k_2n_2 - 1\}).$$

It follows from Proposition 2.4 and (3.18) that

$$\begin{aligned} |x_1(k_1n_1 - 1) - x_2(0)| &\le |x_1(k_1n_1 - 1) - x_1(k_1n_1)| + |x_1(k_1n_1) - x_2(0)| \\ &\le c_1(1 + |\alpha_1|) + 1 + |\alpha_1|, \end{aligned}$$

$$\begin{aligned} &|x(k_1n_1 + k_2n_2 - 1) - x(k_1n_1 + k_2n_2)| \\ &\quad \le |x_2(k_2n_2 - 1) - x_2(k_2n_2)| + |x_2(k_2n_2) - x(0) - k_1m_1 - k_2m_2| \\ &\quad \le c_1(1 + |\alpha_2|) + 1 + |\alpha_2|. \end{aligned}$$

From these relations, the definition of x, and (1.2) we conclude that

$$\begin{aligned} &I(0, k_1n_1 + k_2n_2, x) - I(0, k_1n_1, x_1) - I(0, k_2n_2, x_2) \\ &\quad \le I(k_1n_1 - 1, k_1n_1, x) + I(k_1n_1 + k_2n_2 - 1, k_1n_1 + k_2n_2, x) \\ &\qquad - I(k_1n_1 - 1, k_1n_1, x_1) - I(k_2n_2 - 1, k_2n_2, x_2) \\ &\quad \le 2\delta_0 + 2\sup\{|f(t, w, p)| : (t, w, p) \in R^3,\ |p| \le (c_1 + 1)(1 + |\alpha_1| + |\alpha_2|)\}. \end{aligned}$$

This inequality and the relation $x_i \in \mathcal{M}_{\text{per}}(\alpha_i)$ for $i = 1, 2$ yield

$$\begin{aligned} E(\beta\alpha_1 + (1 - \beta)\alpha_2) &\le I(0, k_1n_1 + k_2n_2, x)/(q_2n_1n_2N) \\ &\le (Nq_2n_1n_2)^{-1}[2\delta_0 + 2\sup\{|f(t, w, p)| : (t, w, p) \in R^3, \\ &\qquad |p| \le (c_1 + 1)(1 + |\alpha_1| + |\alpha_2|)\} \\ &\qquad + k_1n_1E(\alpha_1) + E(\alpha_2)k_2n_2] \\ &= \beta E(\alpha_1) + (1 - \beta)E(\alpha_2) + (Nq_2n_1n_2)^{-1} \\ &\qquad \cdot [2\delta_0 + 2\sup\{|f(t, w, p)| : (t, w, p) \in R^3, \\ &\qquad\qquad |p| \le (c_1 + 1)(1 + |\alpha_1| + |\alpha_2|)\}]. \end{aligned}$$

Since this relation holds for any sufficiently large integer N such that $N/q_2n_1n_2$ is an integer, (3.19) is established.

We will next prove that strict inequality holds in (3.19), which will conclude the proof.

We may assume without loss of generality that $\alpha_1 < \alpha_2$. By Proposition 2.3 x_1 and x_2 satisfy Euler's equation. It follows from Propositions 2.4 and 2.11 that there is $\tau_0 \in R^1$ such that

$$x_2(t) > x_1(t) \quad \text{for every } t \in R^1,$$
$$t > \tau_0, \quad x_2(t) < x_1(t) \quad \text{for every } t \in R^1,\ t < \tau_0.$$

There exists $y \in W^{1,1}_{\text{loc}}(R^1)$ such that

$$\begin{aligned} y(t + q_2 n_1 n_2) &= y(t) + q_1 n_2 m_1 + (q_2 - q_1) m_2 n_1 \quad \text{for every } t \in R^1, \\ y(t) &= x_1(t) \quad \text{for every } t \in [\tau_0 - q_1 n_1 n_2, \tau_0], \\ y(t) &= x_2(t) \quad \text{for every } t \in [\tau_0, \tau_0 + (q_2 - q_1) n_1 n_2]. \end{aligned} \tag{3.20}$$

Assume that

$$\begin{aligned} &I(\tau_0 - q_1 n_1 n_2, \tau_0 + (q_2 - q_1) n_1 n_2, y)/q_2 n_1 n_2 \\ &\quad = E([q_1 n_2 m_1 + (q_2 - q_1) m_2 n_1]/q_2 n_1 n_2). \end{aligned}$$

Then y and x_1 satisfy the Euler–Lagrange equation, and by (3.20) $y(t) = x_1(t)$ for every $t \in R^1$. It follows from this and the definition of $y(\cdot)$ that $x_1(t) = x_2(t)$ for every t, contradicting $\alpha_1 \neq \alpha_2$. The resulting contradiction proves that

$$\begin{aligned} E(\beta\alpha_1 + (1 - \beta)\alpha_2) &< (q_2 n_1 n_2)^{-1} I(\tau_0 - q_1 n_1 n_2, \tau_0 + (q_2 - q_1) n_1 n_2, y) \\ &= \beta E(\alpha_1) + (1 - \beta) E(\alpha_2). \end{aligned}$$

The proposition is proved. □

PROPOSITION 3.5. *$E_f(\alpha) \to +\infty$ as $|\alpha| \to \infty$.*

Proof. Let $\alpha \in R$, $x \in \mathcal{M}(\alpha)$. Propositions 2.4 and 3.3 imply that for each $T > 0$

$$|x(T) - x(0) - \alpha T| \leq c_1(1 + |\alpha|), \tag{3.21}$$

$$\left| \int_0^T f(t, x(t), x'(t))\, dt - E_f(\alpha) T \right| \leq \Gamma_f(\alpha). \tag{3.22}$$

It follows from (1.2) that we have for every $T > 0$

$$\int_0^T f(t, x(t), x'(t))\, dt \geq \int_0^T [-c_0 + \delta_0 |x'(t)| - 1]\, dt \geq -(c_0 + 1)T + \delta_0 |x(T) - x(0)|. \tag{3.23}$$

We obtain from (3.21) that

$$|x(T) - x(0)| \geq |\alpha| T - c_1(1 + |\alpha|),$$

which, in view of (3.22) and (3.23),yields

$$T E_f(\alpha) \geq -\Gamma_f(\alpha) - (c_0 + 1)T + \delta_0 |\alpha| T - \delta_0 c_1 (1 + |\alpha|).$$

Dividing the last inequality by T and then letting $T \to \infty$ we obtain

$$E_f(\alpha) \geq |\alpha| \delta_0 - c_0 - 1.$$

This completes the proof of the proposition. □

Theorem 1.1 now follows from Propositions 3.1–3.5.

4. Discrete reformulation of (P). For $x, y \in R^1$ we set

$$v(x, y) = \inf \left\{ \int_0^1 f(t, u(t), u'(t))\, dt : u \in W^{1,1}(0, 1),\ u(0) = x,\ u(1) = y \right\}. \tag{4.1}$$

It is easy to verify that

$$
\begin{aligned}
&v(x+1, y+1) = v(x, y), \qquad x, y \in R^1;\\
&\inf\{v(x, y) : x, y \in R^1\} \geq -c_0;\\
&\qquad\qquad v(x, y) \to +\infty \quad \text{as } |x - y| \to \infty.
\end{aligned} \tag{4.2}
$$

PROPOSITION 4.1. *The function $v : R^2 \to R^1$ is continuous.*

Proof. The lower semicontinuity of v follows from Propositions 2.1 and 2.2, and we will establish the upper semicontinuity of v. Let $a, b \in R^1$; $y \in W^{1,1}(0, 1)$; $y(0) = a$; $y(1) = b$; $v(a, b) = I(0, 1, y)$; $\{a_i\}_{i=1}^{\infty}$, $\{b_i\}_{i=1}^{\infty} \subset R^1$, $a_i \to a$, $b_i \to b$ as $i \to \infty$; and actually $y \in C^2[0, 1]$. For an integer $i \geq 1$ we define

$$\Phi_i(t) = h_i t + d_i \quad \text{for } t \in [0, 1],$$

where $d_i = a_i - a$, $h_i = b_i - b - d_i$. Clearly

$$y(0) + \Phi_i(0) = a_i, \quad y(1) + \Phi_i(1) = b_i \quad \text{for } i \geq 0,$$

$$y(t) + \Phi_i(t) \to y(t), \ y'(t) + \Phi_i'(t) \to y'(t) \text{ uniformly in } [0, 1] \text{ as } i \to \infty.$$

Therefore

$$
\begin{aligned}
\limsup_{i\to\infty} v(a_i, b_i) &\leq \limsup_{i\to\infty} \int_0^1 f(t, y(t) + \Phi_i(t), y'(t) + \Phi_i'(t))\, dt\\
&= \int_0^1 f(t, y(t), y'(t))\, dt = v(a, b).
\end{aligned}
$$

This completes the proof of the proposition. □

Set

$$\mu = \inf\left\{\liminf_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} v(z_i, z_{i+1}) : \{z_i\}_{i=0}^{\infty} \subset R^1\right\}, \tag{4.3}$$

which is the minimal long-run average cost. Similarly to results in [9], with slight changes in the proofs, we obtain the following two results.

PROPOSITION 4.2. *There exists a constant M_0 such that*

(i) *for every sequence $\{z_i\}_{i=0}^{\infty} \subset R^1$ and every integer $N \geq 1$*

$$\sum_{i=0}^{N-1} [v(z_i, z_{i+1}) - \mu] \geq -M_0;$$

(ii) *for every initial value z_0 there is a sequence $\{z_i^*\}_{i=0}^{\infty}$ with $z_0^* = z_0$ that satisfies*

$$\left|\sum_{i=0}^{N-1} [v(z_i^*, z_{i+1}^*) - \mu]\right| \leq 4M_0$$

for all $N \geq 1$;

(iii) *for every sequence $\{z_i\}_{i=0}^{\infty} \subset R^1$ the sequence*

$$\left\{\sum_{i=0}^{N-1} [v(z_i, z_{i+1}) - \mu]\right\}_{N=1}^{\infty}$$

either is bounded or diverges to infinity.

PROPOSITION 4.3. *We define*

$$\pi(a) = \inf\left\{\liminf_{N\to\infty}\sum_{i=0}^{N-1}[v(z_i, z_{i+1}) - \mu] : \{z_i\}_{i=0}^{\infty} \subset R^1,\ z_0 = a\right\},$$

$$\theta(a, b) = v(a, b) - \mu + \pi(b) - \pi(a)$$

for $a, b \in R^1$. *Then the functions* $\pi : R^1 \to R^1$ *and* $\theta : R^2 \to R^1$ *are continuous*;

$$\pi(a + 1) = \pi(a), \quad \theta(a + 1, b + 1) = \theta(a, b) \quad \textit{for all } a, b, \in R^1;$$

the function θ *is nonnegative*; *and the set*

$$F(a) = \{b \in R^1 : \theta(a, b) = 0\}$$

is nonempty for every $a \in R^1$.

THEOREM 4.1. (i) *For each* $x \in W^{1,1}_{\mathrm{loc}}([0, \infty))$ *and each* $T > 0$

$$\int_0^T [f(t, x(t), x'(t)) - \mu]\, dt \geq -M_0 - c_0 - |\mu|.$$

(ii) *There exists a unique number* α_f *such that*

$$E_f(\alpha_f) = \inf\left\{E_f(\beta) : \beta \in R^1\right\} = \mu.$$

Proof. Assertion (i) follows from Proposition 4.2 and (1.2). We will prove assertion (ii). By Propositions 3.4 and 3.5 there exists a unique number α_f such that

$$E_f(\alpha_f) = \inf\{E_f(\beta) : \beta \in R^1\}.$$

Choose $x \in \mathcal{M}(\alpha_f)$. It follows from assertion (i) and Proposition 3.3 that

$$E_f(\alpha_f) \geq \mu. \tag{4.4}$$

There exists a sequence $\{z_i\}_{i=0}^{\infty} \subset R^1$ such that $\theta(z_i, z_{i+1}) = 0$ for $i = 0, 1, \ldots$. By (4.2) the sequence $\{z_i - z_{i+1}\}_{i=0}^{\infty}$ is bounded. For any $i \in \{0, 1, \ldots\}$ there is an integer k_i such that $|z_i - z_0 + k_i| < 1$. For $i \in \{0, 1, \ldots\}$ and $p \in \{-i, -i + 1, \ldots\}$ we set $y_p^i = z_{p+i} + k_i$, $Y^i = \{y_p^i\}_{p=-i}^{\infty}$.

It is easy to verify that for every integer p the set

$$\{y_p^i : i \in Z,\ i \geq \max\{0, -p\}\}$$

is bounded. Therefore there exist a sequence $\{a_p\}_{p=-\infty}^{\infty} \subset R^1$ and a subsequence $\{Y^{i_s}\}_{s=1}^{\infty}$ such that $y_p^{i_s} \to a_p$ as $s \to \infty$ for any integer p. Evidently $\theta(a_p, a_{p+1}) = 0$ for any integer p. Recalling Proposition 2.2 we can find an $x \in W^{1,1}_{\mathrm{loc}}(R^1)$ such that

$$x(p) = a_p, \qquad I(p, p + 1, x) = U(p, p + 1, x(p), x(p + 1)) = v(x(p), x(p + 1))$$

for any integer p.

It is easy to verify that x is a minimal solution of (P) and

$$\sup\left\{\left|\int_i^j [f(t, x(t), x'(t)) - \mu]\, dt\right| : i, j \in Z\right\} < \infty. \tag{4.5}$$

By Proposition 3.2 the minimal solution x has a rotation number α. It follows from (4.5) and Proposition 3.3 that

$$\mu = E_f(\alpha) \geq E_f(\alpha_f),$$

and in view of (4.4) we obtain $E_f(\alpha_f) = \mu$ and $\alpha_f = \alpha$. This completes the proof of the theorem. □

5. Minimal solutions with irrational rotation number. To establish Theorem 1.2 we need to study in more details the structure of minimal solutions with irrational rotation number. Some of the results in this section were established in [16], while others were presented in [15] without proof. We remark that their proof is based on the approach developed by Aubry and Le Daeron [2] for discrete-time systems (see also [19]).

Let x be a minimal solution of (P) that has an irrational rotation number α. The minimal solution x is called *regular* if for every pair of integers j, k

$$x(t+j) - k - x(t) > 0 \quad \text{for all } t \in R^1 \text{ iff } \alpha j - k > 0.$$

We will see in Proposition 5.4 that every minimal solution with irrational rotation number is regular. At this stage we can merely establish the existence of a regular minimal solution with a prescribed irrational rotation number.

Propositions 2.4, 2.6, and 2.10 imply the following result.

PROPOSITION 5.1. *Let α be an irrational number and $\{\alpha_i\}_{i=1}^{\infty}$ be a sequence of rational numbers such that $\alpha_i \to \alpha$ as $i \to \infty$. Assume that $x_i \in \mathcal{M}_{\text{per}}(\alpha_i)$, $|x_i(0)| \leq 1, i = 1, 2, \ldots,$ and that the sequence $\{x_i\}_{i=1}^{\infty}$ converges with first derivatives uniformly on any compact set to a function $x \in W^{1,1}_{\text{loc}}(R^1)$. Then $x \in \mathcal{M}(\alpha)$ is a regular minimal solution.*

COROLLARY 5.1. *For every irrational number α there is a regular minimal solution $x \in \mathcal{M}(\alpha)$.*

Assume that α is an irrational number and $x \in \mathcal{M}(\alpha)$ is regular. Using the regular minimal solution x and following [2], [15], [16], and [19] we will obtain an explicit expression for all minimal solutions with rotation number α. Using this explicit expression we will prove that every minimal solution with rotation number α is regular.

We define a function $U^x(t, \theta)$ by

$$U^x(t, \theta) = x(t+j) - k, \tag{5.1}$$

where $t \in R^1$ and θ is any number of the form $\alpha t + \alpha j - k$ for some integers j and k (see [15], [16]). Clearly the function $U^x(t, \theta)$ is strictly monotone in θ on the dense set on which it is defined. One can extend U^x to functions U^x_+, U^x_- by

$$U^x_+(t, \theta) = \lim_{\theta' \to \theta+} U^x(t, \theta'), \; U^x_-(t, \theta) = \lim_{\theta'' \to \theta-} U^x(t, \theta''), \tag{5.2}$$

where θ' (resp., θ'') are decreasing (resp., increasing) sequences taken from the dense set on which U^x is defined. Clearly the functions $U^x_{\pm}(t, \theta) - \theta$ have period 1 in t, θ,

$$U^x_{\pm}(t+1, \theta) = U^x_{\pm}(t, \theta), \qquad U^x_{\pm}(t, \theta+1) = U^x_{\pm}(t, \theta) + 1. \tag{5.3}$$

Moreover, using Propositions 2.4 and 2.6 one obtains that

$$U^x_{\pm}(t, \alpha t + \beta) \in \mathcal{M}(\alpha) \tag{5.4}$$

for every choice of β.

A minimal solution $u \in \mathcal{M}(\alpha)$ is called *recurrent* if there exist sequences of integers $\{k_p\}_{p=1}^{\infty}$, $\{j_p\}_{p=1}^{\infty}$ such that $|k_p| + |j_p| \to \infty$ as $p \to \infty$, and u is the limit of a sequence of translates $\{u_p\}_{p=1}^{\infty} \subset \mathcal{M}(\alpha)$ in the C^1-topology on compact sets in R^1 where

$$u_p(t) = u(t + j_p) - k_p, \qquad t \in R^1,\ p = 1, 2, \ldots.$$

We have the following result (see [16, Thm. 6.5]).

PROPOSITION 5.2. *The solutions $U_{\pm}^x(t, \alpha t + \beta)$ are recurrent for every β.*

The validity of Proposition 5.2 follows from (5.2) and (5.3).

In the next proposition we will show that all the minimal solutions y that are of the form $y(t) = U_{\pm}^x(t, \alpha t + \beta)$ for some $\beta \in R^1$ have the same limit point set on the torus.

PROPOSITION 5.3. *There exists a closed set $H(\alpha) \subset R^1$ such that for every number β the set $H(\alpha)$ consists of all the points z that are of the form*

$$z = \lim_{p \to \infty} U_+^x(t_p, \alpha t_p + \beta) - i_p$$

for some sequence of integers $\{i_p\}_{p=1}^{\infty}$ and a sequence $\{t_p\}_{p=1}^{\infty} \subset R^1$ with $|t_p| \to \infty$ as $p \to \infty$. Moreover, for every β the set $H(\alpha)$ consists of all the points z that are of the form

$$z = \lim_{p \to \infty} U_-^x(t_p, \alpha t_p + \beta) - i_p$$

for some sequence of integers $\{i_p\}_{p=1}^{\infty}$ and a sequence $\{t_p\}_{p=1}^{\infty} \subset R^1$ with $|t_p| \to \infty$ as $p \to \infty$.

Proof. Let $\beta_1, \beta_2 \in R^1$, $\beta_1 \neq \beta_2$, $\{i_p\}_{p=1}^{\infty}$ be a sequence of integers, and $\{t_p\}_{p=1}^{\infty} \subset R^1$ be a sequence of numbers such that

$$|t_p| \to \infty \quad \text{as } p \to \infty, \quad U_+^x(t_p, \alpha t_p + \beta_2) - i_p \to z \quad \text{as } p \to \infty.$$

It follows from (5.2) that for every integer $p \geq 1$ there exist integers k_p and j_p such that $|j_p| \geq 10p|t_p|$, $\alpha j_p - k_p > \beta_2 - \beta_1$, and

$$U_+^x(t_p, \alpha t_p + \beta_2) > U_+^x(t_p, \alpha t_p + \beta_1 + \alpha j_p - k_p) - p^{-1}.$$

It follows from (5.3) that

$$\begin{aligned} 0 &< U_-^x(t_p + j_p, \alpha(t_p + j_p) + \beta_1) - k_p - U_+^x(t_p, \alpha t_p + \beta_2) \\ &\leq U_+^x(t_p + j_p, \alpha(t_p + j_p) + \beta_1) - k_p - U_+^x(t_p, \alpha t_p + \beta_2) \leq p^{-1} \end{aligned}$$

for any natural number p. Therefore

$$|j_p + t_p| \to \infty, \quad \text{as } p \to \infty, \tag{5.5}$$

$$U_+^x(t_p + j_p, \alpha(t_p + j_p) + \beta_1) - k_p - i_p \to z \quad \text{as } p \to \infty, \tag{5.6}$$

$$U_-^x(t_p + j_p, \alpha(t_p + j_p) + \beta_1) - k_p - i_p \to z \quad \text{as } p \to \infty. \tag{5.7}$$

Analogously we can show that if $\{i_p\}_{p=1}^{\infty}$ is a sequence of integers and $\{t_p\}_{p=1}^{\infty}$ is a sequence of numbers that satisfy

$$|t_p| \to \infty, \quad U_-^x(t_p, \alpha t_p + \beta_2) - i_p \to z \quad \text{as } p \to \infty,$$

then there exist sequences of integers $\{j_p\}_{p=1}^{\infty}$, $\{k_p\}_{p=1}^{\infty}$ such that (5.5), (5.6), and (5.7) hold. This completes the proof of the proposition. □

The following result was presented in [15] without proof. It asserts that every minimal solution with an irrational rotation number α is regular.

PROPOSITION 5.4. *Assume that y is a minimal solution of* (P), *α is the rotation number of y, and j_1, j_2, k_1, k_2 are integers satisfying $\alpha j_1 - k_1 > \alpha j_2 - k_2$. Then*

$$y(t + j_1) - k_1 > y(t + j_2) - k_2 \quad \textit{for all } t \in R^1.$$

Proof. Set $\beta_i = \alpha j_i - k_i$, $i = 1, 2$; $y_i(t) = y(t + j_i) - k_i$ for every $t \in R^1$, $i = 1, 2$. By Proposition 3.2

$$\sup\{|y_2(t) - y_1(t)| : t \in R^1\} < \infty.$$

We will show next that $\limsup_{t\to\infty}[y_1(t) - y_2(t)] > 0$ and $\limsup_{t\to-\infty}[y_1(t) - y_2(t)] > 0$. It follows from Proposition 3.2, (5.3), and (5.4) that there exist numbers $\ell_2 > \ell_1$ for which

$$U_-^x(t, \alpha t + \ell_1) \le y(t) \le U_+^x(t, \alpha t + \ell_2) \quad \text{for every } t \in R^1.$$

For $p = 1, 2$ we have

$$U_-^x(t + j_p, \alpha(t + j_p) + \ell_1) - k_p \le y_p(t) \le U_+^x(t + j_p, \alpha(t + j_p) + \ell_2) - k_p, \qquad t \in R^1;$$

$$U_-^x(t, \alpha t + \beta_p + \ell_1) \le y_p(t) \le U_+^x(t, \alpha t + \beta_p + \ell_2) \quad \text{for every } t \in R^1;$$

$$\sup\{y_p(t) - U_-^x(t, \alpha t + \beta_p + \ell_1) : t \in R^1\}$$
$$\le \sup\{U_+^x(t, \alpha t + \beta_p + \ell_2) - U_-^x(t, \alpha t + \beta_p + \ell_1) : t \in R^1\} < \infty, \qquad p = 1, 2.$$

Clearly

$$\begin{aligned}\limsup_{t\to\infty}[y_1(t) - y_2(t)] &\ge \limsup_{t\to\infty}[y_1(t) - U_-^x(t, \alpha t + \beta_1 + \ell_1)] \\ &\quad - \limsup_{t\to\infty}[y_2(t) - U_-^x(t, \alpha t + \beta_2 + \ell_1)] \\ &\quad + \inf\{U_-^x(t, \alpha t + \beta_1 + \ell_1) - U_-^x(t, \alpha t + \beta_2 + \ell_1) : t \in R^1\} \\ &= \inf\{U_-^x(t, \alpha t + \beta_1 + \ell_1) - U_-^x(t, \alpha t + \beta_2 + \ell_1) : t \in R^1\}\end{aligned} \tag{5.8}$$

since the first two terms in the left-hand side of (5.8) cancel each other. Analogously it is easy to show that

$$\begin{aligned}&\limsup_{t\to-\infty}[y_1(t) - y_2(t)] \ge \\ &\quad \inf\{U_-^x(t, \alpha t + \beta_1 + \ell_1) - U_-^x(t, \alpha t + \beta_2 + \ell_1) : t \in R^1\}.\end{aligned} \tag{5.9}$$

We will show that

$$\inf\{U_-^x(t, \alpha t + \beta_1 + \ell_1) - U_-^x(t, \alpha t + \beta_2 + \ell_1) : t \in R^1\} > 0.$$

Assume the converse. Then there exists a sequence $\{t_p\}_{p=1}^{\infty} \subset R^1$ such that

$$U_-^x(t_p, \alpha t_p + \beta_1 + \ell_1) - U_-^x(t_p, \alpha t_p + \beta_2 + \ell_1) \to 0 \quad \text{as } p \to \infty.$$

For $t \in R^1$ we set

$$[t] = \inf\{p \in \{0, \pm 1, \ldots\} : p \le t\}, \qquad \{t\} = t - [t].$$

Evidently

$$U_-^x(\{t_p\}, \alpha\{t_p\} + \{\alpha[t_p]\} + \beta_1 + \ell_1) - U_-^x(\{t_p\}, \alpha\{t_p\} + \{\alpha[t_p]\} + \beta_2 + \ell_1) \to 0 \quad \text{as } p \to \infty. \tag{5.10}$$

We may assume without loss of generality that

$$\{t_p\} \to \tau, \quad \{\alpha[t_p]\} \to h \quad \text{as } p \to \infty, \tag{5.11}$$

and the sequence $\{\alpha[t_p]\}_{p=1}^{\infty}$ is either nonincreasing or nondecreasing. By (5.4) and Proposition 2.6 one of the conditions below is fulfilled:

$$U_-^x(t, \alpha t + \{\alpha[t_p]\} + \beta_j + \ell_1) \to U_-^x(t, \alpha t + h + \beta_j + \ell_1) \quad \text{as } p \to \infty, \ j = 1, 2,$$

uniformly on any compact subset of R^1, or

$$U_-^x(t, \alpha t + \{\alpha[t_p]\} + \beta_j + \ell_1) \to U_+^x(t, \alpha t + h + \beta_j + \ell_1) \quad \text{as } p \to \infty, \ j = 1, 2,$$

uniformly on any compact subset of R^1.

By (5.10) and (5.11) in the first case we have

$$\begin{aligned} U_-^x(\tau, \alpha\tau + h + \beta_1 + l_1) &= \lim_{p\to\infty} U_-^x(\{t_p\}, \alpha\{t_p\} + h + \beta_1 + l_1) \\ &= \lim_{p\to\infty} U_-^x(\{t_p\}, \alpha\{t_p\} + \{\alpha[t_p]\} + \beta_1 + l_1) \\ &= \lim_{p\to\infty} U_-^x(\{t_p\}, \alpha\{t_p\} + \{\alpha[t_p]\} + \beta_2 + l_1) \\ &= \lim_{p\to\infty} U_-^x(\{t_p\}, \alpha\{t_p\} + h + \beta_2 + l_1) \\ &= U_-^x(\tau, \alpha\tau + h + \beta_2 + l_1). \end{aligned}$$

In the second case we obtain analogously

$$U_+^x(\tau, \alpha\tau + h + \beta_1 + \ell_1) = U_+^x(\tau, \alpha\tau + h + \beta_2 + \ell_1).$$

On the other hand, the functions $U_\pm^x(t, \theta)$ are strictly monotone in θ and $\beta_1 > \beta_2$. The obtained contradiction proves that

$$\limsup_{t\to-\infty}[y_1(t) - y_2(t)] > 0, \limsup_{t\to\infty}[y_1(t) - y_2(t)] > 0.$$

It follows from Propositions 2.11 and 2.12 that $y_1(t) > y_2(t)$ for all $t \in R^1$. The proposition is proved. □

Let x be a minimal solution of (P) with an irrational rotation number α. By Proposition 5.4 x is regular and belongs to $\mathcal{M}(\alpha)$. Consider the functions $U_\pm^x(t, \theta)$, $U^x(t, \theta)$ defined by (5.1), (5.2). It follows from Proposition 2.12 that one of the conditions below holds.

(i) $U_+^x(t, \alpha t) = x(t)$ for every $t \in R^1$.

(ii) $U_-^x(t, \alpha t) = x(t)$ for every $t \in R^1$.

(iii) $U_-^x(t, \alpha t) < x(t) < U_+^x(t, \alpha t)$ for every $t \in R^1$.

PROPOSITION 5.5. *Let x be a minimal solution of* (P), *and let the rotation number of x be an irrational number α. If x is recurrent, then one of the conditions below holds:*

$$U_+^x(t, \alpha t) = x(t) \quad \textit{for every } t \in R^1$$

or

$$U_-^x(t, \alpha t) = x(t) \quad \textit{for every } t \in R^1.$$

Proof. There exist sequences of integers $\{k_p\}_{p=1}^{\infty}$ and $\{j_p\}_{p=1}^{\infty}$ such that

$$|k_p| + |j_p| \to \infty, \quad x(j_p) - k_p \to x(0) \quad \text{as } p \to \infty.$$

It is easy to verify that $x(0) \in \{U_+^x(0, 0), U_-^x(0, 0)\}$. This completes the proof of the proposition. □

The following important result was presented in [15] without proof. It establishes that actually the functions U_+^x and U_-^x do not depend on x.

PROPOSITION 5.6. *Assume that α is an irrational number, $x, y \in \mathcal{M}(\alpha)$, and consider the functions $U^x(t, \theta)$, $U^y(t, \theta)$, $U_\pm^x(t, \theta)$, $U_\pm^y(t, \theta)$ (see* (5.1), (5.2)).

Then there exists a number δ such that

$$U_\pm^x(t, \theta) = U_\pm^y(t, \theta + \delta) \quad \textit{for each } \theta, t \in R^1.$$

Proof. Consider numbers β, γ and define

$$\sigma(t) = U_+^x(t, \alpha t + \beta) - U_+^y(t, \alpha t + \gamma), \qquad t \in R^1.$$

By Proposition 2.4

$$\sup\{|\sigma(t)| : t \in R^1\} < \infty.$$

We will show that either $\sigma(t) \geq 0$ for all $t \in R^1$ or $\sigma(t) \leq 0$ for all $t \in R^1$. Assume to the contrary that there exist numbers S, T satisfying

$$\sigma(T) > h, \qquad \sigma(S) < -h,$$

where h is a positive number. Since the functions $U_+^x(t, \theta)$, $U_+^y(t, \theta)$ are right-continuous in θ, there exists $\epsilon > 0$ such that

$$U_+^x(T, \alpha T + \beta + z) - U_+^y(T, \alpha T + \gamma + z) > h,$$
$$U_+^x(S, \alpha S + \beta + z) - U_+^y(S, \alpha S + \gamma + z) < -h$$

for every $z \in (0, \epsilon)$.

There exist sequences of integers $\{p_i\}_{i=1}^{\infty}$, $\{q_i\}_{i=1}^{\infty}$ such that $\{p_i\alpha + q_i\}_{i=1}^{\infty}$ is a decreasing sequence that converges to zero. We may assume without loss of generality that

$$p_i\alpha + q_i \in (0, \epsilon) \quad \text{for every } i = 1, 2, \ldots.$$

Clearly $|p_i| \to \infty$ as $i \to \infty$. For every natural number i we have

$$\begin{aligned}
&U_+^x(T, \alpha T + \beta + p_i\alpha) - U_+^y(T, \alpha T + \gamma + p_i\alpha) \\
&\quad = U_+^x(T, \alpha T + \beta + p_i\alpha + q_i) - U_+^y(T, \alpha T + \gamma + p_i\alpha + q_i) > h, \\
&\quad U_+^x(S, \alpha S + \beta + p_i\alpha) - U_+^y(S, \alpha S + \gamma + p_i\alpha) \\
&\qquad = U_+^x(S, \alpha S + \beta + p_i\alpha + q_i) - U_+^y(S, \alpha S + \gamma + p_i\alpha + q_i) < -h, \\
&U_+^x(T + p_i, \alpha(T + p_i) + \beta) - U_+^y(T + p_i, \alpha(T + p_i) + \gamma) > h, \\
&U_+^x(S + p_i, \alpha(S + p_i) + \beta) - U_+^y(S + p_i, \alpha(S + p_i) + \gamma) < -h.
\end{aligned}$$

These relations are contradictory to Proposition 2.11. Therefore, one of the conditions below holds:

$$\sigma(t) \geq 0 \quad \text{for every } t \in R^1,$$
$$\sigma(t) \leq 0 \quad \text{for every } t \in R^1.$$

Assume that the first condition holds. It follows from the properties of the functions U_+^x, U_+^y (see (5.3)) that

$$U_+^x(t, \alpha t + \beta + \theta) \geq U_+^y(t, \alpha t + \gamma + \theta)$$

for every $t \in R^1$ and $\theta \in \{\alpha p - q : p, q \text{ are integers}\}$.

Since the functions U_+^x, U_+^y are right-continuous in θ, we conclude that

$$U_+^x(t, \beta + \theta) \geq U_+^y(t, \gamma + \theta) \quad \text{for each } t, \theta \in R^1. \tag{5.12}$$

We have thus proved that for every numbers β, γ either (5.12) or the following relation holds:

$$U_+^x(t, \beta + \theta) \leq U_+^y(t, \gamma + \theta) \quad \text{for each } t, \theta \in R^1.$$

Let E be defined by

$$E = \{\delta : U_+^y(t, \theta + \delta) \geq U_+^x(t, \theta) \text{ for each } t, \theta \in R^1\}.$$

Clearly $E \neq \emptyset$ and

$$\delta_0 = \inf\{\delta : \delta \in E\} > -\infty.$$

It is easy to verify that

$$U_\pm^y(t, \theta + \delta_0) \geq U_\pm^x(t, \theta) \quad \text{for each } t, \theta \in R^1. \tag{5.13}$$

Let $\delta < \delta_0$. Then there exist $t(\delta), \theta(\delta) \in R^1$ such that

$$U_+^x(t(\delta), \theta(\delta)) > U_+^y(t(\delta), \theta(\delta) + \delta).$$

It follows from this that

$$U_+^x(t, \theta) \geq U_+^y(t, \theta + \delta) \quad \text{for each } t, \theta \in R^1.$$

Since this relation holds for every $\delta < \delta_0$, we conclude that

$$U_\pm^y(t, \theta + \delta_0) \leq U_\pm^x(t, \theta) \quad \text{for each } t, \theta \in R^1.$$

Together with (5.13) this implies that

$$U_\pm^y(t, \theta + \delta_0) = U_\pm^x(t, \theta) \quad \text{for each } t, \theta \in R^1.$$

The proposition is proved. □

For every irrational α we set $U_\pm^\alpha = U_\pm^x$ where $x \in \mathcal{M}(\alpha)$. Propositions 5.5 and 5.6 imply the following result.

PROPOSITION 5.7. *Let y be a minimal solution with an irrational rotation number α. If y is recurrent, then there exists a unique number β such that one of the conditions below holds:*

$$y(t) = U_+^\alpha(t, \alpha t + \beta) \quad \textit{for every } t \in R^1,$$
$$y(t) = U_-^\alpha(t, \alpha t + \beta) \quad \textit{for every } t \in R^1.$$

If y is not recurrent, then there exists a number β such that

$$U_-^\alpha(t, \alpha t + \beta) < y(t) < U_+^\alpha(t, \alpha t + \beta) \quad \text{for every } t \in R^1.$$

PROPOSITION 5.8. *Assume that y is a minimal solution with a rotation number s/r, which is an irreducible fraction. Then there exist $z_+, z_- \in \mathcal{M}_{\text{per}}(s/r)$ such that*

$$\sup\{|y(p+t) - z_+(p+t)|, |y'(p+t) - z_+'(p+t)| : t \in [0,1]\} \to 0$$

as $p \to +\infty$ on the integers and

$$\sup\{|y(p+t) - z_-(p+t)|, |y'(p+t) - z_-'(p+t)| : t \in [0,1]\} \to 0$$

as $p \to -\infty$ on the integers.

Proof. Define a minimal solution u as

$$u(t) = y(t+r) - s, \qquad t \in R^1. \tag{5.14}$$

We may assume without loss of generality that $u \neq y$. By Proposition 2.11 there exists a number $t_0 > 0$ such that one of the conditions below holds:

$$y(t) \geq u(t) \quad \text{for every } t \geq t_0,$$
$$y(t) \leq u(t) \quad \text{for every } t \geq t_0.$$

Assume that

$$y(t) \leq u(t) \quad \text{for every } t \geq t_0. \tag{5.15}$$

(The other case may be treated analogously.) We set

$$b(t) = y(t) - st/r, \qquad t \in R^1.$$

By Proposition 3.2

$$|b(t) - b(0)| \leq c_1(1 + |s/r|) + 6 \quad \text{for every } t \in R^1.$$

It follows from (5.14) and (5.15) that

$$b(t+r) \geq b(t) \quad \text{for every } t \in R^1,\ t \geq t_0.$$

Therefore, for any $t \in R^1$ the sequence $\{b(t+kr)\}_{k=0}^\infty$ converges to a number $h(t)$. Evidently $h(t+r) = h(t)$ for every $t \in R^1$. Set

$$z_+(t) = h(t) + st/r \quad \text{for every } t \in R^1.$$

It is easy to see that

$$z_+(t+r) = z_+(t) + s \quad \text{for every } t \in R^1,$$
$$z_+(t) = \lim_{k \to \infty} [y(t+kr) - ks], \quad \text{where } k \text{ is an integer.}$$

Analogously we can prove that for every $t \in R^1$ there exists

$$z_-(t) = \lim_{i \to -\infty} [y(t+ir) - is],$$

where i is an integer satisfying $z_-(t+r) = z_-(t) + s$ for $t \in R^1$. For an integer k we set

$$y_k(t) = y(t+kr) - ks \quad \text{for } t \in R^1.$$

It follows from Proposition 2.6 that $z_+, z_- \in \mathcal{M}_{\text{per}}(s/r)$, and

$$y_k \to z_+ \quad \text{as } k \to +\infty, \qquad y_k \to z_- \quad \text{as } k \to -\infty$$

in the C^1-topology on compact sets in R^1. This completes the proof. □

6. Proof of Theorem 1.2. Consider the continuous function $v : R^2 \to R^1$, where

$$v(x, y) = \inf\left\{\int_0^1 f(t, u(t), u'(t))\, dt : u \in W^{1,1}(0, 1),\ u(0) = x,\ u(1) = y\right\}, \tag{6.1}$$

and set

$$\mu = \inf\left\{\liminf_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} v(z_i, z_{i+1}) : \{z_i\}_{i=0}^{\infty} \subset R^1\right\}. \tag{6.2}$$

Let the functions $\pi : R^1 \to R^1$ and $\theta : R^2 \to R^1$ be as defined in Proposition 4.3. Theorem 4.1 implies the following result.

PROPOSITION 6.1. *For each $x \in W^{1,1}_{\mathrm{loc}}([0, \infty))$ the function*

$$T \to \int_0^T [f(t, x(t), x'(t)) - \mu]\, dt, \qquad T \in R^1,\ T > 0,$$

is either bounded or diverges to $+\infty$ as $T \to +\infty$.

A function $x \in W^{1,1}_{\mathrm{loc}}([0, \infty))$ is called a *good configuration* if

$$\sup\left\{\left|\int_0^T [f(t, x(t), x'(t)) - \mu]\, dt\right| : T \in R^1,\ T > 0\right\} < \infty.$$

For $x \in W^{1,1}_{\mathrm{loc}}([0, \infty))$ denote by $\Omega(x)$ the set of all real numbers z such that there exist sequences of integers $\{i_p\}_{p=1}^{\infty}$ and $\{k_p\}_{p=1}^{\infty}$ such that

$$x(k_p) - i_p \to z, \quad k_p \to +\infty \quad \text{as } p \to \infty.$$

Evidently $\Omega(x)$ is a closed set.

PROPOSITION 6.2. *Assume that $E_f(\alpha_f) = \min\{E_f(\alpha) : \alpha \in R^1\}$, where α_f is irrational and $x \in W^{1,1}_{\mathrm{loc}}([0, \infty))$ is a good configuration. Then*

$$\{y(0) : y \in \mathcal{M}(\alpha_f)\} \supset \Omega(x) \supset \{U_+^{\alpha_f}(0, \tau) : \tau \in R^1\} \cup \{U_-^{\alpha_f}(0, \tau) : \tau \in R^1\}.$$

Proof. Let $z \in \Omega(x)$. We will first show that $z \in \{y(0) : y \in \mathcal{M}(\alpha_f)\}$. There exist sequences of integers $\{i_p\}_{p=1}^{\infty}$, $\{k_p\}_{p=1}^{\infty}$ such that $k_p \to +\infty$, $x(k_p) - i_p \to z$ as $p \to \infty$. We may assume that $1 \le k_p < k_{p+1}$ $(p = 1, \dots)$. For a natural number p we define $x_p \in W^{1,1}(-k_p, k_p)$ as

$$x_p(t) = x(t + k_p) - i_p, \qquad t \in [-k_p, k_p].$$

It follows from Theorem 4.1 that for every natural number N the sequence

$$\left\{\int_{-N}^{N} f(t, x_p(t), x_p'(t))\, dt : p \in \{1, 2, \dots\},\ k_p \ge N\right\}$$

is bounded, and by (1.2) the sequence

$$\{x_p' : p \in \{1, 2, \dots\},\ k_p \ge N\}$$

is bounded in $L^2[-N, N]$. We can assume, by extracting a subsequence and reindexing, that for some $Y \in W^{1,1}_{\mathrm{loc}}(R^1)$

$$x_p' \to Y' \text{ weakly in } L^2[-N, N] \text{ and } x_p \to Y \text{ uniformly in } [-N, N] \text{ as } p \to \infty$$

for each natural number N.

It is easy to verify that $Y(0) = z$, $\theta(x(i), x(i+1)) \to 0$ as $i \to \infty$ and

$$\theta(Y(i), Y(i+1)) = 0 \quad \text{for every } i = 0, \pm 1, \pm 2, \ldots. \tag{6.3}$$

Since x is a good configuration, it follows from Proposition 4.2 that

$$\int_i^{i+1} f(t, x(t), x'(t))\, dt - v(x(i), x(i+1)) \to 0 \quad \text{as } i \to +\infty.$$

Together with Proposition 2.1 this implies that

$$\int_i^{i+1} f(t, Y(t), Y'(t))\, dt = v(Y(i), Y(i+1)) \quad (i = 0, \pm 1, \pm 2, \ldots).$$

It follows from this relation and (6.3) that $Y \in W^{1,1}_{\mathrm{loc}}(R^1)$ is a minimal solution. Evidently for each integers i, j satisfying $i < j$

$$\left| \int_i^j f(t, Y(t), Y'(t))\, dt - (j-i)\mu \right| \le 2 \sup\{|\pi(t)| : t \in R^1\}.$$

By Proposition 3.3 and Theorem 4.1 α_f is the rotation number of Y. Therefore, $z \in \{y(0) : y \in \mathcal{M}(\alpha_f)\}$. Since z is an arbitrary element of $\Omega(x)$, we conclude that

$$\Omega(x) \subset \{y(0) : y \in \mathcal{M}(\alpha_f)\}.$$

Next we will establish that

$$\Omega(x) \supset \{U_+^{\alpha_f}(0, \tau) : \tau \in R^1\} \cup \{U_-^{\alpha_f}(0, \tau) : \tau \in R^1\}.$$

By Proposition 5.6 we may assume without loss of generality that

$$U_\pm^{\alpha_f} = U_\pm^Y.$$

Let

$$r \in \{U_+^{\alpha_f}(0, \tau) : \tau \in R^1\} \cup \{U_-^{\alpha_f}(0, \tau) : \tau \in R^1\}.$$

There exist sequences of integers $\{g_p\}_{p=1}^\infty$, $\{s_p\}_{p=1}^\infty$ such that

$$Y(g_p) - s_p \to r \quad \text{as } p \to \infty.$$

It follows from the definition of Y that $Y(g_p) - s_p \in \Omega(x)$ for each natural number p. Since $\Omega(x)$ is closed, we conclude that $r \in \Omega(x)$. This completes the proof of the proposition. □

PROPOSITION 6.3. *Assume that α_f is an irrational number, $\mu = E_f(\alpha_f)$, $x, y \in W^{1,1}_{\mathrm{loc}}([0, \infty))$, $x(0) = y(0)$, and*

$$\int_i^{i+1} f(t, x(t), x'(t))\, dt = v(x(i), x(i+1)),\ \theta(x(i), x(i+1)) = 0 \quad \textit{for } i = 0, 1, \ldots. \tag{6.4}$$

Then

$$\liminf_{i \to +\infty} \int_0^i [f(t, x(t), x'(t)) - f(t, y(t), y'(t))]\, dt \le 0,$$

where i is an integer. Moreover if

$$\liminf_{i\to+\infty} \int_0^i [f(t, x(t), x'(t)) - f(t, y(t), y'(t))]\, dt = 0,$$

where i is an integer, then

$$\int_i^{i+1} f(t, y(t), y'(t))\, dt = v(y(i), y(i+1)),$$

$$\theta(y(i), y(i+1)) = 0 \quad \textit{for } i = 0, 1, \ldots.$$

Proof. It is easy to verify that x is a good configuration. We may assume without loss of generality that y is a good configuration. There exists

$$\beta_f \in \{U_+^{\alpha_f}(0, \tau) : \tau \in [0, 1]\} \cup \{U_-^{\alpha_f}(0, \tau) : \tau \in [0, 1]\}$$

such that

$$\pi(\beta_f) \geq \pi(U_\pm^{\alpha_f}(0, \tau) \quad \text{for every } \tau \in R^1. \tag{6.5}$$

We will show that there exist sequences of integers $\{G_p\}_{p=1}^\infty$, $\{n_p\}_{p=1}^\infty$, and $\{\tilde{n}_p\}_{p=1}^\infty$ such that

$$x(G_p) - \tilde{n}_p \to \beta_f, \quad G_p \to \infty \quad \text{as } p \to \infty$$

and the following relation holds:

$$\lim_{p\to\infty} [y(G_p) - n_p] \in \{U_\pm^{\alpha_f}(0, t) : t \in R^1\}.$$

By Proposition 6.2 there exist sequences of integers $\{i_p\}_{p=1}^\infty$, $\{k_p\}_{p=1}^\infty$ such that $1 \leq k_p < k_{p+1}$ $(p = 1, \ldots)$, $k_p \to +\infty$, $y(k_p) - i_p \to U_+^{\alpha_f}(0, 0)$ as $p \to \infty$.

We can assume by extracting a subsequence of reindexing that there exist $z_0 \in \Omega(x)$ and a sequence of integers $\{\tilde{i}_p\}_{p=1}^\infty$ such that

$$x(k_p) - \tilde{i}_p \to z_0 \quad \text{as } p \to \infty.$$

For a natural number p we define $x_p, y_p \in W^{1,1}(-k_p, k_p)$ as

$$x_p(t) = x(t + k_p) - \tilde{i}_p, \quad y_p(t) = y(t + k_p) - i_p \quad (t \in [-k_p, k_p]). \tag{6.6}$$

It follows from Theorem 4.1 that for every natural number N the sequences

$$\left\{\int_{-N}^N f(t, x_p(t), x_p'(t))\, dt : p \in \{1, 2, \ldots\},\ k_p \geq N\right\},$$

$$\left\{\int_{-N}^N f(t, y_p(t), y_p'(t))\, dt : p \in \{1, 2, \ldots\},\ k_p \geq N\right\}$$

are bounded, and then (1.2) implies that the sequences

$$\{x_p' : p \in \{1, 2, \ldots\},\ k_p \geq N\}, \qquad \{y_p' : p \in \{1, 2, \ldots\},\ k_p \geq N\}$$

are bounded in $L^2[-N, N]$ for every natural number N. We can assume, by extracting a subsequence and reindexing, that for some $X, Y \in W^{1,1}_{\mathrm{loc}}(R^1)$,

$$x_p' \to X', \ y_p' \to Y' \text{ weakly in } L^2[-N, N] \text{ as } p \to \infty, \tag{6.7}$$

$$x_p \to X, \ y_p \to Y \text{ uniformly in } [-N, N] \text{ as } p \to \infty \tag{6.8}$$

for each natural number N.

It is easy to see that

$$X(0) = z_0, Y(0) = U_+^{\alpha_f}(0, 0). \tag{6.9}$$

Since y is a good configuration, we have $\theta(y(i), y(i+1)) \to 0$ as $i \to \infty$ and

$$\theta(Y(i), Y(i+1)) = 0 \quad \text{for } i = 0, \pm 1, \ldots. \tag{6.10}$$

Clearly

$$\theta(X(i), X(i+1)) = 0 \qquad (i = 0, \pm 1, \ldots). \tag{6.11}$$

Since y is a good configuration, it follows from Proposition 4.2 that

$$\int_i^{i+1} f(t, y(t), y'(t))\, dt - v(y(i), y(i+1)) \to 0 \text{ as } i \to +\infty.$$

It follows from this relation, (6.4), and Proposition 2.1 that

$$\int_i^{i+1} f(t, X(t), X'(t))\, dt = v(X(i), X(i+1)),$$

$$\int_i^{i+1} f(t, Y(t), Y'(t))\, dt = v(Y(i), Y(i+1)) \qquad (i = 0, \pm 1, \ldots).$$

In view of (6.10) and (6.11) this implies that X and Y are minimal solutions. Evidently for each integers i, j satisfying $i < j$

$$\left| \int_i^j f(t, X(t), X'(t))\, dt - (j-i)\mu \right|, \ \left| \int_i^j f(t, Y(t), Y'(t))\, dt - (j-i)\mu \right|$$
$$\leq 2 \sup\{|\pi(t)| : t \in R^1\}.$$

By Proposition 3.3 and Theorem 4.1 we conclude that $X, Y \in \mathcal{M}(\alpha_f)$. It follows from Proposition 5.7 and (6.9) that

$$Y(t) = U_+^{\alpha_f}(t, \alpha_f t) \quad \text{for every } t \in R^1. \tag{6.12}$$

By Proposition 5.6

$$\{U_+^X(0, \tau) : \tau \in R^1\} \cup \{U_-^X(0, \tau) : \tau \in R^1\}$$
$$= \{U_+^{\alpha_f}(0, \tau) : \tau \in R^1\} \cup \{U_-^{\alpha_f}(0, \tau) : \tau \in R^1\}.$$

Therefore, there exist sequences of integers $\{g_p\}_{p=1}^{\infty}, \{s_p\}_{p=1}^{\infty}$ such that

$$X(g_p) - s_p \to \beta_f \quad \text{as } p \to \infty,$$

and the sequence $\{\alpha_f g_p - s_p\}$ converges as $p \to \infty$. For $p = 1, 2, \ldots$ we have

$$Y(g_p) - s_p = U_+^{\alpha_f}(g_p, \alpha_f g_p) - s_p = U_+^{\alpha_f}(0, \alpha_f g_p - s_p).$$

We may assume without loss of generality that the sequence $\{Y(g_p) - s_p\}_{p=1}^{\infty}$ converges to a number

$$\beta \in \{U_+^{\alpha_f}(0, \tau) : \tau \in R^1\} \cup \{U_-^{\alpha_f}(0, \tau) : \tau \in R^1\} \tag{6.13}$$

as $p \to \infty$. Let p be a natural number. It follows from the definition of X, Y (see (6.6) and (6.8)) that

$$X(g_p) = \lim_{q \to \infty} x_q(g_p) = \lim_{q \to \infty} [x(g_p + k_q) - \tilde{i}_q],$$
$$Y(g_p) = \lim_{q \to \infty} y_q(g_p) = \lim_{q \to \infty} [y(g_p + k_q) - i_q].$$

There exists a natural number q_p such that $k_{q_p} \geq p + 1 + 2|g_p|$,

$$|X(g_p) - x(g_p + k_{q_p}) - \tilde{i}_{q_p}| \leq \frac{1}{p},$$
$$|Y(g_p) - y(g_p + k_{q_p}) - i_{q_p}| \leq \frac{1}{p}.$$

We set

$$G_p = g_p + k_{q_p}, \quad n_p = s_p + i_{q_p}, \quad \tilde{n} = s_p + \tilde{i}_{q_p} \quad \text{for } p = 1, 2, \ldots.$$

It is easy to verify (see (6.13) and (6.5)) that

$$\begin{gathered} G_p \geq 1 \quad \text{for } p = 1, 2, \ldots, \quad G_p \to \infty \quad \text{as } p \to \infty, \\ x(G_p) - \tilde{n}_p \to \beta_f, \quad y(G_p) - n_p \to \beta \quad \text{as } p \to \infty, \\ \lim_{p \to \infty} \pi(y(G_p)) = \pi(\beta) \leq \pi(\beta_f) = \lim_{p \to \infty} \pi(x(G_p)). \end{gathered} \tag{6.14}$$

It follows from (6.4) that for $p = 1, 2, \ldots$ we have

$$\begin{aligned} &\int_0^{G_p} [f(t, x(t), x'(t)) - f(t, y(t), y'(t))]\, dt \\ &= \pi(x(0)) - \pi(x(G_p)) - \sum_{i=0}^{G_p - 1} \left[\int_i^{i+1} f(t, y(t), y'(t))\, dt - v(y(i), y(i+1)) \right] \\ &\quad - \sum_{i=0}^{G_p - 1} \theta(y(i), y(i+1)) - (\pi(y(0)) - \pi(y(G_p))) \\ &\leq \pi(y(G_p)) - \pi(x(G_p)). \end{aligned}$$

Together with (6.14) this relation implies the validity of the proposition. □

Theorem 1.2 follows from Propositions 2.2, 4.3, and 6.3 and Theorem 4.1.

7. Lagrange multipliers for the discrete-time problem. Let R^n be the Euclidean n-dimensional space with the norm $\|x\| = \max\{|x_i| : i = 1, \dots, n\}$. Let $v : R^n \times R^n \mapsto R^1$ be lower semicontinuous function, which we call the *value function*. A sequence $\bar{x} = \{x_k\}_{k=0}^{\infty}$ in R^n is called a *program*, and with every program $\bar{x}$ we associate the *cost flow* $\{D_N(\bar{x})\}_{N=1}^{\infty}$, where

$$D_N(\bar{x}) = \sum_{k=0}^{N-1} v(x_k, x_{k+1}). \tag{7.1}$$

We consider the problem of minimizing $D_N(\bar{x})$ as $N \to \infty$ in various senses, e.g., of minimizing the functional $\bar{x} \mapsto \liminf_{N\to\infty} \frac{1}{N} D_N(\bar{x})$.

Denote $I = \{0, \pm 1, \pm 2, \dots\}$. We consider value functions $v(\cdot, \cdot)$ that satisfy

$$v(x + m, y + m) = v(x, y) \quad \text{for every } x, y \in R^n \text{ and } m \in I^n. \tag{7.2}$$

We will furthermore assume that

$$\inf\{v(x, y) : x, y \in R^n\} > -\infty; \tag{7.3}$$

and if we denote

$$K = \{(x, y) \in R^n \times R^n : x_i \le y_i \le x_i + 1,\ i = 1, \dots, n\},$$

then

$$a \equiv \sup\{v(x, y) : (x, y) \in K\} < \infty. \tag{7.4}$$

Finally assume that there exists a number $\Gamma > 0$ such that

$$\inf\{v(x, y) : x, y \in R^n,\ |x - y| > \Gamma\} > a. \tag{7.5}$$

We may assume that $\Gamma > n$.

PROPOSITION 7.1. *For every program $\bar{x} = \{x_k\}_{k=0}^{\infty}$ there exists a program $\bar{y} = \{y_k\}_{k=0}^{\infty}$ that satisfies $y_0 = x_0$,*

$$D_N(\bar{y}) \le D_N(\bar{x}) \quad \textit{for every } N \ge 1, \tag{7.6}$$

and

$$|y_{k+1} - y_k| \le \Gamma \quad \textit{for every } k \ge 0. \tag{7.7}$$

Proof. Suppose that for some $l \ge 0$ we have $|x_{l+1} - x_l| > \Gamma$. Then there exists an $m \in I^n$ such that $(x_l, x_{l+1} - m) \in K$, and then by (7.4) we have that

$$v(x_l, x_{l+1} - m) \le a. \tag{7.8}$$

We define the program $\bar{y}$ as

$$y_k = \begin{cases} x_k & \text{if } \ 0 \le k \le l, \\ x_k - m & \text{if } \ k \ge l + 1. \end{cases}$$

Then by (7.2)

$$v(y_k, y_{k+1}) = v(x_k, x_{k+1}) \quad \text{for every } k \ne l,$$

while by (7.5) and (7.8)

$$v(y_l, y_{l+1}) \leq a < v(x_l, x_{l+1}).$$

This completes the proof of the proposition. □

By Proposition 2.1 we may restrict attention only to programs that satisfy (7.7). More generally we say that $\bar{x}$ is a *program with bounded increments* if there exists a constant $b > 0$ such that $|x_{k+1} - x_k| \leq b$ for every $k \geq 0$.

DEFINITION 7.1. *For a program with bounded increments $\bar{x}$ we define the set of* rotation vectors *as the set of all limit points of the sequence $\{x_k/k\}_{k=1}^{\infty}$. We denote this set by $A(\bar{x})$.*

It is clear that for every program with bounded increments $\bar{x}$ the set $A(\bar{x})$ is nonempty and compact. In the sequel we will consider only programs with bounded increments.

With every program $\bar{x}$ we associate the cost growth rate $\liminf_{N\to\infty} \frac{1}{N} D_N(\bar{x})$ and have the following result.

PROPOSITION 7.2. *For every program $\bar{x}$ there exists a program $\bar{y}$ such that $y_0 = x_0$ and the limit $\lim_{N\to\infty} \frac{1}{N} D_N(\bar{y})$ exists and satisfies*

$$\lim_{N\to\infty} \frac{1}{N} D_N(\bar{y}) \leq \liminf_{N\to\infty} \frac{1}{N} D_N(\bar{x}). \tag{7.9}$$

Proof. Suppose that $\limsup_{N\to\infty} \frac{1}{N} D_N(\bar{x}) > \liminf_{N\to\infty} \frac{1}{N} D_N(\bar{x})$, and denote

$$d = \inf_{k\geq 0} \left\{ \liminf_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} v(x_{k+j}, x_{k+j+1}) \right\}.$$

Then clearly $d \leq \liminf_{N\to\infty} \frac{1}{N} D_N(\bar{x})$. There exists an $\epsilon_0 > 0$ such that for every $0 < \epsilon < \epsilon_0$ there exists a finite sequence $\{x_{j+1}, \ldots, x_{j+l}\}$ with arbitrarily large l such that

$$\frac{1}{l} \sum_{k=j+1}^{j+l} v(x_k, x_{k+1}) < d + \epsilon. \tag{7.10}$$

(Clearly (7.10) will also hold for smaller integers $l' < l$ with another finite sequence, e.g., if l' is a divisor of l.) Let $\{\epsilon_i\}_{i=0}^{\infty}$ be a decreasing sequence, $\epsilon_i > 0$, $\epsilon_i \to 0$ as $i \to \infty$. For every ϵ_i let j_i and l_i correspond to ϵ_i as j and l corresponded above to ϵ and such that

$$\frac{1}{l_i} \sum_{k=j_i+1}^{j_i+l_i} v(x_k, x_{k+1}) < d + \epsilon_i. \tag{7.11}$$

We denote $N_p = \sum_{i=1}^{p} l_i$ for every $p \geq 1$; define a program $\bar{y}$ as

$$\{y_0, y_1, \ldots, y_{l_1}\} = \{x_0, x_{j_1+1} - m_1, \ldots, x_{j_1+l_1} - m_1\}$$

for some $m_1 \in I^n$ such that $(x_0, x_{j_1} - m_1) \in K$, and for every $p \geq 2$ define

$$(y_{N_{p-1}+1}, \ldots, y_{N_p}) = (x_{j_p+1} - m_p, \ldots, x_{j_p+l_p} - m_p)$$

for some $m_p \in I^n$ such that $(y_{N_{p-1}}, x_{j_p+1} - m_p) \in K$. As mentioned above, we can choose $\{l_i\}_{i=1}^{\infty}$ such that $l_i/N_i \to 0$ as $i \to \infty$. It follows from the periodicity of $v(\cdot, \cdot)$, the definition of d, and from (7.1) that $\bar{y}$ satisfies

$$\lim_{N\to\infty} \frac{1}{N} D_N(\bar{y}) = d.$$

Since $d \leq \liminf_{N\to\infty} \frac{1}{N} D_N(\bar{x})$, this completes the proof of the proposition. □

We next define the function $\alpha \mapsto \Phi(\alpha)$. For every $\alpha \in R^n$ let

$$\mathcal{P}_\alpha = \{\bar{x} : \alpha \in A(\bar{x})\},$$

and define

$$\Phi(\alpha) = \inf\left\{\liminf_{j\to\infty} \frac{1}{N_j}\sum_{k=0}^{N_j-1} v(x_k, x_{k+1}) : \bar{x} \in \mathcal{P}_\alpha,\ \frac{x_{N_j}}{N_j} \to \alpha\right\},$$

where for every $\bar{x} \in \mathcal{P}_\alpha$ the infimum is over all the subsequences $\{N_j\}_{j=1}^\infty$ such that $x_{N_j}/N_j \to \alpha$ as $j \to \infty$.

PROPOSITION 7.3. *The function $\Phi(\cdot)$ is lower semicontinuous.*

Proof. Let $\alpha_i \to \alpha$, and let $\bar{x}^i$ be such that for an increasing sequence of integers $\{n_j(i)\}_{j=1}^\infty$ we have

$$\frac{1}{n_j}x^i_{n_j} \to \alpha_i \quad \text{and} \quad \frac{1}{n_j}\sum_{k=0}^{n_j-1} v(x^i_k, x^i_{k+1}) \to \Phi(\alpha_i) + \epsilon_i$$

with $\epsilon_i \to 0$ as $i \to \infty$. Moreover, we can choose the programs $\{\bar{x}^i\}$ such that $\{x^i_1\}_{i=1}^\infty$ is bounded. (For simplicity we wrote above n_j instead of $n_j(i)$.) We construct an increasing sequence of integers as

$$N_1 = n_1(1),$$
$$N_2 = n_j(2) \quad \text{for some } j \text{ such that } N_2 > N_1$$

and generally for $k > 1$ let

$$N_{k+1} = n_j(k+1) \quad \text{for some } j \text{ such that } N_{k+1} \geq \sum_{i=1}^k N_i.$$

We now define a program $\bar{y}$ as

$$\{y_1, \ldots, y_{N_1}\} = \{x^1_1, \ldots, x^1_{N_1}\},$$
$$\{y_{N_1+1}, \ldots, y_{N_1+N_2}\} = \{x^2_1 + m_2, \ldots, x^2_{N_2} + m_2\}$$

for some $m_2 \in I^n$ such that $(y_{N_1}, y_{N_1} - (x^2_1 + m_2)) \in K$, and more generally we denote $M_k = \sum_{j=1}^k N_j$ and define for every $k \geq 2$

$$\{y_{M_{k+1}}, \ldots, y_{M_k}\} = \{x^k_1 + m_k, \ldots, x^k_{N_k} + m_k\}$$

for some $m_k \in I^n$ such that $(y_{M_{k-1}}, y_{M_{k-1}} - (x^k_1 + m_k)) \in K$. It is easy to see, in view of the condition $N_{k+1} > M_k$ and the boundedness of $\{x^i_1\}_{i=1}^\infty$, that

$$y_{M_k}/M_k \to \alpha \quad \text{as } k \to \infty$$

and

$$\liminf_{k\to\infty} \frac{1}{M_k}\sum_{j=1}^{M_k-1} v(y_j, y_{j+1}) \leq \liminf_{i\to\infty} \Phi(\alpha_i),$$

implying that $\Phi(\alpha) \leq \liminf_{i\to\infty} \Phi(\alpha_i)$. This concludes the proof. □

The following result is implied by the proof of Proposition 7.3:

COROLLARY 7.1. *For every $\alpha \in R^n$ there exists a program $\bar{x}$ such that for some increasing sequence of integers $\{N_j\}_{j=1}^{\infty}$ we have*

$$x_{N_j}/N_j \to \alpha \quad \text{and} \quad N_j^{-1} \sum_{k=0}^{N_j-1} v(x_k, x_{k+1}) \to \Phi(\alpha) \quad \text{as } j \to \infty.$$

The following result is basic in our study of programs with a prescribed rotation vector. For results of a similar type, see Mather [13].

THEOREM 7.1. *The function $\Phi(\cdot)$ is convex on R^n.*

Proof. For every α, $\Phi(\alpha)$ is finite as can be seen by considering the sequence $x_k = k\alpha$, $k \geq 1$. We will prove that it is convex.

Let $\alpha_1, \alpha_2 \in R^n$ and $\epsilon > 0$ be given. We then can find arbitrarily large integers N and M and finite sequences $\{x_k\}_{k=1}^{N}$ and $\{y_k\}_{k=1}^{M}$ such that $|x_1| \leq 1$, $|y_1| \leq 1$, and

$$\left|\frac{x_N}{N} - \alpha_1\right| < \epsilon, \qquad \left|\frac{y_M}{M} - \alpha_2\right| < \epsilon,$$

$$\frac{1}{N}\sum_{k=1}^{N} v(x_k, x_{k+1}) < \Phi(\alpha_1) + \epsilon, \qquad \frac{1}{M}\sum_{k=1}^{M} v(y_k, y_{k+1}) < \Phi(\alpha_2) + \epsilon. \tag{7.12}$$

Given a number $0 < t < 1$ we can find integers p and q such that

$$\left|\frac{pN}{pN + qM} - t\right| < \epsilon. \tag{7.13}$$

We construct a sequence $\{z_k\}_{k=1}^{pN+qM}$ as

$$\{z_1, \ldots, z_N\} = \{x_1, \ldots, x_N\},$$

and for every $j = 1, \ldots, p-1$ we let

$$\{z_{jN+1}, \ldots, z_{(j+1)N}\} = \{x_1 + m_j, \ldots, x_N + m_j\}, \tag{7.14}$$

where $m_j \in I^n$ is such that $(z_{jN}, z_{jN} - x_1 - m_j) \in K$. We then define

$$\{z_{pN+1}, \ldots, z_{pN+M}\} = \{y_1 + l_1, \ldots, y_M + l_1\} \tag{7.15}$$

for some $l_1 \in I^n$ such that $(z_{pN}, z_{pN} - y_1 - l_1) \in K$. Moreover, for every $1 \leq j \leq q-1$ we define

$$\{z_{pN+jM+1}, \ldots, z_{pN+(j+1)M}\} = y_1 + l_{j+1}, \ldots, y_M + l_{j+1}\}, \tag{7.16}$$

where $l_{j+1} \in I^n$ is such that $(z_{pN+jM}, z_{pN+jM} - y_1 - l_{j+1}) \in K$. It is easy to see from (7.12) and the definitions of z in (7.14) and (7.15) that

$$|z_{pN+qM} - pN\alpha_1 - qM\alpha_2| < (pN + qM)\epsilon + p + q,$$

implying

$$\left|(pN + qM)^{-1} z_{pN+qM} - \left(\frac{pN}{pN+qM}\alpha_1 + \frac{qM}{pN+qM}\alpha_2\right)\right| < \epsilon + \frac{1}{N} + \frac{1}{M}.$$

In view of (7.13) this implies

$$|(pN+qM)^{-1}a_{pN+qM} - (t\alpha_1 + (1-t)\alpha_2)| < (1+|\alpha_1|+|\alpha_2|)\epsilon + \frac{1}{N} + \frac{1}{M}.$$

A similar computation for the cost expression yields

$$\sum_{k=1}^{pN+qM-1} v(z_k, z_{k+1}) - pN\Phi(\alpha_1) - qM\Phi(\alpha_2) < (pN+qM)\epsilon + (p+q+1)C_1$$

for some constant $C_1 > 0$, implying that

$$(pN+qM)^{-1}\sum_{k=1}^{pN+qM-1} v(z_k, z_{k+1}) - t\Phi(\alpha_1) - (1-t)\Phi(\alpha_2) \le C_2\left(\epsilon + \frac{1}{N} + \frac{1}{M}\right) \tag{7.17}$$

for some constant $C_2 > 0$.

Since (7.16) and (7.17) hold for arbitrarily small $\epsilon > 0$ and arbitrarily large N and M, it follows that

$$\liminf_{\alpha \to t\alpha_1 + (1-t)\alpha_2} \Phi(\alpha) \le t\Phi(\alpha_1) + (1-t)\Phi(\alpha_2). \tag{7.18}$$

Since $\Phi(\cdot)$ is lower semicontinuous, (7.18) implies that $\Phi(\cdot)$ is convex, concluding the proof of the theorem. □

For every $\alpha \in R^n$ we consider the functional J_α defined by

$$J_\alpha(\bar{x}) = \infty \quad \text{if } \alpha \notin A(\bar{x}),$$

$$J_\alpha(\bar{x}) = \inf\left\{\liminf_{j\to\infty} \frac{1}{N}\sum_{k=0}^{N_j-1} v(x_k, x_{k+1}),\ \frac{x_{N_j}}{M_j} \to \alpha \text{ as } j \to \infty\right\} \quad \text{if } \alpha \in A(\bar{x}),$$

where the infimum is over all the increasing sequences of integers $\{N_j\}_{j=1}^{\infty}$ for which $x_{N_j}/N_j \to \alpha$.

By Theorem 7.1 $\Phi(\cdot)$ is convex and let its epigraph be the set

$$\text{epi}\Phi = \{(\alpha, \beta) \in R^{n+1} : \beta \ge \Phi(\alpha),\ \alpha \in R^n\}.$$

Let $\alpha \in R^n$ be such that $(\alpha, \Phi(\alpha))$ is an exposed point of epiΦ; namely, there exists a $\lambda \in R^n$ such that

$$\Phi(\alpha') > \Phi(\alpha) + \lambda \cdot (\alpha' - \alpha) \quad \text{for every } \alpha' \ne \alpha. \tag{7.19}$$

Clearly if $\Phi(\cdot)$ is strictly convex, then every $\alpha \in R^N$ is such that $(\alpha, \Phi(\alpha))$ is an exposed point of epiΦ.

We will need the following result about the existence of programs with minimal-cost growth rate.

THEOREM 7.2. *There exists a minimizer $\bar{z}^*$ to the functional*

$$\bar{z} \mapsto \liminf_{N\to\infty} \frac{1}{N}\sum_{k=0}^{N-1} v(z_k, z_{k+1}),$$

and it is such that $\lim_{N\to\infty} \frac{1}{N}\sum_{k=0}^{N-1} v(z_k^*, z_{k+1}^*)$ *exists.*

Proof. The proof is essentially the same as that of Theorem 3.1 in Leizarowitz [9], and we will not repeat it here. □

THEOREM 7.3. *Suppose that* $(\alpha, \Phi(\alpha))$ *is an exposed point of* epiΦ. *Then the problem*

$$\text{minimize } \bar{x} \mapsto J_\alpha(\bar{x})$$

has a minimizer $\bar{x}^*$ *that has definite limits both for the rotation vector and the average cost, so that the following limits exist:*

$$\lim_{k\to\infty} \frac{x_k^*}{k} = \alpha \quad \text{and} \quad \liminf_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} v(x_k^*, x_{k+1}^*) = \Phi(\alpha). \tag{7.20}$$

Actually $\bar{x}$ *is a minimizer of the problem*

$$\text{minimize } \liminf_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} v_\alpha(x_k, x_{k+1}),$$

where λ *satisfies* (7.19) *and* $v_\alpha(\cdot,\cdot)$ *is defined by*

$$v_\alpha(x, y) = v(x, y) + \lambda \cdot (x - y). \tag{7.21}$$

If $\Phi(\cdot)$ *is strictly convex, then the assertions of the theorem and* (7.20) *hold for every* $\alpha \in R^n$.

Proof. It is easy to verify that $v_\alpha(\cdot,\cdot)$ satisfies all the assumptions that we had for $v(\cdot,\cdot)$. Let $\bar{x}^*$ be a minimizer of the functional

$$\bar{x} \mapsto \liminf_{N\to\infty} \frac{1}{N} \sum_{k=0}^{\infty} v_\alpha(x_k, x_{k+1}),$$

which, by Theorem 7.2, exists and is such that $\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} v_\alpha(x_k^*, x_{k+1}^*)$ exists. Denote it by c, and we claim that $c = \Phi(\alpha) - \lambda \cdot \alpha$ and that assertions of the theorem hold true for the program $\bar{x}^*$.

Let $\{N_k\}_{k=1}^\infty$ be an increasing sequence of integers, and suppose that $x_{N_k}/N_k \to \alpha'$ for some $\alpha' \in R^n$ (otherwise we consider a subsequence of $\{N_k\}_{k=1}^\infty$). Then by definition of $\Phi(\cdot)$ we have

$$\liminf_{k\to\infty} \frac{1}{N_k} \sum_{k=0}^{N_k-1} v_\alpha(x_k^*, x_{k+1}^*) \geq \Phi(\alpha') - \lambda \cdot \alpha',$$

which, by (7.19), exceeds $c = \Phi(\alpha) - \lambda \cdot \alpha$ whenever $\alpha' \neq \alpha$. However, by Corollary 7.1 there exists a program $\bar{x}$ for which

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{k=0}^{\infty} v_\alpha(x_k, x_{k+1}) = \alpha.$$

It thus follows by Theorem 7.2 that (7.20) holds for $\bar{x}^*$, and $\bar{x}^*$ is a minimizer of $J_\alpha(\cdot)$ since $J_\alpha(\bar{x}^*) = \Phi(\alpha)$. This completes the proof of the theorem. □

The next result follows immediately from Theorem 7.3. We now restrict attention only to programs $\bar{x}$ for which the limit

$$\mu(\bar{x}) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} v(x_k, x_{k+1})$$

exists. We call this limit, wherever it exists, the *cost growth rate* of the program $\bar{x}$. If we wish to minimize the functional

$$\bar{x} \mapsto \liminf_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} v(x_k, x_{k+1}),$$

then by Proposition 7.2 we may consider for the minimization only programs $\bar{x}$ for which $\mu(\bar{x})$ is well defined.

THEOREM 7.4. *Suppose that* $(\alpha, \Phi(\alpha))$ *is an exposed point of* $\operatorname{epi}\Phi$. *Then the problem*

$$\text{minimize } \{\mu(\bar{x}) : \alpha \in A(\bar{x})\}$$

has a minimizer $\bar{x}^*$ *such that*

$$\lim_{N\to\infty} \bar{x}^*_N/N = \alpha \quad \text{and} \quad \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} v(x^*_k, x^*_{k+1}) = \phi(\alpha).$$

Namely, the minimal cost growth rate over all the programs with rotation vector α *is attained by a program with a single rotation vector* α. *If* Φ *is strictly convex, then the assertion of the theorem holds for every* $\alpha \in R^n$.

Remark. For continuous-time control problems we discretize time and reduce the problem to the one we studied above. There is then a natural correspondence between continuous-time trajectories on one hand and discrete-time programs on the other hand. The rotation vectors and the cost growth rate are preserved under this correspondence, which enables an easy extension of our results to the continuous-time situation.

REFERENCES

[1] Z. ARTSTEIN AND A. LEIZAROWITZ, *Tracking periodic signals with the overtaking criterion*, IEEE Trans. Automat. Control, 30 (1985), pp. 1122–1126.

[2] S. AUBRY AND P. Y. LE DAERON, *The discrete Frenkel-Kontorova model and its extensions* I. *Exact results for the ground states*, Phys. D, 8 (1983), pp. 381–422.

[3] V. BANGERT, *Geodesic rays, Busemann functions and monotone twist maps*, Calc. Variations Partial Differential Equations, 2 (1994), pp. 49–63.

[4] W. A. BROCK AND A. HAURIE, *On existence of overtaking optimal trajectories over on infinite time horizon*, Math. Oper. Res., 1 (1976), pp. 337–346.

[5] M. L. BYLAYI AND L. V. POLTEROVICH *Geodesic flows on the two-dimensional torus and phase transitions "commensurability-noncommensurability,"* Functional Anal. Appl., 20 (1986), pp. 260–266.

[6] D. A. CARLSON, *The existence of catching-up optimal solutions for a class if infinite horizon optimal control problems with time delay*, SIAM J. Control Optim., 28 (1990), pp. 402–422.

[7] ———, *On the existence of sporadically catching-up optimal solutions for infinite horizon optimal control problems*, J. Optim. Theory Appl., 53 (1987), pp. 219–235.

[8] D. A. CARLSON, A. HAURIE, AND A. LEIZAROWITZ, *Infinite Horizon Optimal Control*, Springer-Verlag, Berlin, 1991.

[9] A. LEIZAROWITZ, *Infinite horizon autonomous systems with unbounded cost*, Appl. Math. Optim., 13 (1985), pp. 19–43.

[10] ———, *Existence of overtaking optimal trajectories for problems with convex integrands*, Math. Oper. Res., 10 (1985), pp. 450–461.

[11] ———, *Optimal trajectories of infinite horizon deterministic control systems*, Appl. Math. Optim., 19 (1989), pp. 11–32.

[12] J. N. MATHER, *Existence of quasi-periodic orbits for twist homeomorphisms of the annulas*, Topology, 21 (1982), pp. 457–467.

[13] ———, *Minimal measures*, Comment. Math. Helv., 64 (1989), pp. 375–394.

[14] C. B. MORREY, *Multiple Integrals in the Calculas of Variations*, Springer-Verlag, 1966.

[15] J. MOSER, *Recent developments in the theory of Hamiltonian systems*, SIAM Rev., 28 (1986), pp. 459–485.

[16] J. MOSER, *Minimal solutions of variational problems on a torus*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 3 (1986), pp. 229–272.

[17] W. SENN, *Strikte Konvexität für variationsprobleme auf dem n-dimensionalen torus*, Manuscripta Math., 71 (1991), pp. 45–65.

[18] YA. G. SINAI, *Commensurate-incommensurate phase transitions in one-dimensional chains*, J. Statist. Phys., 29 (1982), pp. 401–425.

[19] A. J. ZASLAVSKI, *Ground states in Frenkel-Kontorova model*, Math. USSR-Izv., 29 (1987), pp. 323–354.

[20] ———, *Optimal programs on infinite horizon* I, SIAM J. Control Optim., 33 (1995), pp. 1643–1660.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1135–1150, July 1996

002

COPOSITIVITY AND THE MINIMIZATION OF QUADRATIC FUNCTIONS WITH NONNEGATIVITY AND QUADRATIC EQUALITY CONSTRAINTS*

J. C. PREISIG†

Abstract. The problem of finding the minimum value of a quadratic function on a set defined by nonnegativity and quadratic equality constraints is analyzed. The difficulty in finding the solution to this problem is primarily due to the fact that the feasible region is nonconvex. An algorithm that requires the Hessian of the quadratic constraint function be strictly copositive is developed for finding the minimal value of the quadratic objective function. The problem of finding this global minima can be mapped into the problem of determining whether or not a particular matrix is copositive. This result is equivalent to earlier results characterizing the solutions to a large class of fractional programming problems. A more efficient algorithm for finding solutions that satisfy the Kuhn–Tucker necessary conditions is developed, and its convergence behavior is analyzed. This algorithm requires that the Hessians of the quadratic constraint and objective functions be both positive semidefinite and strictly copositive.

Key words. copositivity, fractional programming, nonlinear programming

AMS subject classifications. 90C30, 90C32, 90C90

1. Introduction.

This paper addresses the problem of finding the minimal value of a quadratic function with nonnegativity and quadratic equality constraints (Problem 1) and the problem of finding the vector $\mathbf{w}$ that yields this minimal value (Problem 2).

PROBLEM 1. $\lambda_{\text{opt}} = \min_{\substack{w\geq 0 \\ \mathbf{w}^t B\mathbf{w}=b>0}} \mathbf{w}^t A\mathbf{w}.$

PROBLEM 2. $\mathbf{w}_{\text{opt}} = \arg\min_{\substack{w\geq 0 \\ \mathbf{w}^t B\mathbf{w}=b>0}} \mathbf{w}^t A\mathbf{w}.$

Without any loss of generality, the matrices A and B can be assumed to be symmetric and the constant b can be set equal to one. The matrix property of copositivity is used extensively throughout the paper. The following definition of this property is used.

DEFINITION 1. *A matrix Q is copositive if $\mathbf{w} \geq 0$ implies that $\mathbf{w}^t Q\mathbf{w} \geq 0$. A matrix Q is strictly copositive if $\mathbf{w} \geq 0$ and $\mathbf{w} \neq 0$ implies that $\mathbf{w}^t Q\mathbf{w} > 0$.*

Throughout this paper, uppercase letters denote matrix quantities, boldface lowercase letters denote vector quantities, and the superscript t denotes transpose. The development of algorithms to solve Problems 1 and 2 was motivated by an array signal processing problem. This motivating application is described in §2. Section 3 presents an algorithm for solving Problem 1, while §5 presents an algorithm for finding a solution that satisfies the Kuhn–Tucker necessary conditions for Problem 2. The results in §3 are related to results from the field of fractional programming in §4. Finally, a numerical analysis of the efficiency of the two algorithms is presented in §6.

2. An array processing application.

The application that motivated the development of the algorithms described herein involves the processing of signals received at an array of sensors to determine the location of the source of each of the signals received [8]. For each possible source location denoted by ϕ, the array processor estimates the average power in the received signal, which is emitted by a source at that location. Ideally this estimate equals zero if no source is present at the location ϕ. The estimate is denoted by $\hat{\sigma^2}(\phi)$ and is referred to as the ambiguity function. The array processor detailed in [8] creates the ambiguity function by first calculating the sample cross-spectral correlation matrix of the received signal ($\hat{R}$). Then for each possible source location of interest, the processor calculates the signal replica vector,

*Received by the editors July 9, 1993; accepted for publication (in revised form) February 6, 1995. This research was supported by Office of Naval Research grants N00014-90-J-1452 and N00014-91-J-1246. This paper is WHOI Contribution 8427.

†Department of Applied Ocean Physics and Engineering, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, and Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115.

denoted by $\mathbf{q}(\phi)$, that is associated with that location. The ambiguity function for that location is then given by

$$\hat{\sigma^2}(\phi) = \frac{1}{\mathbf{q}(\phi)^{\mathrm{h}} \hat{R}^{-1} \mathbf{q}(\phi)},$$

where the superscript h denotes Hermitian.

The signal replica vector mentioned above (i.e., $\mathbf{q}(\phi)$) is a quantitative description of the spatial structure that would characterize any signal emitted by a source at location ϕ as it is received at the array of sensors used by the processor. This spatial structure allows the processor to differentiate between signals emitted by sources at different locations. Unfortunately, the replica vector is highly dependent on the characteristics of the propagation medium between the location of the source and the array of sensors. Thus, if the processor does not have detailed and accurate environmental information as is often the case in problems involving acoustic signals that have propagated through the ocean, the processor may not be able to calculate $\mathbf{q}(\phi)$ accurately for each location.

The array processor in [8] addresses this problem by creating a set of allowable replica vectors, denoted by $\mathbf{Q}(\phi)$, for each location. This set is defined as

$$\mathbf{Q}(\phi) = \left\{ \mathbf{q} \mid \exists w_1, \ldots, w_M;\ w_i \geq 0;\ \mathbf{q} = \sum_{i=1}^{M} w_i \mathbf{q}_i(\phi) \text{ and } |\ \mathbf{q}\ |^2 = 1 \right\}, \tag{2.1}$$

where the vectors $\mathbf{q}_1(\phi)$ through $\mathbf{q}_M(\phi)$ are prototype replica vectors for the location ϕ. That is, $\mathbf{Q}(\phi)$ is the set of all vectors with a norm of one that are expressible as a nonnegative linear combination of the prototype vectors. The selection of the prototype replica vectors depends on the range of the environmental conditions over which the processor is expected to operate. Given the prototype vectors for each location, the processor calculates the ambiguity function as

$$\hat{\sigma^2}(\phi) = \max_{\mathbf{q} \in \mathbf{Q}(\phi)} \frac{1}{\mathbf{q}^{\mathrm{h}} \hat{R}^{-1} \mathbf{q}}. \tag{2.2}$$

Let the matrix $\tilde{Q}(\phi) = \left[\mathbf{q}_1(\phi), \ldots, \mathbf{q}_M(\phi)\right]$. Then (2.1) can be rewritten as

$$\mathbf{Q}(\phi) = \left\{ \mathbf{q} \mid \exists \mathbf{w} \in \mathbb{R}^M;\ \mathbf{w} \geq 0;\ \mathbf{q} = \tilde{Q}(\phi)\mathbf{w} \text{ and } \mathbf{w}^{\mathrm{t}} \tilde{Q}^{\mathrm{h}}(\phi) \tilde{Q}(\phi) \mathbf{w} = 1 \right\}. \tag{2.3}$$

Let $A(\phi) = \mathrm{Real}(\tilde{Q}(\phi)^{\mathrm{h}} \hat{R}^{-1} \tilde{Q}(\phi))$ and $B(\phi) = \mathrm{Real}(\tilde{Q}(\phi)^{\mathrm{h}} \tilde{Q}(\phi))$. Then using (2.3), the problem in (2.2) can be expressed as

$$\hat{\sigma^2}(\phi)^{-1} = \min_{\substack{\mathbf{w} \geq 0 \\ \mathbf{w}^{\mathrm{t}} B(\phi) \mathbf{w} = 1}} \mathbf{w}^{\mathrm{t}} A(\phi) \mathbf{w}.$$

Thus, the inverse of the ambiguity function is the solution to Problem 1 and the associated weights are the solutions to Problem 2.

3. Finding the global minimum. This section addresses the problem of finding the global minimum of a quadratic function subject to a quadratic equality constraint and a nonnegativity constraint (Problem 1). No additional restrictions are placed on the matrix A. However, the matrix B must be strictly copositive. In this section, this problem is shown to be NP-complete and an algorithm is developed for solving the problem.

The following four lemmas, the proofs of which follow directly from the definitions of copositivity and strict copositivity, are required in proving the main result of this section.

LEMMA 3.1. *Let B be a strictly copositive, symmetric matrix. Then the set $\mathbf{w} \geq 0$, $\mathbf{w}^t B\mathbf{w} = 1$ is a compact set.*

LEMMA 3.2. *Let A and B be symmetric matrices, and assume that B is strictly copositive. Then $\exists$ a constant λ_0 s.t.*

$(A - \lambda_0 B)$ *is copositive but not strictly copositive,*
$\forall \lambda < \lambda_0 \quad (A - \lambda B)$ *is strictly copositive, and*
$\forall \lambda > \lambda_0 \quad (A - \lambda B)$ *is not copositive.*

Letting the symbol $\mathbf{e}$ denote the column vector of all ones, the third lemma is as follows.

LEMMA 3.3. $\exists \mathbf{w}_0 \geq 0$, $\mathbf{w}_0 \neq 0$ *s.t.*

$$\mathbf{w}_0^t (A - \lambda_0 B)\, \mathbf{w}_0 = 0$$

and

$$\mathbf{w}_0 = \arg \min_{\substack{\mathbf{w} \geq 0 \\ \mathbf{e}^t \mathbf{w} = 1}} \mathbf{w}^t (A - \lambda_0 B)\, \mathbf{w},$$

where λ_0 is as defined in Lemma 3.2.

The final lemma necessary to prove the main result of this section is as follows.

LEMMA 3.4.

$$\forall \lambda < \lambda_0 \qquad \min_{\substack{\mathbf{w} \geq 0 \\ \mathbf{e}^t \mathbf{w} = 1}} \mathbf{w}^t (A - \lambda B)\, \mathbf{w} > 0$$

and

$$\forall \lambda > \lambda_0 \qquad \min_{\substack{\mathbf{w} \geq 0 \\ \mathbf{e}^t \mathbf{w} = 1}} \mathbf{w}^t (A - \lambda B)\, \mathbf{w} < 0,$$

where λ_0 is as defined in Lemma 3.2.

The main theorem and result of this section relates the constant λ_0 defined above to the solution to Problem 1.

THEOREM 3.5. *Let A and B be symmetric matrices, B be strictly copositive,*

$$\mathbf{w}_{\text{opt}} = \arg \min_{\substack{\mathbf{w} \geq 0 \\ \mathbf{w}^t B \mathbf{w} = 1}} \mathbf{w}^t A \mathbf{w},$$

$$\lambda_{\text{opt}} = \mathbf{w}_{\text{opt}}^t A \mathbf{w}_{\text{opt}},$$

and λ_0 be as defined in Lemma 3.2.

Then $\lambda_{\text{opt}} = \lambda_0$.

Proof. From Lemma 3.1, the feasible region $\{\mathbf{w} \mid \mathbf{w} \geq 0, \mathbf{w}^t B \mathbf{w} = 1\}$ is compact. In addition, the function $\mathbf{w}^t A \mathbf{w}$ is continuous. Therefore, $\mathbf{w}_{\text{opt}}$ exists as defined above. Let $\tilde{\mathbf{w}} = \mathbf{w}_{\text{opt}}/(\mathbf{e}^t \mathbf{w}_{\text{opt}})$. Then $\tilde{\mathbf{w}} \geq 0$, $\mathbf{e}^t \tilde{\mathbf{w}} = 1$, and $\tilde{\mathbf{w}}^t (A - \lambda_{\text{opt}} B) \tilde{\mathbf{w}} = 0$. Then, Lemma 3.4 implies that $\lambda_{\text{opt}} \geq \lambda_0$.

Let $\mathbf{w}_0$ be as defined in Lemma 3.3, and let $\tilde{\mathbf{w}} = \mathbf{w}_0 / \sqrt{\mathbf{w}_0^t B \mathbf{w}_0}$. Then $\tilde{\mathbf{w}}^t (A - \lambda_0 B) \tilde{\mathbf{w}} = 0$, $\tilde{\mathbf{w}}^t B \tilde{\mathbf{w}} = 1$, and $\tilde{\mathbf{w}} \geq 0$. Combining the first two equalities yields $\tilde{\mathbf{w}}^t A \tilde{\mathbf{w}} = \lambda_0$. Noting that $\tilde{\mathbf{w}}$ is contained in the feasible region of the minimization problem that defines $\mathbf{w}_{\text{opt}}$ and λ_{opt}, this last equality implies that $\lambda_{\text{opt}} \leq \lambda_0$.

Thus, $\lambda_{\text{opt}} = \lambda_0$. □

Utilizing Theorem 3.5, the following bisection algorithm can be used to find the solution to Problem 1.

1. Select λ^- and λ^+ such that $A-\lambda^- B$ is strictly copositive and $A-\lambda^+ B$ is not strictly copositive. Various simple strategies can be used to make this initial selection. (For example, choose λ^+ so that at least one element on the main diagonal of $A-\lambda^+ B$ is ≤ 0. If A is strictly copositive, let $\lambda^- = 0$. Otherwise, let $\lambda^- = \frac{\lambda_{\min}|\mathbf{w}_{\min}|^2}{\gamma_{\min}}$, where $\lambda_{\min}$ is the minimum eigenvalue of A, $\mathbf{w}_{\min} = \arg\min_{\substack{\mathbf{w}\geq 0 \\ \mathbf{e}^t\mathbf{w}=1}} \mathbf{w}^t B\mathbf{w}$, and $\gamma_{\min} = \mathbf{w}_{\min}^t B\mathbf{w}_{\min}$.) Select some tolerance, $\epsilon > 0$.

2. Let $\lambda = (\lambda^+ + \lambda^-)/2$.

3. If $A-\lambda B$ is strictly copositive, set $\lambda^- = \lambda$. Otherwise, $\lambda^+ = \lambda$.

4. If $A-\lambda^+ B$ is copositive or if $\lambda^+ - \lambda^- < \epsilon$, then let $\lambda_{\text{opt}} = \lambda^+$ and terminate the algorithm. Otherwise, go to step 2.

At termination, λ_{opt} will be greater than or equal to the solution to Problem 1 and the magnitude of the difference between the two will be no greater than ϵ.

Finding the solution to Problem 1 to within a particular tolerance can be mapped into determining whether or not a matrix is copositive. In addition, the problem of determining whether or not a matrix is copositive can be mapped into solving Problem 1. (Given any symmetric matrix Q, divide it into two symmetric matrices A and B such that $Q = A - B$ and B has all positive elements. Then B is strictly copositive. Then, if the solution to Problem 1 for this A and B equals one, Q is copositive but not strictly copositive. If the solution is less than one, Q is strictly copositive. Otherwise, Q is not copositive.) Since the problem of determining whether or not a matrix is copositive is NP-complete [7], determining the solution to Problem 1 is also NP-complete. Even though the problem of determining whether or not a matrix is copositive is NP-complete, there are a number of procedures available for making this determination [2, 6, 11, 12].

4. Relationship to fractional programming. Since B is assumed to be strictly copositive, Problem 1 can be recast as a fractional programming problem. Then, Theorem 3.5 can be shown to be equivalent to a theorem in [3, 9, 10] that relates the solution of a parametric optimization problem to the solution of a fractional programming problem.

To recast Problem 1 as a fractional programming problem, we use the following lemma.

LEMMA 4.1. *Let* $\lambda_{\text{opt}} = \min_{\substack{\mathbf{w}\geq 0 \\ \mathbf{w}^t B\mathbf{w}=1}} \mathbf{w}^t A\mathbf{w}$ *and* $\lambda_1 = \min_{\substack{\mathbf{w}\geq 0 \\ \mathbf{e}^t\mathbf{w}=1}} \frac{\mathbf{w}^t A\mathbf{w}}{\mathbf{w}^t B\mathbf{w}}$.

Then $\lambda_{\text{opt}} = \lambda_1$.

Proof. Select $\mathbf{w}_1$ s.t. $\mathbf{w}_1 \geq 0$, $\mathbf{w}_1^t B\mathbf{w}_1 = 1$, and $\lambda_{\text{opt}} = \mathbf{w}_1^t A\mathbf{w}_1$. Let $\tilde{\mathbf{w}}_1 = \frac{\mathbf{w}_1}{\mathbf{e}^t\mathbf{w}_1}$. Then

$$\tilde{\mathbf{w}}_1 \geq 0, \quad \mathbf{e}^t\tilde{\mathbf{w}}_1 = 1, \quad \text{and} \quad \lambda_{\text{opt}} = \frac{\tilde{\mathbf{w}}_1^t A\tilde{\mathbf{w}}_1}{\tilde{\mathbf{w}}_1^t B\tilde{\mathbf{w}}_1}.$$

Therefore, $\lambda_1 \leq \lambda_{\text{opt}}$. Select $\mathbf{w}_2$ such that

$$\mathbf{w}_2 \geq 0, \quad \mathbf{e}^t\mathbf{w}_2 = 1, \quad \text{and} \quad \lambda_1 = \frac{\mathbf{w}_2^t A\mathbf{w}_2}{\mathbf{w}_2^t B\mathbf{w}_2}.$$

Let

$$\tilde{\mathbf{w}}_2 = \frac{\mathbf{w}_2}{\sqrt{\mathbf{w}_2^t B\mathbf{w}_2}}.$$

Then

$$\tilde{\mathbf{w}}_2 \geq 0, \quad \tilde{\mathbf{w}}_2^t B\tilde{\mathbf{w}}_2 = 1, \quad \text{and} \quad \lambda_1 = \frac{\mathbf{w}_2^t A\mathbf{w}_2}{\mathbf{w}_2^t B\mathbf{w}_2} = \frac{\tilde{\mathbf{w}}_2^t A\tilde{\mathbf{w}}_2}{\tilde{\mathbf{w}}_2^t B\tilde{\mathbf{w}}_2} = \tilde{\mathbf{w}}_2^t A\tilde{\mathbf{w}}_2.$$

Therefore, $\lambda_{\text{opt}} \leq \lambda_1$.

Therefore, $\lambda_{\text{opt}} = \lambda_1$. □

The result from [3, 9, 10] that relates the solution of a parametric programming problem to the fractional programming problem in Lemma 4.1 can be expressed as

Let $F(\lambda) = \min_{\substack{\mathbf{w}\geq 0 \\ \mathbf{e}^t\mathbf{w}=1}} \mathbf{w}^t (A - \lambda B)\,\mathbf{w}$, and let $\lambda_2 \in \mathbb{R}$ be the unique zero of $F(\lambda)$. Then $\lambda_2 = \min_{\substack{\mathbf{w}\geq 0 \\ \mathbf{e}^t\mathbf{w}=1}} \frac{\mathbf{w}^t A\mathbf{w}}{\mathbf{w}^t B\mathbf{w}}$.

A constraint such as $\mathbf{e}^t\mathbf{w} = 1$ in Lemma 4.1 is necessary to satisfy the requirements of this result from [3, 9, 10] that the feasible region be compact and that the function $F(\lambda)$ have a unique zero. From Lemmas 3.3 and 3.4, $\lambda_0 = \lambda_2$ is the unique zero of $F(\lambda)$. Since Lemma 4.1 relates the solutions of Problem 1 and the fractional programming problem, the results in [3, 9, 10] and Theorem 3.5 state equivalent characterizations of the solution to Problem 1.

References [3, 9, 10] present an algorithm and some modifications for finding the unique zero of the function $F(\lambda)$. However, each iteration of this algorithm or its modifications requires computing $F(\lambda)$ for a new value of λ. Since $F(\lambda)$ is in general the minima of a nonconvex function, this may be a difficult minimization problem. In fact, this problem will be at least as difficult as determining whether or not the matrix $A - \lambda B$ is strictly copositive, copositive but not strictly copositive, or not copositive as required by the algorithm detailed in §3.

5. Finding a Kuhn–Tucker point. The optimization problem arising in the application for which the algorithms developed herein were developed (see §2) often has a dimensionality of approximately 200. Experience has shown that for problems of such high dimensionality, the approach outlined in §3 is too inefficient to be useful. An alternative procedure detailed in this section is efficient enough to handle such problems. Results presented in Section 6 compare the performance of the two algorithms.

Rather than finding the global solution to Problem 1, the algorithm in this section finds a point that satisfies the Kuhn–Tucker necessary conditions for Problem 2. The algorithm is iterative and at each iteration solves a quadratic programming problem (quadratic objective function, nonnegativity and linear equality constraints). At the nth iteration, the linear equality constraint of the quadratic programming problem is derived by replacing the left-hand vector $\mathbf{w}$ in the quadratic equality constraint in Problem 2 with the solution from the previous iteration, $\mathbf{w}_{n-1}$. As $\mathbf{w}_n$ converges to a solution, the linear equality constraint approaches the quadratic equality constraint in Problem 2. The proof of the convergence properties of the algorithm requires that A and B be positive semidefinite and strictly copositive. These requirements will be introduced at the appropriate times in the analysis that follows.

The Kuhn–Tucker necessary conditions for $\mathbf{w}_{\text{opt}}$ to be a solution to Problem 2 are

$$\begin{aligned} \exists\, \mathbf{v},\ \lambda \quad \text{s.t.} \quad A\mathbf{w}_{\text{opt}} - \lambda B\mathbf{w}_{\text{opt}} - \mathbf{v} &= 0, \\ \mathbf{v},\ \mathbf{w}_{\text{opt}} &\geq 0, \\ \mathbf{v}^t\mathbf{w}_{\text{opt}} &= 0, \\ \mathbf{w}_{\text{opt}}^t B\mathbf{w}_{\text{opt}} &= 1. \end{aligned}$$

The algorithm presented and analyzed in this section can be shown to either

a) terminate with a solution satisfying the Kuhn–Tucker necessary conditions or
b) generate an infinite sequence of solutions that contains at least one convergent subsequence for which the limit point satisfies the Kuhn–Tucker necessary conditions.

In addition, it can be shown that the objective function ($\mathbf{w}^t A\mathbf{w}$) strictly decreases with each iteration.

The iterative algorithm is as follows.

1. Select a convergence tolerance ($\epsilon > 0$) and initial $\mathbf{w}$ ($\mathbf{w}_0$ s.t. $\mathbf{w}_0 \geq 0$ and $\mathbf{w}_0^t B \mathbf{w}_0 = 1$). Set $n = 1$.

2. $\tilde{\mathbf{w}}_n = \arg\min_{\substack{\mathbf{w} \geq 0 \\ \mathbf{w}_{(n-1)}^t B \mathbf{w} = 1}} \mathbf{w}^t A \mathbf{w}$.

3. $\mathbf{w}_n = \dfrac{\tilde{\mathbf{w}}_n}{\sqrt{\tilde{\mathbf{w}}_n^t B \tilde{\mathbf{w}}_n}}$.

4. If $\cos^2(\mathbf{w}_{n-1}, \mathbf{w}_n; B) > (1 - \epsilon)$, then $\mathbf{w}_{\text{opt}} = \mathbf{w}_n$. Otherwise, set $n = n + 1$ and go to step 2.

Here, $\cos^2$ is defined as

$$0 \leq \cos^2(\mathbf{w}_{n-1}, \mathbf{w}_n; B) \triangleq \frac{| \mathbf{w}_{n-1}^t B \mathbf{w}_n |^2}{(\mathbf{w}_{n-1}^t B \mathbf{w}_{n-1})(\mathbf{w}_n^t B \mathbf{w}_n)} \leq 1.$$

The analysis of this algorithm begins with the following theorem, which states that the objective function is a strictly decreasing function of n.

THEOREM 5.1. *Let A and B be symmetric matrices, A be copositive, B be strictly copositive, and $\tilde{\mathbf{w}}_n$ and $\mathbf{w}_n$ be as defined in the algorithm detailed earlier in this section.*

If $\cos^2(\mathbf{w}_{n-1}, \mathbf{w}_n; B) < 1$, *then* $\mathbf{w}_n^t A \mathbf{w}_n < \mathbf{w}_{n-1}^t A \mathbf{w}_{n-1}$.

Proof. B is strictly copositive. Therefore $\forall n$ $\mathbf{w}_{n-1}^t B \mathbf{w}_{n-1}, \mathbf{w}_n^t B \mathbf{w}_n > 0$. In addition, $\mathbf{w}_{n-1}, \tilde{\mathbf{w}}_n, \mathbf{w}_n \geq 0$, and $\mathbf{w}_{n-1}^t B \mathbf{w}_{n-1} = \mathbf{w}_{n-1}^t B \tilde{\mathbf{w}}_n = 1$. $\mathbf{w}_{n-1}$ and $\tilde{\mathbf{w}}_n$ are therefore both in the feasible region of the minimization problem in step 2 of the algorithm. Since $\tilde{\mathbf{w}}_n$ is a solution to the problem in step 2 and A is copositive,

$$0 \leq \tilde{\mathbf{w}}_n^t A \tilde{\mathbf{w}}_n \leq \mathbf{w}_{n-1}^t A \mathbf{w}_{n-1}.$$

In addition,

$$\cos^2(\mathbf{w}_{n-1}, \mathbf{w}_n; B) = \cos^2(\mathbf{w}_{n-1}, \tilde{\mathbf{w}}_n; B) = \frac{| \mathbf{w}_{n-1}^t B \tilde{\mathbf{w}}_n |^2}{(\mathbf{w}_{n-1}^t B \mathbf{w}_{n-1})(\tilde{\mathbf{w}}_n^t B \tilde{\mathbf{w}}_n)} = \frac{1}{(\tilde{\mathbf{w}}_n^t B \tilde{\mathbf{w}}_n)}.$$

Assume that $\cos^2(\mathbf{w}_{n-1}, \mathbf{w}_n; B) < 1$. Then

$$\mathbf{w}_n^t A \mathbf{w}_n = \frac{\tilde{\mathbf{w}}_n^t A \tilde{\mathbf{w}}_n}{\tilde{\mathbf{w}}_n^t B \tilde{\mathbf{w}}_n} = \tilde{\mathbf{w}}_n^t A \tilde{\mathbf{w}}_n \cos^2(\mathbf{w}_{n-1}, \mathbf{w}_n; B) < \tilde{\mathbf{w}}_n^t A \tilde{\mathbf{w}}_n \leq \mathbf{w}_{n-1}^t A \mathbf{w}_{n-1},$$

completing the proof. □

The fact that the objective function is a descent function not only shows that successive solutions result in lower values of the objective function but will also be used later to prove the convergence result summarized above. The final intermediate result needed to prove the convergence result is a proof that the mapping carried out by steps 2 and 3 of the algorithm is closed. The definition of a closed mapping is as follows (see [1]).

DEFINITION 2. *Let X and Y be nonempty closed sets in E_p and E_q, respectively. Let $\mathrm{F} : X \to Y$ be a point-to-set map. The map* A *is said to be closed if*

$$x_k \in X, \qquad x_k \to x_0,$$
$$y_k \in \mathrm{F}(x_k), \qquad y_k \to y_0$$

imply that $y_0 \in \mathrm{F}(x_0)$. The map F *is said to be closed on $Z \subset X$ if it is closed at each point in Z.*

The mapping that must be shown to be closed is given by

$$\tilde{\mathbf{y}}_n = \arg\min_{\substack{y \geq 0 \\ \mathbf{x}_n^t B \mathbf{y} = 1}} \mathbf{y}^t A \mathbf{y}, \tag{5.1}$$

$$F(\mathbf{x}_n) \triangleq \mathbf{y}_n = \frac{\tilde{\mathbf{y}}_n}{\sqrt{\tilde{\mathbf{y}}_n^t B \tilde{\mathbf{y}}_n}}. \tag{5.2}$$

The following definitions will be useful in the proof that $F(\mathbf{x}_n)$ is a closed mapping:

$$A_{\min} \triangleq \min_{\substack{\mathbf{x}\geq 0 \\ \mathbf{x}^t\mathbf{x}=1}} \mathbf{x}^t A\mathbf{x},$$

$$A_{\max} \triangleq \max_{\substack{\mathbf{x}\geq 0 \\ \mathbf{x}^t\mathbf{x}=1}} \mathbf{x}^t A\mathbf{x},$$

$$B_{\min} \triangleq \min_{\substack{\mathbf{x}\geq 0 \\ \mathbf{x}^t\mathbf{x}=1}} \mathbf{x}^t B\mathbf{x},$$

$$B_{\max} \triangleq \max_{\substack{\mathbf{x}\geq 0 \\ \mathbf{x}^t\mathbf{x}=1}} \mathbf{x}^t B\mathbf{x}.$$

The following lemma uses these definitions to bound quantities of interest.

LEMMA 5.2. *Let A and B be symmetric, strictly copositive matrices. Assume that $\mathbf{x}_n \geq 0$ and $\mathbf{x}_n^t B\mathbf{x}_n = 1$. Define $\tilde{\mathbf{y}}_n$ and $\mathbf{y}_n$ as in* (5.1) *and* (5.2), *respectively. Then*

$$0 < \left(\frac{A_{\min}B_{\min}}{A_{\max}B_{\max}}\right)^{\frac{1}{2}} \leq (\tilde{\mathbf{y}}_n^t B\tilde{\mathbf{y}}_n)^{-\frac{1}{2}} = \mathbf{x}_n^t B\mathbf{y}_n \leq 1.$$

Proof. A and B are finite and strictly copositive matrices. Therefore,

$$0 < A_{\min} \leq A_{\max} < \infty,$$
$$0 < B_{\min} \leq B_{\max} < \infty.$$

Therefore,

$$0 < \left(\frac{A_{\min}B_{\min}}{A_{\max}B_{\max}}\right)^{\frac{1}{2}}. \tag{5.3}$$

By definition

$$\mathbf{x}_n^t A\mathbf{x}_n \leq |\,\mathbf{x}_n\,|^2 A_{\max},$$
$$\mathbf{x}_n^t B\mathbf{x}_n = 1 \geq |\,\mathbf{x}_n\,|^2 B_{\min}.$$

Combined with the fact that A is strictly copositive this implies that

$$0 < \mathbf{x}_n^t A\mathbf{x}_n = \frac{\mathbf{x}_n^t A\mathbf{x}_n}{\mathbf{x}_n^t B\mathbf{x}_n} \leq \frac{|\,\mathbf{x}_n\,|^2 A_{\max}}{|\,\mathbf{x}_n\,|^2 B_{\min}} = \frac{A_{\max}}{B_{\min}}. \tag{5.4}$$

The definition of $B_{\max}$ and the fact that B is strictly copositive and $\mathbf{x}_n^t B\tilde{\mathbf{y}}_n = 1$ imply that

$$0 < \tilde{\mathbf{y}}_n^t B\tilde{\mathbf{y}}_n \leq |\,\tilde{\mathbf{y}}_n\,|^2 B_{\max}. \tag{5.5}$$

The definition of $A_{\min}$ and the fact that $\mathbf{x}_n$ is in the feasible region of the minimization problem in (5.1) imply that

$$|\,\tilde{\mathbf{y}}_n\,|^2 A_{\min} \leq \tilde{\mathbf{y}}_n^t A\tilde{\mathbf{y}}_n \leq \mathbf{x}_n^t A\mathbf{x}_n. \tag{5.6}$$

Combining (5.5) and (5.6) yields

$$0 < \tilde{\mathbf{y}}_n^t B\tilde{\mathbf{y}}_n \leq \mathbf{x}_n^t A\mathbf{x}_n \frac{B_{\max}}{A_{\min}}. \tag{5.7}$$

Combining (5.3), (5.4), and (5.7) yields

$$0 < \left(\frac{A_{\min}B_{\min}}{A_{\max}B_{\max}}\right)^{\frac{1}{2}} \leq (\tilde{\mathbf{y}}_n^t B\tilde{\mathbf{y}}_n)^{-\frac{1}{2}}. \tag{5.8}$$

Combining the facts that $\mathbf{x}_n^t B\tilde{\mathbf{y}}_n = 1$, $\mathbf{y}_n = \frac{\tilde{\mathbf{y}}_n}{\sqrt{\tilde{\mathbf{y}}_n^t B\tilde{\mathbf{y}}_n}}$, and $\tilde{\mathbf{y}}_n^t B\tilde{\mathbf{y}}_n > 0$ implies that

$$(\tilde{\mathbf{y}}_n^t B\tilde{\mathbf{y}}_n)^{-\frac{1}{2}} = \mathbf{x}_n^t B\mathbf{y}_n. \tag{5.9}$$

The facts that $\mathbf{x}_n^t B\mathbf{x}_n = \mathbf{y}_n^t B\mathbf{y}_n = 1$ and that $\cos^2(\mathbf{x}_n, \mathbf{y}_n; B) = \frac{|\mathbf{x}_n^t B\mathbf{y}_n|^2}{(\mathbf{x}_n^t B\mathbf{x}_n)(\mathbf{y}_n^t B\mathbf{y}_n)} \leq 1$ imply that $\mathbf{x}_n^t B\mathbf{y}_n \leq 1$. Combining this with (5.8) and (5.9) yields

$$0 < \left(\frac{A_{\min}B_{\min}}{A_{\max}B_{\max}}\right)^{\frac{1}{2}} \leq (\tilde{\mathbf{y}}_n^t B\tilde{\mathbf{y}}_n)^{-\frac{1}{2}} = \mathbf{x}_n^t B\mathbf{y}_n \leq 1. \qquad \square$$

The following theorem establishes that the mapping is closed.

THEOREM 5.3. *Assume that A and B are symmetric, strictly copositive matrices and that A is positive semidefinite. Then the mapping $F(x_n)$ defined above is closed on the set* $\mathbf{x} \geq 0$, $\mathbf{x}^t B\mathbf{x} = 1$.

Proof. Assume that $\forall n$, $\mathbf{x}_n \geq 0$, $\mathbf{x}_n^t B\mathbf{x}_n = 1$; that $\mathbf{x}_n \to \mathbf{x}_o$; and that $F(\mathbf{x}_n) = \mathbf{y}_n \to \mathbf{y}_o$. Then

$$0 < \left(\frac{A_{\min}B_{\min}}{A_{\max}B_{\max}}\right)^{\frac{1}{2}} \leq \mathbf{x}_n^t B\mathbf{y}_n \quad \forall n$$

implies that $\mathbf{x}_n^t B\mathbf{y}_n \to \mathbf{x}_o^t B\mathbf{y}_o > 0$.

By (5.2) and Lemma 5.2,

$$\mathbf{y}_n = \frac{\tilde{\mathbf{y}}_n}{\sqrt{\tilde{\mathbf{y}}_n^t B\tilde{\mathbf{y}}_n}} = \tilde{\mathbf{y}}_n(\mathbf{x}_n^t B\mathbf{y}_n).$$

Therefore, $\tilde{\mathbf{y}}_n = \frac{\mathbf{y}_n}{\mathbf{x}_n^t B\mathbf{y}_n}$. Therefore, $\mathbf{x}_o^t B\mathbf{y}_o > 0$ implies that

$$\tilde{\mathbf{y}}_n = \frac{\mathbf{y}_n}{\mathbf{x}_n^t B\mathbf{y}_n} \to \frac{\mathbf{y}_o}{\mathbf{x}_o^t B\mathbf{y}_o} \triangleq \tilde{\mathbf{y}}_o.$$

$\tilde{\mathbf{y}}_n$ satisfies the Kuhn–Tucker necessary and sufficient conditions (sufficient since A is positive semidefinite) for the problem in (5.1), which are

$$\exists\, \tilde{\mathbf{v}}_n,\ \tilde{\lambda}_n \quad \text{s.t.}$$

$$A\tilde{\mathbf{y}}_n - \tilde{\lambda}_n B\mathbf{x}_n - \tilde{\mathbf{v}}_n = 0, \tag{5.10}$$

$$\tilde{\mathbf{v}}_n,\ \tilde{\mathbf{y}}_n \geq 0,$$

$$\tilde{\mathbf{v}}_n^t\tilde{\mathbf{y}}_n = 0, \tag{5.11}$$

$$\mathbf{x}_n^t B\tilde{\mathbf{y}}_n = 1. \tag{5.12}$$

$\forall n$ $\tilde{\mathbf{y}}_n \geq 0$ implies that $\tilde{\mathbf{y}}_o \geq 0$. $\forall n$ $\mathbf{x}_n^t B\tilde{\mathbf{y}}_n = 1$ implies that $\mathbf{x}_o^t B\tilde{\mathbf{y}}_o = 1$.

Left multiply (5.10) by $\tilde{\mathbf{y}}_n$. Then, substituting in (5.11) and (5.12) yields $\tilde{\lambda}_n = \tilde{\mathbf{y}}_n^t A\tilde{\mathbf{y}}_n$. Therefore, $\tilde{\mathbf{y}}_n \to \tilde{\mathbf{y}}_o$ implies that $\tilde{\lambda}_n \to \tilde{\mathbf{y}}_o^t A\tilde{\mathbf{y}}_o \triangleq \tilde{\lambda}_o$.

Then, $\tilde{\mathbf{v}}_n = A\tilde{\mathbf{y}}_n - \tilde{\lambda}_n B\mathbf{x}_n \to A\tilde{\mathbf{y}}_o - \tilde{\lambda}_o B\mathbf{x}_o \triangleq \tilde{\mathbf{v}}_o$. In addition, $\forall n\ \tilde{\mathbf{v}}_n \geq 0$ implies that $\tilde{\mathbf{v}}_o \geq 0$. $\forall n\ \tilde{\mathbf{v}}_n^t \tilde{\mathbf{y}}_n = 0$ implies that $\tilde{\mathbf{v}}_o^t \tilde{\mathbf{y}}_o = 0$.

Combining the above yields

$$\begin{aligned} A\tilde{\mathbf{y}}_o - \tilde{\lambda}_o B\mathbf{x}_o - \tilde{\mathbf{v}}_o &= 0, \\ \tilde{\mathbf{v}}_o,\ \tilde{\mathbf{y}}_o &\geq 0, \\ \tilde{\mathbf{v}}_o^t \tilde{\mathbf{y}}_o &= 0, \\ \mathbf{x}_o^t B\tilde{\mathbf{y}}_o &= 1. \end{aligned}$$

Therefore, $\tilde{\mathbf{y}}_o$ satisfies the Kuhn–Tucker necessary and sufficient conditions for the problem in (5.1) and is a solution to that problem. Then (5.2) and Lemma 5.2 imply that

$$\mathbf{y}_o = \frac{\tilde{\mathbf{y}}_o}{\sqrt{\tilde{\mathbf{y}}_n^t B\tilde{\mathbf{y}}_n}}.$$

Therefore, $\mathbf{y}_o \in F(\mathbf{x}_o)$. Therefore the mapping is closed on $\mathbf{x} \geq 0$, $\mathbf{x}^t B\mathbf{x} = 1$. □

Given the results established in Theorems 5.1 and 5.3, the iteration of steps 2 and 3 of the algorithm satisfies the necessary conditions of the Convergence Theorem in [1] (see Theorem 7.2.3). The algorithm will therefore either terminate in a finite number of steps with a solution or generate an infinite sequence such that every convergent subsequence has a limit satisfying the termination conditions. From Lemma 3.1, the feasible region containing any infinite sequence of solutions is compact. Since all points in the sequence of solutions are contained in this compact set, any infinite sequence of solutions will contain at least one convergent subsequence.

The final point left to establish is that the solution at termination or the limit point of at least one convergent subsequence of solutions satisfies the Kuhn–Tucker necessary conditions for Problem 2. In establishing this result, it is assumed that the termination criterion ϵ is arbitrarily close to zero. Therefore, if the algorithm terminates, $\cos^2(\mathbf{w}_{n_o-1}, \mathbf{w}_{\text{opt}}; B) = 1$ where n_o is the index at termination.

The following theorem establishes the fact that when the algorithm terminates with a solution, that solution is a Kuhn–Tucker point of Problem 2.

THEOREM 5.4. *Assume that B is a symmetric, positive semidefinite, and strictly copositive matrix. Then if the algorithm detailed at the beginning of this section terminates with a solution, that solution is a Kuhn–Tucker point of Problem* 2.

Proof. Let n_o be the index at termination. Then $\tilde{\mathbf{w}}_{n_o}$ satisfies the Kuhn–Tucker necessary conditions for the minimization problem in step 2 of the algorithm. That is,

$$\begin{aligned} &\exists\ \tilde{\mathbf{v}}_o,\ \tilde{\lambda}_o \quad \text{s.t.} \\ A\tilde{\mathbf{w}}_{n_o} - \tilde{\lambda}_o B\mathbf{w}_{n_o-1} - \tilde{\mathbf{v}}_o &= 0, \\ \tilde{\mathbf{v}}_o,\ \tilde{\mathbf{w}}_{n_o} &\geq 0, \\ \tilde{\mathbf{v}}_o^t \tilde{\mathbf{w}}_{n_o} &= 0, \\ \mathbf{w}_{n_o-1}^t B\tilde{\mathbf{w}}_{n_o} &= 1. \end{aligned} \tag{5.13}$$

The facts that B is strictly copositive and that $\mathbf{w}_{\text{opt}} = \tilde{\mathbf{w}}_{n_o}/\sqrt{\tilde{\mathbf{w}}_{n_o}^t B\tilde{\mathbf{w}}_{n_o}}$ imply that

$$\begin{aligned} &\exists\ \mathbf{v}_o,\ \lambda_o' \quad \text{s.t.} \\ A\mathbf{w}_{\text{opt}} - \lambda_o' B\mathbf{w}_{n_o-1} - \mathbf{v}_o &= 0, \\ \mathbf{v}_o,\ \mathbf{w}_{\text{opt}} &\geq 0, \\ \mathbf{v}_o^t \mathbf{w}_{\text{opt}} &= 0, \\ \mathbf{w}_{\text{opt}}^t B\mathbf{w}_{\text{opt}} &= 1, \end{aligned} \tag{5.14}$$

where

$$\mathbf{v}_o \triangleq \frac{\tilde{\mathbf{v}}_o}{\sqrt{\tilde{\mathbf{w}}_{n_o}^t B \tilde{\mathbf{w}}_{n_o}}} \text{ and } \lambda_o' \triangleq \frac{\tilde{\lambda}_o}{\sqrt{\tilde{\mathbf{w}}_{n_o}^t B \tilde{\mathbf{w}}_{n_o}}}.$$

Since B is positive semidefinite, it can be decomposed as $B = \tilde{B}^t \tilde{B}$. Then, $\cos^2(\mathbf{w}_{n_o-1}, \mathbf{w}_{opt}; B) = 1$ can be rewritten as $\cos^2(\tilde{B}\mathbf{w}_{n_o-1}, \tilde{B}\mathbf{w}_{opt}; I) = 1$. Therefore

$$\tilde{B}\mathbf{w}_{n_o-1} = a\tilde{B}\mathbf{w}_{opt} \tag{5.15}$$

for some constant a. Right multiplying both sides of (5.15) by $\tilde{B}^t$ yields $B\mathbf{w}_{n_o-1} = a\, B\mathbf{w}_{opt}$. Substituting this into (5.14) and letting $\lambda_o = a\lambda_o'$ result in

$$\begin{aligned} &\exists\ \mathbf{v}_o,\ \lambda_o \quad \text{s.t.} \\ A\mathbf{w}_{opt} - \lambda_o B\mathbf{w}_{opt} - \mathbf{v}_o &= 0, \\ \mathbf{v}_o,\ \mathbf{w}_{opt} &\geq 0, \\ \mathbf{v}_o^t \mathbf{w}_{opt} &= 0, \\ \mathbf{w}_{opt}^t B\mathbf{w}_{opt} &= 1. \end{aligned}$$

These are the Kuhn–Tucker necessary conditions for Problem 2. Therefore, $\mathbf{w}_{opt}$ is a Kuhn–Tucker point of Problem 2. □

The final theorem establishes an analogous result for the case where the algorithm does not terminate.

THEOREM 5.5. *Assume that B is a symmetric, positive semidefinite, and strictly copositive matrix. Then, if the algorithm detailed at the beginning of this section generates an infinite sequence of solutions, then there exists a convergent subsequence of solutions for which the limit point satisfies the Kuhn–Tucker necessary conditions for Problem* 2.

Proof. Define the vector

$$\mathbf{z}_n \triangleq \begin{bmatrix} \mathbf{w}_{n-1} \\ \mathbf{w}_n \end{bmatrix}$$

and the mapping

$$G(\mathbf{z}_n) \triangleq \mathbf{z}_{n+1} = \begin{bmatrix} \mathbf{w}_n \\ F(\mathbf{w}_n) \end{bmatrix}.$$

The fact that F is closed implies that G is also closed. In addition, the function $q(\mathbf{z}_n) = \mathbf{w}_n^t A \mathbf{w}_n$ has already been shown to be a strictly decreasing function of n. Therefore, the convergence results derived for the iteration of the mapping F also apply to the iteration of the mapping G. The same termination criterion,

$$\cos^2(\mathbf{z}_n; B) \triangleq \cos^2(\mathbf{w}_{n-1}, \mathbf{w}_n; B) > (1 - \epsilon) \to 1,$$

is used for both iterations.

Assume that for particular A, B and initial point $\mathbf{w}_o$ the iteration of F generates an infinite sequence. Then for this same A, B, and initial point, the iteration of G generates an infinite sequence. The set of allowable $\mathbf{z}_n$ is a compact set, so this infinite sequence $\mathbf{z}_n$ contains

at least one convergent subsequence whose limit satisfies the termination criterion. (Note that not every convergent subsequence of $\mathbf{w}_n$ corresponds to a convergent subsequence of $\mathbf{z}_n$ but that at least one convergent subsequence of $\mathbf{w}_n$ corresponds to a convergent subsequence of $\mathbf{z}_n$. In addition, every convergent subsequence of $\mathbf{z}_n$ corresponds to a convergent subsequence of $\mathbf{w}_n$.) Denote this limit by

$$\mathbf{z}_{\text{opt}} = \begin{bmatrix} \mathbf{w}'_{\text{opt}} \\ \mathbf{w}_{\text{opt}} \end{bmatrix}.$$

Then $\mathbf{w}_{\text{opt}} = F(\mathbf{w}'_{\text{opt}})$ and $\cos^2(\mathbf{z}_{\text{opt}}; B) = \cos^2(\mathbf{w}'_{\text{opt}}, \mathbf{w}_{\text{opt}}; B) = 1$. Therefore, the results of Theorem 5.4 can be applied by letting $\mathbf{w}_{n_0-1} = \mathbf{w}'_{\text{opt}}$ in the proof. Therefore, $\mathbf{w}_{\text{opt}}$ satisfies the Kuhn–Tucker necessary conditions for Problem 2. □

This completes the characterization of the solutions generated by the algorithm detailed at the beginning of this section. To summarize the constraints on the matrices A and B in order for all of these results to hold, A and B must be positive semidefinite and strictly copositive matrices.

6. Numerical results. The major advantage of the iterative algorithm in §5 (subsequently referred to as the iterative algorithm) with respect to the global optimization routine in §3 (subsequently referred to as the copositivity algorithm) is its greater numerical efficiency. The numerical results presented in this section demonstrate the relative efficiency of the two algorithms.

The programs that implement the algorithms were written for the Matlab software package and executed on a Sun Sparcstation2 workstation. When compared to compiled code written in Fortran or C, the Matlab code is very slow. In addition, the Sparcstation2 processor is much slower than many other available processors. Therefore, these results should be used to evaluate the relative efficiency of the two algorithms and the rate of growth of the execution time as a function of problem size. The absolute execution times shown here are not indicative of what could be achieved with programs written in Fortran or C.

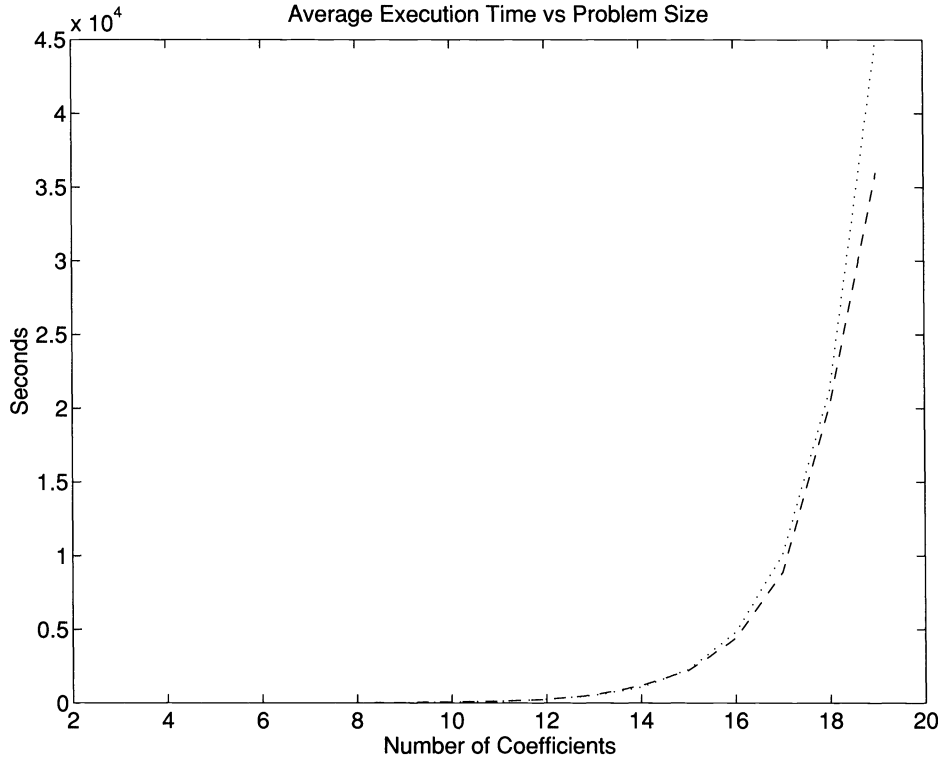
Both algorithms were tested using randomly generated positive semidefinite matrices. A number of trials were run for successively larger problem sizes. Problem size, denoted by n, is the size of the vector $\mathbf{w}$. For each trial, two $n \times n$ random matrices ($\tilde{A}$ and $\tilde{B}$) were generated by Matlab's random number generator. The individual elements of each matrix were Gaussian random variables with a mean of zero and a variance of one. The A and B matrices used by the algorithms for the trial were then given by $A = \tilde{A}^t\tilde{A}$ and $B = \tilde{B}^t\tilde{B}$. No check was conducted to ensure that the matrices A and B were strictly copositive as is required by the iterative algorithm. For the copositivity algorithm, the tolerance of $\epsilon = 10^{-4}$ was used, while for the iterative algorithm, the convergence tolerance of $\epsilon = 0.9999$ was used.

Figures 6.1 and 6.2 show the average execution times ($T[n]$) for the copositivity and iterative algorithms, respectively, as a function of the problem size (i.e., the size of the vector $\mathbf{w}$). In each case, the dashed line shows the measured average execution time while the dotted line shows a parametric model prediction of the average execution time. For the copositivity algorithm, the exponential growth model of

$$\hat{T}[n] = 0.0321 * 2.1057^n$$

was used where $\hat{T}[n]$ is the predicted average execution time in seconds and n is the problem size. For the iterative algorithm, the following polynomial time model was used:

$$\hat{T}[n] = (0.4524 + 0.0340 * n)^{3.5}.$$



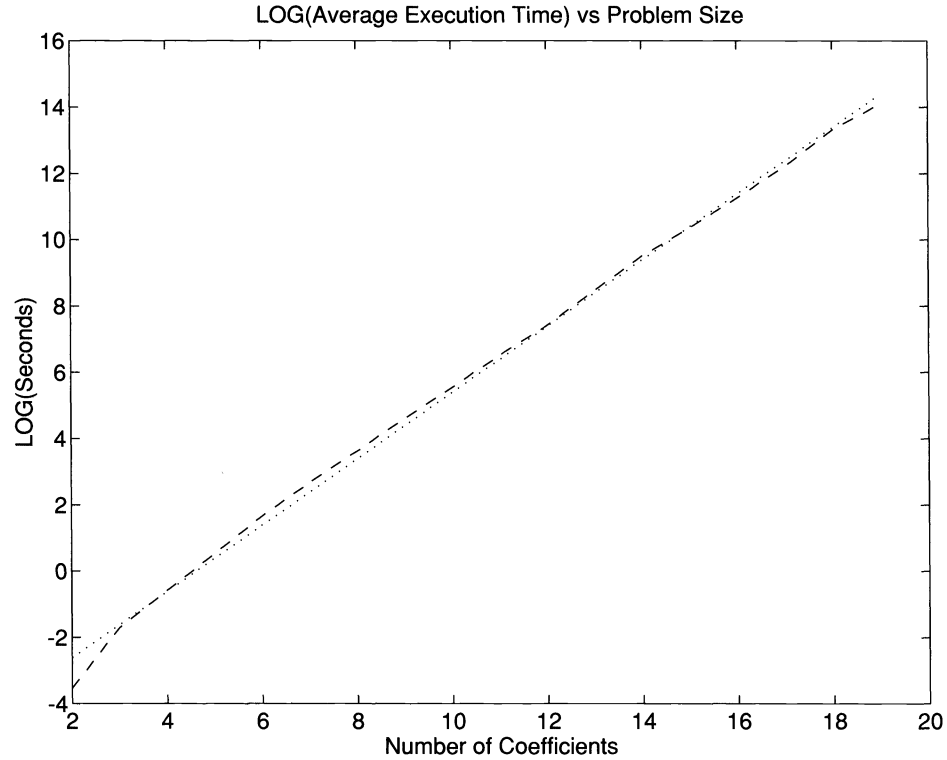
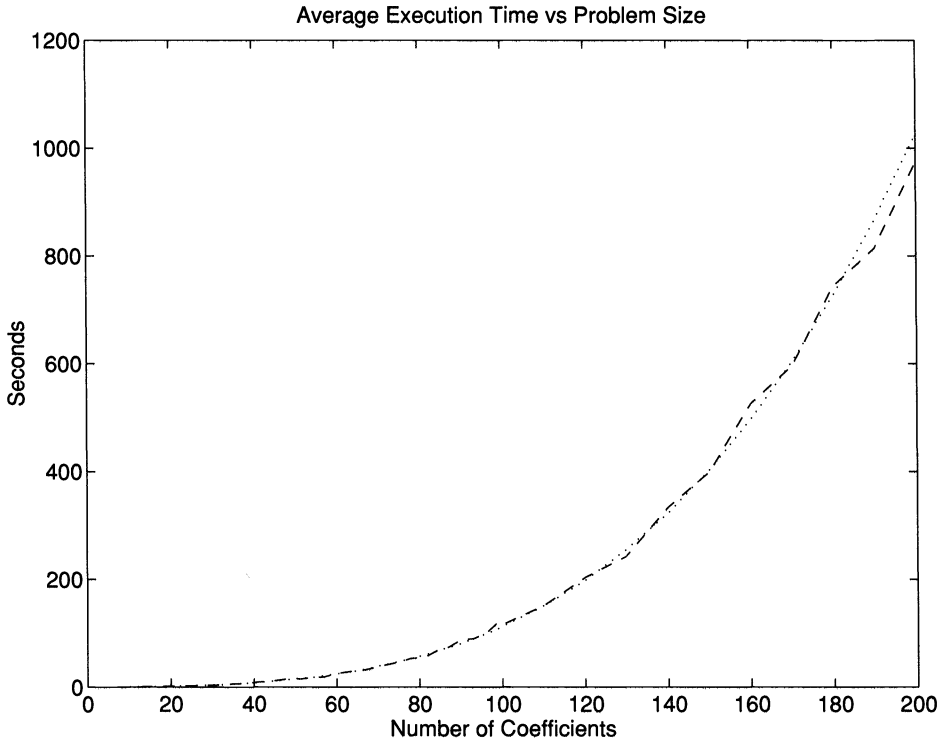


FIG. 6.1. *Average execution times for the copositivity algorithm.*



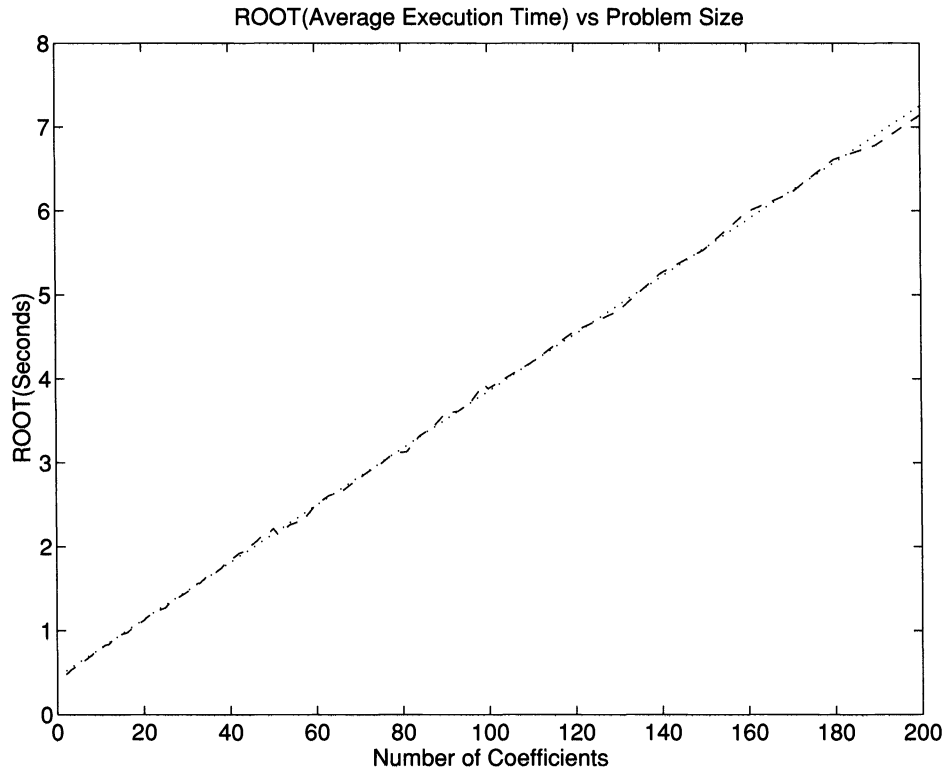


FIG. 6.2. *Average execution times for the iterative algorithm.*

Figures 6.1(a) and 6.2(a) show a strong agreement between the predicted and actual average execution times. In Figure 6.1(b), $\log_{2.1057}(T[n])$ and $\log_{2.1057}(\hat{T}[n]) = n - 4.6182$ are shown while in Figure 6.2(b), $(T[n])^{\frac{1}{3.5}}$ and $(\hat{T}[n])^{\frac{1}{3.5}} = 0.4524 + 0.0340 * n$ are shown. These plots also show strong agreement between the predicted and actual average execution times at all the problem sizes tested. This numerical evidence shows that the copositivity algorithm exhibits exponential growth in execution time while the iterative algorithm exhibits polynomial growth in execution time.

The numerical analysis of the iterative algorithm bears some more discussion. The minimization problem in step 2 of this algorithm is a quadratic programming problem. There are a number of algorithms available for solving this problem, and any of them can be used in implementing the iterative algorithm. The quadratic programming algorithm used to generate the numerical results presented in this section is the complementary pivoting algorithm developed by Lemke [5]. For an explanation and analysis of the complementary pivoting algorithm, see [1, 4, 6]. In general, the complementary pivoting algorithm is guaranteed to terminate with a finite solution if A is positive semidefinite or if A has nonnegative elements and positive diagonal elements (see Theorem 11.2.4 in [1]).

A plot of average execution time per iteration ($\overline{T}[n]$) of the iterative algorithm shows that the majority of the growth in the execution time is due to growth in time per iteration rather than number of iterations. Figures 6.3(a) and (b) show close agreement between the actual average execution time per iteration (dashed line) and that predicted by the model (dotted line)

$$\hat{\overline{T}}[n] = (0.03126 + 0.0179 * n)^{3.23}.$$

Further analysis confirms that the overall polynomial growth term of $n^{3.5}$ for the iterative algorithm can be attributed to a term of $n^{3.23}$ for the growth in average execution time per iteration and a term of $n^{0.27}$ for the growth in average number of iterations required to solve the problem. Further analysis has also shown that approximately 99% of the execution time of each iteration is taken in solving the quadratic programming problem in step 2 of the algorithm. Since quadratic programming algorithms that have better numerical efficiency than Lemke's complementary pivoting algorithm are available, the use of one of these algorithms can make further improvements in the overall numerical efficiency of the iterative algorithm.

7. Conclusions. The algorithms developed above address two problems associated with the minimization of a quadratic objective function subject to nonnegativity and quadratic equality constraints. The first algorithm addresses the problem of finding the global minimum of the quadratic objective function and requires the repeated testing of a particular matrix to determine whether or not it is copositive. The only constraint on the problem is that the Hessian of the quadratic equality constraint (B) be strictly copositive. This problem is NP-complete. The second algorithm addresses the problem of finding a vector that satisfies the Kuhn–Tucker necessary conditions for the minimization problem. The algorithm is shown to converge, and it is further shown that the objective function strictly decreases with each iteration. This algorithm requires the solution of a sequence of quadratic programming problems. The constraints placed on the problem are that the Hessian matrices of both the quadratic objective function (A) and the quadratic constraint function (B) be positive semidefinite and strictly copositive.

Acknowledgments. The author thanks Steven Isabelle and Richard Pawlowicz for their feedback and ideas during the research and the preparation of this paper. In addition, the reviewers are thanked for their constructive input. In particular, their pointing out the relationship of the results herein to previous results in the area of fractional programming was helpful.

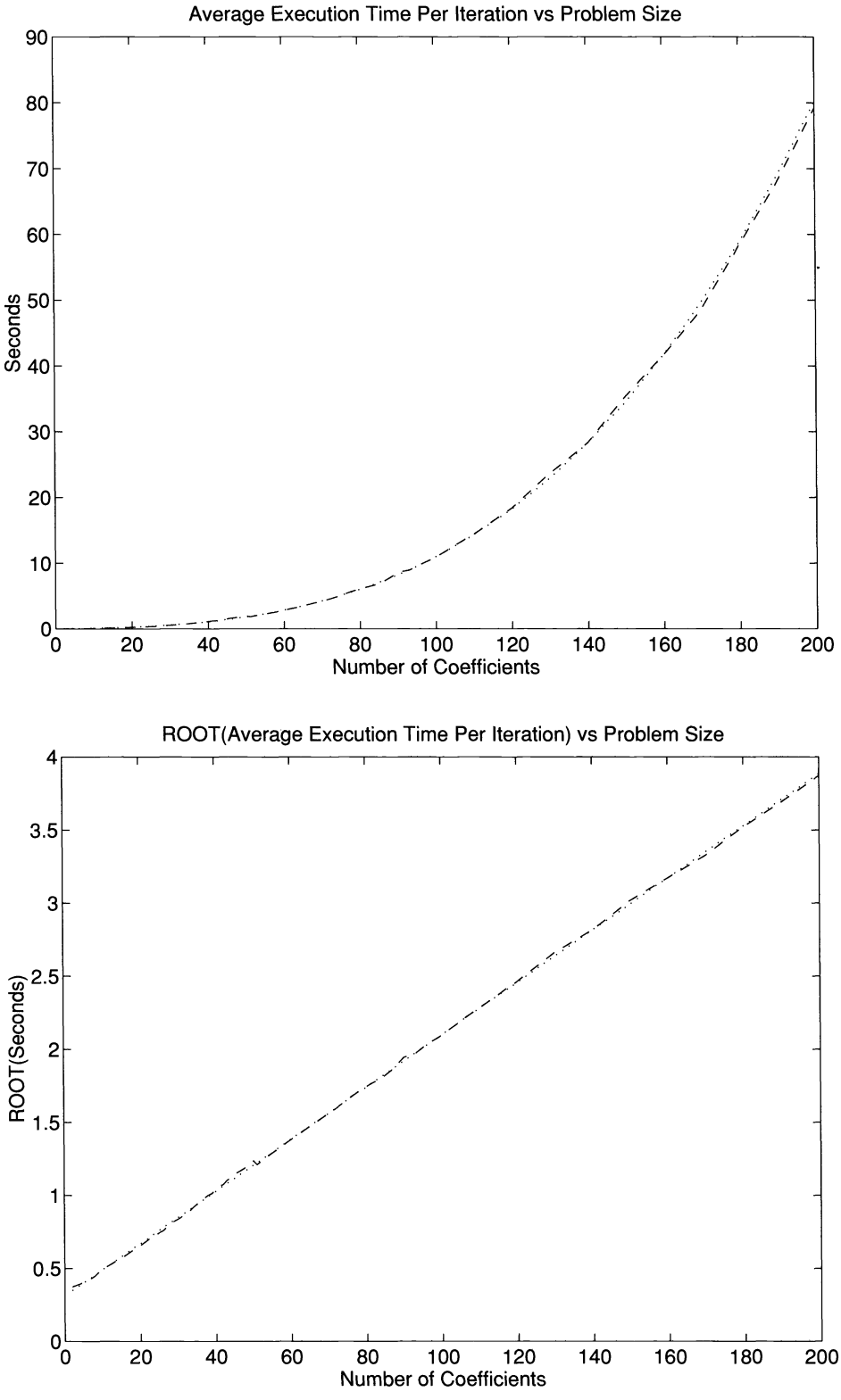


FIG. 6.3. *Average execution time per iteration for iterative algorithm.*

REFERENCES

[1] M. S. BAZARAA AND C. M. SHETTY, *Nonlinear Programming, Theory and Applications*, John Wiley and Sons, New York, 1979.

[2] G. DANNINGER, *A recursive algorithm for determining (strict) copositivity of a symmetric matrix*, in XIV Symposium on Operations Research, Ulm, 1989, Methods Oper. Res. 62, Athenäum/Hain/Hanstein, Königstein, 1990, pp. 45–62.

[3] W. DINKELBACH, *On nonlinear fractional programming*, Management Sci., 13 (1967), pp. 492–498.

[4] B. EAVES, *The linear complementarity problem*, Management Sci., 11 (1971), pp. 612–634.

[5] C. E. LEMKE, *On complementary pivot theory*, in Mathematics of the Decision Sciences, Part 1, G. B. Dantzig and A. F. Veinott, Jr., eds., American Mathematical Society, Providence, RI, 1968, pp. 95–114.

[6] K. G. MURTY, *Linear Complementarity, Linear and Nonlinear Programming*, Heldermann Verlag, Berlin, 1988.

[7] K. G. MURTY AND S. N. KABADI, *Some NP-Complete problems in quadratic and nonlinear programming*, Math. Programming, 39 (1987), pp. 117–129.

[8] J. C. PREISIG, *Robust maximum energy matched field processing*, IEEE Trans. Signal Processing, 42 (1994), pp. 1585–1593.

[9] S. SCHAIBLE, *Fractional programming. II, On Dinkelbach's algorithm*, Management Sci., 22 (1976), pp. 868–873.

[10] ———, *A survey of fractional programming*, in Proc. of NATO ASI, Vancouver, Canada, 1980, Generalized Concavity in Optimization and Economics, Academic Press, New York, 1981, pp. 417–440.

[11] H. VÄLIAHO, *Criteria for copositive matrices*, Linear Algebra Appl., 81 (1986), pp. 19–34.

[12] ———, *Quadratic-programming criteria for copositive matrices*, Linear Algebra Appl., 119 (1989), pp. 163–182.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1151–1171, July 1996

003

PERTURBED OPTIMIZATION IN BANACH SPACES I: A GENERAL THEORY BASED ON A WEAK DIRECTIONAL CONSTRAINT QUALIFICATION*

J. FRÉDÉRIC BONNANS† AND ROBERTO COMINETTI‡

Abstract. Using a directional form of constraint qualification weaker than Robinson's, we derive an implicit function theorem for inclusions and use it for first- and second-order sensitivity analyses of the value function in perturbed constrained optimization. We obtain Hölder and Lipschitz properties and, under a *no-gap* condition, first-order expansions for exact and approximate solutions. As an application, differentiability properties of metric projections in Hilbert spaces are obtained, using a condition generalizing polyhedricity. We also present in the appendix a short proof of a generalization of the convex duality theorem in Banach spaces.

Key words. sensitivity analysis, marginal function, approximate solutions, directional constraint qualification, regularity and implicit function theorems, convex duality

AMS subject classifications. 46A20, 46N10, 47H19, 49K27, 49K40, 58C15, 90C31

1. Introduction.

This paper is the first of a trilogy devoted to sensitivity analysis of parametrized optimization problems of the form

$$(P_u) \qquad \min_x \{f(x,u) : G(x,u) \in K\}$$

where f and G are C^2 mappings from $X \times \mathbb{R}_+$ to $\mathbb{R}$ and Y, respectively, X and Y are Banach spaces, and K is a closed convex subset of Y.

While the theory is fairly complete in the case of finite-dimensional mathematical programming, that is, optimization problems with finitely many equality and inequality constraints, the sensitivity of perturbed optimization problems in Banach spaces is still being developed. Just to mention a couple of recent works related to this topic, see for instance [3, 8, 9, 11, 19, 21, 26] as well as the monographs [10, 13, 18] and references therein.

Loosely speaking, the assumptions that support a complete sensitivity analysis of the value function and optimal solutions are uniqueness of the optimal solution for the unperturbed problem, constraint qualification, existence of Lagrange multipliers, and second-order sufficient optimality conditions.

Concerning constraint qualification, the standard assumption is Robinson's generalization [23] of the Mangasarian–Fromovitz condition [20]. Following the lines of previous works in mathematical programming [2, 5, 7, 12, 14], in this paper we show that sensitivity analysis is still possible under a weak directional form of constraint qualification that takes into account the nature of perturbations. This condition is used to derive a generalization of Robinson's implicit function theorem for systems of inequalities that, in conjunction with a strong second-order sufficient condition, allows us to obtain first- and second-order upper and under estimates of the marginal function. When these two estimates coincide (we give some sufficient conditions for this) the first-order sensitivity of approximate optimal solutions of (P_u) is obtained.

Our second-order expansion includes a term that takes into account the possible curvature of the boundary of K and does not appear in the classical setting of mathematical programming where K is a polyhedral set. This curvature term, studied in [9, 17] in the context of second-order necessary conditions (see also the previous work [4]), leads to a generalization of the

*Received by the editors May 9, 1994; accepted for publication (in revised form) February 15, 1995. This research was supported by the French–Chilean ECOS (Evaluation-Orientation de la Cooperation Scientifique avec le Chili et l'Uruguay) Program and European Community contract 931091CL.

†INRIA-Rocquencourt, B.P. 105, 78153 Rocquencourt, France.

‡Universidad de Chile, Casilla 170/3 Correo 3, Santiago, Chile. The research of this author was partially supported by FONDECYT (contract 1940564) and Fundación Andes.

notion of polyhedric set and to new results on differentiability of metric projections onto convex sets in Hilbert spaces.

We observe that in the case of the trivial perturbation $f(x,u) = f(x,0)$ and $G(x,u) = G(x,0)$ for all u, the directional constraint qualification reduces to Robinson's condition and our upper estimates to the necessary optimality conditions obtained in [9]. Similarly, from our under estimates one can easily derive (new) sufficient conditions for local optimality.

When the strong second-order condition fails, and particularly when the set of Lagrange multipliers is empty, we know that directional differentiability of solutions and of the marginal function may fail [5]. It seems that the directional constraint qualification considered in this paper is too weak to obtain a satisfactory sensitivity analysis in such cases. This motivates a strenghtened form of directional qualification, which is the subject of part II of this work.

Finally, in part III we study the application of both theories to semi-infinite programming, that is to say, optimization problems with X finite dimensional and infinitely many inequality constraints. In that case there is a gap between the upper and lower estimates, so we will fill this gap by computing sharper lower estimates.

We denote the feasible set, optimal value, and solution set of (P_u) as

$$\begin{aligned} F(u) &:= \{x \in X : G(x,u) \in K\}, \\ v(u) &:= \inf\{f(x,u) : x \in F(u)\}, \\ S(u) &:= \{x \in F(u) : f(x,u) = v(u)\}. \end{aligned}$$

Similarly, given an optimization problem (P) we denote by $F(P)$, $v(P)$, and $S(P)$ its feasible set, optimal value, and optimal solution set, respectively.

The set of Lagrange multipliers associated with an optimal solution $x \in S(u)$ is

$$\Lambda_u(x) := \{\lambda \in Y^* : \lambda \in N_K(G(x,u)),\ \mathcal{L}'_x(x,\lambda,u) = 0\}$$

with Y^* denoting the dual space of Y, $N_K(y)$ the normal cone to K at y, and $\mathcal{L}$ the Lagrangian function

$$\mathcal{L}(x,\lambda,u) := f(x,u) + \langle \lambda, G(x,u)\rangle.$$

For the rest of this paper we assume $v(0)$ finite and $S(0)$ nonempty. We also consider a fixed optimal solution $x_0 \in S(0)$ and denote by $\Lambda_0 := \Lambda_0(x_0)$ the corresponding set of multipliers.

Finally, we recall the definition of the first- and second-order tangent sets:

$$T_K(y) := \{h \in Y :\ \text{there exists } o(u) \text{ such that } y + uh + o(u) \in K\},$$

$$T_K^2(y,h) := \left\{k \in Y :\ \text{there exists } o(u^2) \text{ such that } y + uh + \frac{1}{2}u^2 k + o(u^2) \in K\right\}.$$

Throughout this paper $o(u)$ and $o(u^2)$ will be used freely to denote any terms that are negligible compared to u and u^2. Similary, $O(u)$ and $O(u^2)$ denote terms of orders u and u^2.

2. Upper estimates of the value function. We are interested in sensitivity analysis of (P_u), that is to say, the study of differentiability properties of the optimal value function v and the optimal (set-valued) map S. To this end we consider the *linear* and *quadratic* approximating problems:

$$(L) \qquad \min_d \{f'(x_0,0)(d,1) :\ G'(x_0,0)(d,1) \in T_K(G(x_0,0))\},$$

$$(Q) \qquad \min\{v(L_d) :\ d \in S(L)\},$$

$$(L_d) \qquad \min_w \{f'_x(x_0,0)w + \Phi_f(d) :\ G'_x(x_0,0)w + \Phi_G(d) \in T_K^2(d)\},$$

where we have set

$$\Phi_f(d) := f''(x_0, 0)(d, 1)(d, 1),$$
$$\Phi_G(d) := G''(x_0, 0)(d, 1)(d, 1),$$
$$T_K^2(d) := T_K^2(G(x_0, 0), G'(x_0, 0)(d, 1)).$$

The motivation for these approximating problems is the following.

We say that $u \to x_u$ is a *feasible path* if $x_u \in F(u)$ for $u > 0$ small enough and x_u tends to x_0 when $u \downarrow 0$. Suppose that we have a feasible path of the form $x_u = x_0 + ud + o(u)$. A first-order expansion gives $G(x_u, u) = G(x_0, 0) + uG'(x_0, 0)(d, 1) + o(u) \in K$, so d is feasible for (L) and also

$$v(u) \le f(x_u, u) = v(0) + uf'(x_0, 0)(d, 1) + o(u), \tag{1}$$

suggesting that $v(u) \le v(0) + u\ v(L) + o(u)$.

Similarly, if $d \in S(L)$ and $x_u = x_0 + ud + \frac{1}{2}u^2 w + o(u^2)$ is a feasible path, a second-order Taylor expansion of $G(x_u, u)$ shows that $w \in T(L_d)$, and

$$v(u) \le f(x_u, u) = v(0) + u\ v(L) + \frac{1}{2}u^2[f_x'(x_0, 0)w + \Phi_f(d)] + o(u^2), \tag{2}$$

so we may expect $v(u) \le v(0) + u\ v(L) + \frac{1}{2}u^2\ v(Q) + o(u^2)$.

To prove these upper estimates it suffices to show that each $d \in F(L)$ admits an $o(u)$ correction such that $x_0 + ud + o(u) \in F(u)$ and similarly that each $w \in F(L_d)$ admits an $o(u^2)$ correction such that $x_0 + ud + \frac{1}{2}u^2 w + o(u^2) \in F(u)$. The existence of such corrections may be established by using Robinson's regularity theorem [23, Thm. 1], which is based on the constraint qualification condition

$$0 \in \text{int}\,[G(x_0, 0) + G_x'(x_0, 0)X - K]. \tag{CQ}$$

However, this condition does not take into account the specific form of perturbations, so, loosely speaking, it will work uniformly no matter what type of perturbations are being considered. We shall rather use the following refinement of Robinson's regularity theorem proved in Appendix B, which allows us to discriminate those perturbations for which sensitivity analysis can be carried out.

THEOREM B.5. *Let us assume the* directional constraint qualification

$$0 \in \text{int}\,[G(x_0, 0) + G'(x_0, 0)X \times (0, \infty) - K]. \tag{DCQ}$$

Then for each trajectory $x_u = x_0 + O(u)$ there exist constants $c \ge 0, u_0 > 0$ and a second trajectory y_u such that $G(y_u, u) \in K$ and

$$\|y_u - x_u\| \le c\ d(G(x_u, u), K)$$

for all $u \in [0, u_0]$.

It may not be apparent that (CQ) implies (DCQ). To see this we remark (see Appendix B) that the latter is equivalent to

$$0 \in \text{int}\,[G(x_0, 0) + G'(x_0, 0)X \times [0, \infty) - K]. \tag{DCQ}'$$

PROPOSITION 2.1. *Suppose (DCQ) holds. Then* $\limsup_{u \downarrow 0}[v(u) - v(0)]/u \le v(L)$ *and when $v(L) > -\infty$ we have the first-order upper estimate*

$$v(u) \le v(0) + u\ v(L) + o(u). \tag{3}$$

Also, $\limsup_{u\downarrow 0} 2[v(u) - v(0) - u\ v(L)]/u^2 \le v(Q)$ *and when* $v(Q) > -\infty$ *the following second-order upper estimate holds:*

$$v(u) \le v(0) + u\ v(L) + \frac{1}{2}u^2\ v(Q) + o(u^2). \tag{4}$$

Proof. Let d be feasible for (L). Applying Theorem B.5 with $x_u = x_0 + ud$ we find a feasible trajectory y_u such that $\|y_u - x_u\| \le c\ d(G(x_u, u), K) = o(u)$. Then $y_u = x_0 + ud + o(u)$ and the first-order estimate follows from (1).

To prove the second-order estimate, let $d \in S(L)$ and $w \in F(L_d)$. Applying Theorem B.5 with $x_u = x_0 + ud + \frac{1}{2}u^2 w$ we get a feasible trajectory y_u with $\|y_u - x_u\| \le c\ d(G(x_u, u), K) = o(u^2)$. Then $y_u = x_0 + ud + \frac{1}{2}u^2 w + o(u^2)$ and the conclusion follows from (2). □

The above upper estimates are only meaningful if $v(L) < +\infty$ and $v(Q) < +\infty$. Let us then prove the following result.

PROPOSITION 2.2. *Assuming (DCQ) we have* $v(L) < +\infty$. *Moreover, in this case* $v(Q) < +\infty$ *if and only if there exists* $d \in S(L)$ *such that* $T_K^2(d) \ne \phi$.

Proof. Using *(DCQ)* we may find $t > 0$ and $d \in X$ with $G'(x_0, 0)(d, t) \in K - G(x_0, 0)$. Then d/t is feasible for (L) and consequently $v(L) < +\infty$.

Clearly $v(Q) < +\infty$ requires $T_K^2(d) \ne \phi$ for some $d \in S(L)$.

To prove the converse we fix $k \in T_K^2(d)$ so that, according to [9, Prop. 3.1],

$$k + \mathbb{R}_+[T_K(G(x_0, 0)) - G'(x_0, 0)(d, 1)] \subset T_K^2(d). \tag{5}$$

Using *(DCQ)* we find $\mu > 0$ with $\mu[k - \Phi_G(d)] \in G(x_0, 0) + G'(x_0, 0)X \times (0, \infty) - K$, and then for some $z \in X$ and $t > 0$ we get

$$\begin{aligned}\Phi_G(d) &\in k + \frac{1}{\mu}[K - G(x_0, 0) - G'(x_0, 0)(z, t)] \\ &\in k - G'_x(x_0, 0)\frac{z - td}{\mu} + \frac{1}{\mu}[T_K(G(x_0, 0)) - tG'(x_0, 0)(d, 1)].\end{aligned}$$

Letting $w := (z - td)/\mu$ and using (5) we deduce that

$$G'_x(x_0, 0)w + \Phi_G(d) \in k + \frac{t}{\mu}[T_K(G(x_0, 0)) - G'(x_0, 0)(d, 1)] \subset T_K^2(d),$$

proving that (L_d) is feasible and then $v(Q) \le v(L_d) < +\infty$. □

3. Differentiability of the value function and suboptimal trajectories. To find lower estimates of the cost and sufficient conditions for the existence of the right derivative $v'(0)$, we use convex duality theory to get the following characterization for $v(L)$.

PROPOSITION 3.1. *Assume (DCQ). Then* $v(L) = v(D)$ *and* $S(D) \ne \phi$, *where*

$$\max\{\mathcal{L}'_u(x_0, \lambda, 0) : \lambda \in \Lambda_0\}. \tag{D}$$

Moreover, $v(L) > -\infty$ *if and only if* $\Lambda_0 \ne \phi$, *in which case* $S(D)$ *is a nonempty weak* compact subset of* Λ_0.

Proof. This is a consequence of the convex duality theorem of Appendix A, Theorem A.2, applied to problem (L) with the perturbation function

$$\varphi(d, y) := \begin{cases} f'(x_0, 0)(d, 1) & \text{if } G'(x_0, 0)(d, 1) + y \in T_K(G(x_0, 0)), \\ +\infty & \text{otherwise.} \end{cases}$$

Indeed, from (*DCQ*) we get

$$\begin{aligned} Y &= T_K(G(x_0, 0)) - G'(x_0, 0)X \times (0, \infty) \\ &= \mathbb{R}_+ \bigcup_{d \in X} [T_K(G(x_0, 0)) - G'(x_0, 0)(d, 1)], \end{aligned}$$

so $\mathbb{R}_+ \cup_d \operatorname{dom} \varphi(d, \cdot) = Y$ and Theorem A.2 can be used to deduce

$$v(L) = -\min_{\lambda} \varphi^*(0, \lambda). \tag{6}$$

A straightforward computation shows that

$$\varphi^*(x^*, \lambda) = \begin{cases} -\mathcal{L}'_u(x_0, \lambda, 0) & \text{if } \lambda \in N_K(G(x_0, 0)), \mathcal{L}'_x(x_0, \lambda, 0) = x^*, \\ +\infty & \text{otherwise,} \end{cases}$$

which combined with (6) yields the desired conclusions. □

We state our next results using suboptimal paths. We say that x_u is an $o(u)$-*optimal trajectory* if it is a feasible path and $v(u) = f(x_u, u) + o(u)$.

Existence of $o(u)$- and $o(u^2)$-optimal paths requires finiteness of $v(u)$. Conversely, when the latter holds, one may always find $o(u)$ or $o(u^2)$ approximate solutions of (P_u). The fact that these paths do converge to x_0 as u tends to 0 can be proved in a number of particular situations (see for instance [6, 12]).

In addition, we shall either assume Hölder and Lipschitz stability of these suboptimal paths (these assumptions will be discussed in §6) or we shall suppose that problem (P_0) is convex in the sense that for all $y \in K$ and $\lambda \in N_K(y)$ the mapping $\mathcal{L}(\cdot, \lambda, 0)$ is convex. The next result, under the convexity assumption, extends that given by Gol'stein [15].

PROPOSITION 3.2. *Suppose that (DCQ) holds, there exists an $o(u)$-optimal trajectory x_u, and either (P_0) is convex or $x_u = x_0 + o(\sqrt{u})$. Then v is right differentiable at 0 with $v'(0) = v(L)$. Moreover, when $\Lambda_0 \neq \phi$ we have*

$$v(u) = v(0) + u\, v(L) + o(u).$$

Proof. If $\Lambda_0 = \phi$ we have $v(L) = -\infty$ and the result follows immediately from Proposition 2.1. Otherwise, by Proposition 3.1 we may take $\lambda \in S(D) \subset \Lambda_0$ so that

$$\begin{aligned} v(u) - v(0) &= f(x_u, u) - f(x_0, 0) + o(u) \\ &\geq \mathcal{L}(x_u, \lambda, u) - \mathcal{L}(x_0, \lambda, 0) + o(u). \end{aligned}$$

Since $\mathcal{L}'_x(x_0, \lambda, 0) = 0$, when (P_0) is convex we get $\mathcal{L}(x_0, \lambda, 0) \leq \mathcal{L}(x_u, \lambda, 0)$ and when $x_u = x_0 + o(\sqrt{u})$ a second-order expansion gives $\mathcal{L}(x_0, \lambda, 0) = \mathcal{L}(x_u, \lambda, 0) + o(u)$. In both cases we obtain

$$v(u) - v(0) \geq \mathcal{L}(x_u, \lambda, u) - \mathcal{L}(x_u, \lambda, 0) + o(u)$$

and, since x_u tends to x_0, deduce that

$$\liminf_{u \downarrow 0} \frac{v(u) - v(0)}{u} \geq \mathcal{L}'_u(x_0, \lambda, 0) = v(D) = v(L),$$

which combined with Proposition 2.1 yields the desired conclusions. □

As a further consequence we establish a relation between the solution set $S(L)$ and the right derivatives of suboptimal trajectories.

PROPOSITION 3.3. *With the assumptions of Proposition* 3.2 *we have:*

(a) *$S(L)$ is the set of all weak accumulation points of $(x_u - x_0)/u$, where x_u ranges over all possible $o(u)$-optimal trajectories.*

(b) *If $S(L) \neq \phi$, then there exists an $o(u)$-optimal trajectory such that $x_u = x_0 + O(u)$. The converse holds if X is reflexive.*

(c) *If x_u is chosen as in* (b), *then $\Lambda_u(x_u)$ is uniformly bounded for u small. Moreover, if $\lambda_u \in \Lambda_u(x_u)$, then every weak* accumulation point of λ_u belongs to $S(D)$.*

Proof. (a) Let x_u be an $o(u)$-optimal trajectory and $u_k \downarrow 0$ be such that $(x_{u_k} - x_0)/u_k \rightharpoonup d$. Then we have $[G(x_{u_k}, u_k) - G(x_0, 0)]/u_k \rightharpoonup G'(x_0, 0)(d, 1)$ and, since $T_K(G(x_0, 0))$ is weakly closed, deduce that $G'(x_0, 0)(d, 1) \in T_K(G(x_0, 0))$, proving that $d \in F(L)$. Similarly, $[f(x_{u_k}, u_k) - f(x_0, 0)]/u_k \to f'(x_0, 0)(d, 1)$ and then

$$v(u_k) = f(x_{u_k}, u_k) + o(u_k) = v(0) + u_k f'(x_0, 0)(d, 1) + o(u_k),$$

so Proposition 2.1 implies $f'(x_0, 0)(d, 1) \leq v(L)$, which shows $d \in S(L)$.

Conversely, let $d \in S(L)$ and apply Theorem B.5 to the trajectory $x_u = x_0 + ud$ to find $y_u = x_0 + ud + o(u) \in F(u)$. Proposition 3.2 then implies

$$f(y_u, u) = f(x_0, 0) + uf'(x_0, 0)(d, 1) + o(u) = v(0) + u\, v(L) + o(u) = v(u) + o(u),$$

proving that y_u is an $o(u)$-optimal trajectory with $(y_u - x_0)/u \rightharpoonup d$ (notice that the limit can be taken in the strong sense as well).

(b) The argument developed in (a) shows that $S(L) \neq \phi$ implies the existence of $o(u)$-optimal trajectories with $x_u = x_0 + O(u)$. Conversely, if such a trajectory exists, then by reflexivity we may find a sequence $u_k \downarrow 0$ such that $(x_{u_k} - x_0)/u_k$ converges weakly. From (a) the limit belongs to $S(L)$, which is then nonempty.

(c) Let $\lambda_u \in \Lambda_u(x_u)$ and select $r_u \in B_Y$ with $\|\lambda_u\|/2 \leq \langle r_u, -\lambda_u \rangle$. From relation (17) in Lemma B.4, for all u small enough there exist $d_u \in B_X$ and $k_u \in K$ such that

$$u\varepsilon r_u = G(x_u, u) + umG'_x(x_0, 0)d_u - k_u$$

where $\varepsilon > 0$ and $m > 0$ are given constants. Taking the product with $-\lambda_u$ we get

$$\begin{aligned}
\frac{\varepsilon}{2}\|\lambda_u\| &\leq m\langle \lambda_u, G'_x(x_0, 0)d_u \rangle \\
&\leq m\|G'_x(x_0, 0) - G'_x(x_u, u)\|\|\lambda_u\| + m\langle \lambda_u, G'_x(x_u, u)d_u \rangle \\
&\leq \frac{\varepsilon}{4}\|\lambda_u\| - mf'_x(x_u, u)d_u \\
&\leq \frac{\varepsilon}{4}\|\lambda_u\| + m(\|f'_x(x_0, 0)\| + 1)
\end{aligned}$$

for u small, and the desired uniform bound on $\Lambda_u(x_u)$ follows.

Now let $\lambda := \lim_k \lambda_{u_k}$ be a weak* accumulation point of λ_u where $u_k \downarrow 0$. Then

$$\begin{aligned}
&\forall y \in K \qquad \langle \lambda, y - G(x_0, 0) \rangle = \lim_k \langle \lambda_{u_k}, y - G(x_{u_k}, u_k) \rangle \leq 0, \\
&\forall d \in X \qquad \mathcal{L}'_x(x_0, \lambda, 0)d = \lim_k \mathcal{L}'_x(x_{u_k}, \lambda_{u_k}, u_k)d = 0,
\end{aligned}$$

proving that $\lambda \in \Lambda_0 = F(D)$. To show λ is also optimal for (D) we observe that

$$\begin{aligned}
v(u) &\leq f(x_u, u) \\
&\leq f(x_u, u) - \langle \lambda_u, G(x_0, 0) - G(x_u, u) \rangle \\
&= v(0) + \mathcal{L}(x_u, \lambda_u, u) - \mathcal{L}(x_0, \lambda_u, 0) \\
&= v(0) + u\mathcal{L}'_u(x_0, \lambda_u, 0) + o(u + \|x_u - x_0\|).
\end{aligned}$$

Dividing by u and passing to the limit in the subsequence u_k we get $\mathcal{L}'_u(x_0, \lambda, 0) \geq v'(0) = v(D)$, so $\lambda \in S(D)$. □

Remark. In part (a) above we also showed that $S(L)$ is the set of all *strong* limits of differential quotients of the type $(x_{u_k} - x_0)/u_k$ with $u_k \downarrow 0$ and even the set of *continuous* strong limits

$$d := \lim_{u \downarrow 0} \frac{x_u - x_0}{u},$$

where now x_u ranges over all $o(u)$-optimal trajectories for which this limit exists.

4. Second-order expansion of the value function. In this section we supplement Proposition 2.1 by deriving second-order lower estimates for the value function. The next simple result shows that (4) is a sharp bound.

PROPOSITION 4.1. *Suppose (DCQ) holds and assume there exists an $o(u^2)$-optimal path x_u that admits an expansion of the form $x_u = x_0 + ud_0 + \frac{1}{2}u^2 w_0 + o(u^2)$. Then $d_0 \in S(Q)$, $w_0 \in S(L_{d_0})$, and we have*

$$v(u) = v(0) + u\, v(L) + \frac{1}{2}u^2\, v(Q) + o(u^2).$$

Proof. Propositions 3.2 and 3.3(a) imply $v'(0) = v(L)$ and $d_0 \in S(L)$. On the other hand, a second-order expansion of $G(x_u, u)$ shows that $w_0 \in F(L_{d_0})$ and also

$$\begin{aligned} v(u) &= f(x_u, u) + o(u^2) \\ &= f(x_0, 0) + uf'(x_0, 0)(d_0, 1) + \frac{1}{2}u^2[f'_x(x_0, 0)w_0 + \Phi_f(d_0)] + o(u^2) \\ &= v(0) + u\, v(L) + \frac{1}{2}u^2[f'_x(x_0, 0)w_0 + \Phi_f(d_0)] + o(u^2), \end{aligned}$$

which combined with Proposition 2.1 gives the desired conclusions. □

Unfortunately this result is of more theoretical than practical interest since we must ensure a priori the existence of a second-order expansion of x_u. While it is possible to find conditions giving a first-order expansion (see §6), we dispose of no analogue for the second-order case. To overcome this difficulty we tackle the second-order lower estimates using duality theory as was done in the previous section for the first order. Let us then dualize problem (L_d).

PROPOSITION 4.2. *Suppose (DCQ) holds. Then $v(L_d) = v(D_d)$ where*

$$(D_d) \qquad \max\{\mathcal{L}''(x_0, \lambda, 0)(d, 1)(d, 1) - \sigma(\lambda, T_K^2(d)) : \lambda \in S(D)\},$$

and $\sigma(\lambda, T_K^2(d)) := \sup\{\langle \lambda, k\rangle : k \in T_K^2(d)\}$ is the support function of $T_K^2(d)$. Moreover the solution set $S(D_d)$ is nonempty.

Proof. The case $T_K^2(d) = \phi$ being trivial, we shall assume $T_K^2(d) \neq \phi$ (notice that in this case $d \in F(L)$). Let us consider problem (L_d) with the perturbation function

$$\varphi(w, y) := \begin{cases} f'_x(x_0, 0)w + \Phi_f(d) & \text{if } G'_x(x_0, 0)w + \Phi_G(d) + y \in T_K^2(d), \\ +\infty & \text{otherwise.} \end{cases}$$

To apply Theorem A.2 we must check that $\mathbb{R}_+ \cup_w \operatorname{dom} \varphi(w, \cdot) = Y$. To this end we fix $k \in T_K^2(d)$ and use property (5) to get

$$\begin{aligned} \bigcup_w \operatorname{dom} \varphi(w, \cdot) &= T_K^2(d) - G'_x(x_0, 0)X - \Phi_G(d) \\ &\supset k + T_K(G(x_0, 0)) - G'(x_0, 0)X \times (0, \infty) - \Phi_G(d) \\ &= Y, \end{aligned}$$

the last equality since (*DCQ*) implies $T_K(G(x_0, 0)) - G'(x_0, 0)X \times (0, \infty) = Y$.

We may then use the convex duality theorem to deduce

$$v(L_d) = -\min_{\lambda} \varphi^*(0, \lambda)$$

and a straightforward computation to obtain

$$\varphi^*(0, \lambda) = \begin{cases} \sigma(\lambda, T_K^2(d)) - \mathcal{L}''(x_0, \lambda, 0)(d, 1)(d, 1) & \text{if } \mathcal{L}'_x(x_0, \lambda, 0) = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

To complete the proof we note that if λ satisfies $\mathcal{L}'_x(x_0, \lambda, 0) = 0$, we may have $\sigma(\lambda, T_K^2(d)) < +\infty$ only if $\lambda \in S(D)$ (and $d \in S(L)$). Indeed, if $\sigma(\lambda, T_K^2(d)) < +\infty$, property (5) shows that

$$\langle \lambda, h - G'(x_0, 0)(d, 1) \rangle \leq 0 \quad \text{for all } h \in T_K(G(x_0, 0)).$$

This implies $\lambda \in N_K(G(x_0, 0))$; hence $\lambda \in \Lambda_0$, and also $\langle \lambda, G'(x_0, 0)(d, 1) \rangle \geq 0$ so that

$$f'(x_0, 0)(d, 1) \leq \mathcal{L}'(x_0, \lambda, 0)(d, 1) = \mathcal{L}'_u(x_0, \lambda, 0).$$

Since $\lambda \in F(D)$ and $d \in F(L)$, this inequality proves that $\lambda \in S(D)$ and $d \in S(L)$. □

With this result we have the following min-max characterization of $v(Q)$:

$$v(Q) = \min_{d \in S(L)} \max_{\lambda \in S(D)} \mathcal{L}''(x_0, \lambda, 0)(d, 1)(d, 1) - \sigma(\lambda, T_K^2(d)).$$

The term $\sigma(\lambda, T_K^2(d))$ above will be referred to as the "σ-term" for short and is related, loosely speaking, to the curvature of the set K (see also [9, 17]). Neglecting this σ-term we obtain second-order lower estimates that, however, may not be sharp. To be precise, let us consider the function

$$\Gamma(d) := \max_{\lambda \in S(D)} \mathcal{L}''(x_0, \lambda, 0)(d, 1)(d, 1)$$

and the optimization problems

$$(\tilde{Q}) \qquad \min\{\Gamma(d) : d \in S(L)\},$$

$$(\tilde{Q}_\varepsilon) \qquad \min\{\Gamma(d) : d \in S_\varepsilon(L)\},$$

where $S_\varepsilon(L)$ is the set of approximate solutions of (L)

$$S_\varepsilon(L) := \{d \in F(L) : f'(x_0, 0)(d, 1) \leq v(L) + \varepsilon\}.$$

To obtain meaningful second-order lower bounds we must assume that $v(L) > -\infty$. By Proposition 3.3 this amounts to $\Lambda_0 \neq \phi$, in which case $S(D)$ is a weak* compact subset of Λ_0.

PROPOSITION 4.3. *Suppose (DCQ) holds, $\Lambda_0 \neq \phi$, and assume there exists an $o(u^2)$-optimal path x_u such that $x_u = x_0 + O(u)$. Then, for each $\varepsilon > 0$ we have*

$$v(u) \geq v(0) + u\, v(L) + \frac{1}{2} u^2\, v(\tilde{Q}_\varepsilon) + o(u^2). \tag{7}$$

Moreover, if any of the following conditions hold:

(a) *the path may be expanded as $x_u = x_0 + u d_0 + o(u)$,*

(b) *X is reflexive and Γ is weakly l.s.c. at each $d_0 \in S(L)$,*

then the previous lower bound may be strengthened to

$$v(u) \geq v(0) + u\, v(L) + \frac{1}{2}u^2\, v(\tilde{Q}) + o(u^2). \tag{8}$$

Proof. For each $\lambda \in S(D)$ we have

$$\begin{aligned} v(u) &= f(x_u, u) + o(u^2) \\ &\geq f(x_u, u) + \langle \lambda, G(x_u, u) - G(x_0, 0)\rangle + o(u^2) \\ &= v(0) + \mathcal{L}(x_u, \lambda, u) - \mathcal{L}(x_0, \lambda, 0) + o(u^2) \\ &= v(0) + u\, v(L) + \frac{1}{2}\mathcal{L}''(x_0, \lambda, 0)(x_u - x_0, u)(x_u - x_0, u) + o(u^2) \end{aligned} \tag{9}$$

and the small term $o(u^2)$ may be chosen uniform in λ since $S(D)$ is bounded.

Applying Theorem B.4 to the mapping $\tilde{G}(x, u) := G(x_0, 0) + G'(x_0, 0)(x - x_0, u)$ we find a path y_u with $\tilde{G}(y_u, u) \in K$ and

$$\|y_u - x_u\| \leq c\, d(\tilde{G}(x_u, u), K) \leq c\, \|\tilde{G}(x_u, u) - G(x_u, u)\| = o(u). \tag{10}$$

Replacing in (9) we find

$$v(u) \geq v(0) + u\, v(L) + \frac{1}{2}\mathcal{L}''(x_0, \lambda, 0)(y_u - x_0, u)(y_u - x_0, u) + o(u^2)$$

with $o(u^2)$ still independent of λ. Thus, letting $d_u := (y_u - x_0)/u$ and taking the supremum in λ over the bounded set $S(D)$, we get

$$v(u) \geq v(0) + u\, v(L) + \frac{1}{2}u^2\Gamma(d_u) + o(u^2). \tag{11}$$

But $\tilde{G}(y_u, u) \in K$ implies $d_u \in F(L)$, and the equality $v(u) = f(x_u, u) + o(u^2) = f(y_u, u) + o(u)$ implies that for each $\varepsilon > 0$ the vector d_u belongs to $S_\varepsilon(L)$ for u small, so (7) follows immediately from (11).

Let us next choose $u_k \downarrow 0$, realizing the lower limit $\liminf_u 2[v(u) - v(0) - uv(L)]/u^2$. When (a) holds we have $d_{u_k} \to d_0$, while in case (b) we may assume (by eventually passing to a subsequence) that d_{u_k} converges weakly to some d_0. In both cases Proposition 3.3 implies $d_0 \in S(L)$ and using (11) (and the l.s.c. of Γ) we get

$$v(u_k) \geq v(0) + u_k\, v(L) + \frac{1}{2}u_k^2\Gamma(d_0) + o(u_k^2), \tag{12}$$

from which (8) follows. □

5. Asymptotic expansions of suboptimal solutions. In this section we prove the analogue of Proposition 3.3 for the second-order problem (Q). Roughly speaking, the solution set $S(Q)$ is the set of right derivatives of $o(u^2)$-optimal paths.

This result is obtained under a strong assumption, namely, that there exists no gap between the upper and lower estimates (4) and (8). This *no-gap* condition is not true in general—we will see in part III that semi-infinite programming does not satisfy this property—but is still valid for a large class of applications, one of which will be considered in §7.

The next result gives sufficient conditions for having no gap.

PROPOSITION 5.1. (a) *For $\lambda \in S(D)$ and $d \in S(L)$ one has $\sigma(\lambda, T_K^2(d)) \leq 0$.*

(b) *If $d \in S(L)$ and $0 \in T_K^2(d)$, then $\sigma(\lambda, T_K^2(d)) = 0$ for all $\lambda \in S(D)$.*

(c) *If $0 \in T_K^2(d)$ for all d in a (strongly) dense subset of $S(L)$, then $v(Q) = v(\tilde{Q})$.*

Proof. For all $\lambda \in S(D)$ and $d \in S(L)$ we have $\langle \lambda, G'(x_0, 0)(d, 1)\rangle = 0$. Moreover, since $\lambda \in N_K(G(x_0, 0))$, for each $k \in T_K^2(d)$ we get

$$\langle \lambda, G(x_0, 0) + uG'(x_0, 0)(d, 1) + \frac{1}{2}u^2 k + o(u^2) - G(x_0, 0)\rangle \leq 0,$$

from which $\langle \lambda, k\rangle \leq 0$ and (a) follows.

Property (b) is obvious from (a). To prove (c) we notice that (a) implies $v(Q) \geq v(\tilde{Q})$, so we must only show the converse inequality. To this end it suffices to assume $S(L) \neq \phi$, in which case $S(D)$ is weak* compact and then Γ is strongly continuous. The required inequality follows using (b). □

Note that $0 \in T_K^2(y, h)$ when K is polyhedral in the sense that $T_K(y) = \mathbb{R}_+(K - y)$. This is the case for optimization problems with equality constraints and finitely many inequality constraints, where $K = \{0\} \times \mathbb{R}_-^p$. Thus, the condition "$0 \in T_K^2(d)$ for all d in a dense subset of $S(L)$" may be interpreted as a generalization of polyhedrality which, in a certain sense, rules out any curvature of K. We shall refer to this condition as *extended polyhedricity* (see also the discussion at the end of §7).

COROLLARY 5.2. *Let the hypothesis of Proposition* 4.3(b) *be satisfied, and suppose that the extended polyhedricity condition holds. Then* $v(Q) = v(\tilde{Q})$ *and we have*

$$v(u) = v(0) + u\, v(L) + \frac{1}{2}u^2\, v(Q) + o(u^2).$$

The previous results raise the question whether a second-order expansion compatible with curvature may hold. In this sense, we mention that the *sharp* lower estimate

$$v(u) \geq v(0) + u\, v(L) + \frac{1}{2}u^2\, v(Q) + o(u^2) \tag{13}$$

holds under assumption (a) of Proposition 4.3 and the additional hypothesis:

(H) For all sequences $u_n \downarrow 0$ and $y_n = y + u_n h + o(u_n) \in K$, there exists $k_n \in T_K^2(y, h)$ with $y_n = y + u_n h + \frac{1}{2}u_n^2 k_n + o(u_n^2)$.

The proof is similar to that of Proposition 4.3 and is left to the reader. In the case of assumption (b) in Proposition 4.3, (H) must be suitably modified in terms of weakly convergent sequences.

While (H) is not always satisfied, we observe that it holds whenever $0 \in T_K^2(y, h)$. To see that (H) is in fact more general than the latter one may consider the set $K = \{(x, y) \in \mathbb{R}^2 : y \geq x^2\}$ that satisfies (H) but $0 \notin T_K^2(y, h)$. Unfortunately, we do not know an easy way to check (H) in the general case. Nevertheless, in part III of this work we obtain sufficient conditions for obtaining the sharp lower estimate (13) in semi-infinite programming problems.

The next result links $S(Q)$ with the asymptotic behavior of suboptimal paths. Part (b) is a converse of Proposition 4.1.

PROPOSITION 5.3. *Suppose (DCQ) holds,* $\Lambda_0 \neq \phi$, *there exists an* $o(u^2)$*-optimal path* x_u *such that* $x_u = x_0 + O(u)$, *and suppose in addition that* $v(Q) = v(\tilde{Q})$ *and* Γ *is weakly l.s.c. at every* $d \in S(L)$. *Then:*

(a) $S(Q) \subset S(\tilde{Q})$ *and for every* $o(u^2)$*-optimal path* z_u, *the weak accumulation points of* $(z_u - x_0)/u$ *belong to* $S(\tilde{Q})$.

(b) *If* X *is reflexive,* $d_0 \in S(Q)$, *and* $w_0 \in S(L_{d_0})$, *then there exists an* $o(u^2)$*-optimal path of the form* $z_u = x_0 + ud_0 + \frac{1}{2}u^2 w_0 + o(u^2)$.

Proof. (a) Since $v(Q) = v(\tilde{Q})$ and the cost of (Q) dominates the cost of $(\tilde{Q})$, we deduce $S(Q) \subset S(\tilde{Q})$. If d_0 is the weak limit of $(z_{u_k} - x_0)/u$, reasoning as in the proof of (9) and using (4) we obtain $v(Q) \geq \Gamma(d_0)$. But Proposition 3.3 implies $d_0 \in S(L)$, so $\Gamma(d_0) \geq v(\tilde{Q}) = v(Q)$ and then $d_0 \in S(\tilde{Q})$.

(b) Using Theorem B.4 we may find a feasible path $z_u = x_0 + ud_0 + \frac{1}{2}u^2 w_0 + o(u^2)$. Expanding $f(z_u, u)$ we get

$$\begin{aligned} f(z_u, u) &= f(x_0, 0) + uf'(x_0, 0)(d_0, 1) + \frac{1}{2}u^2[f'_x(x_0, 0)w_0 + \Phi_f(d_0)] + o(u^2) \\ &= v(0) + uv(L) + \frac{1}{2}u^2 v(Q) + o(u^2) \\ &= v(0) + uv(L) + \frac{1}{2}u^2 v(\tilde{Q}) + o(u^2) \\ &\leq v(u) + o(u^2), \end{aligned}$$

where the last inequality follows from Proposition 4.3. This shows that z_u is $o(u^2)$-optimal and the proof is complete. □

Remark. In the next section we check that, under some reasonable hypothesis, every $o(u^2)$-optimal path satisfies $x_u = x_0 + O(u)$. When X is reflexive this implies the existence of weak accumulation points of $(x_u - x_0)/u$, so that $S(\tilde{Q})$ is nonempty. We also observe that when $0 \in T_K^2(d)$ for all $d \in S(L)$, the cost function in (Q) and $(\tilde{Q})$ coincide so that $S(Q) = S(\tilde{Q})$.

6. Hölder and Lipschitz properties of suboptimal paths. We discuss next the Hölder and Lipschitz stability properties of suboptimal paths assumed in the previous sections. The results we present are simple variants of known results (e.g., [8, 12, 14, 26]). The essential difference lies in the use of the weaker directional regularity condition (DCQ) and the extension to the infinite-dimensional setting.

Typically, the stability properties follow from different second-order sufficient optimality conditions. More precisely, for each set $\Omega \subset \Lambda_0$ we consider the second-order condition

$$SOC(\Omega) \qquad \text{There exist } \alpha, \eta > 0 \text{ s.t. } \max_{\lambda \in \Omega} \mathcal{L}''_x(x_0, \lambda, 0)dd \geq \alpha \quad \forall\, d \in C_\eta,$$

where

$$C_\eta = \{d \in X : \|d\| = 1,\ f'_x(x_0, 0)d \leq \eta,\ G'_x(x_0, 0)d \in T_K(G(x_0, 0)) + \eta B_Y\}.$$

When the space X is finite dimensional, or more generally when C_η is strongly compact for some $\eta > 0$, this condition is equivalent to the positive definiteness requirement:

$$SOC'(\Omega) \qquad \text{For each } d \in C_0 \text{ we have } \max_{\lambda \in \Omega} \mathcal{L}''_x(x_0, \lambda, 0)dd > 0,$$

where only the *critical cone* C_0 needs to be considered. Also, when (CQ) holds, one can replace C_η by a smaller set (see [8]).

PROPOSITION 6.1. *Assume (DCQ), $\Lambda_0 \neq \phi$, and suppose $SOC(\Omega)$ holds for some bounded $\Omega \subset \Lambda_0$. Then for each $O(u)$-optimal path x_u we have $x_u = x_0 + O(\sqrt{u})$.*

Proof. By contradiction suppose there exists $u_k \downarrow 0$ such that $\lim_k \tau_k^2/u_k = +\infty$, where $\tau_k := \|x_{u_k} - x_0\|$.

Then $\lim_k u_k/\tau_k = 0$ and letting $d_k := (x_{u_k} - x_0)/\tau_k$ we have $G(x_{u_k}, u_k) = G(x_0, 0) + \tau_k G'_x(x_0, 0)d_k + o(\tau_k)$ so that $G'_x(x_0, 0)d_k \in T_K(G(x_0, 0)) + \eta B_Y$ for k large. On the other hand, since x_u is an $O(u)$-optimal path and using Proposition 2.1, we may find a constant M such that for u small

$$f(x_u, u) \leq v(0) + Mu, \tag{14}$$

and since $f(x_{u_k}, u_k) = f(x_0, 0) + \tau_k f'_x(x_0, 0)d_k + o(\tau_k)$, we deduce $f'_x(x_0, 0)d_k \leq \eta$ for all k large enough. The previous argument shows that $d_k \in C_\eta$ for large k.

Now, using (14), for each $\lambda \in \Omega$ we have

$$\mathcal{L}(x_u, \lambda, u) - \mathcal{L}(x_0, \lambda, 0) \leq f(x_u, u) - f(x_0, 0) \leq Mu,$$

and since $\mathcal{L}'_x(x_0, \lambda, 0) = 0$, a second-order expansion of f and G leads to

$$\frac{1}{2}\mathcal{L}''(x_0, \lambda, 0)(x_u - x_0, u)(x_u - x_0, u) \leq [M - \mathcal{L}'_u(x_0, \lambda, 0)]u + (1 + \|\lambda\|)o(\|x_u - x_0\|^2 + u^2)$$

with the small term $o(\|x_u - x_0\|^2 + u^2)$ not depending on λ. Since Ω is bounded, we deduce that

$$\max_{\lambda\in\Omega} \mathcal{L}''(x_0, \lambda, 0)(d_k, u_k/\tau_k)(d_k, u_k/\tau_k) \leq M' \frac{u_k}{\tau_k^2} + M'' \frac{o(\tau_k^2 + u_k^2)}{\tau_k^2}$$

for some constants M' and M'', from which we get

$$\limsup_{k\to\infty} \max_{\lambda\in\Omega} \mathcal{L}''_x(x_0, \lambda, 0)d_k d_k \leq 0,$$

contradicting $SOC(\Omega)$. □

COROLLARY 6.2. *Assume $\Lambda_0 \neq \phi$ and any of the two following conditions:*

(a) *(CQ) and $SOC(\Lambda_0)$,*

(b) *(DCQ), C_η is strongly compact for some $\eta > 0$, and $SOC'(\Lambda_0)$.*

Then for each $O(u)$-optimal path x_u we have $x_u = x_0 + O(\sqrt{u})$.

Proof. In case (a) the set $\Omega := \Lambda_0$ is bounded and the result follows at once from the previous proposition.

In case (b) the set C_0 is compact and then, letting $\Lambda_0^k := \Lambda_0 \cap B(0, k)$, we get

$$\lim_{k\uparrow\infty} \min_{d\in C_0} [\max_{\lambda\in\Lambda_0^k} \mathcal{L}''_x(x_0, \lambda, 0)dd] = \min_{d\in C_0} [\max_{\lambda\in\Lambda_0} \mathcal{L}''_x(x_0, \lambda, 0)dd] > 0.$$

Hence, for k large $SOC'(\Omega)$ holds with $\Omega := \Lambda_0^k$, and we may conclude again using the previous proposition. □

The preceding results are not as strong as to ensure the property $x_u = x_0 + o(\sqrt{u})$ needed in Proposition 3.2. Let us then prove a Lipschitz stability result, valid for general Banach spaces, that can be used to check the hypothesis of both Proposition 4.3 and Proposition 3.2.

PROPOSITION 6.3. *Suppose (DCQ), $\Lambda_0 \neq \phi$, and assume $SOC(\Omega)$ holds for $\Omega := S(D)$. Suppose also that $v(Q) < +\infty$. Then, for each $O(u^2)$-optimal path x_u we have $x_u = x_0 + O(u)$.*

Proof. The proof is similar to that of Proposition 6.1. We proceed by contradiction assuming $\lim_k \tau_k/u_k = +\infty$ for a given sequence $u_k \downarrow 0$ and $\tau_k := \|x_{u_k} - x_0\|$, so that $d_k := (x_{u_k} - x_0)/\tau_k$ belongs to C_η for k large.

Since x_u is an $O(u^2)$-optimal path, using Proposition 2.1 we may find a constant M such that for u small

$$f(x_u, u) \leq v(0) + u\ v(L) + Mu^2,$$

and then for each $\lambda \in S(D)$ we have

$$\mathcal{L}(x_u, \lambda, u) - \mathcal{L}(x_0, \lambda, 0) \leq f(x_u, u) - f(x_0, 0) \leq u\ \mathcal{L}'_u(x_0, \lambda, 0) + Mu^2.$$

Expanding f and G we get

$$\mathcal{L}''(x_0, \lambda, 0)(d_k, u_k/\tau_k)(d_k, u_k/\tau_k) \leq 2M \left(\frac{u_k}{\tau_k}\right)^2 + \frac{o(\tau_k^2 + u_k^2)}{\tau_k^2}$$

with the small term $o(\tau_k^2 + u_k^2)$ not depending on λ (here we use the boundedness of $S(D)$). The conclusion follows as in Proposition 6.1. □

7. Directional differentiability of metric projections. In this section we use the preceding results to compute the directional derivatives of projections onto convex sets in Hilbert spaces. More precisely, the problem is to study the right differentiability of the unique optimal solution of

$$(P_u) \qquad \min\left\{\frac{1}{2}\|x - y_u\|^2 : x \in K\right\},$$

where K is a closed convex subset of a Hilbert space H and $u \to y_u$ is a smooth mapping from $\mathbb{R}_+$ to H. Let us consider the slightly more general format

$$(P'_u) \qquad \min\left\{\frac{1}{2}\|x - y_u\|^2 : G(x, u) \in K\right\},$$

assuming that $G(\cdot, 0)$ is a linear mapping $G(x, 0) = Ax$ and that (DCQ) and $\Lambda_0 \neq \phi$ hold. Notice that these properties are satisfied when we have (CQ), which is obviously the case if A is surjective and particularly if $G(x, 0) = x$ as in (P_u).

Since $G(x, 0)$ is linear, we have $\mathcal{L}''_x(x_0, \lambda, 0) = I$, so $SOC(\Omega)$ is automatically satisfied for $\Omega = S(D)$ and problem $(\tilde{Q})$ is strongly convex. In particular, $S(\tilde{Q})$ is reduced to a singleton.

PROPOSITION 7.1. *Suppose (DCQ), $\Lambda_0 \neq \phi$, and the extended polyhedricity condition. Then the unique solution x_u of (P'_u) may be expanded as*

$$x_u = x_0 + ud_0 + o(u)$$

where d_0 is the unique solution of $(\tilde{Q})$.

Proof. Propositions 6.3 and 5.3(a) imply that $d_u := (x_u - x_0)/u$ converges weakly to d_0, the unique solution of $(\tilde{Q})$. Now, using the second-order bound (4), the equality $v(Q) = v(\tilde{Q}) = \Gamma(d_0)$, and inequality (9), we deduce that

$$\limsup_{u \downarrow 0} \Gamma(d_u) \leq \Gamma(d_0).$$

Since Γ is strongly convex, we conclude that d_u converges strongly to d_0, completing the proof. □

In the special case $G(x, u) = x$ and $y_u = y_0 + uh_0$; that is, when we study directional differentiability of the projection onto K at y_0 in the direction h_0, the set $S(L)$ is just the critical cone

$$S(L) = C_0 = \{d \in T_K(x_0) : d \perp (y_0 - x_0)\},$$

so the problem $(\tilde{Q})$ reduces to

$$\min\{\|d - h_0\|^2 : d \in C_0\}.$$

Hence we get as an immediate consequence the following result.

COROLLARY 7.2. *Assuming the extended polyhedricity condition, the projection x_u of $y_0 + uh_0$ onto K can be expanded as*

$$x_u = x_0 + ud_0 + o(u),$$

where d_0 is the projection of h_0 onto C_0.

Among the papers studying differentiability properties of metric projections we mention [11, 16, 19, 22, 26, 27]. A common hypothesis in these studies is that K has to be *polyhedric* in this sense that for each $x \in K$ and every $\lambda \in N_K(x)$ one has

$$T_K(x) \cap \lambda^\perp = \overline{\mathbb{R}_+(K-x) \cap \lambda^\perp}.$$

Since $S(L) = T_K(x_0) \cap (y_0 - x_0)^\perp$ and $0 \in T_K^2(x_0, d)$ whenever $d \in \mathbb{R}_+(K - x_0)$, the extended polyhedricity condition is in fact a generalization of polyhedricity. Notice that this hypothesis always holds when $y_0 \in K$ since then $C_0 = T_K(G(x_0, 0))$, which was the case studied in [27]. Another extension of polyhedricity is considered in [3].

8. Conclusion and further problems. We have shown that a satisfactory sensitivity analysis for perturbed problems of the form

$$(P_u) \qquad \min_x \{f(x, u) : G(x, u) \in K\}$$

may be obtained under directional constraint qualification conditions that are weaker than the standard Robinson's condition.

Since the results are scattered throughout the paper, we provide a summarized (though necessarily incomplete) version of the main results obtained in the paper. The precise meaning of the stated assumptions and notation is made clear in the preceding sections of the paper, to which the reader is referred.

THEOREM 8.1. *Let the functions f, G defining (P_u) be of class C^2, and suppose X is a reflexive Banach space. Let x_0 be an optimal solution for (P_0) at which the following assumptions are satisfied:*

(i) *directional constraint qualification (DCQ),*
(ii) *existence of multipliers $\Lambda_0 \neq \emptyset$,*
(iii) *second-order sufficient condition SOC(Ω) for $\Omega = S(D)$,*
(iv) *existence of an $o(u^2)$-optimal trajectory,*
(v) *extended polyhedricity,*
(vi) *$d \to \mathcal{L}''_x(x_0, \lambda, 0)dd$ is weakly lower semicontinuous for all $\lambda \in S(D)$.*

Then:

(a) *The optimal value function may be expanded as*

$$v(u) = v(0) + uv(L) + \frac{1}{2}u^2 v(\tilde{Q}) + o(u^2),$$

where (L) and $(\tilde{Q})$ are the linear and quadratic approximating optimizing problems.

(b) *The optimal solutions of (L) are the same as the weak accumulation points of the differential quotients $(x_u - x_0)/u$ where x_u ranges over the set of all possible $o(u)$-optimal trajectories.*

(c) *Every $o(u^2)$-optimal path z_u satisfies $z_u = z_0 + O(u)$, and the weak accumulation points of $(z_u - z_0)/u$ are optimal solutions for $(\tilde{Q})$.*

We remark that a key ingredient for attaining these results is the generalization of Robinson's implicit function theorem presented in Appendix B, which is based on the weak directional constraint qualification condition (*DCQ*).

The main results of this paper are limited to problems for which there is existence of multipliers and satisfying the strong second-order sufficient condition stated as (iii) above, which ensure the existence of suboptimal paths of the form $x_u = x_0 + O(u)$.

In the setting of finite-dimensional mathematical programming we know [5] that this type of expansion may fail. For instance, when $\Lambda_0 \neq \phi$ but only the weak second-order

condition holds, suboptimal paths may only satisfy $x_u = x_0 + O(\sqrt{u})$ and it may happen that $v'(0) < v(L)$. On the other hand, when $\Lambda_0 = \phi$ we may even have $v(u) = v(0) + O(\sqrt{u})$.

It seems that (*DCQ*) is too weak to extend these results to the general framework discussed in the present paper. Theorem B.4 may not be used since it requires the a priori bound $x_u = x_0 + O(u)$, and its refinement Theorem B.1 may only handle those paths such that $\|x_u - x_0\| \leq \gamma\sqrt{u}$ for a sufficiently small γ.

These remarks lead us to consider a strenghtened form of directional constraint qualification, well suited to the analysis of problems of the form

$$\min\{f(x,u) : G_1(x,u) \in K_1,\ G_2(x,u) \in K_2\}$$

where K_1 and K_2 are closed convex subsets of some Banach spaces with $\text{int}(K_2) \neq \phi$. This study will be the subject of part II of this work.

Appendix A. The convex duality theorem in Banach spaces. This short appendix presents a short proof of the convex duality theorem of Robinson [24]. This result is a generalization of [25, Thm. 18(c)] (see also [1, Thm. 1.1]). We include it since the version we present is more directly applicable to the dualization of problems (L) and (L_d) in the previous sections and also since the method of proof is very simple. The basic argument is the following lemma due to Robinson [24] (also used in Appendix B) for which we provide a simplified proof too.

Given a subset $C \subset X \times Y$ we denote by C_X and C_Y the projections of C onto X and Y, respectively.

LEMMA A.1. *Let X, Y be two normed spaces with X complete. Let $C \subset X \times Y$ be a closed convex set with C_X bounded. Then*

$$\text{int}\,(C_Y) = \text{int}\,(\overline{C_Y}).$$

Proof. It clearly suffices to show $\text{int}\,(\overline{C_Y}) \subset C_Y$; that is, given $\overline{y} \in \text{int}\,(\overline{C_Y})$ we must find $\overline{x} \in X$ such that $(\overline{x}, \overline{y}) \in C$. To this end let us take $\varepsilon > 0$ with $B(\overline{y}, \varepsilon) \subset \overline{C_Y}$ and choose an arbitrary point $(x_0, y_0) \in C$ from which we generate a sequence $(x_k, y_k) \in C$ using the following "algorithm."

while $(y_k \neq \overline{y})$ *do*

- Let $\alpha_k = \varepsilon / \|y_k - \overline{y}\|$ so that $w := \overline{y} + \alpha_k(\overline{y} - y_k) \in B(\overline{y}, \varepsilon) \subset \overline{C_Y}$.
- Take $(u, v) \in C$ with $\|v - w\| \leq \frac{1}{2}\|y_k - \overline{y}\|$ and define

$$(x_{k+1}, y_{k+1}) := \frac{\alpha_k}{1+\alpha_k}(x_k, y_k) + \frac{1}{1+\alpha_k}(u, v) \in C.$$

endwhile.

If the algorithm stops, then we have $y_k = \overline{y}$ and we may take $\overline{x} = x_k$. Otherwise, the generated sequence satisfies

$$\text{(i)} \quad \|x_{k+1} - x_k\| = \frac{\|x_k - u\|}{1+\alpha_k} \leq \frac{\text{diam}(C_X)}{\varepsilon}\|y_k - \overline{y}\|,$$

$$\text{(ii)} \quad \|y_{k+1} - \overline{y}\| = \frac{\|v - w\|}{1+\alpha_k} \leq \frac{1}{2}\|y_k - \overline{y}\|.$$

From (ii) it follows that $\|y_k - \overline{y}\| \leq \|y_0 - \overline{y}\|/2^k$. This implies that $y_k \to \overline{y}$ and also, in combination with (i), that (x_k) is a Cauchy sequence. The completeness of X gives the existence of a limit $\overline{x}$ for (x_k), and the closedness of C implies $(\overline{x}, \overline{y}) \in C$ as required. □

We may now proceed by proving the convex duality theorem.

THEOREM A.2. *Let* $\theta(y) := \inf\{\varphi(x, y) : x \in X\}$, *where* $\varphi : X \times Y \to \mathbb{R} \cup \{+\infty\}$ *is a closed proper convex function with* X, Y *Banach spaces and* $\mathbb{R}_+ \cup_x \operatorname{dom} \varphi(x, \cdot) = Y$. *Then* θ *is continuous in a neighborhood of* 0 *and* $\theta(0) < +\infty$.

In particular $\theta(0) = \theta^{**}(0)$, *which can be written as*

$$\inf_{x \in X} \varphi(x, 0) = - \min_{y^* \in Y^*} \varphi^*(0, y^*), \tag{15}$$

and the solution set of the minimum on the right is $\partial\theta(0)$, *which is nonempty and weak*-compact when* $\theta(0)$ *is finite, and the whole space* Y^* *when* $\theta(0) = -\infty$.

Proof. Since θ is convex, the continuity near 0 is equivalent to θ being bounded above in a certain neighborhood of 0. To show this, let $\alpha \in \mathbb{R}$ and $x_0 \in X$ be such that $\varphi(x_0, 0) < \alpha$ and consider the closed convex set

$$C = \{(x, y) : \varphi(x, y) \le \alpha;\ \|x\| \le \|x_0\| + 1\}$$

that is nonempty and has C_X bounded.

Since $\theta(y) \le \alpha$ for all $y \in C_Y$, it suffices to show that C_Y is a neighborhood of 0. From Lemma A.1 this amounts to $0 \in \operatorname{int}(\overline{C_Y})$, which, by Baire's lemma, is a consequence of the fact that C_Y is absorbing as we show next. For any $y \in Y$ there exist $t > 0$ and $x \in X$ with $\varphi(x, ty) < +\infty$, so for $\varepsilon > 0$ small enough we have

$$\begin{aligned} &\|(1 - \varepsilon)x_0 + \varepsilon x\| \le \|x_0\| + 1, \\ &\varphi((1 - \varepsilon)(x_0, 0) + \varepsilon(x, ty)) \le (1 - \varepsilon)\varphi(x_0, 0) + \varepsilon\varphi(x, ty) \le \alpha, \end{aligned}$$

showing that $\varepsilon t y \in C_Y$ for all $\varepsilon > 0$ small.

We observe that $\theta^*(y^*) = \varphi^*(0, y^*)$ so that (15) is just a rewriting of $\theta(0) = \theta^{**}(0)$. From this we also get that $\partial\theta(0)$ is the solution set of $\min \varphi^*(0, y^*)$, and the last claim is a well-known fact in convex analysis (see [25]). □

Appendix B. Regularity theorems under directional constraint qualification conditions. Throughout this section we suppose that $G : X \times \mathbb{R}_+ \to Y$ is a $\mathcal{C}^2$ mapping and the spaces X, Y are Banach. Also $K \subset Y$ is a closed convex set and $x_0 \in X$ is such that $G(x_0, 0) \in K$ and satisfies the constraint qualification

$$0 \in \operatorname{int}\big[G(x_0, 0) + G'(x_0, 0)X \times (0, \infty) - K\big]. \tag{DCQ}$$

We begin by stating the equivalence.

PROPOSITION B.1. *Condition* (*DCQ*) *is equivalent to*

$$0 \in \operatorname{int}\,[G(x_0, 0) + G'(x_0, 0)X \times [0, \infty) - K]. \tag{DCQ}'$$

Proof. Clearly (*DCQ*) implies (*DCQ*)′. Conversely, suppose (*DCQ*)′ holds and choose $\varepsilon > 0$ with

$$\varepsilon B_Y \subset [G(x_0, 0) + G'(x_0, 0)X \times [0, \infty) - K].$$

Let $\delta > 0$ be such that $\delta[B_Y - G'_u(x_0, 0)] \subset \varepsilon B_Y$. Then

$$\delta B_Y \subset [G(x_0, 0) + G'(x_0, 0)X \times [\delta, \infty) - K],$$

from which (*DCQ*) follows. □

THEOREM B.2. *Let x_u be a trajectory such that $\|x_u - x_0\| \leq \gamma\sqrt{u}$, and suppose $d(G(x_u, u), K) \leq mu$ for some constants γ, m and all $u \geq 0$ close to 0. If γ is small enough, we can find constants $c \geq 0$, $u_0 > 0$ and a trajectory y_u with*

$$G(y_u, u) \in K,$$
$$\|y_u - x_u\| \leq \frac{c}{u}(u + \|x_u - x_0\|)d(G(x_u, u), K),$$

for all $u \in (0, u_0]$.

Our proof will be based on the following couple of lemmas.

LEMMA B.3. *Under assumption (DCQ), there exist $\varepsilon > 0$, $\alpha \geq 1$, and $\overline{u} > 0$ such that for all $u \in [0, \overline{u}]$*

$$2u\varepsilon B_Y \subset G(x_0, 0) + uG'_u(x_0, 0) + u\alpha G'_x(x_0, 0)B_X - K.$$

Proof. Letting $A_k := G(x_0, 0) + kG'(x_0, 0)B_X \times [0, 1] - K \cap kB_Y$, condition (*DCQ*) gives

$$0 \in \operatorname{int} \bigcup \{A_k : k \in \mathbb{N}\};$$

thus the completeness of Y implies $0 \in \operatorname{int}(\overline{A_k})$ for some $k \in \mathbb{N}$. But the set A_k can be expressed as the projection over the fourth component of the closed convex set

$$C := \left\{(x, y, t, G(x_0, 0) + G'(x_0, 0)(x, t) - y) : \|x\| \leq k, \|y\| \leq k, y \in K, t \in [0, k]\right\},$$

and since the projection of C onto its first three components is bounded, Lemma A.1 gives $\operatorname{int}(A_k) = \operatorname{int}(\overline{A_k})$. Therefore we may find $\varepsilon > 0$ such that

$$2\varepsilon k B_Y \subset G(x_0, 0) + kG'_u(x_0, 0) - k[0, 1]G'_u(x_0, 0) + kG'_x(x_0, 0)B_X - K,$$

which multiplied by u/k and rearranged becomes

$$2u\varepsilon B_Y \subset G(x_0, 0) + uG'_u(x_0, 0) + uG'_x(x_0, 0)B_X - S, \tag{16}$$

where

$$S := \left(1 - \frac{u}{k}\right) G(x_0, 0) + [0, 1]uG'_u(x_0, 0) + \frac{u}{k}K.$$

Now, (*DCQ*) implies $G'_u(x_0, 0) = [y - G(x_0, 0) - G'_x(x_0, 0)d]/\delta$ for some $y \in K$, $d \in X$, and $\delta > 0$, so

$$S = \bigcup_{\lambda \in [0,1]} \left[\left(1 - \frac{\lambda u}{\delta} - \frac{u}{k}\right) G(x_0, 0) + \frac{\lambda u}{\delta} y + \frac{u}{k}K\right] - \frac{\lambda u}{\delta} G'_x(x_0, 0)d.$$

Since K is convex, we deduce that $S \subset K - [0, 1]\frac{u}{\delta}G'_x(x_0, 0)d$ for all $u \leq \overline{u} := \delta k/(\delta + k)$, which combined with (16) yields the desired conclusion for $\alpha := 1 + \|d\|/\delta$. □

In the next lemma we denote

$$M = \sup\{\|G''(x, u)\| : \|x - x_0\| \leq 1, 0 \leq u \leq \overline{u}\}.$$

LEMMA B.4. *Let $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ be such that $\varphi(u) \leq \frac{1}{4}\sqrt{u\varepsilon/M}$ for all u sufficiently small. Then there exists $u_0 > 0$ such that, for each trajectory x_u with*

$$\|x_u - x_0\| \leq \varphi(u) \quad \forall u \in [0, u_0],$$

one has for all $u \in [0, u_0]$

$$u\varepsilon B_y \subset G(x_u, u) + (\alpha u + \|x_u - x_0\|)G'_x(x_0, 0)B_X - K. \tag{17}$$

Moreover, we can associate to x_u *another trajectory* y_u *such that for all* $u \in (0, u_0]$

$$\text{(i)} \quad d(G(y_u, u), K) \le \frac{1}{2} d(G(x_u, u), K),$$

$$\text{(ii)} \quad \|y_u - x_u\| \le \frac{2}{u\varepsilon}(\alpha u + \|x_u - x_0\|)d(G(x_u, u), K).$$

Proof. The hypothesis on $\varphi(u)$ ensures the existence of $u_0 \in (0, \overline{u}]$ such that

$$8M[\alpha u + \varphi(u)]^2 \le u\varepsilon \le M \quad \forall\, u \in [0, u_0]. \tag{18}$$

To show (17) we observe from (18) that $\|x_u - x_0\| \le \varphi(u) \le 1$, and then letting $b := G(x_u, u) - G(x_0, 0) - G'(x_0, 0)(x_u - x_0, u)$ we have

$$\|b\| \le M(u + \|x_u - x_0\|)^2 \le M[\alpha u + \varphi(u)]^2 \le u\varepsilon. \tag{19}$$

Thus, Lemma B.3 gives

$$u\varepsilon B_Y - b \subset 2u\varepsilon B_Y \subset G(x_0, 0) + uG'_u(x_0, 0) + u\alpha G'_x(x_0, 0)B_X - K,$$

and then

$$u\varepsilon B_Y \subset G(x_u, u) + G'_x(x_0, 0)[-(x_u - x_0) + u\alpha B_X] - K,$$

from which (17) follows at once.

Let us construct next the trajectory y_u for $u \in (0, u_0]$.

If $G(x_u, u) \in K$ we just take $y_u = x_u$ so that (i) and (ii) hold trivially.

Otherwise we choose r such that $G(x_u, u) + r \in K$ and

$$\|r\| \le 2d(G(x_u, u), K), \tag{20}$$

and we use (17) to select d with $\|d\| \le \alpha u + \|x_u - x_0\|$ such that

$$u\varepsilon \frac{r}{\|r\|} \in G(x_u, u) + G'_x(x_0, 0)d - K. \tag{21}$$

With these choices we define $y_u = x_u + \beta d$, where $\beta := \|r\|/(u\varepsilon + \|r\|) < 1$.

Property (ii) follows immediately from (20) and the inequality

$$\|y_u - x_u\| = \beta\|d\| \le \frac{\|r\|}{u\varepsilon}(\alpha u + \|x_u - x_0\|).$$

To check property (i) we observe that $\|d\| \le \alpha u + \varphi(u)$ and then, using (18),

$$\|y_u - x_0\| \le \alpha u + 2\varphi(u) \le 1.$$

Then we can apply the mean value theorem to find $\xi \in]x_u, y_u[$ with

$$\begin{aligned} \|G(y_u, u) - G(x_u, u) - \beta G'_x(x_0, 0)d\| &\le \|G'_x(\xi, u) - G'_x(x_0, 0)\|\, \|d\|\, \beta \\ &\le M(u + \|\xi - x_0\|)\|d\|\frac{\|r\|}{u\varepsilon} \\ &\le \frac{2M}{u\varepsilon}[\alpha u + \varphi(u)]^2\|r\| \\ &\le \frac{1}{2}d(G(x_u, u), K), \end{aligned} \tag{22}$$

where we have used the bound $u + \|\xi - x_0\| \le 2[\alpha u + \varphi(u)]$, (18), and (20).

Now, from (21) we get

$$G(x_u, u) + \beta G'_x(x_0, 0)d \in (1-\beta)G(x_u, u) + \beta u\varepsilon \frac{r}{\|r\|} + \beta K,$$

since $1 - \beta = \beta u\varepsilon/\|r\|$, we deduce

$$[G(x_u, u) + \beta G'_x(x_0, 0)d \in (1-\beta)(G(x_u, u) + r) + \beta K \subset K,$$

which combined with (22) yields (i). □

Proof of Theorem B.2. Let $\varphi(u) := e^{4m/\varepsilon}(\alpha u + \|x_u - x_0\|)$ and suppose that $\gamma < \frac{1}{4}e^{-4m/\varepsilon}\sqrt{\frac{\varepsilon}{M}}$ so that Lemma B.4 can be used to find u_0.

Starting with $y_u^0 := x_u$ we shall construct recursively a sequence y_u^k such that for all $u \in (0, u_0]$ one has

$$\text{(i)} \quad d(G(y_u^k, u), K) \le \frac{1}{2} d(G(y_u^{k-1}, u), K),$$

$$\text{(ii)} \quad \|y_u^k - y_u^{k-1}\| \le \frac{2}{u\varepsilon}(\alpha u + \|y_u^{k-1} - x_0\|) d(G(y_u^{k-1}, u), K).$$

To prove the existence of such a sequence it suffices to check inductively that

$$\text{(iii)} \quad \|y_u^k - x_0\| \le \varphi(u) \quad \forall\, u \in (0, u_0],$$

so that Lemma B.4 can be used to find the next term y_u^{k+1}. Since (iii) obviously holds for $k = 0$, we only need to prove the inductive step. Suppose $y_0, y_1, \ldots, y_k$ are such that (i) and (ii) hold; then for every $u \in (0, u_0]$ we have

$$\begin{aligned} \|y_u^k - y_u^{k-1}\| &\le \frac{2}{u\varepsilon}(\alpha u + \|y_u^{k-1} - x_0\|) \frac{d(G(x_u, u), K)}{2^{k-1}} \\ &\le \frac{2m}{\varepsilon 2^{k-1}}(\alpha u + \|y_u^{k-1} - x_0\|), \end{aligned} \tag{23}$$

so that letting $a_k := \alpha u + \|y_u^k - x_0\|$ we get

$$a_k \le a_{k-1} + \|y_u^k - y_u^{k-1}\| \le \left(1 + \frac{2m}{\varepsilon 2^{k-1}}\right) a_{k-1}.$$

It follows that

$$\ln a_k \le \ln a_{k-1} + \ln\left(1 + \frac{2m}{\varepsilon 2^{k-1}}\right) \le \ln a_{k-1} + \frac{2m}{\varepsilon 2^{k-1}}$$

and then recursively

$$\ln a_k \le \ln a_0 + \frac{2m}{\varepsilon}\left(\frac{1}{2^0} + \frac{1}{2^1} + \cdots + \frac{1}{2^{k-1}}\right) \le \ln a_0 + \frac{4m}{\varepsilon},$$

from which we obtain the desired conclusion (iii) as

$$\|y_u^k - x_0\| \le a_k \le a_0 e^{4m/\varepsilon} = \varphi(u).$$

The existence of the sequence (y_k) being established, we may use the previous bound $a_k \le \varphi(u)$ and (23) to obtain

$$\|y_u^k - y_u^{k-1}\| \le \frac{2\varphi(u) d(G(x_u, u), K)}{u\varepsilon 2^{k-1}}, \tag{24}$$

which shows that $(y_u^k)_{k\in\mathbb{N}}$ is a Cauchy sequence for each $u \in (0, u_0]$.

Let $y_u := \lim_{k\uparrow\infty} y_u^k$. From (i) we deduce that $G(y_u, u) \in K$, while (24) implies

$$\|y_u - x_u\| \le \frac{4\varphi(u)}{u\varepsilon} d(G(x_u, u), K),$$

proving the theorem with $c := \frac{4}{\varepsilon} e^{4m/\varepsilon}$. □

A careful analysis of the previous proof shows that the result is still valid if G is supposed of class C^1 (or merely strictly differentiable at $(x_0, 0)$) provided we restrict to the case of trajectories $x_u = x_0 + O(u)$. More precisely we have the following theorem.

THEOREM B.5. *Let $G : X \times \mathbb{R} \to Y$ be strictly differentiable at $(x_0, 0)$ and $K \subset Y$ be a closed convex set. Suppose that $G(x_0, 0) \in K$ and (DCQ) holds. Then for each trajectory $x_u = x_0 + O(u)$ there exist constants $c \ge 0$, $u_0 > 0$ and a second trajectory y_u such that*

$$G(y_u, u) \in K,$$
$$\|y_u - x_u\| \le c\, d(G(x_u, u), K),$$

for all $u \in [0, u_0]$.

Proof. It is clear that the result will follow from Theorem B.2, which is applicable since $x_u = x_0 + O(u)$ implies $x_u = x_0 + o(\sqrt{u})$ and $d(G(x_u, u), K) = O(u)$.

However, we must check that Theorem B.2 remains valid under the weaker C^1 assumption on G and the stronger $x_u = x_0 + O(u)$ condition. To this end all we need is to modify Lemma B.4. More specifically, it suffices to adjust the arguments leading to the bounds (19) and (22), which is easily accomplished by fixing $\ell \in \mathbb{R}$ and $u_0 \in (0, \overline{u}]$ such that $\varphi(u) \le \ell u$ for all $u \in [0, u_0]$ and then reducing u_0 so that

$$\|G(y, v) - G(x, u) - G'(x_0, 0)(y - x, v - u)\| \le \frac{\varepsilon}{4(\alpha + \ell)} (\|y - x\| + |v - u|)$$

for each $u, v \in [0, u_0]$ and every $x, y \in B(x_0, (\alpha + 2\ell)u_0)$. □

As a corollary of the preceding result we obtain the following *directional* version of Robinson–Ursescu's regularity theorem for convex multifunctions.

THEOREM B.6. *Let $M : X \longrightarrow 2^Y$ be a multifunction with closed convex graph. Let $y_0 \in M(x_0)$ and let y_u be a C^1 trajectory with $y(0) = y_0$ and*

$$(RU) \qquad 0 \in \operatorname{int}[M(X) - y(0) - (0, \infty)y'(0)].$$

Then for each trajectory $x_u = x_0 + O(u)$ one has

$$d(x_u, M^{-1}(y_u)) \le c\, d(y_u, M(x_u))$$

for a given constant c and all $u \ge 0$ sufficiently small.

Proof. The result follows as a direct application of Theorem B.5 to the function $G(x, u) = (x, y_u)$ and the closed convex set $K = \operatorname{graph}(M)$. □

APPLICATION. As a particular case of the previous result let us consider $y_0 \in M(x_0)$ and suppose that $d \in Y$ is such that

$$[0 \in \operatorname{int}[M(X) - y_0 - (0, \infty)d].$$

Then, for each trajectory $x_u = x_0 + O(u)$ there exists $\tilde{x}_u$ such that

$$y_0 + ud \in M(\tilde{x}_u),$$
$$\|\tilde{x}_u - x_u\| \le c\, d(y_u, M(x_u)).$$

In particular, letting $x_u \equiv x_0$ we obtain the existence of a trajectory $\tilde{x}_u = x_0 + O(u)$ with $y_0 + ud \in M(\tilde{x}_u)$.

REFERENCES

[1] H. ATTOUCH AND H. BRÉZIS, *Duality for the sum of convex functions in general Banach spaces*, in Aspects of Mathematics and Its Applications, J. A. Barroso, ed., Elsevier, Amsterdam, 1986, pp. 125–133.

[2] A. AUSLENDER AND R. COMINETTI, *First and second order sensitivity analysis of nonlinear programs under directional constraint qualification conditions*, Optimization, 21 (1990), pp. 351–363.

[3] L. BARBET, *Etude de sensibilité différentielle dans un problème d'optimisation paramétré avec contraintes en dimension infinie*, Thesis, Université de Poitiers, 1992.

[4] A. BEN-TAL AND J. ZOWE, *A unified theory of first and second order conditions for extremum problems in topological vector spaces*, Math. Prog. Study, 19 (1982), pp. 39–76.

[5] J. F. BONNANS, *Directional derivatives of optimal solutions in smooth nonlinear programming*, J. Optim. Theory Appl., 73 (1992), pp. 27–45.

[6] J. F. BONNANS AND E. CASAS, *Optimal control of semilinear multistate systems with state constraints*, SIAM J. Control Optim., 27 (1989), pp. 446–455.

[7] J. F. BONNANS, A. D. IOFFE, AND A. SHAPIRO, *Expansion of exact and approximate solutions in nonlinear programming*, in Proc. French-German Conference in Optimization, W. Oettli and D. Pallaschke, eds., Lecture Notes in Economics and Math. Systems, Springer-Verlag, New York, 1992, pp. 103–117.

[8] J. F. BONNANS AND A. SHAPIRO, *Sensitivity analysis of parametrized programs under cone constraints*, SIAM J. Control Optim., 30 (1992), pp. 1409–1422.

[9] R. COMINETTI, *Metric regularity, tangent sets and second order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.

[10] A. DONTCHEV AND T. ZOLEZZI, *Well-posed Optimization Problems*, Springer-Verlag, Berlin, 1993.

[11] S. FITZPATRICK AND R. PHELPS, *Differentiability of metric projections in Hilbert space*, Trans. Amer. Math. Soc., 270 (1982), pp. 483–501.

[12] J. GAUVIN AND R. JANIN, *Directional behaviour of optimal solutions in nonlinear mathematical programming*, Math. Oper. Res., 13 (1988), pp. 629–649.

[13] J. GAUVIN, *Théorie de la programmation mathématique non convexe*, Centre de Recherches Mathématiques, Montreal, 1992. (English translation, *Theory of Nonconvex Programming*, Centre de Recherches Mathématiques, 1993.)

[14] B. GOLLAN, *On the marginal function in nonlinear programming*, Math. Oper. Res., 9 (1984), pp. 208–221.

[15] E. G. GOL'STEIN, *Theory of Convex Programming*, Moscow (1970); English transl. Math. Monographs 36, Amer. Math. Soc., Providence, RI, 1972.

[16] A. HARAUX, *How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities*, J. Math. Soc. Japan, 29 (1977), pp. 615–631.

[17] H. KAWASAKI, *An envelope-like effect of inifinitely many inequality constraints on second order necessary conditions for minimization problems*, Math. Prog., 41 (1988), pp. 73–96.

[18] E. S. LEVITIN, *Perturbation Theory in Mathematical Programming and Its Applications*, J. Wiley, New York, 1994.

[19] K. MALANOWSKI, *Second-order conditions and constraint qualifications in stability and sensitivity analysis of solutions to optimization problems in Hilbert spaces*, Appl. Math. Optim., 25 (1992), pp. 51–79.

[20] O. MANGASARIAN AND S. FROMOVITZ, *The Fritz-John necessary optimality condition in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 7 (1967), 37–47.

[21] H. MAURER AND J. ZOWE, *First and second order necessary and sufficient optimality conditions for infinite dimensional programming problems*, Math. Prog., 16 (1979), pp. 98–110.

[22] F. MIGNOT, *Contrôle dans les inéquations variationnelles elliptiques*, J. Funct. Anal., 22 (1976), 25–39.

[23] S. M. ROBINSON, *Stability theorems for systems of inequalities. Part II: differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.

[24] ———, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.

[25] R. T. ROCKAFELLAR, *Conjugate Duality and Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math., SIAM, Philadelphia, PA, 1974.

[26] A. SHAPIRO, *Sensitivity analysis of nonlinear programs and differentiability properties of metric projections*, SIAM J. Control Optim., 26 (1988), pp. 628–645.

[27] E. ZARANTONELLO, *Projections on convex sets in Hilbert space and spectral theory*, in Contrib. Nonlinear Funct. Anal., 27, Math. Res. Center, Univ. of Wisconsin, 1971, pp. 237–424.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1172–1189, July 1996

004

PERTURBED OPTIMIZATION IN BANACH SPACES II: A THEORY BASED ON A STRONG DIRECTIONAL CONSTRAINT QUALIFICATION*

J. FRÉDÉRIC BONNANS† AND ROBERTO COMINETTI‡

Abstract. We study the sensitivity of the optimal value and optimal solutions of perturbed optimization problems in two cases. The first one is when multipliers exist but only the weak (and not the strong) second-order sufficient optimality condition is satisfied. The second case is when no Lagrange multipliers exist. To deal with these pathological cases, we are led to introduce a directional constraint qualification stronger than in part I of this paper, which reduces to the latter in the important case of equality-inequality constrained problems. We give sharp upper estimates of the cost based on paths varying as the square root of the perturbation parameter and, under a *no-gap* condition, obtain the first term of the expansion for the cost. When multipliers exist we study the expansion of approximate solutions as well. We show in the appendix that the strong directional constraint qualification is satisfied for a large class of problems, including regular problems in the sense of Robinson.

Key words. sensitivity analysis, marginal function, square root expansion, approximate solutions, directional constraint qualification

AMS subject classifications. 46N10, 47H19, 49K27, 49K40, 58C15, 90C31

1. Introduction.

This paper is the second in a trilogy (see [4, 5]) devoted to the analysis of parametric optimization problems of the form

$$(P_u) \qquad \min_x \{f(x,u) : G(x,u) \in K\}$$

with X and Y Banach spaces, K a closed convex subset of Y, and $f(x,u)$, $G(x,u)$ mappings of class $\mathcal{C}^2$ from $X \times \mathbb{R}$ into $\mathbb{R}$ and Y, respectively. We denote the feasible set, value function, and set of solutions of (P_u) as

$$\begin{aligned} F(u) &:= \{x \in X \ : \ G(x,u) \in K\}, \\ v(u) &:= \inf\{f(x,u) \ : \ x \in F(u)\}, \\ S(u) &:= \{x \in F(u) \ : \ f(x,u) = v(u)\}, \end{aligned}$$

respectively. Similarly $v(P)$, $F(P)$, $S(P)$ will respectively denote the optimal value, feasible set, and solution set of an optimization problem (P).

Our aim is to study the expansion of $v(u)$ and possibly $S(u)$ in the vicinity of a local solution x_0 of (P_0). Such sensitivity analysis usually relies (among other assumptions) upon stability properties of the feasible set $F(u)$ that follow from so-called *constraint qualification conditions*. In part I of this work (see [4]) our study was based on the following generalization of Gollan's constraint qualification (see [1, 10]):

$$(DCQ) \qquad 0 \in \operatorname{int}\left[G(x_0,0) + G'(x_0,0)X \times (0,\infty) - K\right],$$

which is a *directional* version of Robinson's condition [14]

$$(CQ) \qquad 0 \in \operatorname{int}\left[G(x_0,0) + G'_x(x_0,0)X - K\right].$$

Under (DCQ) we obtained the following upper estimate of the optimal value:

$$(1.1) \qquad v'_+(0) \le v(L),$$

*Received by the editors May 9, 1994; accepted for publication (in revised form) February 15, 1995.

†INRIA-Rocquencourt, B.P. 105, 78153 Rocquencourt, France.

‡Universidad de Chile, Casilla 170/3 Correo 3, Santiago, Chile. The research of this author was partially supported by Fundación Andes FONDECYT contract 1940564.

where $v'_+(0)$ and $v'_-(0)$ denote the upper and lower Dini derivatives of the value function:

$$v'_+(0) := \limsup_{u \downarrow 0} \frac{v(u) - v(0)}{u},$$

$$v'_-(0) := \liminf_{u \downarrow 0} \frac{v(u) - v(0)}{u},$$

and (L) is the problem with linearized data:

$$(L) \qquad \min_d \{f'(x_0, 0)(d, 1) : \ G'(x_0, 0)(d, 1) \in T_K(G(x_0, 0))\}.$$

Using duality theory we proved that $v(D) = v(L) < \infty$, where (D) is the problem

$$(D) \qquad \max\{\mathcal{L}'_u(x_0, \lambda, 0) : \lambda \in \Lambda_0\},$$

with $\mathcal{L}$ the Lagrangian and Λ_0 the set of multipliers associated with x_0, that is to say, denoting by $N_K(y)$ the cone of outward normals at a point $y \in K$,

$$\mathcal{L}(x, \lambda, u) := f(x, u) + \langle \lambda, G(x, u) \rangle,$$

$$\Lambda_0 := \{\lambda \in Y^* : \lambda \in N_K(G(x_0, 0)); \ \mathcal{L}'_x(x_0, \lambda, 0) = 0\}.$$

Define a *path* as a mapping $u \to x_u$ from $\mathbb{R}_+$ to X, with $x_u \to x_0$ when $u \downarrow 0$. The path is said to be feasible if $G(x_u, u) \in K$ for u small enough. Under a strong second-order condition on the Lagrangian it can be shown [4] that any $o(u^2)$-optimal path x_u, i.e., a feasible path x_u such that $f(x_u, u) \leq v(u) + o(u^2)$, satisfies $x_u = x_0 + O(u)$. In this case $v'(0)$ exists, being equal to $v(L)$, and some estimates for the second-order variation of $v(u)$ can be obtained. In fact, under suitable conditions we proved that

$$(1.2) \qquad v(u) = v(0) + u \ v(L) + \frac{1}{2} u^2 \ v(\tilde{Q}) + o(u^2),$$

where $(\tilde{Q})$ is a subproblem involving the expansion of orders 1 and 2 of the data at $(x_0, 0)$. A remarkable property in this case is that every weak limit of $(x_u - x_0)/u$, with x_u an $o(u^2)$-optimal path, belongs to $S(\tilde{Q})$.

The available perturbation theory for nonlinear programming shows that this is not the end of the story. Under the directional qualification hypothesis of Gollan [10] and the weak second-order sufficient condition, it appears (see [9] by Gauvin and Janin) that $v'(0)$ exists but may be strictly less than $v(L)$. In that case, a path of $O(u)$-optimal solutions satisfies only $x_u = x_0 + O(\sqrt{u})$. One can still formulate (see Bonnans, Ioffe, and Shapiro [6]) a subproblem (M) such that $v'(0) = v(M)$ and $S(M)$ coincides with the limit points of $(x_u - x_0)/\sqrt{u}$ where x_u ranges over the set of all possible $o(u)$-optimal paths. For this it is necessary to assume the existence of at least one multiplier. A similar theory for the case when no multiplier exists was developed in [3] by Bonnans; here the variation of the cost as well that of the solutions is of order $O(\sqrt{u})$.

The aim of this paper is to extend these two theories to the Banach space setting. Our main results are Theorem 3.9 and Theorem 4.6.

The first one, valid under the weak second-order sufficient condition, provides a first-order expansion of the form

$$v(u) = v(0) + uv(\tilde{D}) + o(u),$$

where $(\tilde{D})$ is a problem involving the expansion of orders 1 and 2 of the data. Moreover, it shows that every weak limit point of $(x_u - x_0)/\sqrt{u}$, with x_u an $o(u)$-optimal path, solves $(\tilde{D})$.

The second one is concerned with problems where no Lagrange multipliers exist. In this case we obtain a square root expansion of the form

$$v(u) = v(0) + \sqrt{u}v(\hat{D}) + o(\sqrt{u}),$$

where $(\hat{D})$ is another linear-quadratic approximating problem.

To prove these results we need a constraint qualification that is still directional but, apparently, stronger than (*DCQ*). Specifically, in addition to (*DCQ*) we need a restorability property that, roughly speaking, asserts that to certain almost feasible *square root* paths (i.e., paths satisfying $x_u = x_0 + O(\sqrt{u})$), one can associate a sufficiently close feasible path. In the case of nonlinear programming, that stronger hypothesis (*SDCQ*) reduces to the condition of Gollan (see [1, 10]) used in [9, 3, 6], so we recover the main results of these three references. Let us mention that square root paths have already been used for sensitivity analysis in a Banach space setting (see [2, 11, 12]). However, our qualification condition is weaker than those in these references.

As in part I of this work, in our extension to the Banach space setting, an additional difficulty related to the possible curvature of the convex K appears. To be more precise, let us recall the definition of first- and second-order tangent sets:

$$T_K(y) := \{h \in Y : \text{ there exists } o(t) \text{ such that } y + th + o(t) \in K\},$$

$$T_K^2(y, h) := \left\{k \in Y : \text{ there exists } o(t^2) \text{ such that } y + th + \frac{1}{2}t^2k + o(t^2) \in K\right\}.$$

The fact that in general 0 does not belong to the set $T_K^2(y, h)$ may cause a *gap* between the upper and lower estimates for the cost. Some cases when the curvature makes no contribution to the second-order variation of the cost were analyzed in part I, yielding the expansion (1.2) under a condition of generalized polyhedricity. The results in this paper are obtained under similar assumptions.

The paper is organized as follows. In §2 we describe the strong directional constraint qualification (*SDCQ*). Then in §3 we develop a perturbation theory assuming the set of multipliers Λ_0 to be nonempty, whereas §4 deals with the case when Λ_0 is empty. In both cases we obtain sharp upper estimates as well as some lower estimates of the cost and, under a *no-gap* condition, obtain the first term in the expansion of the cost. Finally in the appendix we discuss sufficient conditions for the strong directional constraint qualification (*SDCQ*).

2. The strong directional qualification condition. Our upper estimates are based on paths that vary as the square root of the perturbation parameter. Specifically, we consider paths satisfying, for given d, w in X, the two conditions

$$x_u = x_0 + \sqrt{u}d + uw + o(u), \tag{2.3}$$

$$\operatorname{dist}(G(x_u, u), K) = o(u). \tag{2.4}$$

Note that we can express (2.4) using the concept of a second-order tangent set. Namely, if x_u satisfies (2.3), then the expansion

$$G(x_u, u) = G(x_0, 0) + \sqrt{u}G_x'(x_0, 0)d + u\left[G'(x_0, 0)(w, 1) + \frac{1}{2}G_x''(x_0, 0)dd\right] + o(u)$$

shows that (2.4) is equivalent to

$$\Psi_G(w, d) \in T_2^K(d), \tag{2.5}$$

where we have set

$$T_2^K(d) := \frac{1}{2}T_K^2(G(x_0,0), G'_x(x_0,0)d),$$

$$\Psi_G(w,d) := G'(x_0,0)(w,1) + \frac{1}{2}G''_x(x_0,0)dd,$$

$$\Psi_f(w,d) := f'(x_0,0)(w,1) + \frac{1}{2}f''_x(x_0,0)dd.$$

Remark. The set $T_2^K(d)$ should not be confused with the set

$$T_K^2(d) := T_K^2(G(x_0,0), G'(x_0,0)(d,1))$$

defined in part I of this paper and which will not be used here.

DEFINITION 1. *We say that x_0 is* restorable *(with respect to G and K) if, given a path x_u satisfying* (2.3) *and* (2.4), *for $\gamma < 1$ close to 1 one can find $w_\gamma \in X$ with $w_\gamma \to w$ and feasible paths of the form*

$$x_u^\gamma = x_0 + \gamma\sqrt{u}d + uw_\gamma + o(u). \tag{2.6}$$

We say that the strong directional constraint qualification *(SDCQ) holds at x_0 if x_0 is restorable and the weak directional constraint qualification (DCQ) holds.*

We discuss some sufficient conditions for *(SDCQ)* in the appendix at the end of this paper. We show in particular that for equality-inequality constrained problems (i.e., when $K = \{0\} \times K_2$ with K_2 a closed convex cone with nonempty interior), property *(SDCQ)* is equivalent to *(DCQ)*. In fact, it may be that the restorability property is always a consequence of *(DCQ)*, but we do not have a proof nor a counterexample for this.

Before proceeding with the sensitivity analysis we summarize in the next lemma four general properties that will be of constant use throughout the paper. Here $\sigma(\lambda, T_2^K(d)) := \sup\{\langle\lambda, k\rangle : k \in T_2^K(d)\}$ denotes the support function of $T_2^K(d)$.

LEMMA 2.1. *For every $d \in X$ we have the following.*

(P1) $T_2^K(d) + T_K(G(x_0,0)) - \mathbb{R}_+G'_x(x_0,0)d \subset T_2^K(d)$.

(P2) If (DCQ) holds, then $0 \in \operatorname{int}[T_K(G(x_0,0)) - G'(x_0,0)X \times \{1\}]$.

(P3) $T_2^K(\gamma d) = \gamma^2 T_2^K(d)$ *for all* $\gamma > 0$.

(P4) If $T_2^K(d) \neq \phi$, *then the following are equivalent:*

(a) $\sigma(\lambda, T_2^K(d)) \leq 0$.

(b) $\sigma(\lambda, T_2^K(d))$ *is finite.*

(c) $\lambda \in N_K(G(x_0,0))$ *and* $\langle\lambda, G'_x(x_0,0)d\rangle = 0$.

Proof. Properties *(P1)* and *(P2)* are straightforward consequences of [8, Prop. 3.1] and [4, Lem. B.3], respectively, while *(P3)* is an easy exercise.

Let us prove *(P4)*. Since $T_2^K(d) \neq \phi$, the implication (a) $\Rightarrow$ (b) is straightforward. Also, the nonemptyness of $T_2^K(d)$ implies $G'_x(x_0,0)d \in T_K(G(x_0,0))$ and then (b) $\Rightarrow$ (c) follows from property *(P1)*. To prove (c) $\Rightarrow$ (a) let us pick $y \in T_2^K(d)$ and choose $y_t \to y$ with $z_t := G(x_0,0) + tG'_x(x_0,0)d + t^2y_t \in K$. Using (c) we deduce

$$0 \geq \langle\lambda, z_t - G(x_0,0)\rangle = \langle\lambda, tG'_x(x_0,0)d + t^2y_t\rangle = t^2\langle\lambda, y_t\rangle,$$

so that $\langle\lambda, y\rangle = \lim\langle\lambda, y_t\rangle \leq 0$, proving (a). □

3. Perturbation analysis assuming the existence of multipliers. In this section we study the case when $\Lambda_0 \neq \phi$. First we give an upper estimate of $v'_+(0)$, which we can express as a supremum of a certain function over Λ_0. We then rely on second-order conditions to obtain lower estimates for $v'_-(0)$ and to investigate the coincidence of both estimates.

3.1. Sharp first-order upper estimates of the cost. Let C_0 denote the cone of critical directions at x_0, i.e.,

$$C_0 := \{d \in X : \ f'_x(x_0, 0)d \leq 0; \ G'_x(x_0, 0)d \in T_K(G(x_0, 0))\}.$$

When $\Lambda_0 \neq \phi$ one has in fact $f'_x(x_0, 0)d = 0$ for all $d \in C_0$. To a path satisfying (2.3) and (2.4) is associated the constraint (2.5), whereas $\Psi_f(w, d)$ is the first term of the expansion of the cost. This leads to the problem

$$(L^d) \qquad \inf_{w \in X} \{\Psi_f(w, d) : \Psi_G(w, d) \in T_2^K(d)\},$$

which admits the dual

$$(D^d) \qquad \sup_{\lambda \in \Lambda_0} \left\{ \mathcal{L}'_u(x_0, \lambda, 0) + \frac{1}{2}\mathcal{L}''_x(x_0, \lambda, 0)dd - \sigma(\lambda, T_2^K(d)) \right\}.$$

We also consider the problem

$$(\tilde{L}) \qquad \inf_d \{v(L^d) : d \in C_0\},$$

which plays a role in the following upper estimate of the cost.

THEOREM 3.1. *Assume Λ_0 to be nonempty and (SDCQ). Then*

$$v'_+(0) \leq v(\tilde{L}) = \inf_{d \in C_0} v(D^d) \leq v(L) < \infty.$$

In particular, if $v(\tilde{L})$ is finite, then

$$v(u) \leq v(0) + uv(\tilde{L}) + o(u).$$

The theorem is an immediate consequence of the next two lemmas. The first one gives the primal upper estimate of $v'_+(0)$.

LEMMA 3.2. *Assuming (SDCQ) we have*

$$v'_+(0) \leq v(\tilde{L}) \leq v(L) < \infty.$$

Proof. Let $d \in C_0$ and take a feasible $w \in F(L^d)$. Using the restorability property we may find $w_\gamma \to w$ and feasible paths of the form

$$x_u^\gamma = x_0 + \gamma\sqrt{u}d + uw_\gamma + o(u).$$

Expanding $f(x_u^\gamma, u)$ and using the fact that d is critical, it follows that

$$v(u) \leq f(x_u^\gamma, u) \leq f(x_0, 0) + u\Psi_f(w_\gamma, \gamma d) + o(u)$$

so that $v'_+(0) \leq \Psi_f(w_\gamma, \gamma d)$. Passing to the limit when $\gamma \uparrow 1$ we deduce that $v'_+(0) \leq \Psi_f(w, d)$, and taking the infimum over $w \in F(L^d)$ and $d \in C_0$ we get

$$v'_+(0) \leq v(\tilde{L}).$$

We conclude by noting that for $d = 0$ problem (L^d) reduces to problem (L) and that $v(L) < \infty$ by [4, Prop. 2.2]. □

Let us prove next the dual expression for $v(\tilde{L})$.

LEMMA 3.3. *Assume Λ_0 to be nonempty and (SDCQ). For each $d \in C_0$ we have the following.*

(i) $v(D^d) \le v(L^d)$.

(ii) *If* (L^d) *is feasible, then for all* $\gamma \in (0,1)$, $v(D^{\gamma d}) = v(L^{\gamma d}) \in \mathbb{R}$ *and* $S(D^{\gamma d})$ *is nonempty and bounded.*

(iii) *If* (L^d) *is infeasible, then* $v(D^{\gamma d}) = \infty$ *for all* $\gamma > 1$.

(iv) $\limsup_{\gamma \uparrow 1} v(D^{\gamma d}) \le v(D^d)$.

As a consequence we obtain

$$v(\tilde{L}) = \inf_{d \in C_0} v(D^d). \tag{3.7}$$

Proof. Let us begin by showing that (3.7) is a consequence of (i)–(iv). The inequality $v(\tilde{L}) \ge \inf_{d \in C_0} v(D^d)$ is obvious from (i). To show the converse inequality it suffices to check that $v(D^d) \ge v(\tilde{L})$ for those $d \in C_0$ such that $v(D^d) < \infty$. By (iii) this implies $(L^{\gamma d})$ is feasible for each $\gamma \in (0,1)$, and then (ii) gives $v(D^{\gamma d}) = v(L^{\gamma d}) \ge v(\tilde{L})$ for all $\gamma \in (0,1)$. We conclude by letting $\gamma \uparrow 1$ and using (iv).

We now prove properties (i)–(iv).

(i) It suffices to show that if w and λ are feasible for (L^d) and (D^d), respectively, then the dual cost is not greater than the primal one. From the primal constraint it follows that

$$\sigma(\lambda, T_2^K(d)) \ge \langle \lambda, \Psi_G(w,d) \rangle,$$

which implies

$$\begin{aligned} \Psi_f(w,d) &\ge \Psi_f(w,d) + \langle \lambda, \Psi_G(w,d) \rangle - \sigma(\lambda, T_2^K(d)) \\ &= \mathcal{L}_u'(x_0, \lambda, 0) + \frac{1}{2}\mathcal{L}_x''(x_0, \lambda, 0)dd - \sigma(\lambda, T_2^K(d)), \end{aligned}$$

as was to be proved.

(ii) We first claim that $v(L^d)$ and $v(D^d)$ are finite and equal with $S(D^d)$ nonempty and bounded, whenever

$$Y = \mathbb{R}_+ \left[T_2^K(d) - G'(x_0,0)X \times \{1\} - \frac{1}{2}G_x''(x_0,0)dd \right]. \tag{3.8}$$

To motivate this relation, let us consider the family of problems obtained by perturbing additively the constraint of (L^d), that is, $\min_{w \in X} \varphi(w,y)$ with

$$\varphi(w,y) := \begin{cases} \Psi_f(w,d) & \text{if } \Psi_G(w,d) + y \in T_2^K(d), \\ \infty & \text{otherwise.} \end{cases}$$

Property (3.8) amounts to $Y = \mathbb{R}_+ \cup_w \operatorname{dom} \varphi(w,\cdot)$, so we may apply the convex duality theorem of part I [4, Thm. A.2] to deduce

$$v(L^d) = \inf_{w \in X} \varphi(w,0) = -\min_{\lambda \in Y^*} \varphi^*(0,\lambda) \tag{3.9}$$

as well as the boundedness and nonemptiness of the set of dual solutions. Now we compute

$$\begin{aligned} \varphi^*(0,\lambda) &= \sup_{w \in X, y \in Y} \{\langle \lambda, y \rangle - \Psi_f(w,d) : \Psi_G(w,d) + y \in T_2^K(d)\} \\ &= \sup_{w \in X} \left\{ \sigma(\lambda, T_2^K(d)) - \mathcal{L}'(x_0, \lambda, 0)(w,1) - \frac{1}{2}\mathcal{L}_x''(x_0, \lambda, 0)dd \right\}. \end{aligned}$$

Maximizing over w we deduce that $\varphi^*(0,\lambda) = \infty$ if $\mathcal{L}_x'(x_0, \lambda, 0) \ne 0$, and then using $(P4)$ we get

$$\varphi^*(0,\lambda) = \begin{cases} \sigma(\lambda, T_2^K(d)) - \mathcal{L}_u'(x_0, \lambda, 0) - \frac{1}{2}\mathcal{L}_x''(x_0, \lambda, 0)dd & \text{if } \lambda \in \Lambda_0, \\ \infty & \text{otherwise.} \end{cases}$$

This and (3.9) imply the equality $v(L^d) = v(D^d)$. Moreover, since the dual is attained, property $(P4)$ shows that this common value is finite. This proves our claim.

In view of the previous discussion, to prove (ii) it suffices to check that for each $\gamma \in (0, 1)$ property (3.8) holds with d replaced by $d_\gamma := \gamma d$. To see this let us choose a feasible $w \in F(L^d)$, that is,

$$G'(x_0, 0)(w, 1) + \frac{1}{2}G''_x(x_0, 0)dd \in T_2^K(d).$$

Multiplying by γ^2 and using $(P3)$ we deduce that

$$G'(x_0, 0)(\gamma^2 w, \gamma^2) + \frac{1}{2}G''_x(x_0, 0)d_\gamma d_\gamma \in T_2^K(d_\gamma).$$

From this and $(P1)$ we get

$$T_K(G(x_0, 0)) - G'(x_0, 0)X \times \{1 - \gamma^2\} \subset T_2^K(d_\gamma) - G'(x_0, 0)X \times \{1\} - \frac{1}{2}G''_x(x_0, 0)d_\gamma d_\gamma,$$

which multiplied by $\mathbb{R}_+$ and using $(P2)$ yields (3.8) for d_γ as required.

(iii) Let $\gamma > 1$ and set $d_\gamma := \gamma d$ as before. If $T_2^K(d)$ is empty, by $(P3)$ so is $T_2^K(d_\gamma)$ and then $\sigma(\lambda, T_2^K(d_\gamma)) = -\infty$, hence $v(D^{\gamma d}) = \infty$.

Let us then assume $T_2^K(d)$ to be nonempty. Since (L^d) is infeasible, the convex set $T_2^K(d) - G'(x_0, 0)X \times \{1\}$ does not contain $\frac{1}{2}G''_x(x_0, 0)dd$. But $(P1)$ and $(P2)$ show that this convex set has a nonempty interior, so that the Hahn–Banach theorem gives a nonzero $\mu \in Y^*$ that separates the set and the point, that is,

$$\left\langle \mu, G'(x_0, 0)(w, 1) + \frac{1}{2}G''_x(x_0, 0)dd \right\rangle \geq \sigma(\mu, T_2^K(d)) \quad \text{for all } w \in X. \tag{3.10}$$

This inequality and property $(P4)$ imply $\mu \in N_K(G(x_0, 0))$. Also, taking the infimum over $w \in X$ we deduce $\mu \circ G'_x(x_0, 0) = 0$ (that is to say, μ is a *singular multiplier*, as defined in the next section) so that for each $\lambda \in \Lambda_0$ and $t > 0$ we have $\lambda + t\mu \in \Lambda_0$. Since $S(D)$ is bounded (see [4, Prop. 3.1]), it follows that

$$\langle \mu, G'_u(x_0, 0)\rangle < 0.$$

With these observations property (3.10) reduces to

$$\Xi(\mu, d) := \left\langle \mu, G'_u(x_0, 0) + \frac{1}{2}G''_x(x_0, 0)dd \right\rangle - \sigma(\mu, T_2^K(d)) \geq 0,$$

which multiplied by γ^2 and using $(P3)$ gives

$$\Xi(\mu, d_\gamma) \geq (1 - \gamma^2)\langle \mu, G'_u(x_0, 0)\rangle > 0. \tag{3.11}$$

Let us fix $\lambda \in \Lambda_0$. Since $\Xi(\cdot, d_\gamma)$ is positively homogeneous and concave, and since $\lambda + t\mu \in \Lambda_0$, it follows that

$$\begin{aligned} v(D^{\gamma d}) &\geq f'_u(x_0, 0) + \frac{1}{2}f''_x(x_0, 0)d_\gamma d_\gamma + \Xi(\lambda + t\mu, d_\gamma) \\ &\geq f'_u(x_0, 0) + \frac{1}{2}f''_x(x_0, 0)d_\gamma d_\gamma + \Xi(\lambda, d_\gamma) + t\Xi(\mu, d_\gamma). \end{aligned}$$

To conclude we observe that $(P4)$ implies the finiteness of $\Xi(\lambda, d_\gamma)$, so that letting $t \uparrow \infty$ and using (3.11) we get $v(D^{\gamma d}) = \infty$.

(iv) Using $(P3)$ we obtain

$$\begin{aligned} v(D^{\gamma d}) &= \sup_{\lambda \in \Lambda_0} \left\{ \mathcal{L}'_u(x_0, \lambda, 0) + \frac{\gamma^2}{2} \mathcal{L}''_x(x_0, \lambda, 0)dd - \gamma^2 \sigma(\lambda, T_2^K(d)) \right\} \\ &\leq \sup_{\lambda \in \Lambda_0} \{(1-\gamma^2)\mathcal{L}'_u(x_0, \lambda, 0) + \gamma^2 v(D^d)\} \\ &= (1-\gamma^2)v(L) + \gamma^2 v(D^d). \end{aligned}$$

As $v(L) < \infty$, passing to the limit with $\gamma \uparrow 1$ we get the desired inequality. □

When (CQ) holds, for every $d \in C_0$ problem (L^d) is feasible and then $v(D^d) = v(L^d)$. Otherwise the previous lemma shows that $v(D^{\gamma d}) = v(L^{\gamma d})$ except for at most an exceptional value γ_0. The optimal values are finite for $\gamma < \gamma_0$ and equal to $+\infty$ for $\gamma > \gamma_0$. The following lemma shows that $\gamma_0 = 0$ iff $T_2^K(d)$ is empty. It will be useful in §4 as well.

LEMMA 3.4. *Assume (DCQ) and suppose $T_2^K(d)$ is not empty. Then letting $d_\gamma := \gamma d$ we have $F(L^{d_\gamma}) \neq \phi$ for all $\gamma > 0$ sufficiently small.*

Proof. Taking $k \in T_2^K(d)$ and using $(P2)$ we get

$$\frac{\gamma^2}{2} G''_x(x_0, 0)dd - \gamma^2 k \in T_K(G(x_0, 0)) - G'(x_0, 0)X \times \{1\}$$

for all $\gamma > 0$ sufficiently small. Then, using $(P1)$ and $(P3)$ we deduce

$$\frac{1}{2} G''_x(x_0, 0)d_\gamma d_\gamma \in T_2^K(d_\gamma) - G'(x_0, 0)X \times \{1\},$$

so we may find $w \in X$ with $\Psi_G(w, d_\gamma) \in T_2^K(d_\gamma)$. □

We end this section by giving a condition under which the upper estimate of Theorem 3.1 coincides with $v(L)$. Using $(P4)$, it is easy to see that this condition is satisfied in particular if (P_0) is *convex* in the sense that for all $y \in K$ and $\lambda \in N_K(y)$, the mapping $\mathcal{L}(\cdot, \lambda, 0)$ is convex. In that case the right-derivative $v'(0)$ is actually equal to $v(L)$ (see [4, Prop. 3.2]).

PROPOSITION 3.5. *Assume $(SDCQ)$. Then $v(\tilde{L}) = v(L)$ whenever*

$$\inf_{d \in C_0} \sup_{\lambda \in S(D)} \left\{ \frac{1}{2} \mathcal{L}''_x(x_0, \lambda, 0)dd - \sigma(\lambda, T_2^K(d)) \right\} \geq 0.$$

Proof. By Lemma 3.3 and using the equality $v(L) = v(D)$ we get

$$\begin{aligned} v(\tilde{L}) &= \inf_{d \in C_0} v(D^d) \\ &\geq \inf_{d \in C_0} \sup_{\lambda \in S(D)} \left\{ \mathcal{L}'_u(x_0, \lambda, 0) + \frac{1}{2} \mathcal{L}''_x(x_0, \lambda, 0)dd - \sigma(\lambda, T_2^K(d)) \right\} \\ &\geq v(L) + \inf_{d \in C_0} \sup_{\lambda \in S(D)} \left\{ \frac{1}{2} \mathcal{L}''_x(x_0, \lambda, 0)dd - \sigma(\lambda, T_2^K(d)) \right\} \\ &\geq v(L), \end{aligned}$$

and we conclude with Lemma 3.2. □

3.2. Lower estimates and expansion of solutions. We derive next some lower estimates for $v'_-(0)$. As $v'_-(0) \le v'_+(0) \le v(\tilde{L})$ whenever (*SDCQ*) holds, this is only of interest if $v(\tilde{L}) > -\infty$. We give conditions that imply $v'_-(0) > -\infty$, based on a result of part I (see [4, Prop. 6.1]) that we recall for the convenience of the reader.

For each set $\Omega \subset \Lambda_0$ we consider the second-order condition

$$SOC(\Omega) \qquad \text{There exist } \alpha, \epsilon > 0 \text{ s.t. } \max_{\lambda\in\Omega} \mathcal{L}''_x(x_0, \lambda, 0)dd \ge \alpha\|d\|^2 \quad \forall\, d \in C_\epsilon,$$

where

$$C_\epsilon := \{d \in X : f'_x(x_0, 0)d \le \epsilon\|d\|,\ G'_x(x_0, 0)d \in T_K(G(x_0, 0)) + \epsilon\|d\|B_Y\}.$$

Note that for $\epsilon = 0$ the *extended critical cone* C_ϵ reduces to the critical cone C_0.

PROPOSITION 3.6. *Assume (DCQ) and suppose SOC(Ω) holds for some bounded $\Omega \subset \Lambda_0$. Then, for each $O(u)$-optimal path x_u, we have $x_u = x_0 + O(\sqrt{u})$.*

Now consider the function

$$\Pi(d) := \sup_{\lambda\in\Lambda_0} \left\{ \mathcal{L}'_u(x_0, \lambda, 0) + \frac{1}{2}\mathcal{L}''_x(x_0, \lambda, 0)dd \right\}$$

and the problems

$$(\tilde{D}) \qquad \min\{\Pi(d) : d \in C_0\},$$

$$(\tilde{D}_\epsilon) \qquad \min\{\Pi(d) : f'_x(x_0, 0)d \le \epsilon,\ G'_x(x_0, 0)d \in T_K(G(x_0, 0))\}.$$

Note that $v(\tilde{D}_\epsilon)$ is a nonincreasing function of ϵ; in particular, $\lim_{\epsilon\downarrow 0} v(\tilde{D}_\epsilon) \le v(\tilde{D})$. Moreover, from $(P4)$ we get $\Pi(d) \le v(D^d)$, so under the conditions of Theorem 3.1 we deduce that

$$\lim_{\epsilon\downarrow 0} v(\tilde{D}_\epsilon) \le v(\tilde{D}) \le v(\tilde{L}). \tag{3.12}$$

PROPOSITION 3.7. *Assume (DCQ), the existence of an $o(u)$-optimal path, and SOC(Ω) for some bounded $\Omega \subset \Lambda_0$. Then $v'_-(0) > -\infty$ and*

(i) *if (CQ) holds, then for each $\varepsilon > 0$ we have*

$$v'_-(0) \ge v(\tilde{D}_\varepsilon); \tag{3.13}$$

(ii) *if any of the following conditions hold:*

(a) *the path may be expanded as $x_u = x_0 + \sqrt{u}d_0 + o(\sqrt{u})$,*

(b) *X is reflexive and $d \to \mathcal{L}''_x(x_0, \lambda, 0)dd$ is weakly lower semicontinuous at each $d \in C_0$,*

then the previous lower bound may be strengthened to

$$v'_-(0) \ge v(\tilde{D}). \tag{3.14}$$

Proof. Let x_u be an $o(u)$-optimal path. By Proposition 3.6 $d_u := (x_u - x_0)/\sqrt{u}$ stays bounded as $u \downarrow 0$, and then for each $\lambda \in \Lambda_0$ we have

$$\begin{aligned} v(u) &= f(x_u, u) + o(u) \\ &\ge v(0) + \mathcal{L}(x_u, \lambda, u) - \mathcal{L}(x_0, \lambda, 0) + o(u) \\ &\ge v(0) + u\left[\mathcal{L}'_u(x_0, \lambda, 0) + \frac{1}{2}\mathcal{L}''_x(x_0, \lambda, 0)d_u d_u\right] + o_\lambda(u), \end{aligned} \tag{3.15}$$

with $\|o_\lambda(u)\|/u \to 0$ uniformly when λ varies over bounded sets. From this and the boundedness of d_u, it follows that $v'_-(0) > -\infty$.

To prove (i) we apply Robinson's theorem [14] to the mapping $\tilde{G}(x) := G(x_0, 0) + G'_x(x_0, 0)(x - x_0)$ in order to find $\tilde{x}_u = x_u + o(\sqrt{u})$ such that $\tilde{G}(\tilde{x}_u) \in K$. Then, by suitably modifying the small term $o_\lambda(u)$, in (3.15) we can replace d_u by $\tilde{d}_u := (\tilde{x}_u - x_0)/\sqrt{u}$. Moreover, under (CQ) we know that Λ_0 is bounded so that taking the supremum over λ we get

$$v(u) \geq v(0) + u\Pi(\tilde{d}_u) + o(u),$$

from which (3.13) follows.

To show (ii), let us choose $u_k \downarrow 0$ realizing the lower limit $v'_-(0)$. When (a) holds we have $d_{u_k} \to d_0$, while in case (b) we may assume that $d_{u_k} \rightharpoonup d_0$. In both cases, $d_0 \in C_0$, and using (3.15) we get

$$v'_-(0) \geq \mathcal{L}'_u(x_0, \lambda, 0) + \frac{1}{2}\mathcal{L}''_x(x_0, \lambda, 0)d_0 d_0,$$

where in case (b) we use the weak lower semicontinuity of $\mathcal{L}''_x(x_0, \lambda, 0)dd$. Taking the supremum over $\lambda \in \Lambda_0$ we conclude (3.14). □

We now analyze under which conditions the gap between the estimate of Theorem 3.1 and (3.14) is null. We start with sufficient conditions for the equality between the optimal values of the subproblems giving the upper and lower estimates. We define *extended polyhedricity of the second kind* (for problem (P_0) at point x_0) as

$$0 \in T_2^K(d) \text{ for all } d \text{ in a dense subset of } C_0.$$

We note that in the definition of *extended polyhedricity* given in part I, the set $S(L)$ was considered instead of C_0. If the constraints are unperturbed, then $S(L) = C_0$ and both definitions coincide.

PROPOSITION 3.8. *Assume Λ_0 nonempty and $(SDCQ)$. If one of the two following conditions hold:*

(a) $0 \in T_2^K(d)$ *for all d in C_0,*

(b) (CQ) *and extended polyhedricity of the second kind hold,*

then $v(\tilde{L}) = v(\tilde{D})$ and $S(\tilde{L}) \subset S(\tilde{D})$.

Proof. From $(P4)$ it follows that when $0 \in T_2^K(d)$ we have $\sigma(\lambda, T_2^K(d)) = 0$ for all $\lambda \in \Lambda_0$, and then $\Pi(d) = v(D^d)$. Consider now a minimizing sequence $\{d^k\}$ for $(\tilde{D})$ satisfying $\sigma(\lambda, T_2^K(d^k)) = 0$. The existence of such a sequence is obvious in case (a); while in case (b) it is a consequence of the fact that, due to (CQ), $\Pi(d)$ is continuous. Along this sequence we have, by Theorem 3.1, $\Pi(d^k) = v(D^{d^k}) \geq v(\tilde{L})$. It follows that $v(\tilde{L}) \leq v(\tilde{D})$. Reminding (3.12), we get $v(\tilde{L}) = v(\tilde{D})$. The inclusion $S(\tilde{L}) \subset S(\tilde{D})$ follows easily from this. □

The following is our main result in this section. It provides a formula for the derivative of the marginal value function $v'(0)$ and analyzes the behavior of paths of approximate solutions, for problems with existence of multipliers and satisfying the weak (but not the strong) second-order sufficient optimality condition.

THEOREM 3.9. *Assume X reflexive, the existence of an $o(u)$-optimal path, $\mathcal{L}''_x(x_0, \lambda, 0)dd$ weakly lower semicontinuous and one of the two hypotheses below.*

(i) (CQ), $SOC(\Lambda_0)$, *and extended polyhedricity of the second kind;*

(ii) $(SDCQ)$, $SOC(\Omega)$ *for some bounded $\Omega \subset \Lambda_0$, and $0 \in T_2^K(d)$ for all d in C_0.*

Then:

(a) *There exists $v'(0) = v(\tilde{L}) = v(\tilde{D})$, and $S(\tilde{L}) \subset S(\tilde{D})$.*

(b) *For every $o(u)$-optimal path x_u, the weak accumulation points of $(x_u - x_0)/\sqrt{u}$ belong to $S(\tilde{D})$.*

(c) *If $d_0 \in S(\tilde{L})$ and $w_0 \in S(L^{d_0})$, then there exists an $o(u)$-optimal path of the form $x_u = x_0 + \sqrt{u}d_0 + o(\sqrt{u})$.*

Proof. (a) This follows by combining Theorem 3.1 and Propositions 3.7 and 3.8.

(b) Let d_0 be a weak limit point of $(x_u - x_0)/\sqrt{u}$. Expanding the Lagrangian as in (3.15) we get $v(\tilde{D}) = v'(0) \geq \Pi(d_0)$. As d_0 is feasible for $v(\tilde{D})$, d_0 is a solution of $v(\tilde{D})$.

(c) Using (*SDCQ*) let us select $w_\gamma \to w_0$ and feasible paths of the form $x_u^\gamma = x_0 + \gamma\sqrt{u}d_0 + uw_\gamma + o_\gamma(u)$, with (for each γ) $\|o_\gamma(u)\|/u \to 0$ when $u \to 0$. Take $\gamma_k \uparrow 1$ and choose a strictly decreasing sequence $u_k \downarrow 0$ such that

$$\|o_{\gamma_k}(u)\| \leq \frac{u}{k} \qquad \forall u \in [0, u_k]$$

from which we construct the feasible path

$$x_u = x_u^{\gamma_k} \qquad \forall u \in [u_{k+1}, u_k).$$

Then we have

$$\|x_u - x_0 - \sqrt{u}d_0\| \leq \sqrt{u}(1 - \gamma_k)\|d_0\| + u\|w_{\gamma_k}\| + \frac{u}{k} \qquad \forall u \in [u_{k+1}, u_k)$$

from which we get $x_u = x_0 + \sqrt{u}d_0 + o(\sqrt{u})$. Also, a second-order expansion implies that for $u \in [u_{k+1}, u_k)$ we have

$$f(x_u, u) = f(x_0, 0) + u\left[f'(x_0, 0)(w_{\gamma_k}, 1) + \frac{1}{2}f_x''(x_0, 0)d_0d_0\right] + o(u)$$

so that

$$\begin{aligned} f(x_u, u) &= f(x_0, 0) + u\Psi_f(w_0, d_0) + o(u) \\ &= v(0) + uv(\tilde{L}) + o(u) = v(u) + o(u). \end{aligned}$$

The conclusion follows. □

4. Perturbation analysis assuming nonexistence of multipliers.

4.1. Preliminaries. In this section we analyze the situation when the set of multipliers Λ_0 is empty, extending the theory of perturbed singular nonlinear programs of [3]. The qualitative behavior is radically different from the case studied in §3, so we are led to introduce some new objects. Indeed, if Λ_0 is empty we have $v(L) = -\infty$ and by part I it follows that $v'(0) = -\infty$.

We will check that, under suitable second-order assumptions, the variation of the cost is of order $O(\sqrt{u})$. This leads us to define, analogously to the Dini derivatives, the following quantities:

$$v^\#(0) := \limsup_{u \downarrow 0} \frac{v(u) - v(0)}{\sqrt{u}},$$

$$v_\#(0) := \liminf_{u \downarrow 0} \frac{v(u) - v(0)}{\sqrt{u}}.$$

We define the singular Lagrangian, the set of singular multipliers (at x_0, for problem (P_0)), and the set of *normalized* singular multipliers as

$$\begin{aligned} \hat{\mathcal{L}}(x, \lambda, u) &:= \langle \lambda, G(x, u) \rangle, \\ \Lambda^s &:= \{\lambda \in Y^* \setminus \{0\} : \ \lambda \in N_K(G(x_0, 0)), \ \hat{\mathcal{L}}_x'(x_0, \lambda, 0) = 0\}, \\ \Lambda_N^s &:= \{\lambda \in \Lambda^s : \|\lambda\| \leq 1\}. \end{aligned}$$

The next proposition shows that Λ_0 and Λ^s are both empty only in some very special situations.

PROPOSITION 4.1. *If both Λ_0 and Λ^s are empty, then the set*

$$\mathcal{A} := \mathbb{R}_+[K - G(x_0, 0)] - G'_x(x_0, 0)X$$

is dense in Y but not equal to Y.

Proof. If $\mathcal{A} = Y$ we know that $\Lambda_0 \neq \phi$ [13, 14]. Suppose next that $\mathcal{A}$ is not dense in Y and select $y \in Y$ not belonging to the closure of $\mathcal{A}$. By the Hahn–Banach theorem there exists $\lambda \in Y^* \setminus \{0\}$ such that

$$\langle \lambda, y\rangle > \langle \lambda, t[k - G(x_0, 0)] - G'_x(x_0, 0)w\rangle \quad \text{for all } w \in X,\ k \in K,\ t > 0.$$

Taking the supremum over $w \in X$, we get $\lambda \circ G'_x(x_0, 0) = 0$, and letting $t \uparrow \infty$ we deduce $\langle \lambda, k - G(x_0, 0)\rangle \leq 0$ for all $k \in K$, so $\lambda \in N_K(G(x_0, 0))$ and then $\Lambda^s \neq \phi$. □

4.2. Upper estimate of the cost. To obtain upper estimates for $v^{\#}(0)$ we consider the following optimization problems:

$$(\hat{L}) \qquad \min_{d \in C_0} \left\{ f'_x(x_0, 0)d : \frac{1}{2}G''_x(x_0, 0)dd \in T_2^K(d) - G'(x_0, 0)X \times \{1\} \right\}$$

and

$$(\hat{D}) \qquad \min_{d \in C_0} \left\{ f'_x(x_0, 0)d : \frac{1}{2}G''_x(x_0, 0)dd \in \overline{T_2^K(d) - G'(x_0, 0)X \times \{1\}} \right\}.$$

Problem $(\hat{L})$ will give an upper estimate of the value function whereas $(\hat{D})$, which has the same optimal value as $(\hat{L})$, will provide a comparison with the lower estimate of $v_{\#}(0)$. We remark that, when Λ^s is not empty, problem $(\hat{D})$ is equivalent to

$(\hat{D}')$

$$\min_{d \in C_0} \left\{ f'_x(x_0, 0)d : \hat{\mathcal{L}}'_u(x_0, \lambda, 0) + \frac{1}{2}\hat{\mathcal{L}}''_x(x_0, \lambda, 0)dd \leq \sigma(\lambda, T_2^K(d)),\ \text{for all } \lambda \in \Lambda^s \right\}.$$

To prove this equivalence it suffices to check that the constraints in $(\hat{D})$ and $(\hat{D}')$ coincide, which follows from the next result applied with $y = G'_u(x_0, 0) + \frac{1}{2}G''_x(x_0, 0)dd$.

PROPOSITION 4.2. *If $\Lambda^s \neq \phi$, then the following are equivalent.*

(a) $y \in \overline{T_2^K(d) - G'_x(x_0, 0)X}$.

(b) $\langle \lambda, y\rangle \leq \sigma(\lambda, T_2^K(d))$ *for all* $\lambda \in \Lambda^s$.

Proof. Both (a) and (b) are false if $T_2^K(d)$ is empty, so we may assume the contrary. The implication (a) $\Rightarrow$ (b) is straightforward and the converse follows by a separation argument. Indeed, if (a) fails we may find a *strictly* separating hyperplane, that is, $\lambda \in Y^* \setminus \{0\}$ and $\alpha \in \mathbb{R}$ such that

$$\langle \lambda, y\rangle > \alpha \geq \langle \lambda, k - G'_x(x_0, 0)w\rangle$$

for all $k \in T_2^K(d)$, $w \in X$. Taking the supremum over $w \in X$ it follows that $\lambda \circ G'_x(x_0, 0) = 0$, and then taking the supremum over k we deduce that

$$\langle \lambda, y\rangle > \alpha \geq \sigma(\lambda, T_2^K(d)). \tag{4.16}$$

Using this and $(P4)$ we get $\lambda \in N_K(G(x_0, 0))$, so $\lambda \in \Lambda^s$ and (4.16) contradicts (b). □

We now state the upper estimate.

THEOREM 4.3. *If (SDCQ) holds, then*

$$v^{\#}(0) \leq v(\hat{L}) = v(\hat{D}) \leq 0,$$

so when $v(\hat{L})$ is finite, we have

$$v(u) \le v(0) + \sqrt{u}v(\hat{L}) + o(\sqrt{u}).$$

In addition, $v(\hat{L}) < 0$ iff there exists a direction d such that $f'_x(x_0,0)d < 0$ and $T_2^K(d) \neq \phi$.

Proof. We begin by showing $v^{\#}(0) \le v(\hat{L}) \le 0$. Let $d \in F(\hat{L})$ and select $w \in X$ such that $G'(x_0,0)(w,1) + \frac{1}{2}G''_x(x_0,0)dd \in T_2^K(d)$. Using the restorability property we may find feasible paths of the form $x_u^\gamma = x_0 + \gamma\sqrt{u}d + uw_\gamma + o(u)$ with $w_\gamma \to w$ as $\gamma \uparrow 1$. Expanding f it follows that

$$v(u) \le f(x_u^\gamma, u) = f(x_0,0) + \gamma\sqrt{u}\, f'_x(x_0,0)d + o(\sqrt{u}),$$

from which we deduce

$$v^{\#}(0) \le \gamma f'_x(x_0,0)d.$$

Letting $\gamma \uparrow 1$ and then taking the infimum over $d \in F(\hat{L})$ we get $v^{\#}(0) \le v(\hat{L})$. Moreover, $(P2)$ implies $0 \in F(\hat{L})$, so $v(\hat{L}) \le 0$.

We prove next $v(\hat{L}) = v(\hat{D})$. Since clearly $v(\hat{D}) \le v(\hat{L})$, it suffices to show that $v(\hat{L}) \le f'_x(x_0,0)d$ for each $d \in F(\hat{D})$. Let $d \in F(\hat{D})$ and select sequences $k_n \in T_2^K(d)$, $w_n \in X$ such that $\frac{1}{2}G''_x(x_0,0)dd = \lim_n[k_n - G'(x_0,0)(w_n,1)]$. Using $(P2)$ we find that given any $t > 0$ we will have for all n large enough

$$\frac{1}{2}tG''_x(x_0,0)dd - tk_n + tG'(x_0,0)(w_n,1) \in T_K(G(x_0,0)) - G'(x_0,0)X \times \{1\},$$

which rearranged gives

$$\frac{1}{2}\frac{t}{1+t}G''_x(x_0,0)dd \in \frac{t}{1+t}k_n + T_K(G(x_0,0)) - G'(x_0,0)X \times \{1\}. \tag{4.17}$$

Letting $d_t := \sqrt{t/(1+t)}d$ and using $(P1)$ and $(P3)$ we deduce that

$$\frac{1}{2}G''_x(x_0,0)d_td_t \in T_2^K(d_t) - G'(x_0,0)X \times \{1\}.$$

Hence $d_t \in F(\hat{L})$ and then

$$v(\hat{L}) \le f'_x(x_0,0)d_t.$$

Letting t tend to $+\infty$ we conclude that $v(\hat{L}) \le f'_x(x_0,0)d$, as required.

We conclude by proving the sufficient condition for $v(\hat{L}) < 0$ (the necessity is evident). If $d \in X$ is such that $f'_x(x_0,0)d < 0$ and $T_2^K(d) \neq \phi$, from Lemma 3.4 we get $\alpha d \in F(\hat{L})$ for all $\alpha > 0$ sufficiently small, so that $v(\hat{L}) \le \alpha f'_x(x_0,0)d < 0$. □

Remark. From the estimate (1.1) we already know that $v_{\#}(0) \le 0$. Henceforth Theorem 4.3 improves the upper estimate of the cost only if $v(\hat{L}) < 0$.

4.3. Lower estimates and expansion of solutions. As in the case when $\Lambda_0 \neq \phi$, we will give a lower estimate of the cost that is sharp when the contribution of the curvature of K happens to be null.

We consider the *singular* second-order conditions

$$\text{there exist } \alpha, \epsilon > 0 \text{ s.t. } \sup_{\lambda \in \Lambda_N^s} \hat{\mathcal{L}}''_x(x_0,\lambda,0)dd \ge \alpha\|d\|^2 \quad \forall\, d \in C_\epsilon. \tag{SSOC}$$

PROPOSITION 4.4. *If (SSOC) holds, then for each $O(\sqrt{u})$-optimal path x_u we have $x_u = x_0 + O(\sqrt{u})$.*

Proof. Let x_u be an $O(\sqrt{u})$-optimal path and let $\beta_u := \|x_u - x_0\|$, $d_u := (x_u - x_0)/\beta_u$. For each $\lambda \in \Lambda_N^s$ we have

$$\begin{aligned} 0 &\geq \hat{\mathcal{L}}(x_u, \lambda, u) - \hat{\mathcal{L}}(x_0, \lambda, 0) \\ &= u\hat{\mathcal{L}}'_u(x_0, \lambda, 0) + \frac{\beta_u^2}{2}\hat{\mathcal{L}}''_x(x_0, \lambda, 0)d_u d_u + o(u) + o(\beta_u^2). \end{aligned}$$

The small terms $o(u)$ and $o(\beta_u^2)$ may be chosen independent of $\lambda \in \Lambda_N^s$, so we may take the supremum to deduce that

$$\beta_u^2 \max_{\lambda \in \Lambda_N^s} \mathcal{L}''_x(x_0, \lambda, 0)d_u d_u \leq O(u) + o(\beta_u^2). \tag{4.18}$$

If for some sequence $u_n \downarrow 0$ one has $\beta_{u_n}^2/u_n \uparrow \infty$, then for n large enough d_{u_n} is in C_ϵ. With (*SSOC*) and (4.18), we obtain a contradiction. □

To obtain the desired lower estimate for $v_\#(0)$ we consider a *relaxed* version of problem $(\hat{D})$, namely,

$$(\hat{R}) \qquad \min_{d \in C_0} \left\{ f'_x(x_0, 0)d : \frac{1}{2}G''_x(x_0, 0)dd \in \overline{T_K(G(x_0, 0)) - G'(x_0, 0)X \times \{1\}} \right\}.$$

As for problem $(\hat{D})$, when Λ^s is not empty one may use Proposition 4.2 (with $d = 0$) to derive the following equivalent formulation for $(\hat{R})$:

$$(\hat{R}') \qquad \min_{d \in C_0} \left\{ f'_x(x_0, 0)d : \hat{\mathcal{L}}'_u(x_0, \lambda, 0) + \frac{1}{2}\hat{\mathcal{L}}''_x(x_0, \lambda, 0)dd \leq 0 \text{ for all } \lambda \in \Lambda^s \right\}.$$

Comparing with $(\hat{D}')$ and using $(P4)$, we see that $F(\hat{D}') \subset F(\hat{R}')$. As these two problems have the same cost, it follows that

$$v(\hat{R}) = v(\hat{R}') \leq v(\hat{D}') = v(\hat{D}). \tag{4.19}$$

PROPOSITION 4.5. *Assume there exists an $o(\sqrt{u})$-optimal path x_u. If (SSOC) is satisfied, then $v_\#(0) > -\infty$. Moreover, if any of the two following properties hold:*

(a) *the path may be expanded as $x_u = x_0 + \sqrt{u}d_0 + o(\sqrt{u})$,*

(b) *X is reflexive and for each $\lambda \in \Lambda^s$ the mapping $d \to \hat{\mathcal{L}}''_x(x_0, \lambda, 0)dd$ is weakly lower semicontinuous at every $d_0 \in C_0$,*

then

$$v_\#(0) \geq v(\hat{R}). \tag{4.20}$$

Proof. By Proposition 4.4 we have $x_u = x_0 + O(\sqrt{u})$ and then

$$v(u) = f(x_u, u) + O(\sqrt{u}) = f(x_0, 0) + O(\sqrt{u}),$$

so $v_\#(0) > -\infty$.

Now let us choose $u_n \downarrow 0$ realizing the lower limit $v_\#(0)$, and let $d_n := (x_{u_n} - x_0)/\sqrt{u_n}$. When (a) holds we have $d_n \to d_0$, while in case (b) we may assume that $d_n \rightharpoonup d_0$ for some $d_0 \in X$. In both cases, $d_0 \in C_0$ and we have

$$v_\#(0) = f'_x(x_0, 0)d_0.$$

On the other hand for all $\lambda \in \Lambda^s$

$$\begin{aligned} 0 &\geq \hat{\mathcal{L}}(x_u, \lambda, u) - \hat{\mathcal{L}}(x_0, \lambda, 0) \\ &= u\hat{\mathcal{L}}'_u(x_0, \lambda, 0) + \frac{u}{2}\hat{\mathcal{L}}''_x(x_0, \lambda, 0)d_u d_u + o(u), \end{aligned}$$

so, using in case (b) the lower semicontinuity of $\hat{\mathcal{L}}''_x(x_0, \lambda, 0)dd$ we get

$$0 \geq \hat{\mathcal{L}}'_u(x_0, \lambda, 0) + \frac{1}{2}\hat{\mathcal{L}}''_x(x_0, \lambda, 0)d_0 d_0.$$

It follows that $d_0 \in F(\hat{R}')$. Combining with (4.19) we get

$$v(\hat{R}) = v(\hat{R}') \leq f'_x(x_0, 0)d_0 = v_\#(0),$$

as was to be proved. □

Let us put together the different bounds obtained so far. If (*SDCQ*) and the assumptions of Proposition 4.5 hold, then

$$v(\hat{R}) = v(\hat{R}') \leq v_\#(0) \leq v^\#(0) \leq v(\hat{D}') = v(\hat{D}) = v(\hat{L}) \leq 0.$$

In our next statement, which is our main result for problems with nonexistence of multipliers, we give a condition for all these optimal values to be equal. This gives the first term of the expansion of the optimal value $v(u)$.

THEOREM 4.6. *Assume the existence of an $O(\sqrt{u})$-optimal path x_u, (SSOC), X reflexive, the lower semicontinuity of $d \to \hat{\mathcal{L}}''_x(x_0, \lambda, 0)dd$ for each $\lambda \in \Lambda^s$, (SDCQ), and finally*

$$0 \in T_2^K(d) \quad \text{for all } d \in C_0.$$

Then $v(\hat{R}) = v(\hat{D})$, $S(\hat{R}) = S(\hat{D})$, and

$$v(u) = v(0) + \sqrt{u}\, v(\hat{D}) + o(\sqrt{u}). \tag{4.21}$$

Proof. The equivalence between $(\hat{R})$ and $(\hat{D})$ follows by noting that when $0 \in T_2^K(d)$ (see [8, Prop. 3.1])

$$T_2^K(d) = \overline{T_K(G(x_0, 0)) - \mathbb{R}_+ G'_x(x_0, 0)d},$$

from which we deduce

$$\overline{T_2^K(d) - G'(x_0, 0)X \times \{1\}} = \overline{T_K(G(x_0, 0)) - G'(x_0, 0)X \times \{1\}}.$$

The expansion of $v(u)$ then follows from Theorem 4.3 and Proposition 4.5. □

5. Appendix: Checking the strong directional constraint qualification. We give some sufficient conditions for checking (*SDCQ*) in the case of *decomposed* constraints of the form: $Y := Y_1 \times Y_2$ with Y_1 and Y_2 Banach spaces and $K := K_1 \times K_2$ with K_1 and K_2 closed convex subsets of Y_1 and Y_2. We denote by $G = (G_1, G_2)$ the components of G, and we consider the decomposed directional constraint qualification:

$$(DDCQ) \quad \begin{cases} \text{(i)} & 0 \in \operatorname{int}[G_1(x_0, 0) + G'_1(x_0, 0)X \times \{0\} - K_1], \\ \text{(ii)} & \text{there exists } \bar{w} \in X \text{ such that } G'_1(x_0, 0)(\bar{w}, 1) \in \operatorname{Rec}(K_1) \text{ and} \\ & G_2(x_0, 0) + \alpha G'_2(x_0, 0)(\bar{w}, 1) \in \operatorname{int} K_2 \text{ for some } \alpha > 0, \end{cases}$$

where $\mathrm{Rec}(K_1)$ denotes the recession cone of K_1, that is,

$$\mathrm{Rec}(K_1) := \limsup_{t\to\infty} \frac{K_1}{t}.$$

To illustrate this condition, let us mention two particular cases. The first one is when $K_2 = Y_2$ so that the constraint is only with K_1. Then (*DDCQ*) reduces to Robinson's condition [14]. The second case is when $K_1 = \{0\}$. Then (*DDCQ*) (i) amounts to the surjectivity of $G'_{1x}(x_0, 0)$ and (*DDCQ*) appears as a natural generalization of Gollan's condition [10] used in the aforementioned literature devoted to nonlinear programming.

THEOREM 5.1. *(DDCQ) implies (SDCQ).*

Proof. We first prove that x_0 is restorable. Let x_u be a path satisfying (2.3) and (2.4). Choose $w_\gamma := \gamma^2 w + (1-\gamma^2)\bar{w}$ and consider

$$y_u := x_0 + \gamma\sqrt{u}d + uw_\gamma. \tag{5.22}$$

Expanding in series we get

$$\begin{aligned} G(y_u, u) &= G(x_0, 0) + \gamma\sqrt{u}G'_x(x_0, 0)d + u\Psi_G(w_\gamma, \gamma d) + o(u) \\ &= G(x_0, 0) + \gamma\sqrt{u}G'_x(x_0, 0)d + \gamma^2 u\Psi_G(w, d) \\ &\quad + (1-\gamma^2)uG'(x_0, 0)(\bar{w}, 1) + o(u) \\ &= G(x(\gamma^2 u), \gamma^2 u) + (1-\gamma^2)uG'(x_0, 0)(\bar{w}, 1) + o(u). \end{aligned}$$

Using (*DDCQ*) (ii) and (2.4) we deduce $d(G_1(y_u, u), K_1) = o(u)$. Then (*DDCQ*) (i) allows us to use Robinson's theorem to find a small correction x_u^γ of y_u,

$$x_u^\gamma = x_0 + \gamma\sqrt{u}d + uw_\gamma + o(u), \tag{5.23}$$

such that $G_1(x_u^\gamma, u) \in K_1$.

Expanding $G_2(x_u^\gamma, u)$ as above, we get

$$G_2(x_u^\gamma, u) = G_2(x(\gamma^2 u), \gamma^2 u) + (1-\gamma^2)uG'_2(x_0, 0)(\bar{w}, 1) + o(u), \tag{5.24}$$

so that letting $z := G'_2(x_0, 0)(\bar{w}, 1)$ and using (2.4) we have

$$G_2(x_u^\gamma, u) = t_u + (1-\gamma^2)uz + o(u)$$

for some $t_u \in K_2$, $t_u \to G_2(x_0, 0)$. Moreover, letting $\alpha_u := (1-\gamma^2)u/\alpha$ we may write $G_2(x_u^\gamma, u) = (1-\alpha_u)t_u + \alpha_u r_u$ with

$$r_u = t_u + \alpha z + \alpha o(u)/(1-\gamma^2)u = t_u + \alpha z + o(1).$$

By (*DDCQ*) (i) we have $r_u \in K_2$ for u small; since also $t_u \in K_2$ and $\alpha_u \in (0, 1)$, it follows that $G_2(x_u^\gamma, u) \in K_2$. Hence x_u^γ is a feasible path and x_0 is restorable.

We now check that (*DCQ*) is satisfied. By (*DDCQ*) (i) (see [14]) there exist $\epsilon > 0$ and $\beta > 0$ such that, whenever $y_1 \in Y_1$ satisfies $\|y_1\| < \epsilon$, there exist $\hat{d} \in X$ and $k_1 \in K_1$ such that $\|\hat{d}\| < \beta\|y_1\|$ and

$$G_1(x_0, 0) + G'_1(x_0, 0)(\hat{d}, 0) - k_1 = y_1.$$

Now take d of the form $d = \hat{d} + \alpha\bar{w}$. Then

$$G_1(x_0, 0) + G'_1(x_0, 0)(d, \alpha) - [k_1 + \alpha G'(x_0, 0)(\bar{w}, 1)] = y_1$$

and

$$G_2(x_0, 0) + G_2'(x_0, 0)(d, \alpha) - y_2 = G_2(x_0, 0) + \alpha G_2'(x_0, 0)(\bar{w}, 1) + G_2'(x_0, 0)(\hat{d}, 0) - y_2.$$

We may choose ϵ small so that for all $\|y_1\| < \epsilon$, $\|y_2\| < \epsilon$ we have $\|G_2'(x_0, 0)(\hat{d}, 0) - y_2\|$ small enough to deduce, using (*DDCQ*) (ii), that the left-hand side above is in K_2. From this (*DCQ*) follows easily. □

Remark. We do not know (even for nonlinear programming problems) if the property $d(G(x_0 + \sqrt{u}d + uw, u), K) = o(u)$ together with (*DCQ*) suffices or not to construct a feasible path of the form $x_u = x_0 + \sqrt{u}d_0 + uw + o(u)$ (without γ and w_γ).

Our final result shows that, for the important class of equality-inequality constrained problems, the restorability property is a consequence of the directional constraint qualification condition (*DCQ*). So in this case the strong qualification (*SDCQ*) is equivalent to (*DCQ*).

PROPOSITION 5.2. *If* $K := \{0\} \times K_2$ *with* $\operatorname{int}(K_2)$ *nonempty, then* (*DCQ*), (*SDCQ*), *and* (*DDCQ*) *are equivalent and are satisfied iff the condition* (*EDCQ*) *holds.*

$$(EDCQ) \qquad \begin{cases} \text{(i)} & G_1'(x_0, 0)X \times \{0\} = Y_1, \\ \text{(ii)} & \text{there exists } \bar{w} \in X \text{ such that } G_1'(x_0, 0)(\bar{w}, 1) = 0 \text{ and} \\ & G_2(x_0, 0) + \alpha G_2'(x_0, 0)(\bar{w}, 1) \in \operatorname{int} K_2 \text{ for some } \alpha > 0. \end{cases}$$

Proof. Obviously each of the conditions (*DCQ*), (*SDCQ*), (*DDCQ*), and (*EDCQ*) is a consequence of the one that follows. Therefore it suffices to prove that (*DCQ*) implies (*EDCQ*). From (*DCQ*), $G_1'(x_0, 0)X \times (0, \infty)$ contains a neighborhood of 0. Being a cone, this set is equal to Y_1. In particular there exist $d_0 \in X, \alpha_0 > 0$ such that $G_1'(x_0, 0)(d_0, \alpha_0) = 0$, i.e., $G_u'(x_0, 0) \in G_1'(x_0, 0)X \times \{0\}$. We deduce that

$$Y_1 = G_1'(x_0, 0)X \times (0, \infty) = G_1'(x_0, 0)X \times \{0\},$$

i.e., (*EDCQ*) (i) holds. Now pick $a \in \operatorname{int}(K_2)$ close enough to $G_2(x_0, 0)$ so that there exist $d \in X$ and $\tilde{\alpha} > 0$ such that $(0, a - G_2(x_0, 0)) \in G(x_0, 0) + G'(x_0, 0)(d, \tilde{\alpha}) - K$. It is easily checked that (*EDCQ*) (ii) is satisfied with $\bar{w} := d/\tilde{\alpha}, \alpha := \tilde{\alpha}/2$. □

REFERENCES

[1] A. AUSLENDER AND R. COMINETTI, *First and second order sensitivity analysis of nonlinear programs under directional constraint qualification conditions*, Optimization, 21 (1990), pp. 351–363.

[2] L. BARBET, *Etude de sensibilité différentielle dans un problème d'optimisation paramétré avec contraintes en dimension infinie*, Thesis, Université de Poitiers, 1992.

[3] J. F. BONNANS, *Directional derivatives of optimal solutions in smooth nonlinear programming*, J. Optim. Theory Appl., 73 (1992), pp. 27–45.

[4] J. F. BONNANS AND R. COMINETTI, *Perturbed optimization in Banach spaces* I: *A general theory based on a weak directional constraint qualification*, SIAM J. Control Optim., 34 (1996), pp. 1151–1171.

[5] ———, *Perturbed optimization in Banach spaces* III: *semi-infinite programming*, SIAM J. Control Optim., 34 (1996), to appear.

[6] J. F. BONNANS, A. D. IOFFE, AND A. SHAPIRO, *Expansion of exact and approximate solutions in nonlinear programming*, in Proc. French-German Conference in Optimization, Lecture Notes in Econom. and Math. Systems, W. Oettli and D. Pallaschke, eds., Springer-Verlag, New York, 1992, pp. 103–117.

[7] J. F. BONNANS AND A. SHAPIRO, *Sensitivity analysis of parametrized programs under cone constraints*, SIAM J. Control Optim., 30 (1992), pp. 1409–1422.

[8] R. COMINETTI, *Metric regularity, tangent sets and second order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.

[9] J. GAUVIN AND R. JANIN, *Directional behaviour of optimal solutions in nonlinear mathematical programming*, Math. Oper. Res., 13 (1988), pp. 629–649.

[10] B. GOLLAN, *On the marginal function in nonlinear programming*, Math. Oper. Res., 9 (1984), pp. 208–221.

[11] A. D. IOFFE, *Variational analysis of a composite function: a formula for the lower second order epi-derivative*, J. Math. Anal. Appl., 160 (1990), pp. 379–405.

[12] ———, *Variational analysis of a composite function: perturbations, value function and sensitivity*, preprint, Haifa MT 880.

[13] H. MAURER AND J. ZOWE, *First and second order necessary and sufficient optimality conditions for infinite dimensional programming problems*, Math. Prog., 16 (1979), pp. 98–110.

[14] S. M. ROBINSON, *Stability theorems for systems of inequalities. Part* II: *differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1190–1219, July 1996

005

ON FINITE-GAIN STABILIZABILITY OF LINEAR SYSTEMS SUBJECT TO INPUT SATURATION*

WENSHENG LIU†, YACINE CHITOUR†, AND EDUARDO SONTAG†

Abstract. This paper deals with (global) finite-gain input/output stabilization of linear systems with saturated controls. For neutrally stable systems, it is shown that the linear feedback law suggested by the passivity approach indeed provides stability, with respect to every L^p-norm. Explicit bounds on closed-loop gains are obtained, and they are related to the norms for the respective systems without saturation.

These results do not extend to the class of systems for which the state matrix has eigenvalues on the imaginary axis with nonsimple (size > 1) Jordan blocks, contradicting what may be expected from the fact that such systems are globally asymptotically stabilizable in the state-space sense; this is shown in particular for the double integrator.

Key words. small input saturations, linear systems, finite-gain stability, Lyapunov functions, dissipative systems

AMS subject classifications. 93D25, 93D05, 93D15, 34H05

1. Introduction.

In this work we are interested in those nonlinear systems that are obtained when cascading a linear system with a memory-free input nonlinearity:

$$(\Sigma) \qquad \dot{x} = Ax + B\sigma(u), \quad y = Cx\,.$$

The nonlinearity σ is of a "saturation" type (definitions are given later). Figure 1 shows the type of system being considered, where the linear part has transfer function $W(s)$ and the function σ shown is the standard semilinear saturation (results will apply to more general σ's).

Linear systems with actuator saturation constitute one of the most important classes of nonlinear systems encountered in practice. Surprisingly, until recently few general theoretical results were available regarding global feedback design problems for them. One such general result was given in [14], which showed that global state-space stabilization for such systems is possible under the assumptions that all the eigenvalues of A are in the closed left-hand plane, plus stabilizability and detectability of (A, B, C). (These conditions are best possible, since they are also necessary. The controller consists of an observer followed by a smooth static nonlinearity.) For more recent work, see [20], which showed—based upon techniques introduced in [16] for a particular case—how to simplify the controller that had been proposed in [14]. See also [8] for closely related work showing that such systems can be semiglobally (that is, on compact sets) stabilized by means of linear feedback.

In this paper, we are interested in studying not merely closed-loop *state-space* stability, but also stability with respect to measurement and actuator noise. This is the notion of stability that is often found in input/output studies. The problem is to find a controller C so that the operator $(u_1, u_2) \mapsto (y_1, y_2)$ defined by the standard systems interconnection

$$\begin{aligned} y_1 &= P(u_1 + y_2), \\ y_2 &= C(u_2 + y_1) \end{aligned}$$

is well posed and finite-gain stable, where P denotes the input/output behavior of the original plant Σ. See Fig. 2. (In our main results, we will take for simplicity the initial state to be zero. However, nonzero initial states can be studied as well, and some remarks in that regard

*Received by the editors February 23, 1994; accepted for publication (in revised form) February 24, 1995. This research was supported in part by US Air Force grant AFOSR-91-0346.

†Department of Mathematics, Rutgers University, New Brunswick, NJ 08903 (wliu, chitour, sontag@math.rutgers.edu).

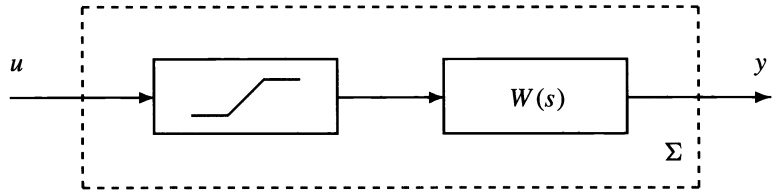


FIG. 1. *Input-saturated linear system.*

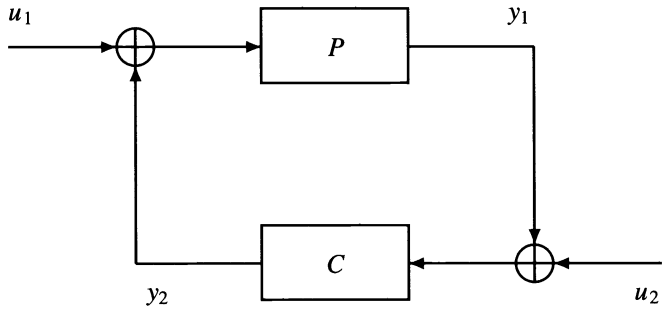


FIG. 2. *Standard closed loop.*

are presented in a latter section of the paper.) Once such input/output stability is achieved, geometric operator-theoretic techniques can be applied; see for instance [3] and the references therein. For other work on computing norms for nonlinear systems in state-space form, see for instance [18] and the references given therein.

We focus on a case which would be trivial if one were only interested in state stability, specifically when the original matrix A is neutrally stable; that is, we focus on the case where all eigenvalues have nonpositive real parts and there are no nontrivial Jordan blocks for eigenvalues in the imaginary axis. (The whole point of [14] and [20] was of course to deal with such possible nontrivial blocks, e.g., multiple integrators.) In this case, a standard passivity approach suggests the appropriate stabilization procedure. For instance, assume that σ is the identity (so the original system is linear), $A + A' \leq 0$, and $C = B'$. Then the system is passive, with storage function $V(x) = \|x\|^2/2$, since integrating the inequality $dV(x(t))/dt \leq y(t)'u(t)$ gives $\int_0^t y(s)'u(s)ds \geq V(x(t)) - V(x(0))$. Thus the negative feedback interconnection with the identity (strictly passive system), that is, $u = -y$, results in finite-gain stability. For this calculation and more discussion on passivity, see for instance [7] and the references given therein. (For the use of the same formulas for just *state-space* stabilization with applications to linear systems with saturations, see [5] and [9]; see also the discussion on the Jurdevic–Quinn method in [13].)

In this paper, we essentially generalize the passivity technique to systems with saturations. We first establish finite-gain stability in the various p-norms, using linear state feedback stabilizers. Then we show how outputs can be incorporated into the framework. Our work is very much in the spirit of the well-known absolute stability area, but we have not been able to find a way to deduce our results from that classical literature.

These results do not extend to the class of systems for which the state matrix has eigenvalues on the imaginary axis with nonsimple (size> 1) Jordan blocks, contradicting what may be expected from the fact that such systems are globally asymptotically stabilizable in the state-space sense; this is shown in particular for the double integrator.

We make one remark on terminology. In the operator approach to nonlinear systems, see, e.g., [19], a "system" is typically defined as a partially defined operator between normed spaces, and "stability" means that the domain of this operator is the entire space. In that context, finite-gain stability is the requirement that the operator be everywhere defined and bounded; the norm of the operator is by definition the gain of the system. In this paper, we use simply the term L^p-stability to mean this stronger finite-gain condition.

The reader is referred to the companion paper [2] for results complementary to those in this paper, dealing with Lipschitz continuity ("incremental gain stability") and continuity of the operators in question. The two papers are technically independent.

Organization of Paper. In §2 we provide definitions and statements of the main results, as well as some related comments. Proofs of the main results are given in §3. Section 4 estimates gains in terms of the corresponding gains for systems without saturation, in particular for $p = 2$ (H_∞-norms). Results regarding nonzero initial states and global asymptotic stability of the origin are collected in §5. Section 6 shows how to enlarge the class of input nonlinearities even more, so as to include nonsaturations as well. The paper closes with §7, which contains the double integrator counterexample.

2. Statements of main results. We introduce now the class of saturation functions to be considered, and state the main results on finite-gain stability. Some remarks are also provided. Proofs are deferred to a later section.

2.1. Saturation functions. We next formally define what we mean by a saturation. Essentially, we ask only that this be a function which has the same sign as its argument, stays away from zero at infinity, is bounded, and is not horizontal near zero.

DEFINITION 1. *We call* $\sigma : \mathbb{R} \to \mathbb{R}$ *a* saturation *function if it satisfies the following two conditions:*

(i) σ *is locally Lipschitz and bounded*;

(ii) $t\sigma(t) > 0$ *if* $t \neq 0$, $\liminf_{t\to 0} \frac{\sigma(t)}{t} > 0$, *and* $\liminf_{|t|\to\infty} |\sigma(t)| > 0$.

For convenience we will simply call a saturation function σ an *S-function*. We say that σ is an $\mathbb{R}^n$*-valued S-function* if $\sigma = (\sigma_1, \ldots, \sigma_n)'$, where each component σ_i is an S-function and

$$\sigma(x) \stackrel{\text{def}}{=} (\sigma_1(x_1), \ldots, \sigma_n(x_n))'$$

for $x = (x_1, \ldots, x_n)' \in \mathbb{R}^n$. Here we use $(\cdots)'$ to denote the transpose of the vector $(\cdots)$.

Remark 1. It follows directly from Definition 1 that most reasonable saturation-type functions are indeed S-functions in our sense. Included are $\arctan(t)$, $\tanh(t)$, and the standard saturation function $\sigma_0(t) = \operatorname{sign}(t) \min\{|t|, 1\}$, i.e.,

$$\sigma_0(t) = \begin{cases} 1 & \text{if } t > 1, \\ t & \text{if } |t| \leq 1, \\ -1 & \text{if } t < -1. \end{cases}$$

Remark 2. It is easy to see that if σ satisfies a bound $|\sigma(t)| \leq M|t|$ for t near zero (in particular if $\sigma(0) = 0$ and (i) in Definition 1 holds), then Condition (ii) in Definition 1 is equivalent to the following condition:

(c) There exist positive numbers a, b, K and a measurable function $\tau : \mathbb{R} \to [a, b]$ such that for all $t \in \mathbb{R}$ we have $|\sigma(t) - \tau(t)t| \leq Kt\sigma(t)$.

It is clear that (c) implies (ii). To see the converse, let $\delta > 0$ be such that $|\sigma(t)| \leq M|t|$ for $|t| \leq \delta$. Then just let

$$\tau(t) = \begin{cases} 1 & \text{if } t = 0, \\ \frac{\sigma(t)}{t} & \text{if } t \in [-\delta, \delta]/\{0\}, \\ \frac{\sigma(\delta)}{\delta} & \text{if } t > \delta, \\ -\frac{\sigma(-\delta)}{\delta} & \text{if } t < -\delta. \end{cases}$$

It is easily verified that there exist positive constants a, b, K such that (c) holds for this τ.

DEFINITION 2. *We say that a constant $K > 0$ is an S-bound for σ if there exist $a, b > 0$ and a measurable function $\tau : \mathbb{R} \to [a, b]$ such that, for all $t \in \mathbb{R}$,*

(i) $b \leq K$,
(ii) $|\sigma(t)| \leq K$,
(iii) $|\sigma(t)| \leq K|t|$,
(iv) $|\sigma(t) - \tau(t)t| \leq Kt\sigma(t)$.

The above discussion shows that such (finite) S-bounds always exist.

A constant $K > 0$ is called an S-bound for an $\mathbb{R}^m$-valued S-function σ if K is an S-bound for each component of σ.

2.2. L^p-Stability. Consider the initialized control system given by

$$\begin{aligned} \dot{x} &= f(x, u), \\ x(0) &= 0, \end{aligned} \tag{1}$$

where the state x and the control u take, respectively, values in $\mathbb{R}^n$ and $\mathbb{R}^m$. We assume that the function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is locally Lipschitz with respect to (x, u). Terminology for systems will be as in any standard reference, such as [13].

Throughout this paper, if ξ is a point in $\mathbb{R}^n$, we use $\|\xi\| = (\sum_{i=1}^n \xi_i^2)^{1/2}$ to denote the usual Euclidean norm. For each matrix S, $\|S\|$ denotes the induced operator norm, and $\|S\|_F$ denotes the Frobenius norm, i.e, $\|S\|_F = \mathrm{Tr}(SS')^{1/2}$, where $\mathrm{Tr}(\cdot)$ denotes trace. Recall that $\|S\| \leq \|S\|_F$.

For each $1 \leq p \leq \infty$ and each integrable (essentially bounded, for $p = \infty$) vector-valued function $x \in L^p([0, \infty), \mathbb{R}^n)$, we let $\|x\|_{L^p}$ denote the usual L^p-norms:

$$\|x\|_{L^p} = \left(\int_0^\infty \|x(t)\|^p dt \right)^{1/p},$$

if $p < \infty$, and

$$\|x\|_{L^\infty} = \text{ess sup}_{0 \leq t < \infty} \|x(t)\|.$$

DEFINITION 3. *Let $1 \leq p \leq \infty$ and $0 \leq M \leq \infty$. We say that (1) has L^p-gain less than or equal to M if for any $u \in L^p([0, \infty), \mathbb{R}^m)$, the solution x of (Σ) corresponding to u is in $L^p([0, \infty), \mathbb{R}^n)$ and satisfies*

$$\|x\|_{L^p} \leq M\|u\|_{L^p}.$$

The infimum of such numbers M will be called the L^p-gain of (Σ). We say that system (Σ) is L^p-stable if its L^p-gain is finite.

By a *neutrally stable* $n \times n$ matrix A we mean one for which all solutions of $\dot{x} = Ax$ are bounded; equivalently, A has no eigenvalues with positive real part and each Jordan block corresponding to a purely imaginary eigenvalue has size 1. Another well-known characterization of such matrices is that they are the ones for which there exists a symmetric positive definite matrix Q such that $A'Q + QA \leq 0$.

We now state our main result.

THEOREM 1. *Let A, B be $n \times n, n \times m$ matrices respectively. Let σ be an $\mathbb{R}^m$-valued S-function. Assume that A is neutrally stable. Then there exists an $m \times n$ matrix F such that the system*

$$\begin{aligned} \dot{x} &= Ax + B\sigma(Fx + u), \\ x(0) &= 0 \end{aligned} \tag{2}$$

is L^p-stable for all $1 \leq p \leq \infty$.

Theorem 1 is an immediate consequence of the more general technical result contained in Theorem 2 below. To state that theorem in great generality, we recall first a standard notion. Let (Σ) $\dot{x} = Ax + Bu$ be a linear system, where x and u take values in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. For each measurable and locally essentially bounded $u : [0, \infty) \to \mathbb{R}^m$ and each $x_0 \in \mathbb{R}^n$, let $x_u(t, x_0)$ be the solution of (Σ) corresponding to u with $x_u(0, x_0) = x_0$. Following the terminology of [6], the *stabilizable subspace* $S(A, B)$ of (A, B) is the subspace of $\mathbb{R}^n$ which consists of all those initial states $x_0 \in \mathbb{R}^n$ for which there is some u so that $x_u(t, x_0) \to 0$ as $t \to \infty$. In other words, $S(A, B)$ is the subspace of $\mathbb{R}^n$ made up of all the states that can be asymptotically controlled to zero (so this includes in particular the reachable subspace). Observe that the pair (A, B) is stabilizable (asymptotically null controllable) iff $S(A, B) = \mathbb{R}^n$.

THEOREM 2. *Let A and B be $n \times n$ and $n \times m$ matrices, respectively. Let $S(A, B)$ be the stabilizable subspace of (A, B). Let σ be an $\mathbb{R}^m$-valued S-function and let $\theta : \mathbb{R}^k \to S(A, B) \subseteq \mathbb{R}^n$ be a locally Lipschitz function such that $\|\theta(\xi)\| \leq \min\{L, L\|\xi\|\}$ for all $\xi \in \mathbb{R}^k$, where $L > 0$ is a constant and $k > 0$ is some integer. Assume that A is neutrally stable. Then there exist an $m \times n$ matrix F and an $\varepsilon > 0$ such that the system*

$$\begin{aligned} \dot{x} &= Ax + B\sigma(Fx + u) + \varepsilon\theta(v), \\ x(0) &= 0 \end{aligned} \tag{3}$$

is L^p-stable for each $1 \leq p \leq \infty$, i.e., there exists for each p a finite constant $M_p > 0$ such that for any $u \in L^p([0, \infty), \mathbb{R}^m)$, $v \in L^p([0, \infty), \mathbb{R}^k)$,

$$\|x\|_{L^p} \leq M_p(\|u\|_{L^p} + \|v\|_{L^p}).$$

The proof is deferred to §3.

Theorem 2 implies Theorem 1 (just take $\theta \equiv 0$) as well as a result dealing with small "nonmatching" state perturbations.

Remark 3. It is possible to make the result even more general by weakening the Lipschitz assumption on θ. Moreover, even the Lipschitz property of σ is not needed. The main problem in dropping this last assumption is that uniqueness of solutions of the closed-loop system is then not guaranteed, so that there is no well-defined input-to-state operator. Nonetheless, one could rephrase all statements by asserting that all possible solutions satisfy the stated bounds. This is consistent with the way stability is defined in some texts on input/output stability, where well-posedness (existence and uniqueness of solutions) is stated as a property independent of stability itself.

2.3. Output stabilization. Consider the initialized linear input/output system

$$\begin{aligned} (\Sigma_{ao})\ \dot{x} &= Ax + B\sigma(u), \\ x(0) &= 0, \\ y &= Ex, \end{aligned}$$

where A, B, and E are, respectively, $n \times n$, $n \times m$, $r \times n$ matrices. Assume that system (Σ_{ao}) is *asymptotically observable* (that is, it is *detectable*). Our main result for input/output systems is as follows.

THEOREM 3. *Assume that system (Σ_{ao}) is asymptotically observable, A is neutrally stable, and the $\mathbb{R}^m$-valued S-function σ is globally Lipschitz. Then there exist an $m \times n$ matrix F and an $n \times r$ matrix L such that the following property holds. Let $1 \leq p \leq \infty$. Pick any $u_1 \in L^p([0, \infty), \mathbb{R}^m)$ and $u_2 \in L^p([0, \infty), \mathbb{R}^r)$, and consider the solution $x = (x_1, x_2)$ of*

$$\dot{x_1} = Ax_1 + B\sigma(y_2 + u_1), \quad y_1 = Ex_1,$$

$$\dot{x_2} = (A + LE)x_2 + B\sigma(Fx_2) - L(y_1 + u_2), \quad y_2 = Fx_2,$$

with $x(0) = 0$. Consider the total output function $y = (y_1, y_2) = (Ex_1, Fx_2)$. Then y is in $L^p([0, \infty), \mathbb{R}^{r+m})$ and

$$\|y\|_{L^p} \leq M_p(\|u_1\|_{L^p} + \|u_2\|_{L^p})$$

for some constant $M_p > 0$.

2.4. Not every feedback stabilizes. One may ask whether *any* F that would stabilize when the saturation is not present would also provide finite gain for (2). Not surprisingly, the answer is negative. In order to give an example, we need first a simple technical remark.

LEMMA 1. *Consider the system $\dot{x} = Ax + B\sigma(Fx + u)$, where the matrix A is assumed to have all eigenvalues in the imaginary axis and where each component of σ is a continuous function whose range contains a neighborhood of the origin (this holds, for instance, if it is an S-function). Furthermore, assume that the pair (A, B) is controllable. Then, given any state $x_0 \in \mathbb{R}^n$, there is some measurable essentially bounded control u steering the origin to x_0 in finite time.*

Proof. Since all eigenvalues of A have zero real part and the pair (A, B) is controllable, for each $\varepsilon > 0$ there is some control v_0 for the system $\dot{x} = Ax + Bu$ so that $|v_0(t)| < \varepsilon$ for all t and v_0 drives in finite time the origin to x_0 (see, e.g., [12]). Considering that the range of σ contains a neighborhood of the origin and using a measurable selection (Fillipov's Theorem), we see that there is a measurable control v which achieves the same transfer, for the system $\dot{x} = Ax + B\sigma(u)$. Now let, along the corresponding trajectory, $u(t) = v(t) - Fx(t)$. It follows that this achieves the desired transfer for $\dot{x} = Ax + B\sigma(Fx + u)$. □

The next two examples show that even if A is neutrally stable, Theorem 1 may not be true if F only satisfies the condition that $A + BF$ is Hurwitz.

Example 1. Let

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad F = -(1/2,\ 1),$$

and any σ so that $\sigma(1/2) = 1$. Then both the origin and $(-1, 0)'$ are equilibrium points of the system

$$\dot{x} = Ax + B\sigma(Fx).$$

By Lemma 1, there is some input u_0 that steers the origin to $(-1, 0)'$ in some finite time T_0. Consider the input u_1 equal to u_0 for $0 \leq t \leq T_0$ and to zero for $t > T_0$. Then if x is the trajectory of (2) corresponding to u_1, we have that $x(t) = (-1, 0)'$ for all $t \geq T_0$. Clearly, for any $1 \leq p < \infty$, $u_1 \in L^p([0, \infty), \mathbb{R})$ and $x \notin L^p([0, \infty), \mathbb{R}^2)$. Therefore, system (3) is not L^p-stable for any $1 \leq p < \infty$. If we use multiple inputs, a different example which includes $p = \infty$ is as follows.

Example 2. Assume that $m = n = 2$. Let

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad F = \begin{pmatrix} -3 & 7 \\ -1 & 2 \end{pmatrix}.$$

Then $A + BF = F$ is Hurwitz. Let $\sigma = (\sigma_0, \sigma_0)'$, where σ_0 is the standard saturation function. Then the system

$$\begin{aligned} \dot{x} &= \sigma(Fx + u), \\ x(0) &= (0, 0)' \end{aligned} \tag{4}$$

is not L^p-stable for any $1 \leq p \leq \infty$. To see this, take a control v on some interval $[0, T]$ that steers $(0, 0)'$ to $(1, 1)'$. Let $u = v$ on $[0, T]$ and $u = (0, 0)'$ on (T, ∞). Let $x = (x_1, x_2)'$ be the solution of (4) corresponding to u. Then on $[T, \infty)$, we have $x_1(t) = x_2(t) = t - T + 1$. Thus (4) is not L^p-stable for any $1 \leq p \leq \infty$. (In fact, the trajectory is not even bounded for a bounded input.)

3. Proofs of the main results. For notational convenience (to avoid having too many negative signs in the formulas) we will prove the main theorem for systems written in the form

$$\begin{aligned} \dot{x} &= Ax - B\sigma(Fx + u) + \varepsilon\theta(v), \\ x(0) &= 0. \end{aligned} \tag{5}$$

A trivial remark is needed before we start.

Remark 4. Assume that $\sigma_1 : \mathbb{R}^{k_1} \to \mathbb{R}^m$ and $\sigma_2 : \mathbb{R}^{k_2} \to \mathbb{R}^n$ each satisfy a growth estimate of the type $\|\sigma_1(u)\| \leq C\|u\|$, $\|\sigma_2(v)\| \leq C\|v\|$ for $u \in \mathbb{R}^{k_1}$, $v \in \mathbb{R}^{k_2}$. It follows from classical linear systems theory that if the system $\dot{x} = Ax$ is globally asymptotically stable—that is, A is a Hurwitz matrix—then the controlled system $\dot{x} = f(x, u, v) = Ax + B\sigma_1(u) + \sigma_2(v)$ is automatically also L^p-stable for all $1 \leq p \leq \infty$. We will be interested in the case in which A is merely stable, but this remark will be used at various points.

We now prove Theorem 2. First note that we can assume that (A, B) is controllable.

3.1. Reduction to the controllable case. Suppose Theorem 2 is already known to be true for controllable (A, B); we show how the general case follows. It is an elementary linear system exercise to show that the stabilizable subspace $S(A, B)$, for any two A, B, is invariant under A; this follows for instance from its characterization as a sum of the reachable subspace and the space of stable modes. Thus the restriction of A to $S(A, B)$ is well defined, and it is again neutrally stable. Now since θ takes values in $S(A, B)$, the trajectories of (5) lie in $S(A, B)$. So we may assume that (A, B) is stabilizable, i.e., $S(A, B) = \mathbb{R}^n$, since otherwise we can restrict ourselves to $S(A, B)$. Then, up to a change of coordinates, we may assume that

$$A = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \qquad B = \begin{pmatrix} B_1 \\ 0 \end{pmatrix},$$

where (A_1, B_1) is controllable and A_1 is neutrally stable. Assume that A_1 is an $r \times r$ matrix and B_1 is an $r \times m$ matrix.

Let $\tilde{\theta} : \mathbb{R}^r \to \mathbb{R}^r$ be given by $\tilde{\theta}(\xi) = (\tilde{\theta}_0(\xi_1), \ldots, \tilde{\theta}_0(\xi_r))'$ for $\xi \in \mathbb{R}^r$, where $\tilde{\theta}_0$ is the standard saturation function, i.e., $\tilde{\theta}_0(t) = \text{sign}\,(t)\,\min\{1, |t|\}$.

By our assumption that the result is known in the controllable case, there exists an $m \times r$ matrix F_1 and $\varepsilon_1 > 0$ so that the system

$$\begin{aligned} \dot{x}_1 &= A_1 x_1 - B_1\sigma(F_1 x_1 + u) + \varepsilon_1\tilde{\theta}(w), \\ x_1(0) &= 0 \end{aligned} \tag{6}$$

is L^p-stable for all $1 \leq p \leq \infty$. Let Γ_p be the L^p-gain of this system, so $\|x_1\|_{L^p} \leq \Gamma_p(\|u\|_{L^p} + \|w\|_{L^p})$ for all $u \in L^p([0,\infty), \mathbb{R}^m)$ and $w \in L^p([0,\infty), \mathbb{R}^r)$.

Since (A, B) is stabilizable, we can find an $m \times n$ matrix E such that $A + BE$ is Hurwitz. Then the system

$$\begin{aligned} \dot{y} &= (A + BE)y + v\,, \\ y(0) &= 0 \end{aligned} \tag{7}$$

is L^p-stable for any $1 \leq p \leq \infty$. Let γ_p be the L^p-gain of (7), so $\|y\|_{L^p} \leq \gamma_p \|v\|_{L^p}$.

Take an $\varepsilon > 0$ such that $\varepsilon L \gamma_\infty \|BE\| \leq \varepsilon_1$. Let $F = (F_1, 0)$. We show that for this choice of F and ε, system (5) is L^p-stable for any $1 \leq p \leq \infty$. For this purpose, let $u \in L^p([0,\infty), \mathbb{R}^m)$, $v \in L^p([0,\infty), \mathbb{R}^k)$. Let x be the solution of (5) corresponding to u, v. Let y be the solution of

$$\begin{aligned} \dot{y} &= (A + BE)y + \varepsilon\theta(v)\,, \\ y(0) &= 0\,. \end{aligned} \tag{8}$$

Then we have $\|y\|_{L^\infty} \leq \varepsilon L \gamma_\infty$ and $\|y\|_{L^p} \leq \varepsilon L \gamma_p \|v\|_{L^p}$ (note that $\|\theta(\xi)\| \leq \min\{L, L\|\xi\|\}$ for all $\xi \in \mathbb{R}^k$). Let $z = x - y$. Then z satisfies

$$\begin{aligned} \dot{z} &= Az - B\sigma(Fz + Fy + u) - BEy\,, \\ z(0) &= 0\,. \end{aligned}$$

Write $z = (z_1, z_2)'$. Then we have $z_2 \equiv 0$ and z_1 satisfies

$$\begin{aligned} \dot{z}_1 &= A_1 z_1 - B_1 \sigma(F_1 z_1 + Fy + u) - B_1 Ey\,, \\ z_1(0) &= 0\,. \end{aligned}$$

Since $\|B_1 Ey\|_{L^\infty} \leq \|B_1 E\|\,\|y\|_{L^\infty} \leq \varepsilon L \gamma_\infty \|B_1 E\| \leq \varepsilon_1$, we have

$$-\frac{B_1 Ey}{\varepsilon_1} = \tilde{\theta}\left(-\frac{B_1 Ey}{\varepsilon_1}\right).$$

Then z_1 satisfies

$$\begin{aligned} \dot{z}_1 &= A_1 z_1 - B_1 \sigma(F_1 z_1 + Fy + u) + \varepsilon_1 \tilde{\theta}\left(\frac{-B_1 Ey}{\varepsilon_1}\right), \\ z_1(0) &= 0\,. \end{aligned}$$

By the L^p-stability of (6) we get that

$$\begin{aligned} \|z\|_{L^p} = \|z_1\|_{L^p} &\leq \Gamma_p\left(\|Fy + u\|_{L^p} + \left\|\frac{B_1 Ey}{\varepsilon_1}\right\|_{L^p}\right) \\ &\leq \Gamma_p\left(\|F\|\,\|y\|_{L^p} + \|u\|_{L^p} + \frac{\|B_1 E\|\,\|y\|_{L^p}}{\varepsilon_1}\right) \\ &\leq \Gamma_p\left(\|u\|_{L^p} + \left(\frac{\|B_1 E\|}{\varepsilon_1} + \|F\|\right)\varepsilon L \gamma_p \|v\|_{L^p}\right). \end{aligned}$$

This shows that (5) is L^p-stable, which concludes the proof that we may assume that (A, B) is controllable.

3.2. Proof of Theorem 2 assuming controllability. From elementary linear algebra, we know that any neutrally stable matrix A is similar to a matrix

$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \tag{9}$$

where A_1 is an $r \times r$ Hurwitz matrix and A_2 is an $(n-r) \times (n-r)$ skew–symmetric matrix. So, up to a change of coordinates, we may assume that A is already in the form (9). In these coordinates, we write

$$B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix},$$

where B_2 is an $(n-r) \times m$ matrix, and we write vectors as $x = (x_1, x_2)'$ and also $\theta = (\theta_1, \theta_2)'$. Consider the feedback law $F = (0, B_2')$. Then system (5), with this choice of F, can be written as

$$\begin{aligned} \dot{x}_1 &= A_1 x_1 - B_1 \sigma(B_2' x_2 + u) + \varepsilon \theta_1(v)\,, \\ \dot{x}_2 &= A_2 x_2 - B_2 \sigma(B_2' x_2 + u) + \varepsilon \theta_2(v)\,, \\ x_1(0) &= 0, \quad x_2(0) = 0\,. \end{aligned} \tag{10}$$

Since A_1 is Hurwitz, it will be sufficient to show that there exists an $\varepsilon > 0$ such that the x_2-subsystem is L^p-stable (we may think of x_2 as an additional input to the first subsystem and apply Remark 4).

The controllability assumption on (A, B) implies that the pair (A_2, B_2) is also controllable. Since A_2 is skew–symmetric, the matrix $\tilde{A} := A_2 - B_2 B_2'$ is Hurwitz. (Just observe that the Lyapunov equation $\tilde{A}' I_{n-r} + I_{n-r} \tilde{A} = -2 B_2 B_2'$ holds, and the pair $(\tilde{A}, B_2)$ is controllable; see [13, Ex. 4.6.7].) Therefore, the theorem is a consequence of the following lemma. This is where the main parts of our argument lie (except for a small technical point, whose proof is deferred to §3.5).

LEMMA 2. *Let σ, θ be as in Theorem 2. Let A be a skew–symmetric matrix. Assume that $\tilde{A} := A - BB'$ is Hurwitz. Then there exists an $\varepsilon > 0$ such that the system*

$$\begin{aligned} \dot{x} &= Ax - B\sigma(B'x + u) + \varepsilon \theta(v)\,, \\ x(0) &= 0 \end{aligned} \tag{11}$$

is L^p-stable for all $1 \le p \le \infty$.

Proof. Assume that $\sigma = (\sigma_1, \ldots, \sigma_m)'$. Let $0 < a \le b < \infty$, $K > 0$ be constants and $\tau_i : \mathbb{R} \to [a, b]$, $i = 1, \ldots, m$, be measurable functions so that the components σ_i of σ satisfy (i)–(iv) in Definition 2 with the respective τ_i's. We may assume that K is large enough such that $K \ge L$. Let

$$\Gamma \stackrel{\text{def}}{=} \min_{i=1,\ldots,m} \liminf_{|\xi| \to \infty} |\sigma_i(\xi)|.$$

Then $\Gamma > 0$. Let $\varepsilon > 0$ satisfy

$$\varepsilon < \frac{\Gamma}{K \gamma_\infty \sqrt{m} \|B\|}\,, \tag{12}$$

where γ_∞ is the L^∞-gain of the initialized linear control system

$$\begin{aligned} \dot{y} &= (A - BB')y + u\,, \\ y(0) &= 0. \end{aligned} \tag{13}$$

By (12) there exists a $\delta \in (0, 1/2]$ such that

$$\varepsilon \leq \frac{(1-2\delta)\Gamma}{K\gamma_\infty \sqrt{m}\|B\|}.$$

Let $u \in L^p([0,\infty), \mathbb{R}^m)$, $v \in L^p([0,\infty), \mathbb{R}^k)$. Let y be the solution of

$$\begin{aligned} \dot{y} &= (A - BB')y + \varepsilon\theta(v), \\ y(0) &= 0. \end{aligned} \tag{14}$$

Let x be the solution of (11) corresponding to u, v and let $z = x - y$. Then z satisfies

$$\begin{aligned} \dot{z} &= Az - B\sigma(B'z + u + B'y) + BB^T y, \\ z(0) &= 0. \end{aligned} \tag{15}$$

Let $\tilde{u} = u + B'y$ and $\tilde{v} = B'y$. Then we get

$$\|\tilde{v}\|_{L^\infty} \leq \|B\|\,\|y\|_{L^\infty} \leq \varepsilon\|B\|\gamma_\infty\|\theta\|_{L^\infty} \leq \frac{(1-2\delta)\Gamma}{\sqrt{m}}. \tag{16}$$

Now (15) can be written as

$$\begin{aligned} \dot{z} &= Az - B\left(\sigma(B'z + \tilde{u}) - \tilde{v}\right), \\ z(0) &= 0. \end{aligned} \tag{17}$$

(We have brought the problem to one of a "matched uncertainty" type, in robust control terms, if we think of $\tilde{v}$ as representing a source of uncertainty.)

Let $\tilde{z}(t) = B'z(t) + \tilde{u}(t)$. For each $1 \leq p < \infty$, consider the function $V_{0,p} : \mathbb{R}^n \to \mathbb{R}$ given by

$$V_{0,p}(x) = \frac{\|x\|^{p+1}}{p+1}.$$

Along the trajectories of (17), we have

$$\begin{aligned} \dot{V}_{0,p}(z(t)) &= -\|z(t)\|^{p-1}z'(t)B\left(\sigma\left(B'z(t) + \tilde{u}(t)\right) - \tilde{v}(t)\right) \\ &= -\|z(t)\|^{p-1}\tilde{z}'(t)\left[\sigma\left(\tilde{z}(t)\right) - \tilde{v}(t)\right] + \|z(t)\|^{p-1}\tilde{u}'(t)\left[\sigma\left(\tilde{z}(t)\right) - \tilde{v}(t)\right]. \end{aligned}$$

Since K is an S-bound for σ and considering (16), we have the following decay estimate:

$$\begin{aligned} \dot{V}_{0,p}(z(t)) \leq{}& -\|z(t)\|^{p-1}\tilde{z}'(t)\left(\sigma\left(\tilde{z}(t)\right) - \tilde{v}(t)\right) \\ &+ \left(K + \frac{(1-2\delta)\Gamma}{\sqrt{m}}\right)\|z(t)\|^{p-1}\|\tilde{u}(t)\|. \end{aligned} \tag{18}$$

We next need to bound the first term in the right-hand side of (18). For that purpose, we will partition $[0,\infty)$ into two subsets. By the definition of Γ, there is some $M_1 \geq 1$ so that

$$\min_{i=1,\ldots,m} \inf_{|\xi|\geq M_1} |\sigma_i(\xi)| \geq (1-\delta)\Gamma.$$

The first subset consists of those t for which $\|\tilde{z}'(t)\| \leq M_1\sqrt{m}$. For such t, trivially,

$$\tilde{z}'(t)\left(\sigma\left(\tilde{z}(t)\right) - \tilde{v}(t)\right) \geq \tilde{z}'(t)\sigma\left(\tilde{z}(t)\right) - M_1\sqrt{m}\|\tilde{v}(t)\|. \tag{19}$$

Next we consider those t for which $\|\tilde{z}'(t)\| > M_1\sqrt{m}$. First we note some general facts about any vector $\xi \in \mathbb{R}^m$ for which

$$\|\xi\| > M_1\sqrt{m}\,. \tag{20}$$

If we pick i_0 so that $|\xi_{i_0}| = \max_{i=1,\dots,m}\{|\xi_i|\}$, then $|\xi_{i_0}| > M_1$, and therefore, by the choice of M_1, $|\sigma_{i_0}(\xi_{i_0})| \geq (1-\delta)\Gamma$. We conclude that if ξ satisfies (20) then

$$\xi'\sigma(\xi) \geq \xi_{i_0}\sigma_{i_0}(\xi_{i_0}) \geq \frac{\|\xi\|}{\sqrt{m}}(1-\delta)\Gamma\,,$$

or equivalently

$$\|\xi\| \leq \frac{\sqrt{m}\,\xi'\sigma(\xi)}{(1-\delta)\Gamma}\,.$$

From this and (16) we have if $\|\tilde{z}(t)\| > M_1\sqrt{m}$,

$$\begin{aligned}
\tilde{z}'(t)\left(\sigma\left(\tilde{z}(t)\right) - \tilde{v}(t)\right) &\geq \tilde{z}'(t)\sigma\left(\tilde{z}(t)\right) - \|\tilde{z}'(t)\|\,\|\tilde{v}(t)\| \\
&\geq \tilde{z}'(t)\sigma\left(\tilde{z}(t)\right) - \frac{\sqrt{m}\|\tilde{v}\|_{L^\infty}}{(1-\delta)\Gamma}\tilde{z}'(t)\sigma\left(\tilde{z}(t)\right) \\
&\geq \left(1 - \frac{1-2\delta}{1-\delta}\right)\tilde{z}'(t)\sigma\left(\tilde{z}(t)\right) \\
&= \frac{\delta}{1-\delta}\tilde{z}'(t)\sigma\left(\tilde{z}(t)\right). \qquad (21)
\end{aligned}$$

Note also that $\frac{\delta}{1-\delta} \leq 1$ for $0 < \delta \leq 1/2$. Combining (19) and (21) we have a common estimate valid for all $t \geq 0$:

$$\tilde{z}'(t)\left(\sigma\left(\tilde{z}(t)\right) - \tilde{v}(t)\right) \geq \frac{\delta}{1-\delta}\tilde{z}'(t)\sigma\left(\tilde{z}(t)\right) - M_1\sqrt{m}\|\tilde{v}(t)\|\,.$$

Using this and (18) we get

$$\begin{aligned}
\dot{V}_{0,p}(z(t)) \leq &-\frac{\delta}{1-\delta}\|z(t)\|^{p-1}\tilde{z}'(t)\sigma\left(\tilde{z}(t)\right) \\
&+\|z(t)\|^{p-1}\left(\left(K + \frac{\Gamma}{\sqrt{m}}\right)\|\tilde{u}(t)\| + M_1\sqrt{m}\|\tilde{v}(t)\|\right). \qquad (22)
\end{aligned}$$

Let $\tau = \operatorname{diag}(\tau_1, \dots, \tau_m)$ with $\tau(\xi) = \operatorname{diag}(\tau_1(\xi_1), \dots, \tau_m(\xi_m))$ for $\xi \in \mathbb{R}^m$. Then $aI \leq \tau(\xi) \leq bI$ for all $\xi \in \mathbb{R}^m$. We have for any $\xi \in \mathbb{R}^m$,

$$\begin{aligned}
\|\tau(\xi)\xi - \sigma(\xi)\| &= \left(\sum_{i=1}^m |\tau_i(\xi_i)\xi_i - \sigma_i(\xi_i)|^2\right)^{1/2} \\
&\leq K\left(\sum_{i=1}^m \xi_i^2(\sigma_i(\xi_i))^2\right)^{1/2} \leq K\xi'\sigma(\xi). \qquad (23)
\end{aligned}$$

Now we rewrite (17) in the form

$$\begin{aligned}
&\dot{z} = \bar{A}(t)z + B\left[\tau\left(\tilde{z}(t)\right)\tilde{z}(t) - \sigma\left(\tilde{z}(t)\right) - \tau\left(\tilde{z}(t)\right)\tilde{u}(t) + \tilde{v}(t)\right], \qquad (24) \\
&z(0) = 0,
\end{aligned}$$

where $\bar{A}(t) = A - B\tau(\tilde{z}(t))B'$. Then $\bar{A}$ satisfies the conditions of Corollary 1 below. Therefore, for each $1 < p < \infty$, there exist a differentiable function $V_{1,p}$ and positive real numbers a_p, b_p, and c_p such that

(P1) $a_p\|x\|^p \leq V_{1,p}(x) \leq b_p\|x\|^p$,

(P2) $\|DV_{1,p}(x)\| \leq c_p\|x\|^{p-1}$,

(P3) $DV_{1,p}(x)\bar{A}(t)x \leq -\|x\|^p$,

for all $x \in \mathbb{R}^n$ and $t \geq 0$. (Note that the constants a_p, b_p, c_p depend only on A, B, a, b.) Moreover $V_{1,p}$ can be chosen so that

(P4) $\limsup_{p\to 1+} c_p = c_1 < \infty$, and the limit $V_{1,1}(x) = \lim_{p\to 1+} V_{1,p}(x)$ exists for all $x \in \mathbb{R}^n$.

Using (23) and (24), we get, for $1 < p < \infty$,

$$\begin{aligned}\frac{dV_{1,p}(z(t))}{dt} &\leq -\|z(t)\|^p + c_p\|B\|\|z(t)\|^{p-1}(\|\tilde{v}(t)\| + b\|\tilde{u}(t)\|)\\ &\quad + c_p\|B\|\|z(t)\|^{p-1}\{\|\tau(\tilde{z}(t))\tilde{z}(t) - \sigma(\tilde{z}(t))\|\}\\ &\leq -\|z(t)\|^p + c_p\|B\|\|z(t)\|^{p-1}(\|\tilde{v}(t)\| + b\|\tilde{u}(t)\|)\\ &\quad + c_pK\|B\|\|z(t)\|^{p-1}\tilde{z}'(t)\sigma(\tilde{z}(t))\,. \qquad (25)\end{aligned}$$

For $1 \leq p < \infty$, let

$$\lambda_p = \frac{K\|B\|c_p(1-\delta)}{\delta}. \qquad (26)$$

(Observe that this constant does not depend on the particular u and v being considered, it depends only on the system and on p.) Finally, consider, for each $1 \leq p < \infty$, the following function:

$$V_p = \lambda_p V_{0,p} + V_{1,p}\,, \qquad (27)$$

where λ_p is given in (26). Using (22), (25), and the fact that $b \leq K$, for $1 < p < \infty$, we have along trajectories of (17),

$$\frac{dV_p(z(t))}{dt} \leq -\|z(t)\|^p + \kappa_p\|z(t)\|^{p-1}(\|\tilde{u}(t)\| + \|\tilde{v}(t)\|)\,, \qquad (28)$$

where

$$\kappa_p = \lambda_p \max\left\{1 + K + \frac{\Gamma}{\sqrt{m}}, \frac{1}{K} + \sqrt{m}M_1\right\}.$$

For any $t \geq 0$, integrating (28) from zero to t, we have

$$V_p(z(t)) + \int_0^t \|z(s)\|^p ds \leq \kappa_p \int_0^t \|z(s)\|^{p-1}(\|\tilde{u}(s)\| + \|\tilde{v}(s)\|)ds\,.$$

When $p = 1$, this inequality is also true as an easy consequence of the Lebesgue dominated convergence theorem (applied to a sequence $\{p^j\}_{j=1}^{\infty}$ decreasing to 1). Thus the inequality is true for all $1 \leq p < \infty$.

Applying Hölder's inequality to $\int_0^t \|z(s)\|^{p-1}(\|\tilde{u}(s)\| + \|\tilde{v}(s)\|)ds$, we conclude that for all $1 \leq p < \infty$ and $t \geq 0$,

$$V_p(z(t)) + \|z\|^p_{L^p[0,t]} \leq \kappa_p\|z\|^{p-1}_{L^p[0,t]}(\|\tilde{u}\|_{L^p} + \|\tilde{v}\|_{L^p})\,. \qquad (29)$$

Since $V_p \geq 0$, we get that $z \in L^p([0,\infty),\mathbb{R}^n)$ and

$$\|z\|_{L^p} \leq \kappa_p(\|\tilde{u}\|_{L^p} + \|\tilde{v}\|_{L^p})\,. \tag{30}$$

Now since $z = x - y$, $\tilde{u} = u + B'y$, $\tilde{v} = B'y$, we have

$$\|\tilde{v}\|_{L^p} \leq \|B\|\,\|y\|_{L^p} \leq \varepsilon K\gamma_p\|B\|\|v\|_{L^p},$$

$$\|\tilde{u}\|_{L^p} \leq \|u\|_{L^p} + \varepsilon K\gamma_p\|B\|\|v\|_{L^p}\,,$$

$$\|z\|_{L^p} \geq \|x\|_{L^p} - \|y\|_{L^p} \geq \|x\|_{L^p} - \varepsilon K\gamma_p\|v\|_{L^p}\,,$$

where γ_p is the L^p-gain of (13). Combining this with (30) we have

$$\|x\|_{L^p} \leq \kappa_p\|u\|_{L^p} + \varepsilon K\gamma_p(1 + 2\kappa_p\|B\|)\|v\|_{L^p}\,.$$

This finishes the proof of the lemma, and hence our main theorem, for the case when $1 \leq p < \infty$.

We now prove the lemma for $p = \infty$. For this, we need to show that system (11) has the *uniform bounded input bounded state* property, i.e., there exists a finite constant M such that $\|x\|_{L^\infty} \leq M(\|u\|_{L^\infty} + \|v\|_{L^\infty})$ for all $u \in L^\infty([0,\infty),\mathbb{R}^m)$ and $v \in L^\infty([0,\infty),\mathbb{R}^k)$. Letting $p = 2$, from (28) we have

$$\frac{dV_2(z(t))}{dt} \leq -\|z(t)\|\,(\|z(t)\| - \kappa_2(\|\tilde{u}\|_{L^\infty} + \|\tilde{v}\|_{L^\infty}))\,. \tag{31}$$

Let $\beta = \|\tilde{u}\|_{L^\infty} + \|\tilde{v}\|_{L^\infty}$. Thus, $\dot{V}_2$ is negative outside the ball of radius $\kappa_2\beta$ centered at the origin. It follows that

$$V_2(z(t)) \leq \sup_{\|\xi\|\leq\kappa_2\beta} V_2(\xi) \leq \frac{\lambda_2\kappa_2^3}{3}\beta^3 + b_2\kappa_2^2\beta^2\,.$$

First assume that $\beta \leq 1$. Then we have

$$a_2\|z(t)\|^2 \leq V_2(z(t)) \leq \left(\frac{\lambda_2\kappa_2^3}{3} + b_2\kappa_2^2\right)\beta^2\,,$$

which implies that

$$\|z\|_{L^\infty} \leq \left\{\frac{\lambda_2\kappa_2^3 + 3b_2\kappa_2^2}{3a_2}\right\}^{1/2}\beta\,.$$

If $\beta > 1$, we have

$$\frac{\lambda_2\|z(t)\|^3}{3} \leq V_2(z(t)) \leq \left(\frac{\lambda_2\kappa_2^3}{3} + b_2\kappa_2^2\right)\beta^3\,.$$

We then get that

$$\|z\|_{L^\infty} \leq \left\{\frac{\lambda_2\kappa_2^3 + 3b_2\kappa_2^2}{\lambda_2}\right\}^{1/3}\beta\,.$$

Let

$$\bar{G}_\infty = \max\left\{\left\{\frac{\lambda_2\kappa_2^3 + 3b_2\kappa_2^2}{3a_2}\right\}^{1/2}, \left\{\frac{\lambda_2\kappa_2^3 + 3b_2\kappa_2^2}{\lambda_2}\right\}^{1/3}\right\}.$$

We have $\|z\|_{L^\infty} \leq \bar{G}_\infty \beta$. Now

$$\beta = \|\tilde{u}\|_{L^\infty} + \|\tilde{v}\|_{L^\infty} \leq \|u\|_{L^\infty} + 2\varepsilon K \gamma_\infty \|B\| \, \|v\|_{L^\infty}$$

and

$$\|z\|_{L^\infty} \geq \|x\|_{L^\infty} - \varepsilon K \gamma_\infty \|v\|_{L^\infty} .$$

We conclude that

$$\|x\|_\infty \leq \bar{G}_\infty \|u\|_{L^\infty} + \varepsilon K \gamma_\infty (1 + 2\bar{G}_\infty \|B\|) \|v\|_{L^\infty} .$$

Now the proof of Lemma 2 is complete. □

3.3. Proof of the output feedback theorem. We now provide a proof of Theorem 3. We will show a somewhat stronger statement, namely, that the state trajectory x also satisfies an estimate as required. The proof will be the usual Luenberger-observer construction, but a bit of care has to be taken because of the nonlinearities.

Asymptotic observability means that there is some $n \times r$ matrix L such that $A + LE$ is Hurwitz. Let F be as in Theorem 2. Let $e = x_1 - x_2$. Then $(x_1, e)'$ satisfies

$$\begin{aligned} \dot{x_1} &= Ax_1 + B\sigma(Fx_1 - Fe + u_1) , \\ \dot{e} &= (A + LE)e + B\,(\sigma(Fx_1 - Fe + u_1) - \sigma(Fx_1 - Fe)) + Lu_2 . \end{aligned}$$

Let $\tilde{v} = \sigma(Fx_1 - Fe + u_1) - \sigma(Fx_1 - Fe)$. Since $\|\tilde{v}(t)\| \leq K\|u_1(t)\|$ (here K is a Lipschitz constant for σ) and $A + LE$ is Hurwitz, we know that e is in $L^p([0, \infty), \mathbb{R}^n)$ and $\|e\|_{L^p} \leq \widehat{M}(\|u_1\|_{L^p} + \|u_2\|_{L^p})$ for some constant $\widehat{M} > 0$. Then the conclusion follows from Theorem 2 applied to the x_1-subsystem.

Note that the conclusion of this theorem can be restated in terms of the finite-gain stability of a standard systems interconnection

$$\begin{aligned} y_1 &= P(u_1 + y_2), \\ y_2 &= C(u_2 + y_1), \end{aligned}$$

where P denotes the input/output behavior of the original system Σ and C is the input/output behavior of the controller with state space x_2 and output y_2.

3.4. Operator stability among different norms. We can actually prove a result stronger than that stated in Theorem 2, namely, that the input-to-state operator $(u, v) \to x$ from $L^p([0, \infty), \mathbb{R}^m) \times L^p([0, \infty), \mathbb{R}^k)$ to $L^p([0, \infty), \mathbb{R}^n)$ is a bounded operator from $L^p([0, \infty), \mathbb{R}^m) \times L^p([0, \infty), \mathbb{R}^k)$ to $L^q([0, \infty), \mathbb{R}^n)$, for any $q \geq p$.

Remark 5. From (29), (30) we get that, for $u \in L^p([0, \infty), \mathbb{R}^m)$, $v \in L^p([0, \infty), \mathbb{R}^k)$, and $t \geq 0$,

$$a_p \|z(t)\|^p \leq V_p(z(t)) \leq \kappa_p \|z\|_{L^p}^{p-1} \| (\|\tilde{u}\|_{L^p} + \|\tilde{v}\|_{L^p}) \leq \kappa_p^p (\|\tilde{u}\|_{L^p} + \|\tilde{v}\|_{L^p})^p ;$$

then, $\|z\|_{L^\infty} \leq C_1(\|\tilde{u}\|_{L^p} + \|\tilde{v}\|_{L^p})$ with $C_1 = \kappa_p a_p^{-1/p}$. Therefore we obtain for $q \geq p$,

$$\|z\|_{L^q}^q \leq \|z\|_{L^\infty}^{q-p} \|z\|_{L^p}^p \leq C_1^{q-p} \kappa_p^p (\|\tilde{u}\|_{L^p} + \|\tilde{v}\|_{L^p})^q . \tag{32}$$

From this one can easily deduce that for any $q \geq p$ the solution x of (11) satisfies

$$\|x\|_{L^q} \leq M_{p,q}(\|u\|_{L^p} + \|v\|_{L^p})$$

for some constants $M_{p,q} > 0$. The same results then hold for the original system in Theorem 2, as is clear from the reduction to (11). That is, for any $u \in L^p([0, \infty), \mathbb{R}^m)$, $v \in L^p([0, \infty), \mathbb{R}^k)$, the solution x of (5) satisfies a similar inequality.

3.5. A remark on robustness of a linear feedback. It is worth pointing out that the same method used to prove Lemma 2 allows us to establish the next proposition, which is a result regarding time-varying multiplicative uncertainties on a linear feedback law $u = -B'x$. For that, we need the following lemma.

LEMMA 3. *Fix two positive real numbers c, d. Let A be an $n \times n$ skew–symmetric matrix, let B be an $n \times m$ matrix, and assume that the pair (A, B) is controllable (or, equivalently, that $A - BB'$ is Hurwitz). Then there is a symmetric positive definite matrix P so that*

$$P(A - BDB') + (A' - BD'B')P \leq -I, \tag{33}$$

for all those $m \times m$ matrices D so that $D + D' \geq cI$ and $\|D\| \leq d$.

Proof. Since (A, B) is controllable, the same is true for (A, rB) for any $r > 0$; thus $A - rBB'$ is Hurwitz for any $r > 0$. Pick $P_1 > 0$ so that $P_1(A - cBB') + (A' - cBB')P_1 = -2I$. We will choose P of the form $P_1 + \beta I$ for a suitable β. Note that

$$2x'P_1(A - BDB')x = -2\|x\|^2 + 2x'P_1B(cI - D)B'x,$$

where the last term has norm bounded above by $C\|x\|\|B'x\|$ for some constant C which depends on c and d. On the other hand,

$$2\beta x'(A - BDB')x = -2\beta x'BDB'x \leq -c\beta\|B'x\|^2.$$

Thus $2x'P(A - BDB')x \leq -2\|x\|^2 + C\|x\|\|B'x\| - \beta c\|B'x\|^2$ and picking β large enough guarantees that this quadratic form is always less than $-\|x\|^2$. □

COROLLARY 1. *Let A and B be as in Lemma 3. Let $c, d > 0$ and $\bar{A}(t) = A - BD(t)B'$, where $D(t)$ is any measurable $m \times m$ matrix such that $D(t) + D'(t) \geq cI$, for almost all t in $[0, \infty)$, and $\sup\{\|D(t)\| : t \in [0, \infty)\} \leq d$. Then for each $1 < p < \infty$, there exist a differentiable function V_p and positive real numbers a_p, b_p, and c_p such that*

(P0) *V_p, a_p, b_p, c_p depend only on A, B, c, d;*

and for all $x \in \mathbb{R}^n$, $t \in [0, \infty)$,

(P1) $a_p\|x\|^p \leq V_p(x) \leq b_p\|x\|^p$;

(P2) $\|DV_p(x)\| \leq c_p\|x\|^{p-1}$;

(P3) $DV_p(x)\bar{A}(t)x \leq -\|x\|^p$.

Moreover, we may choose V_p so that

(P4) $\limsup_{p\to 1+} c_p = c_1 < \infty$, *and the limit $V_1(x) := \lim_{p\to 1+} V_p(x)$ exists for all $x \in \mathbb{R}^n$.*

Proof. Take $V_p(x) = \alpha_p(x'Px)^{p/2}$, where $\alpha_p > 0$ is a proper constant and P is chosen as in Lemma 3. □

As a direct application of Corollary 1, we get Corollary 2.

COROLLARY 2. *Let A be an $n \times n$ skew–symmetric matrix and B be an $n \times m$ matrix. Assume that $A - BB'$ is Hurwitz. Let $D(t)$ be a measurable $m \times m$ matrix with bounded entries. Assume also that there exists a constant $a > 0$ such that $D(t) + D'(t) \geq aI$ for almost all t in $[0, \infty)$. Then the initialized system*

$$(\tilde{\Sigma}) \qquad \begin{aligned} \dot{x} &= \bar{A}(t)x + u, \\ x(0) &= 0, \end{aligned}$$

where $u \in L^p([0, \infty), \mathbb{R}^n)$ and $\bar{A}(t) := A - BD(t)B'$, is L^p-stable for $1 \leq p \leq \infty$, and the L^p-gain depends only on p, a, A, B, and $M = \sup\{\|D(t)\| : t \in [0, \infty)\}$.

Proof. Let V_p be a function satisfying Conditions (P0)–(P3) in Corollary 1 with respect to $\bar{A}$. Along the trajectories of $(\tilde{\Sigma})$, we have

$$\dot{V}_p(x(t)) \leq -\|x(t)\|^p + c_p\|x(t)\|^{p-1}\|u(t)\|,$$

for some $c_p > 0$. The conclusion follows after applying Hölder's inequality. □

4. Comparison with linear gains. From the proof of Lemma 2, we can also obtain explicit bounds for the L^p-gain for (11). For simplicity, we deal only with the case when $\theta \equiv 0$ and we will assume that each component σ_i of σ satisfies a stronger estimate:

$$\forall t \in \mathbb{R}, \quad |\sigma_i(t) - a_i t| \leq K t \sigma_i(t),$$

where $a_i > 0$ are some constants. Of course this implies that $(d\sigma_i(t)/dt)|_{t=0} = a_i$. Specifically, we will compare these bounds with the L^p-gain of the system that is obtained by linearizing (11):

$$\begin{aligned} \dot{x} &= \tilde{A}x - BDu\,, \\ x(0) &= 0\,, \end{aligned} \tag{34}$$

where $\tilde{A} = A - BDB'$ with $D = \operatorname{diag}(a_1, \ldots, a_m)$. (Note that $\tilde{A}$ is Hurwitz.) For the cases $p = 1, 2$ we have the following.

COROLLARY 3. *Let A, B be as in Lemma 2 and σ be as above. Let G_1 and G_2 be, respectively, the L^1- and L^2-gains of the system*

$$\begin{aligned} \dot{x} &= Ax - B\sigma(B'x + u)\,, \\ x(0) &= 0\,. \end{aligned} \tag{35}$$

Let γ_1, γ_2 be, respectively, the L^1- and L^2-gains of (34) *and let $d = \min\{a_1, \ldots, a_m\}$. Then we have*

1. $G_1 \leq (\frac{K^2}{d} + 1)\gamma_1$,
2. $G_2 \leq 2\frac{\sqrt{n}}{d}(K^2 + K)\gamma_2$.

(In the literature, γ_2 is called the "H_∞-norm" of (34) and is usually denoted by $\|W\|_\infty$, where $W(s)$ is the transfer matrix for system (34).)

Proof. For each $u \in L^p([0, \infty), \mathbb{R}^m)$, let x be the solution of (35) corresponding to u. Let $\tilde{x} = B'x + u$.

For the case $p = 1$, consider the derivative of $V = \|x\|^2/2$ along the trajectories of (35). We get

$$\begin{aligned} \dot{V}(x) &= -\tilde{x}'\sigma(\tilde{x}) + u'\sigma(\tilde{x}) \\ &\leq -\tilde{x}'\sigma(\tilde{x}) + K\|u\|\,. \end{aligned}$$

Integrating the above inequality from zero to ∞, we obtain

$$\int_0^\infty \tilde{x}'(s)\sigma(\tilde{x}(s))\,ds \leq K\|u\|_{L^1}. \tag{36}$$

Let

$$v(t) = -\tilde{x}(t) + D^{-1}\sigma(\tilde{x}(t)) + u(t).$$

Then, we have

$$\begin{aligned} \int_0^\infty \|v(s)\|ds &\leq \int_0^\infty \left\{\|D^{-1}\|\,\|D\tilde{x}(s) - \sigma(\tilde{x}(s))\| + \|u(s)\|\right\}ds \\ &\leq \int_0^\infty \left\{\frac{K}{d}\tilde{x}'(s)\sigma(\tilde{x}(s)) + \|u(s)\|\right\}ds \\ &\leq \left(\frac{K^2}{d} + 1\right)\|u\|_{L^1}\,. \end{aligned}$$

Now (35) can be written as

$$\begin{aligned} \dot{x} &= \tilde{A}x - BDv(t)\,, \\ x(0) &= 0\,. \end{aligned}$$

By the definition of γ_1 we have $\|x\|_{L^1} \leq \gamma_1 \|v\|_{L^1} \leq (\frac{K^2}{d} + 1)\gamma_1 \|u\|_{L^1}$. Therefore

$$G_1 \leq \left(\frac{K^2}{d} + 1\right)\gamma_1,$$

and Conclusion 1 is then proved.

Now we show Conclusion 2. Since $\tilde{A}$ is Hurwitz, we take

$$V_2(x) \;=\; \frac{c\|x\|^3}{3} + x'Px,$$

where $c = 2K\|PB\|$ and P is the positive definite symmetric matrix satisfying

$$\tilde{A}'P + P\tilde{A} = -I\,. \tag{37}$$

Then, rewriting (35) as

$$\dot{x} = \tilde{A}x + B\,(D\tilde{x} - \sigma(\tilde{x}) - Du)$$

and proceeding similarly to the proof of Lemma 2, we have

$$\begin{aligned} \dot{V}_2(x) &= -c\|x\|\tilde{x}'\sigma(\tilde{x}) + c\|x\|u'\sigma(\tilde{x}) \\ &\quad -\|x\|^2 + 2x'PB\,(D\tilde{x} - \sigma(\tilde{x}) - Du) \\ &\leq -\|x\|^2 + 2(\|D\| + K^2)\,\|PB\|\,\|x\|\,\|u\|\,. \end{aligned}$$

From this we can get

$$G_2 \leq 2(\|D\| + K^2)\|PB\| \leq 2(K^2 + K)\|PB\|\,. \tag{38}$$

Next we want to compare $\|PB\|$ with γ_2. First, let us compare $\|PBD^{1/2}\|$ with $\widehat{\gamma}_2$, where $\widehat{\gamma}_2$ is the L^2-gain of

$$\begin{aligned} \dot{x} &= \tilde{A}x + BD^{1/2}u\,, \\ x(0) &= 0\,. \end{aligned} \tag{39}$$

Notice that $\widehat{\gamma}_2 \leq \|D^{-1/2}\|\gamma_2$. We now consider the *Hankel norm* $\|W\|_{\text{hankel}}$ for system (39). Note that the matrix P is the *observability Gramian* for (39) (the output is just the state in our case). The *controllability Gramian* for system (39) is defined to be the symmetric matrix $Q \geq 0$ which satisfies

$$\tilde{A}Q + Q\tilde{A}' + BDB' = 0\,. \tag{40}$$

We know that the Hankel norm for (39) is equal to

$$\|W\|_{\text{hankel}} = (\lambda_{\max}(PQ))^{1/2}\,, \tag{41}$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue, cf. [1]. We also know that the H_∞-norm $\widehat{\gamma}_2$ for (39) is related to the Hankel norm by the following inequalities:

$$\widehat{\gamma}_2 \leq (2n+1)\|W\|_{\text{hankel}} \leq (2n+1)\widehat{\gamma}_2\,. \tag{42}$$

Now in our case, since $\tilde{A} = A - BDB'$ and $\tilde{A}' = -A - BDB'$, the controllability Gramian Q is equal to $I/2$. Therefore the Hankel norm for (39) is just

$$\|W\|_{\text{hankel}} = (\lambda_{\max}(P/2))^{1/2} .$$

Since P satisfies

$$(A' - BDB')P + P(A - BDB') + I = PA - AP - BDB'P - PBDB' + I = 0,$$

multiplying both sides by P on the right, we get

$$PAP - APP - BDB'PP - PBDB'P + P = 0. \tag{43}$$

Now taking trace to both sides of (43), we get that

$$\|PBD^{1/2}\|_F^2 = \operatorname{Tr}(P/2).$$

On the other hand we know that $\operatorname{Tr}(P/2)$ is equal to the sum of all the eigenvalues of $P/2$. Therefore $\operatorname{Tr}(P/2) \leq n\lambda_{\max}(P/2)$. Finally we get $\|PB\| \leq \|D^{-1/2}\| \, \|PBD^{1/2}\| \leq \sqrt{n}\|D^{-1/2}\|\lambda_{\max}^{1/2}(P/2) \leq \sqrt{n}\|D^{-1/2}\|\widehat{\gamma}_2 \leq \sqrt{n}\|D^{-1}\|\gamma_2$. Thus

$$G_2 \leq 2\frac{\sqrt{n}}{d}(K^2 + K)\gamma_2 ,$$

and this completes the proof. □

Remark 6. The dimension of the state space does not appear in the bound of the estimate in Conclusion 1 of Corollary 3. We suspect also that the estimate for G_2 should be independent of the dimension of the state space.

5. Nonzero initial states. We now turn to nonzero initial states. We start with an easy observation.

Remark 7. Consider systems as in Theorem 2, but without controls, that is, any system (S) given by $\dot{x} = Ax + B\sigma(Fx)$, where A, B, σ are as in Theorem 2 and F is chosen as in its proof. It is well known that the origin is globally asymptotically stable, assuming for instance controllability of the matrix pair (A, B). It is interesting to see that this fact also can be shown as a consequence of our arguments. From the proof of Theorem 2, it is enough to show that the system $(\widehat{S})$ $\dot{x} = Ax - B\sigma(B'x)$, with A skew–symmetric and (A, B) controllable, is globally asymptotically stable with respect to the origin. But this follows trivially from (28), since we have along the trajectories of $(\widehat{S})$ that $dV_2(x(t))/dt \leq -\|x(t)\|^2$. Thus V_2 is a strict Lyapunov function for this system without controls.

The previous remark suggests the study of relationships between L^p-stability and global asymptotic stability of the origin. We prove below that, even for nonlinear feedback laws, L^p-stability for finite p implies asymptotic stability.

5.1. Relations between state-space stability and L^p-stability. We consider initialized control systems of the type (1). If this system is L^p-stable for some $p \in [1, \infty)$ and if, in addition, f satisfies some growth or regularity assumptions, we are able to draw conclusions regarding the asymptotic behavior of the solutions of

$$\dot{x} = f(x, 0). \tag{44}$$

We next define the various alternative properties of f under which we will be able to obtain several such conclusions:

$(H_{1,p})$: there exist $\alpha \in [0, p]$, $\delta > 0$, $K_1, K_2 \geq 0$, such that for all $x \in \mathbb{R}^n$ with $\|x\| < \delta$ and for all $u \in \mathbb{R}^m$ we have

$$\|f(x,u)\| \leq K_1(\|x\| + \|u\|) + K_2(\|x\|^\alpha + \|u\|^\alpha);$$

$(H_{2,p})$: there exist $\alpha \in [0, p]$, $K_1, K_2 \geq 0$, such that for all $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$ we have

$$\|f(x,u)\| \leq K_1(\|x\| + \|u\|) + K_2(\|x\|^\alpha + \|u\|^\alpha);$$

(H_3) : the function f is differentiable at $(0, 0)$ with $A \stackrel{\text{def}}{=} D_x f(0,0)$ and $B \stackrel{\text{def}}{=} D_u f(0,0)$.

Then we have the following lemma.

LEMMA 4. *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be a locally Lipschitz function. Assume that the system*

$$\dot{x} = f(x,u), \quad x(0) = 0 \tag{45}$$

is L^p-stable for some $p \in [1, \infty)$ with L^p-gain G_p. For each $u \in L^p([0,\infty), \mathbb{R}^m)$, let x_u denote the corresponding solution of (45). *We can make the following conclusions.*

(1) *If f satisfies $(H_{1,p})$, then, for each u, $\lim_{t\to\infty} x_u(t) = 0$.*

(2) *If f satisfies $(H_{2,p})$, then there exists a constant $C > 0$ so that, for each u.*

$$\|x_u\|_{L^\infty} \leq C \max\left(\|u\|_{L^p}, \|u\|_{L^p}^{p/(p+1-\alpha)}\right). \tag{46}$$

(3) *If f satisfies (H_3), then the linearized system*

$$\dot{x} = Ax + Bu, \quad x(0) = 0$$

is L^p-stable with L^p-gain $\gamma_p \leq G_p$ (so, in this case, if (A, B) is controllable, then A must be Hurwitz and the system (44) *is locally exponentially stable).*

Note that if system (45) is L^p-stable, then $f(0,0) = 0$.

Proof. In what follows we write x_u simply as x, when the control is clear from the context.

(1) Assume that the conclusion is not true for some $u \in L^p([0,\infty), \mathbb{R}^m)$. Then there exists $\delta_1 > 0$ so that $\limsup_{t\to\infty} \|x(t)\| \geq 2\delta_1$. Without loss of generality, we may assume that $\delta_1 \leq \min(1, \delta)$.

Take $\varepsilon > 0$ and fix a time $T_0 > 0$ so that

$$\|u\|_{L^p[T_0,\infty)} \leq \varepsilon, \quad \|x\|_{L^p[T_0,\infty)} \leq \varepsilon.$$

Since $\liminf_{t\to\infty} x(t) = 0$, there exist $T_1, T_2 > T_0$ such that

(a) $\frac{\delta_1}{2} \leq \|x(t)\| \leq \delta_1$, for $t \in [T_1, T_2]$;

(b) $\|x(T_2) - x(T_1)\| \geq \frac{\delta_1}{2}$.

Then using $(H_{1,p})$ and applying Hölder's inequality, we obtain

$$\begin{aligned} \frac{\delta_1}{2} \leq \|x(T_2) - x(T_1)\| &\leq \int_{T_1}^{T_2} \|f(x(s), u(s))\| ds \\ &\leq 2K_1\varepsilon(T_2 - T_1)^{(p-1)/p} + 2K_2\varepsilon^\alpha(T_2 - T_1)^{(p-\alpha)/p}, \end{aligned} \tag{47}$$

$$(T_2 - T_1)\left(\frac{\delta_1}{2}\right)^p \leq \int_{T_1}^{T_2} \|x(t)\|^p dt \leq \varepsilon^p. \tag{48}$$

Using (47) and (48), we get

$$\frac{\delta_1}{2} \leq 2\left(\frac{K_1}{(\frac{\delta_1}{2})^{p-1}} + \frac{K_2}{(\frac{\delta_1}{2})^{p-\alpha}}\right)\varepsilon^p.$$

Since ε is arbitrary, we obtain a contradiction.

(2) For each $T > 0$, let $\beta_T = \sup_{t\in[0,T]} \|x(t)\|$ and fix an interval $[T_1, T_2]$ in $[0, T]$ such that

(a) $\frac{\beta_T}{2} \leq \|x(t)\| \leq \beta_T$, for $t \in [T_1, T_2]$;

(b) $\|x(T_2) - x(T_1)\| = \frac{\beta_T}{2}$.

Since $(H_{2,p})$ holds, we obtain, using the L^P-stability of (45) and Hölder's inequality, that

$$\frac{\beta_T}{2} \leq C_1(T_2 - T_1)^{(p-1)/p}\|u\|_{L^p} + C_2(T_2 - T_1)^{(p-\alpha)/p}\|u\|_{L^p}^{\alpha} \tag{49}$$

for appropriate constants C_1, C_2, and

$$(T_2 - T_1)\left(\frac{\beta_T}{2}\right)^p \leq C_3\|u\|_{L^p}^p \tag{50}$$

for some constant $C_3 > 0$. From (49) and (50) we can easily conclude

$$\beta_T \leq C \max\left(\|u\|_{L^p}, \|u\|_{L^p}^{p/(p+1-\alpha)}\right), \tag{51}$$

where $C > 0$ is a constant independent of T. Since T is arbitrary, (46) holds.

(3) For each control u and $\varepsilon \neq 0$, let x_ε be the trajectory of (45) corresponding to εu. Then it is easy to see that $z_\varepsilon(t) \overset{\text{def}}{=} \frac{x_\varepsilon(t)}{\varepsilon}$ converges, for each t as $\varepsilon \to 0$, to the solution $z(t)$ of

$$\dot{z} = Az + Bu, \quad z(0) = 0.$$

We have $\|z_\varepsilon\|_{L^p} \leq G_p\|u\|_{L^p}$. From this we can prove that $\|z\|_{L^p} \leq G_p\|u\|_{L^p}$, which implies that $\gamma_p \leq G_p$; cf. also [18]. □

Remark 8. One can notice that the finiteness of G_p was not used in the proof of Statement (1). Only the fact that inputs in L^p produce state trajectories in L^p is used.

If we assume reachability conditions on (45), together with L^p-stability of the system for some $p \in [1, \infty)$ and a hypothesis as in Lemma 4, we can obtain information on the asymptotic stability of system (45). We will focus on a special class of systems described by (45) and our results are contained in the next lemma.

LEMMA 5. *Let A be an $n \times n$ matrix, B be an $n \times m$ matrix, σ be an $\mathbb{R}^m$-valued S-function, and f be a locally Lipschitz function from $\mathbb{R}^n$ to $\mathbb{R}^m$. We assume that (A, B) is controllable. Consider the system of differential equations*

$$\dot{x} = Ax + B\sigma(f(x)) \tag{52}$$

and the control system

$$\begin{aligned} \dot{x} &= Ax + B\sigma(f(x) + u), \\ x(0) &= 0. \end{aligned} \tag{53}$$

We can make the following conclusions.

(i) *If system* (53) *is L^p-stable for some $p \in [1, \infty)$, then system* (52) *is locally asymptotically stable with respect to the origin*;

(ii) *If the reachable set from zero of* (53) *is equal to $\mathbb{R}^n$ and if system* (53) *is L^p-stable for some $p \in [1, \infty)$, then system* (52) *is globally asymptotically stable with respect to the origin.*

Proof. We first show (i). Note that the system (53) satisfies $(H_{2,p})$ (with $\alpha = 0$). Fix a $u \in L^p([0, \infty), \mathbb{R}^m)$. Let x_u be the solution of (53) corresponding to u. From Lemma 4 we know that $x_u(t) \to 0$ as $t \to \infty$.

To prove stability, we need some elementary reachability results for linear systems. By our assumption we know that the system

$$\dot{x} = Ax + Bu \tag{54}$$

is controllable. Any point $x_0 \in \mathbb{R}^n$ can be reached from zero by trajectories of (54) at time 1. Moreover we can choose a u_{x_0} on $[0, 1]$ that steers zero to x_0 and satisfies $\|u_{x_0}\|_{L^\infty[0,1]} \leq C\|x_0\|$, where $C > 0$ is a constant depending on A, B (cf., e.g., [13]). By a measurable selection it is also true that there is a measurable control v that steers zero to x_0 for the system (S) $\dot{x} = Ax + B\sigma(v)$, provided that x_0 is small enough. Moreover $\|v\|_{L^\infty[0,1]}$ can be made small if $\|x_0\|$ is small. So if we let $u = v(t) - f(x(t))$ on $[0, 1]$, where x is the solution of (S), then u steers zero to x_0 for (S) at time 1. Let U be an open neighborhood of 0. For each $\delta > 0$, let $\theta(\delta) > 0$ be small enough such that, for each x_0 with $\|x_0\| \leq \theta(\delta)$, there exists a u_{x_0} that steers zero to x_0 for (53) with $\|u_{x_0}\|_{L^p[0,1]} < \delta$. If x is the solution of (52) starting at x_0, and if we let $u(t) = u_{x_0}(t)$ on $[0, 1]$ and $u(t) = 0$ on $(1, \infty)$, then the solution x_u of (53) satisfies $x_u(t) = x(t - 1)$ on $[1, \infty)$. By (46) we can take a $\delta > 0$ small enough such that for any x_0 with $\|x_0\| \leq \theta(\delta)$, the solution x of (52) starting at x_0 stays in U. So system (52) is locally stable.

We next show (ii). Local stability follows as in (i). To prove global attraction, note that the reachability assumption implies that any trajectory x of (52) can be seen as a part of a trajectory of (53) corresponding to a control in L^p. Now Lemma 4 provides that $x(t) \to 0$. □

5.2. Dissipation inequality and input-to-state stability. Next we give a slightly different proof of Theorem 2, which results in a weaker statement (we now allow ε to depend on p) but which is somewhat simpler. Moreover, it results in a simple dissipation-type inequality, from which conclusions about nonzero initial states will be evident. We will only sketch the steps, as they parallel those in the previous proofs.

Assume that A is skew–symmetric and $A - BB'$ is Hurwitz. Fix a $1 \leq p < \infty$ first. Let $\tau, a, b, K, V_{0,p}, V_{1,p}$ be as in the proof of Lemma 2. Let

$$\lambda_p = K\|B\|c_p\,,$$
$$\varepsilon_p = \frac{1}{2K\lambda_p}\,.$$

Consider the system

$$\dot{x} = Ax - B\sigma(B'x + u) + \varepsilon_p\theta(v), \tag{55}$$

where the initial states are now arbitrary. Write $\tilde{x}(t) = B'x(t) + u(t)$.

Along the trajectories of (55), we have

$$\begin{aligned}\dot{V}_{0,p}(x(t)) &= -\|x(t)\|^{p-1}\tilde{x}'(t)\sigma(\tilde{x}(t)) \\ &\quad + \|x(t)\|^{p-1}\left(\varepsilon_p x'(t)\theta(v(t)) + u'(t)\sigma(\tilde{x}(t))\right) \\ &\leq -\|x(t)\|^{p-1}\tilde{x}'(t)\sigma(\tilde{x}(t)) \\ &\quad + K\|x(t)\|^{p-1}\|u(t)\| + K\varepsilon_p\|x(t)\|^p. \end{aligned} \tag{56}$$

(Compare this with (22).) Similar to (25) we can get (for $p > 1$)

$$\dot{V}_{1,p}(x(t)) \leq -\|x(t)\|^p + Kc_p\|B\|\,\|x(t)\|^{p-1}\tilde{x}'(t)\sigma(\tilde{x}(t))$$

$$+ c_p K\|x(t)\|^{p-1}\left(\|B\|\,\|u(t)\| + \varepsilon_p\|v(t)\|\right). \tag{57}$$

Again letting $V_p(x) = \lambda_p V_{0,p}(x) + V_{1,p}(x)$, we obtain

$$\dot{V}_p(x(t)) \leq -(1 - K\lambda_p\varepsilon_p)\|x(t)\|^p + \|x(t)\|^{p-1}\left((K+1)\lambda_p\|u(t)\| + c_pK\varepsilon_p\|v(t)\|\right)$$

$$= -\frac{1}{2}\|x(t)\|^p + \|x(t)\|^{p-1}\left((K+1)\lambda_p\|u(t)\| + c_pK\varepsilon_p\|v(t)\|\right).$$

Let

$$\kappa_p = \max\{(K+1)\lambda_p, c_pK\varepsilon_p\}.$$

Thus, for $p > 1$,

$$\dot{V}_p(x(t)) \leq -\frac{1}{2}\|x(t)\|^p + \kappa_p\|x(t)\|^{p-1}(\|u(t)\| + \|v(t)\|). \tag{58}$$

Arguing as in the proof of Lemma 2, we see that this provides L^p-stability provided that $x(0) = 0$. But we also note in this case that it is possible to rewrite (58) in a "dissipation inequality" form, as follows. First, by Young's inequality, we have for any $\alpha, \mu, \nu > 0$ and $p > 1$,

$$\mu^{p-1}\nu \leq \frac{p-1}{p}\alpha^{p/(p-1)}\mu^p + \frac{\nu^p}{p\alpha^p}.$$

Let

$$\alpha_p = \left[\frac{p}{4(p-1)\kappa_p}\right]^{(p-1)/p}.$$

Then (58) can be written as

$$\dot{V}_p(x(t)) \leq -\frac{1}{4}\|x(t)\|^p + \frac{\kappa_p}{p\alpha_p^p}(\|u(t)\| + \|v(t)\|)^p.$$

So if we let $\tilde{V}_p = 4V_p$, $r_p = \frac{4\kappa_p}{p\alpha_p^p}$, we finally conclude, along all solutions of (55),

$$\dot{\tilde{V}}_p(x(t)) \leq -\|x(t)\|^p + r_p(\|u(t)\| + \|v(t)\|)^p. \tag{59}$$

This is sometimes called a *dissipation inequality*; see [7].

Take in particular $p = 2$ and write $V = \tilde{V}_2$. The estimate (59) shows that $V(x(t))$ must decrease if $\|x(t)\|$ is larger than $\sqrt{r_2}$ times the input magnitude. Thus, irrespective of the initial state, the state trajectory is ultimately bounded, assuming that the inputs u and v are bounded, and this asymptotic bound depends on an asymptotic bound on u and v. One way to summarize this conclusion is by means of the estimate

$$\|x(t)\| \leq \beta(\|x(0)\|, t) + \gamma\left(\|(u,v)\|_{L^\infty[0,t]}\right) \tag{60}$$

valid for all $x(0)$, all $t \geq 0$, and all essentially bounded u, v, where γ is a function of class K and β is a class-KL function (that is, $\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is so that for each fixed

$t \geq 0$, $\beta(\cdot, t)$ is a class-K function, and for each fixed $s \geq 0$, $\beta(s, \cdot)$ is decreasing to zero as $t \to \infty$). This is the notion of *ISS stability* discussed in, e.g., [11, 10, 17, 15]; equation (60) is a consequence of (59), which says that V is a *Lyapunov ISS* function. In fact, in our case one can say more about the function γ; namely, it can be taken to be *linear*. Indeed, from the proof in [11, p. 441] one can take any $\gamma \geq \alpha_1^{-1} \circ \alpha_2 \circ \alpha_4$, where $\alpha_4(l) = \sqrt{r_2}l$ and where the α_i's are class-K functions, so that

$$\alpha_1(\|x\|) \leq V(x) \leq \alpha_2(\|x\|)$$

for all $x \in \mathbb{R}^n$. Here we can choose $\alpha_2 = c\alpha_1$, for some $c > 1$, where α_1 is of the form $\alpha(l) = a_1 l^2 + a_2 l^3$ and is thus a convex function. Since for any increasing convex function α and $c > 1$, and any $d > 0$, $\alpha^{-1}(c\alpha(dl)) \leq cdl$ for all l, this gives a linear γ as claimed.

6. More general input nonlinearities. Now we consider a broader class of input nonlinearities, allowing unbounded functions as well. The main result will be extended to this case.

DEFINITION 4. *We call $\Sigma : \mathbb{R} \to \mathbb{R}$ an $\widetilde{S}$-function if it can be written as $\Sigma(t) = \alpha t g(t) + \sigma(t)$, where*

- *$\alpha \geq 0$ is a constant,*
- *$g : \mathbb{R} \to [a, b]$ is measurable and a, b are strictly positive real numbers,*
- *$\sigma : \mathbb{R} \to \mathbb{R}$ is an S-function.*

We say that $\Sigma = (\Sigma_1, \ldots, \Sigma_m)'$ is an $\mathbb{R}^m$-valued $\widetilde{S}$-function if each Σ_i is an $\widetilde{S}$-function. As before if $\xi = (\xi_1, \ldots, \xi_m)' \in \mathbb{R}^m$, then $\Sigma(\xi) = (\Sigma_1(\xi_1), \ldots, \Sigma_m(\xi_m))'$.

With this definition we have the following generalization of Theorem 1.

THEOREM 4. *Let A, B be $n \times n$, $n \times m$ matrices, respectively, and Σ be an $\mathbb{R}^m$-valued $\widetilde{S}$-function. Assume that A is neutrally stable. Then there exists an $m \times n$ matrix F such that the system*

$$\begin{aligned} \dot{x} &= Ax + B\Sigma(Fx + u), \\ x(0) &= 0 \end{aligned} \tag{61}$$

is L^p-stable for all $1 \leq p \leq \infty$.

Proof. As in the proof of Theorem 2, we can assume without loss of generality that A is skew–symmetric and (A, B) is controllable.

Assume that $\Sigma = (\Sigma_1, \ldots, \Sigma_m)'$ with $\Sigma_i(t) = \alpha_i t g_i(t) + \sigma_i(t)$. Let $\sigma = (\sigma_1, \ldots, \sigma_m)'$ and $G = \operatorname{diag}(\alpha_1 g_1, \ldots, \alpha_m g_m)$ with $G(\xi) = \operatorname{diag}(\alpha_1 g_1(\xi_1), \ldots, \alpha_m g_m(\xi_m))$ for $\xi \in \mathbb{R}^m$. Then $\Sigma(\xi) = G(\xi)\xi + \sigma(\xi)$.

The α_i's split into two sets, $\Lambda_1 = \{\alpha_i, \alpha_i > 0\}$ and $\Lambda_2 = \{\alpha_i, \alpha_i = 0\}$. We can assume without loss of generality that

$$\Lambda_1 = \{\alpha_1, \ldots, \alpha_r\} \text{ and } \Lambda_2 = \{\alpha_{r+1}, \ldots, \alpha_m\}, \quad r \leq m.$$

Therefore system (61) becomes

$$\dot{x} = Ax + B \left(\begin{array}{ccc|c} \alpha_1 g_1 & & & \\ & \ddots & \mathrm{O} & \mathrm{O} \\ \mathrm{O} & \ddots & & \\ & & \alpha_r g_r & \\ \hline & \mathrm{O} & & \mathrm{O} \end{array}\right) (Fx + u) + B\sigma(Fx + u).$$

Write $B = (B_1, B_2)$, $F = \binom{F_1}{F_2}$, $u = \binom{u_1}{u_2}$, $\sigma = \binom{\sigma^1}{\sigma^2}$, and let $G_1 = \operatorname{diag}(\alpha_1 g_1, \ldots, \alpha_r g_r)$ with $G_1(\xi) = \operatorname{diag}(\alpha_1 g_1(\xi_1), \ldots, \alpha_r g_r(\xi_r))$ for $\xi \in \mathbb{R}^r$. The sizes of the matrices B_1, B_2, F_1, F_2

are, respectively, $n \times r$, $n \times (m-r)$, $r \times n$, $(m-r) \times n$. As for u_1, u_2, they are, respectively, elements of $\mathbb{R}^r$ and $\mathbb{R}^{m-r}$. The S-functions σ^1, σ^2 are, respectively, $\mathbb{R}^r$- and $\mathbb{R}^{m-r}$-valued. We rewrite (61) as

$$\begin{aligned} \dot{x} &= Ax + B_1 G_1(F_1 x + u_1)(F_1 x + u_1) \\ &\quad + B_1 \sigma^1(F_1 x + u_1) + B_2 \sigma^2(F_2 x + u_2), \\ x(0) &= 0. \end{aligned} \tag{62}$$

Let $R(A, B_1) : \mathbb{R}^{rn} \to \mathbb{R}^n$ be the reachability matrix of (A, B_1). (Here and below we will identify matrices with the corresponding linear maps.)

Let $D = \mathrm{Im}R(A, B_1)$ and $H = D^{\perp}$. We have $D \oplus H = \mathbb{R}^n$. Clearly the subspace D is invariant under A and $\mathrm{Im}(B_1) \subseteq D$. Since A is skew–symmetric, the subspace H is also invariant under A. So there exists an orthogonal $n \times n$ matrix U such that

$$UAU' = \begin{pmatrix} A_1 & O \\ O & A_2 \end{pmatrix}, \tag{63}$$

where A_1 and A_2 are skew–symmetric and are restrictions of A to G and H, respectively. So, up to an orthonormal change of basis, we can assume that A is already of the form (63). According to this decomposition, $D = \mathrm{Im}\ R(A, B_1)$. Let $s = \dim D = \mathrm{rank}\ R(A, B_1)$. Consider now

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad B_1 = \begin{pmatrix} B_{11} \\ B_{12} \end{pmatrix}, \qquad B_2 = \begin{pmatrix} B_{21} \\ B_{22} \end{pmatrix},$$

$$F_1 = (F_{11},\ F_{12}), \qquad F_2 = (F_{21},\ F_{22}).$$

Here, $x_1 \in \mathbb{R}^s$, $x_2 \in \mathbb{R}^{n-s}$ and the sizes of B_{11}, B_{12}, B_{21}, B_{22} and F_{11}, F_{12}, F_{21}, F_{22} are, respectively, $s \times r$, $(n-s) \times r$, $s \times (m-r)$, $(n-s) \times (m-r)$ and $r \times s$, $r \times (n-s)$, $(m-r) \times s$, $(m-r) \times (n-s)$.

Since $\mathrm{Im}B_1 \subset D$, we have $B_{12} = 0$. Now system (62) becomes

$$\begin{aligned} \dot{x}_1 &= A_1 x_1 + B_{11} G_1(F_{11} x_1 + F_{12} x_2 + u_1)(F_{11} x_1 + F_{12} x_2 + u_1) \\ &\quad + B_{11} \sigma^1(F_{11} x_1 + F_{12} x_2 + u_1) + B_{21} \sigma^2(F_{21} x_1 + F_{22} x_2 + u_2), \\ \dot{x}_2 &= A_2 x_2 + B_{22} \sigma^2(F_{21} x_1 + F_{22} x_2 + u_2)\ . \end{aligned}$$

Choose now $F_{12} = F_{21} = 0$, $F_{11} = -B_{11}'$, and $F_{22} = -B_{22}'$. We obtain

$$\begin{aligned} \dot{x}_1 &= \left(A_1 - B_{11} G_1(-B_{11}' x_1 + u_1) B_{11}'\right) x_1 + B_{11} G_1(-B_{11}' x_1 + u_1) u_1 \\ &\quad + B_{11} \sigma^1(-B_{11}' x_1 + u_1) + B_{21} \sigma^2(-B_{22}' x_2 + u_2), \\ \dot{x}_2 &= A_2 x_2 + B_{22} \sigma^2(-B_{22}' x_2 + u_2). \end{aligned}$$

In the above system, replacing $\sigma(\cdot)$ by $-\sigma(-\cdot)$ (still denoted by σ), the system becomes

$$\begin{aligned} \dot{x}_1 &= \left(A_1 - B_{11} G_1(-B_{11}' x_1 + u_1) B_{11}'\right) x_1 + B_{11} G_1(-B_{11}' x_1 + u_1) u_1 \\ &\quad - B_{11} \sigma^1(B_{11}' x_1 - u_1) - B_{21} \sigma^2(B_{22}' x_2 - u_2), \\ \dot{x}_2 &= A_2 x_2 - B_{22} \sigma^2(B_{22}' x_2 - u_2). \end{aligned}$$

Since (A, B) is controllable, (A_2, B_{22}) is also controllable. It follows from Theorem 2 that the x_2-subsystem is L^p-stable for all $1 \le p \le \infty$. So there exists $C_p^1 > 0$ such that $\|x_2\|_{L^p} \le C_p^1 \|u_2\|_{L^p}$.

For $i = 1, \ldots, r$, let $d_i(t) = \sigma_i^1(t)/t$, if $t \neq 0$, and $d_i(t) = 0$, if $t = 0$. Let $\tilde{G}_1(\xi) = \operatorname{diag}(d_1(\xi_1), \ldots, d_r(\xi_r))$. Then we can rewrite the x_1-subsystem as

$$\dot{x}_1 = [A_1 - B_{11}(G_1(-B_{11}'x_1 + u_1) + \tilde{G}_1(B_{11}'x_1 - u_1))B_{11}']x_1 + v,$$

where

$$v = B_{11}G_1(-B_{11}'x_1 + u_1)u_1 + B_{11}\tilde{G}_1(B_{11}'x_1 - u_1)u_1 - B_{21}\sigma^2(B_{22}'x_2 - u_2).$$

We have

$$\|v\| \leq C(\|u_1\| + \|x_2\| + \|u_2\|)$$

for some $C > 0$.

If we let $\widetilde{D}(t) = G_1(-B_{11}'x_1(t) + u_1(t)) + \tilde{G}_1(B_{11}'x_1(t) - u_1(t))$, then the above equation can be written as

$$\dot{x}_1(t) = (A_1 - B_{11}\widetilde{D}(t)B_{11}')x_1(t) + v(t)\,.$$

By definition of an S-function and an $\widetilde{\mathrm{S}}$-function, there exist two real numbers δ_1 and δ_2 such that $0 < \delta_1 \leq \delta_2$, and if we write $\widetilde{D}(t) = \operatorname{diag}(\tilde{d}_1(t), \cdots, \tilde{d}_r(t))$, then

$$\delta_1 \leq \tilde{d}_i(t) \leq \delta_2$$

for $i = 1, \ldots, r$. Since (A, B) is controllable, (A_1, B_{11}) is controllable too. Then it follows from Corollary 2 that

$$\|x_1\|_{L^p} \leq \bar{C}_p^2 \|v\|_{L^p}\,,$$

for some $\bar{C}_p^2 > 0$ depending on A_1, B_{11}, δ_1, δ_2, and p. But we know that

$$\|v\|_{L^p} \leq C(\|u_1\|_{L^p} + \|u_2\|_{L^p} + \|x_2\|_{L^p}) \leq C\|u_1\|_{L^p} + C(1 + C_p^1)\|u_2\|_{L^p}\,.$$

Therefore we have $\|x_1\|_{L^p} \leq C_p^2 \|u\|_{L^p}$ for some constant $C_p^2 > 0$. □

7. Counterexample: The *n*th order scalar integrator. The next result is a negative one, and it concerns systems such as those in equation (2), except that the matrix A is not neutrally stable but instead is assumed to have a nonsimple Jordan block for the zero eigenvalue. In that case, we show that for any possible F which stabilizes the corresponding linear control system

$$\begin{aligned} \dot{x} &= Ax + B(Fx + u)\,, \\ x(0) &= 0\,, \end{aligned}$$

the resulting system (Σ_u) is *not* in general L^p-stable for any $1 \leq p \leq \infty$. We first consider the simplest case, namely the double integrator. The proof is of interest because the origin of the corresponding system without inputs (but *with* the saturation) *is* globally asymptotically stable. Thus the result is quite surprising. In the end we discuss the n-integrator for $n \geq 3$.

PROPOSITION 1. *Let $1 \leq p \leq \infty$. Consider the following 2-dimensional initialized control system:*

$$(S_{a,b}) \qquad \begin{aligned} \dot{x} &= y\,, \\ \dot{y} &= -\sigma(ax + by + u)\,, \\ x(0) &= y(0) = 0\,, \end{aligned}$$

where $a, b > 0$, σ is a scalar S-function, and inputs u belong to $L^p([0, \infty), \mathbb{R})$. Then $(S_{a,b})$ is not L^p-stable.

Proof. Up to a reparameterization of the time and a linear change of variables, it is enough to show that the initialized control system

$$\begin{aligned} \dot{x} &= y\,, \\ \dot{y} &= -\lambda\sigma(x + y + u)\,, \\ x(0) &= y(0) = 0\,, \end{aligned}$$

where $\lambda > 0$, is not L^p-stable. Now replacing $\lambda\sigma$ by σ (note $\lambda\sigma$ is still an S-function) we may assume that $\lambda = 1$. Therefore all we need to show is that the system

$$(S) \qquad \begin{aligned} \dot{x} &= y\,, \\ \dot{y} &= -\sigma(x + y + u)\,, \\ x(0) &= y(0) = 0 \end{aligned}$$

is not L^p-stable. The proof is quite technical, but the idea is not difficult to understand. It is based on the fact that the feedback $u = -y$ makes the system (S) have periodic trajectories, with a control u whose norm is proportional to that of the y-coordinate. But the x-coordinate is the integral of y, so the ratio between the p-norms of x and u can be made to be large for $p < \infty$. (For $p = \infty$, one modifies the argument to reach states of large magnitude.)

Let us first fix a p in $[1, \infty)$. Assume that (S) is L^p-stable. Then the following holds: there exists $C_p > 0$ so that, if $u \in L^p([0, \infty), \mathbb{R})$, then

$$\|y_u^2\|_{L^p} \leq C_p \, \|u\|_{L^p}, \tag{64}$$

where y_u is the second coordinate of (x_u, y_u), the solution of (S) associated to u.

To see this, let $q = 2(p - 1) \geq 0$ and

$$V_q(x, y) = -\frac{xy|y|^q}{q + 1}\,.$$

Then along the trajectory (x_u, y_u) of (S) we have

$$\dot{V}_q = -\frac{1}{q + 1}\,|y_u|^{q+2} + x_u\sigma(x_u + y_u + u)|y_u|^q.$$

Therefore,

$$\dot{V}_q + \frac{1}{q + 1}\,|y_u|^{q+2} \leq K|x_u||y_u|^q\,, \tag{65}$$

where K is an S-bound for σ. From Lemma 4 we know that $\lim_{t\to\infty}(x_u, y_u) = (0, 0)$. Integrating (65) from zero to t and letting $t \to \infty$, we end up with

$$\frac{1}{q + 1}\int_0^\infty |y_u|^{q+2} \leq K\int_0^\infty |x_u||y_u|^q.$$

Therefore, if $p = 1$, we get that $\|y_u^2\|_{L^1} \leq K\|x\|_{L^1}$. If $p > 1$, applying Hölder's inequality, we get

$$\frac{1}{q + 1}\int_0^\infty |y_u|^{2p} \leq K\|x_u\|_{L^p}\left(\int_0^\infty |y_u|^{qp/(p-1)}\right)^{(p-1)/p}.$$

But $q\frac{p}{p-1} = 2(p-1)\frac{p}{p-1} = 2p$. Therefore

$$\|y_u^2\|_{L^p} \leq (2p-1)K\|x_u\|_{L^p}. \tag{66}$$

Since (S) is L^p-stable, $\|x_u\|_{L^p} \leq G_p\|u\|_{L^p}$, where G_p is the L^p-gain of (S). So (64) indeed holds.

Now we construct trajectories of (S) which contradict (64).

We consider the level sets of the following Lyapunov function:

$$V(x, y) = y^2 + G(x),$$

where $G(x) = 2\int_0^x \sigma(s)ds$.

Let $\rho_1 = 2\inf_{|t|\geq 1}|\sigma(t)| > 0$ and define $H : \mathbb{R} \to \mathbb{R}$ by

$$H(x) = \begin{cases} 0 & \text{if } |x| \leq 1, \\ \rho_1(|x|-1) & \text{if } |x| > 1. \end{cases}$$

We have

$$y^2 + H(x) \leq V(x, y) \leq y^2 + 2K|x|. \tag{67}$$

Note that along trajectories of

$$(\widehat{S}) \qquad \begin{aligned} \dot{x} &= y, \\ \dot{y} &= -\sigma(x), \end{aligned}$$

V is constant.

Let us fix a constant $V_0 \geq \max\{1, 2K\}$ and let $x^- < 0$ and $x^+ > 0$ be such that $G(x^+) = G(x^-) = V_0$. Since (S) is controllable, there exist a $T_1 > 0$ and a u_0 in $L^p([0, T_1], \mathbb{R})$ such that $\big(x_{u_0}(T_1), y_{u_0}(T_1)\big) = (0, \sqrt{V_0})$. We can also assume that $u_0(t) = 0$ for $t > T_1$. For $t \geq 0$, consider $(\bar{x}_0(t), \bar{y}_0(t))$, the solution of $(\widehat{S})$ with $(\bar{x}_0(0), \bar{y}_0(0)) = (0, \sqrt{V_0})$. Note that $V(\bar{x}_0(t), \bar{y}_0(t)) \equiv V_0$. Clearly this trajectory is periodic, since it lies in the closed curve $V(x) \equiv V_0$ and there are no equilibria there. Assume that the period is T.

Consider the sequence $\{u_n\}_{n=1}^{\infty}$ of inputs defined as follows:

$$u_n(t) = \begin{cases} u_0(t) & \text{on } [0, T_1], \\ -\bar{y}_0(t - T_1) & \text{on } (T_1, T_1 + nT], \\ 0 & \text{on } (T_1 + nT, \infty). \end{cases}$$

Then if (x_n, y_n) denotes the solution of (S) associated to u_n, we have for $t \in [T_1, T_1 + nT]$,

$$(x_n(t), y_n(t)) = (\bar{x}_0(t - T_1), \bar{y}_0(t - T_1)).$$

In this case (note that $y_n(t) = y_{u_0}(t)$ for $t \in [0, T_1]$ and $y_n(t) = y_{u_0}(t - nT)$ for $t \in [T_1 + nT, \infty)$)

$$\begin{aligned} \int_0^{\infty} |u_n(s)|^p\, ds &= \int_0^{T_1} |u_0(s)|^p\, ds + n\int_0^T |\bar{y}_0(s)|^p\, ds, \\ \int_0^{\infty} |y_n^2(s)|^p\, ds &= \int_0^{\infty} |y_{u_0}^2(s)|^p\, ds + n\int_0^T |\bar{y}_0^2(s)|^p\, ds. \end{aligned}$$

We conclude that

$$\lim_{n\to\infty} \frac{\|y_n^2\|_{L^p}}{\|u_n\|_{L^p}} = \frac{\left(\int_0^T |\bar{y}_0^2(s)|^p ds\right)^{1/p}}{\left(\int_0^T |\bar{y}_0(s)|^p ds\right)^{1/p}} \stackrel{\text{def}}{=} L_{p,V_0}.$$

According to (64), this quotient should be bounded independently of the choice of V_0. We next derive a contradiction by showing that this is not so.

Notice that for any $r \geq 1$, since $\dot{\bar{x}}_0(t) = \bar{y}_0(t)$, we have

$$\int_0^T |\bar{y}_0(s)|^r ds = \int_0^T |\bar{y}_0(s)|^{r-1} |\dot{\bar{x}}_0(s)| ds.$$

Since $V(x, y) = V(x, -y)$, we have

$$\int_0^T |\bar{y}_0(s)|^r ds = \int_0^T |\bar{y}_0(s)|^{r-1} |\dot{\bar{x}}_0(s)| ds = 2 \int_{x^-}^{x^+} |\bar{y}(x)|^{r-1} dx\,, \tag{68}$$

where $|\bar{y}(x)| = \sqrt{V_0 - G(x)}$ for x between x^- and x^+. (Note that the curve $V(x) = V_0$ can be written as the union of the graphs of the functions $y(x) = \pm\sqrt{V_0 - G(x)}$. Thus we can reparameterize the orbit in each of these two parts in terms of the variable x.)

Considering (67), we have $V_0/(2K) \leq |x^-|, x^+ \leq V_0/\rho_1 + 1$. Then it follows from (68) that

$$\int_0^T |\bar{y}_0(s)|^p ds \leq 2V_0^{(p-1)/2}(x^+ - x^-) \leq 4V_0^{(p-1)/2}(V_0/\rho_1 + 1) \leq C_1 V_0^{(p+1)/2}\,,$$

$$\int_0^T |\bar{y}_0^2(s)|^p ds \geq 4 \int_0^{V_0/(2K)} (V_0 - 2Kx)^{p-1/2}\, dx \geq C_2 V_0^{p+1/2}\,,$$

where $C_1, C_2 > 0$ are some constants. Finally, we get $L_{p,V_0} \geq C V_0^{1/2}$ for some $C > 0$. But according to (64), $L_{p,V_0} \leq C_p$. Therefore, for V_0 large enough we get a contradiction. So (S) cannot be L^p-stable for $1 \leq p < \infty$.

We still need to establish the special case $p = \infty$. We use again the level sets of V. Let u^0 on $[0, T_0]$ for some $T_0 > 0$ be an input such that $(x_{u^0}(T_0), y_{u^0}(T_0)) = (0, \sqrt{V_0})$, for some $V_0 > 0$ which will be fixed below.

From $(0, \sqrt{V_0})$, follow the trajectory of

$$(I) \qquad \begin{aligned} \dot{x} &= y, \\ \dot{y} &= \rho_2, \end{aligned}$$

on $[T_0, T_0 + 1]$, where $\rho_2 = -\sigma(-1) > 0$. The trajectory (x, y) of (I), hence, reaches $(\sqrt{V_0} + \rho_2/2, \sqrt{V_0} + \rho_2)$. Let

$$V_1 = (\sqrt{V_0} + \rho_2)^2 + G(\sqrt{V_0} + \rho_2/2) \geq V_0 + 2\rho_2\sqrt{V_0}.$$

Note that also $V_1 \leq V_0 + C(\sqrt{V_0} + 1)$ for some $C > 0$. Furthermore, the trajectory of (I) can be viewed as a trajectory of (S) with $u^1(t) = -1 - x(t) - y(t)$ for $T_0 < t \leq T_0 + 1$. Let

$$u_1 = -u^1(T_0 + 1) = 1 + \sqrt{V_0} + \rho_2/2 + \sqrt{V_0} + \rho_2 = 2\sqrt{V_0} + 3/2\rho_2 + 1.$$

Then, for $T_0 + 1 < t \leq T_1$, follow the trajectory $(\bar{x}, \bar{y})$ of $(\widehat{S})$ from $(\sqrt{V_0} + \rho_2/2, \sqrt{V_0} + \rho_2)$ until the resulting trajectory reaches $(0, \sqrt{V_1})$ at $t = T_1$. This trajectory can also be considered as a trajectory of (S) with $u^1(t) = -\bar{y}(t)$ on $(T_0 + 1, T_1]$. Note that $|u^1(t)| \leq \sqrt{V_1}$ for $T_0 + 1 < t \leq T_1$. Fix V_0 such that $\sqrt{V_1} \leq u_1 \leq 3\sqrt{V_0}$. It is clear that on $[T_0, T_1]$, $|u^1(t)| \leq u_1$.

If we iterate the above construction, we can build three sequences $\{V_n\}_{n=0}^{\infty}$, $\{u_n\}_{n=1}^{\infty}$, and $\{T_n\}_{n=0}^{\infty}$ such that

(1) $V_{n+1} = (\sqrt{V_n} + \rho_2)^2 + G(\sqrt{V_n} + \rho_2/2) \geq V_n + 2\rho_2\sqrt{V_n}$;
(2) $u_n = 2\sqrt{V_{n-1}} + 3/2\rho_2 + 1 \leq 3\sqrt{V_{n-1}}$;
(3) on $[T_n, T_{n+1}]$, there exists an input u^n such that $\sup\{|u^n(t)| : t \in [T_n, T_{n+1}]\} = u_n$ and the trajectory of (S) associated to u^n goes from $(0, \sqrt{V_n})$ to $(0, \sqrt{V_{n+1}})$.

Clearly $\lim_{n\to\infty} V_n = \infty$ and then $\lim_{n\to\infty} u_n = \infty$. Furthermore, let $x_n^- < 0$ be such that $G(x_n^-) = V_n$. Then $|x_n^-| \geq 1/(2K)V_n$ for n large enough, which implies that $\lim_{n\to\infty} |x_n^-| = \infty$.

Let $\{\bar{u}^n\}_{n=0}^{\infty}$ be the sequence of inputs which is equal to the concatenation of $u^0, u^1, \ldots, u^n$ on $[0, T_n]$ and zero for $t > T_n$. For n large enough, we have

$$\|(x_{\bar{u}^n}, y_{\bar{u}^n})\|_\infty \geq |x_n^-|,$$
$$\|\bar{u}^n\|_\infty = u_n .$$

Since $|x_n^-/u_n| \geq 1/(2K)\sqrt{V_n}$ for n large enough, (S) is not L^∞-stable. □

For n integrators and $n > 2$, the proof that L^p-stabilization is not possible is simpler (but the result is far less interesting). We can argue as follows. Let σ be a scalar S-function. It was proved in [4, 21] that, if $n \geq 3$, the n-integrator

$$\begin{aligned} \dot{x}_1 &= x_2 , \\ &\vdots \\ \dot{x}_{n-1} &= x_n , \\ \dot{x}_n &= -\sigma(u) \end{aligned}$$

is not globally asymptotically stabilizable by any possible linear feedback. With this, it follows from Lemma 5 that, if $n \geq 3$, the system

$$\begin{aligned} \dot{x}_1 &= x_2 , \\ &\vdots \\ \dot{x}_{n-1} &= x_n , \\ \dot{x}_n &= -\sigma(Fx + u) , \\ x(0) &= 0 \end{aligned}$$

is not L^p-stable for any $1 \leq p < \infty$ and any row vector F.

Acknowledgment. We wish to thank Malcolm C. Smith for asking questions that led directly to the problems studied in this paper.

REFERENCES

[1] S. P. Boyd and C. H. Barratt, *Linear Controller Design, Limits of Performance*, Prentice–Hall, Englewood Cliffs, NJ, 1991.
[2] Y. Chitour, W. Liu, and E. Sontag, *On the continuity and incremental-gain properties of certain saturated linear feedback loops*, Internat. J. Robust Nonlinear Control, 5 (1995), pp. 413–440.
[3] J. C. Doyle, T. T. Georgiou, and M. C. Smith, *The parallel projection operators of a nonlinear feedback system*, in Proc. IEEE Conf. Dec. and Control, Tucson, AZ, IEEE Publications, Piscataway, NJ, 1992, pp. 1050–1054.
[4] A. T. Fuller, *In the large stability of relay and saturated control systems with linear controllers*, Internat. J. Control, 10 (1969), pp. 457–480.
[5] P.-O. Gutman and P. Hagander, *A new design of constrained controllers for linear systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 22–23.

[6] M. HAUTUS, *(A, B)-Invariant and stabilizability subspaces, a frequency domain description*, Automatica J. IFAC, 16 (1980), pp. 703–707.
[7] D. J. HILL, *Dissipative nonlinear systems: Basic properties and stability analysis*, in Proc. 31st IEEE Conf. Dec. and Control, Tucson, AZ, IEEE Publications, Piscataway, NJ, 1992, pp. 3259–3264.
[8] Z. LIN AND A. SABERI, *Semiglobal Exponential Stabilization of Linear Systems Subject to "Input Saturation" via Linear Feedbacks*, preprint, Washington State Univ., Pullman, WA, 1993.
[9] M. SLEMROD, *Feedback stabilization of a linear control system in Hilbert space*, Math. Control Signals Systems, 2 (1989), pp. 265–285.
[10] L. PRALY AND Z.-P. JIANG, *Stabilization by output feedback for systems with ISS inverse dynamics*, Systems Control Lett., 21 (1993), pp. 19–34.
[11] E. D. SONTAG, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.
[12] ———, *An algebraic approach to bounded controllability of nonlinear systems*, Internat. J. Control, 39 (1984), pp. 181–188.
[13] ———, *Mathematical Control Theory*, Springer-Verlag, New York, 1990.
[14] E. D. SONTAG AND H. J. SUSSMANN, *Nonlinear output feedback design for linear systems with saturating controls*, in Proc. IEEE Conf. Dec. and Control, Honolulu, IEEE Publications, Piscataway, NJ, 1990, pp. 3414–3416.
[15] E. D. SONTAG AND Y. WANG, *On characterizations of the input-to-state stability property*, Systems Control Lett., 24 (1995), pp. 351–359.
[16] A. R. TEEL, *Global stabilization and restricted tracking for multiple integrators with bounded controls*, Systems Control Lett., 18 (1992), pp. 165–171.
[17] J. TSINIAS, *Sontag's 'input to state stability condition' and global stabilization using state detection*, Systems Control Lett., 20 (1993), pp. 219–226.
[18] A. J. VAN DER SCHAFT, *L^2-gain analysis of nonlinear systems and nonlinear state feedback H_∞-control*, IEEE Trans. Automat. Control, 37 (1992), pp. 770–784.
[19] J. C. WILLEMS, *The Analysis of Feedback Systems*, MIT Press, Cambridge, MA, 1971.
[20] Y. YANG, H. J. SUSSMANN, AND E. D. SONTAG, *Stabilization of linear systems with bounded controls*, IEEE Trans. Automat. Control, 39 (1994), pp. 2411–2425.
[21] Y. YANG AND H. J. SUSSMANN, *On the stabilizability of multiple integrators by means of bounded feedback controls*, in Proc. IEEE Conf. Dec. and Control, Brighton, UK, IEEE Publications, Piscataway, NJ, 1991, pp. 70–73.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1220–1234, July 1996

006

ON SOME RELATIONS BETWEEN CHANEY'S GENERALIZED SECOND-ORDER DIRECTIONAL DERIVATIVE AND THAT OF BEN-TAL AND ZOWE*

L. R. HUANG† AND K. F. NG‡

Abstract. For a locally Lipschitz real-valued function f on $\mathbb{R}^n$ and x, u in $\mathbb{R}^n$ our main result implies that if x^* is in Clarke's subdifferential $\partial f(x)$ "coming from the direction u" (in Chaney's sense) such that $x^*(u)$ equals the directional derivative $f'(x; u)$, then Chaney's second-order directional derivative $f''(x; x^*, u)$, when it exists, coincides with the value at x^* of the conjugate function of the Ben-Tal–Zowe second-order directional derivative, provided that this value is finite.

Key words. nonsmooth analysis, generalized second-order directional derivative, conjugate function, subdifferential

AMS subject classifications. Primary, 49J52, 26B25, 26A27; Secondary, 58C20

1. Introduction.

Throughout this paper we consider a locally Lipschitz real-valued function f on $\mathbb{R}^n$. In connection with second-order nonsmooth optimization problems, different kinds of generalized second-order directional derivatives have been introduced, among which are $D^2 f(x; u, w)$ and $f''(x; x^*, u)$, respectively, due to Ben-Tal and Zowe [2, 3] and Chaney [4–8] (their definitions are given in §§2 and 3). Since they do not always exist, one can consider f''_-, f''_+ as in [4–8] and $D^2_- f$, $D^2_+ f$. Our definitions for D^2_-, D^2_+ given in §3 are distinct from those given by Penot [18] and Studniarski [24]. In their framework, some relations between the second-order directional derivative of Ben-Tal and Zowe and the second-order epiderivative of Rockafellar were recently obtained by Penot [19]. As noted in [14] our approach has the advantage that if $D^2_- f = D^2_+ f$ at $(x; u, w)$, then $f'(x; u)$ exists, and our definition degenerates to that of Ben-Tal and Zowe [3]. In the case where f is C^2, it is well known and easy to verify that the second-order derivatives of Chaney and Ben-Tal–Zowe are given by

$$f''(x; x^*, u) = \frac{1}{2}\langle u, \nabla^2 f(x)u\rangle,$$

$$D^2 f(x; u, w) = \frac{1}{2}\langle u, \nabla^2 f(x)u\rangle + x^*(w),$$

where $x^* = \nabla f(x)$ (see (2.1) and (3.5)). This implies the following relationship between the two derivatives:

$$f''(x; x^*, u) = D^2 f(x; u, w) - x^*(w), \quad \forall w. \tag{1.1}$$

This relationship persists in some nonsmooth cases; for example, (1.1) is true if $\partial_u f(x) = \{x^*\}$ and $f''(x; x^*, u)$ exists (see Definition 2.2 and Corollary 4.2). But (1.1) may fail to hold in general. A simple max-function given in §6 shows that the right-hand side of (1.1) may not be a constant function (of w); indeed it may be a (convex and) nonaffine function of w. Based on the works of Ben-Tal and Zowe [3] the following weaker relation was established by Chaney [5]:

$$f''(x; x^*, u) = \inf\{D^2 f(x; u, w) - x^*(w) : w \in \mathbb{R}^n\} (\in \mathbb{R}) \tag{1.2}$$

*Received by the editors April 22, 1994; accepted for publication (in revised form) February 25, 1995. This research was supported by the Research Grant Council of Hong Kong, the Institute of Mathematical Sciences, Chinese University of Hong Kong, and the United College.

†Department of Mathematics, South China Normal University, Guangzhou, People's Republic of China.

‡Department of Mathematics, Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.

for a very special class of nonsmooth functions. Here we show in Theorem 4.5 that (1.2) holds in general provided that both sides of (1.2) exist in $\mathbb{R}$. When f'', $D^2 f$ do not exist, some (inequality) relationships for f''_+, f''_-, D^2_+, and D^2_- are also established. In addition to the references cited above, studies of second-order directional derivatives (based on different definitions or approach) have also been made in [10–13, 16, 20–22].

We end this section with a few standard definitions. Recall that Clarke's generalized upper directional derivative $f^0(x; u)$ of f at x in the direction u and subdifferential $\partial f(x)$ of f at x are defined respectively [9] by

$$f^0(x; u) := \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{1}{t}\{f(y + tu) - f(y)\}$$

and

$$\partial f(x) := \{x^* \in \mathbb{R}^n : x^*(u) \leq f^0(x; u) \text{ for any } u \in \mathbb{R}^n\},$$

while the lower and upper Dini directional derivatives of f at x in the direction u are defined by

$$D_- f(x; u) := \liminf_{t \downarrow 0} \frac{1}{t}\{f(x + tu) - f(x)\}$$

and

$$D_+ f(x; u) := \limsup_{t \downarrow 0} \frac{1}{t}\{f(x + tu) - f(x)\}.$$

If $D_- f(x; u) = D_+ f(x; u)$, then the common value is denoted by $f'(x; u)$, i.e., the directional derivative of f at x in the direction u. We denote the open and closed balls centered at x with radius δ by

$$B(x, \delta) := \{y \in \mathbb{R}^n : \|y - x\| < \delta\} \quad \text{and} \quad B[x, \delta] := \{y \in \mathbb{R}^n : \|y - x\| \leq \delta\},$$

respectively.

2. Chaney's generalized second-order directional derivatives. To begin, let us recall a few definitions from Chaney [5].

DEFINITION 2.1. *Let u be a vector in $\mathbb{R}^n$. Suppose that the sequence (x_k) in $\mathbb{R}^n$ converges to x. We say that (x_k) converges to x in the direction u, denoted by $(x_k) \to_u x$, if $x_k \neq x$ and the sequence $\|u\|((x_k - x)/\|x_k - x\|)$ converges to u.*

DEFINITION 2.2. *Let u be a vector in $\mathbb{R}^n$. Define the subset $\partial_u f(x)$ of $\mathbb{R}^n$ by*

$$\partial_u f(x) := \{x^*: \text{there exist sequences } (x_k) \text{ and } x_k^* \in \partial f(x_k) \text{ such that } (x_k) \to_u x \text{ and } (x_k^*) \to x^*\}.$$

Thus, $\partial_u f(x)$ is nonempty and $\partial_u f(x) \subseteq \partial f(x)$ since the multifunction ∂f is upper semi-continuous [9]; *loosely speaking, $\partial_u f(x)$ consists of all those x^* in $\partial f(x)$ coming from the direction u.*

DEFINITION 2.3. *Let u be a vector in $\mathbb{R}^n$. Suppose that $x^* \in \partial_u f(x)$. Then $f''_-(x; x^*, u)$ is defined to be the infimum of all (extended real) numbers*

$$\liminf_{k \to \infty} \frac{1}{t_k^2}\{f(x_k) - f(x) - x^*(x_k - x)\}$$

taken over all triples of sequences (x_k), (x_k^*), and (t_k) for which

(a) $t_k > 0$ *for each* k *and* (x_k) *converges to* x,
(b) (t_k) *converges to* 0 *and* $((x_k - x)/t_k)$ *converges to* u,
(c) (x_k^*) *converges to* x^* *with* x_k^* *in* $\partial f(x_k)$ *for each* k.

Similarly, $f''_+(x; x^*, u)$ *is defined by the supremum of all (extended real) numbers*

$$\limsup \frac{1}{t_k^2}\{f(x_k) - f(x) - x^*(x_k - x)\},$$

taken over all triples of sequences (x_k), (x_k^*), *and* (t_k) *for which* (a), (b), *and* (c) *all hold. Clearly,* $f''_-(x; x^*, u) \le f''_+(x; x^*, u)$. *Further, if* $f''_-(x; x^*, u) = f''_+(x; x^*, u)$, *then we denote this common value by* $f''(x; x^*, u)$ *and call it Chaney's generalized second-order directional derivative of* f *at* x, x^* *in the direction* u. *It is easy to see that if* f *is a* C^2*-function, then*

$$f''(x; \nabla f(x), u) = \frac{1}{2} u \cdot \nabla^2 f(x) u. \tag{2.1}$$

The following lemma will be used in §4.

LEMMA 2.4. *Let* $x, u \in \mathbb{R}^n$ *and* $g, h: \mathbb{R}^n \to \mathbb{R}$ *be locally Lipschitz at* x. *Let* $f = g + h$. *Then one has*

(i) $x^* \in \partial_u f(x)$ *if and only if* $-x^* \in \partial_u(-f)(x)$. *Moreover, if* $x^* \in \partial_u f(x)$, *then*

$$(-f)''_-(x; -x^*, u) = -f''_+(x; x^*, u).$$

(ii) $\partial_u f(x) \subseteq \partial_u g(x) + \partial_u h(x)$. *Equality holds if* $\partial_u h(x)$ *is a singleton.*
(iii) *The set* $\partial_u h(x)$ *contains vectors* x_+^* *and* x_-^* *such that*

$$x_+^*(u) = D_+h(x; u) \text{ and } x_-^*(u) = D_-h(x; u).$$

(iv) *if* $\partial_u h(x) = \{x^*\}$, *then* $h'(x; u)$ *exists and* $h'(x; u) = x^*(u)$.

Proof. Part (i) follows readily from Definitions 2.2 and 2.3. For part (ii), suppose that $x^* \in \partial_u f(x)$. By Definition 2.2 there exist sequences (x_k) and (x_k^*) such that $(x_k) \to_u x$ and $x_k^* \in \partial f(x_k)$ with $(x_k^*) \to x^*$. Thus, $x_k^* \in \partial g(x_k) + \partial h(x_k)$ by [9, Prop. 2.3.3], so there exist $y_k^* \in \partial g(x_k)$ and $z_k^* \in \partial h(x_k)$ with $y_k^* + z_k^* = x_k^*$ for each k. Since ∂f takes values locally in a compact set [9], by considering a subsequence if necessary we can assume that the sequences (y_k^*) and (z_k^*) are convergent to some $y^* \in \partial g(x)$ and some $z^* \in \partial h(x)$, respectively. By Definition 2.2, $y^* \in \partial_u g(x)$ and $z^* \in \partial_u h(x)$. It follows that $x^* = y^* + z^*$ is in $\partial_u g(x) + \partial_u h(x)$, so the inclusion in (ii) holds. Similarly, by (i), we also have

$$\partial_u g(x) \subseteq \partial_u f(x) + \partial_u(-h)(x) = \partial_u f(x) - \partial_u h(x).$$

Consequently,

$$\partial_u f(x) \subseteq \partial_u g(x) + \partial_u h(x) \subseteq \partial_u f(x) - \partial_u h(x) + \partial_u h(x)$$

and all equalities must hold if $\partial_u h(x)$ is a singleton. Part (iii) is taken from [13, Lem. 2.3], and clearly (iv) follows from (iii). □

Below we give a simple example of a convex (so semismooth [17]) function that satisfies assumption (iv).

Example 2.5. Let $f: \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = |x|$. Then $\partial f(0) = [-1, 1]$, $\partial f(x) = \{1\}$ for $x > 0$, and $\partial f(x) = \{-1\}$ for $x < 0$. Hence $\partial_1 f(0) = \{1\}$ and $\partial_{-1} f(0) = \{-1\}$.

3. Ben-Tal and Zowe's generalized second-order directional derivatives. We begin with the following definition extending that introduced by Ben-Tal and Zowe in [3] who considered the case when $f'(x; u)$ exists.

DEFINITION 3.1. *The Ben-Tal and Zowe lower and upper generalized second-order directional derivatives of f at x in the directions u, w are defined, respectively, by*

$$D^2_- f(x; u, w) := \liminf_{t \downarrow 0} \frac{1}{t^2}\{f(x + tu + t^2 w) - f(x) - t D_+ f(x; u)\} \tag{3.1}$$

and

$$D^2_+ f(x; u, w) := \limsup_{t \downarrow 0} \frac{1}{t^2}\{f(x + tu + t^2 w) - f(x) - t D_- f(x; u)\}. \tag{3.2}$$

If the values in (3.1) *and* (3.2) *are equal, then the common value is simply denoted by $D^2 f(x; u, w)$. We note that it may be finite or infinite.*

For all $x, u \in \mathbb{R}^n$, it is easy to see that $D^2_+ f(x; u, \cdot)$ is a locally Lipschitz function for all $x, u \in \mathbb{R}^n$ (since f is locally Lipschitz), provided that this function is finite-valued. In general, the Lipschitz constant K for f satisfies

$$D^2_+ f(x; u, w_1) \le D^2_+ f(x; u, w_2) + K\|w_1 - w_1\| \quad \forall w_1, w_2 \in \mathbb{R}^n.$$

Similar remarks for $D^2_- f(x; u, \cdot)$ are of course also valid. Note also that

$$D^2_- f(x; u, w) \le D^2_+ f(x; u, w), \tag{3.3}$$

$$-D^2_- f(x; u, w) = D^2_+(-f)(x; u, w). \tag{3.4}$$

Similar but different notions have appeared in the literature, e.g., in Penot [18] and Studniarski [24]; these authors use $D_+ f$ and $D_- f$ in the above definitions to replace $D_- f$ and $D_+ f$ respectively. An advantage of our adoption over the earlier approach is evidenced by the following lemma taken from [14]. We include a proof here for the reader's convenience (for contrast to an earlier approach, see [24, Ex. 3.6]).

LEMMA 3.2. *If $D^2_- f(x; u, w)$ and $D^2_+ f(x; u, w)$ are equal and finite, then $f'(x; u)$ exists.*

Proof. Take a sequence $(t_k) \downarrow 0$ with

$$D^2_+ f(x; u, w) = \lim_{k \to \infty} \frac{1}{t_k^2}\{f(x + t_k u + t_k^2 w) - f(x) - t_k D_- f(x; u)\}.$$

By assumption,

$$\begin{aligned} 0 &= D^2_- f(x; u, w) - D^2_+ f(x; u, w) \\ &\le \liminf_{k \to \infty} \frac{1}{t_k^2}\{f(x + t_k u + t_k^2 w) - f(x) - t_k D_+ f(x; u)\} \\ &\quad - \lim_{k \to \infty} \frac{1}{t_k^2}\{f(x + t_k u + t_k^2 w) - f(x) - t_k D_- f(x; u)\} \\ &= \liminf_{k \to \infty} \frac{1}{t_k}\{D_- f(x; u) - D_+ f(x; u)\} \le 0. \end{aligned}$$

Hence equality holds throughout, and consequently one must have $D_- f(x; u) = D_+(x; u)$. □

Remark. Thus, in the situation of Lemma 3.2, one may use $f'(x; u)$ to replace $D_+ f(x; u)$ and $D_- f(x; u)$ in Definition 3.1; that is, Definition 3.1 coincides with Ben-Tal and Zowe's definition in this special situation.

The following result was proved in [3] for the special case when f is strictly differentiable.

LEMMA 3.3. *Let $x \in \mathbb{R}^n$. If $\partial_u f(x) = \{x^*\}$, then for any $w \in \mathbb{R}^n$ one has*

$$D_-^2 f(x; u, w) = x^*(w) + D_-^2 f(x; u, 0) \quad \text{and} \quad D_+^2 f(x; u, w) = x^*(w) + D_+^2 f(x; u, 0).$$

Proof. By Lebourg's mean value theorem [9, Thm. 2.3.7], for each $t > 0$, there exists $\theta_t \in (0, 1)$ and $x_t^* \in \partial f(x + tu + \theta_t t^2 w)$ such that

$$x_t^*(w) = \frac{1}{t^2}\{f(x + tu + t^2 w) - f(x + tu)\}.$$

Consider any sequence $(t_k) \downarrow 0$ and notice that $x + t_k u + \theta_{t_k} t_k^2 w \to_u x$; it follows that any cluster point of $(x_{t_k}^*)$ must be in $\partial_u f(x) = \{x^*\}$ by Definition 2.2 and the assumption. Since ∂f takes values locally in a compact set [9], we conclude that $x_t^* \to x^*$ as $t \downarrow 0$. Now the first desired equality of this lemma follows from the definition of D_-^2, that is,

$$\begin{aligned} D_-^2 f(x; u, w) &= \liminf_{t \downarrow 0} \frac{1}{t^2}\{t^2 x_t^*(w) + [f(x + tu) - f(x) - tf'(x; u)]\} \\ &= x^*(w) + D_-^2 f(x; u, 0), \end{aligned}$$

where $f'(x; u)$ exists by Lemma 2.4(iv). Similar considerations apply to D_+^2. □

Remark. From the above lemma it is clear that if f is a C^2-function, then

$$D^2 f(x; u, w) = f'(x; w) + \frac{1}{2} u \cdot \nabla^2 f(x) u. \tag{3.5}$$

4. Conjugacy of f'' to $D^2 f$. Recall from convex analysis [23] that if ϕ, ψ are real-valued functions on $\mathbb{R}^n$, then the ("lower" and "upper") conjugate functions ϕ_* and ψ^* are, respectively, defined by

$$\begin{aligned} \phi_*(x^*) &= \inf\{\phi(w) - x^*(w) : w \in \mathbb{R}^n\}, \\ \psi^*(x^*) &= \sup\{\psi(w) - x^*(w) : w \in \mathbb{R}^n\} \end{aligned}$$

for all $x^* \in \mathbb{R}^n$. In this section, we shall show, with reasonable conditions, that if $x^* \in \partial_u f(x)$, then the value of the (lower) conjugate function at x^* of the function $D_-^2 f(x; u, \cdot)$ lies between $f_-''(x; x^*, u)$, $f_+''(x; x^*, u)$ and hence equals $f''(x; x^*, u)$ when the latter exists. Two different sufficient conditions to ensure that this happens are given in Theorems 4.1 and 4.5. Similar results are also discussed for $D_+^2 f(x; u, \cdot)$. The proof for Theorem 4.5 is rather lengthy; we will construct an appropriate triple of sequences satisfying (a), (b), and (c) of Definition 2.3 for the Chaney derivative $f_-''(x; x^*, u)$ from very mild assumptions of the theorem. The proof of Theorem 4.1 is relatively easier, as an appropriate triple of sequences is at hand because of the assumption that $\partial_u f(x)$ is a singleton.

THEOREM 4.1. *Let $u, x \in \mathbb{R}^n$ with $\partial_u f(x) = \{x^*\}$. Then for any $w \in \mathbb{R}^n$, one has*

$$f_-''(x; x^*, u) \le D_-^2 f(x; u, w) - x^*(w) \le D_+^2 f(x; u, w) - x^*(w) \le f_+''(x; x^*, u). \tag{4.1}$$

Consequently,

$$\begin{aligned} f_-''(x; x^*, u) &\le \inf\{D_-^2 f(x; u, w) - x^*(w) : w \in \mathbb{R}^n\} \\ &\le \sup\{D_+^2 f(x; u, w) - x^*(w) : w \in \mathbb{R}^n\} \le f_+''(x; x^*, u). \end{aligned} \tag{4.2}$$

Remark. Both the infimum and the supremum in (4.2) may be infinite (see Examples 6.1 and 6.2).

Proof. Consider any sequence $(t_k) \downarrow 0$ and take $x_k^* \in \partial f(x + t_k u + t_k^2 w)$. Then, as in the proof of Lemma 3.3, we have $x + t_k u + t_k^2 w \to_u x$ and $x_k^* \to x^*$ thanks to the assumption $\partial_u f(x) = \{x^*\}$. Thus, the triple of sequences $(x + t_k u + t_k^2 w)$, (x_k^*), (t_k) satisfies the properties (a), (b), and (c) in Definition 2.3 for $f''_{\pm}(x; x^*, u)$. Hence,

$$\begin{aligned} f''_-(x; x^*, u) &\le \liminf_{k\to\infty} \frac{1}{t_k^2}\{f(x + t_k u + t_k^2 w) - f(x) - x^*(t_k u + t_k^2 w)\} \\ &= \liminf_{k\to\infty} \frac{1}{t_k^2}\{f(x + t_k u + t_k^2 w) - f(x) - t_k x^*(u)\} - x^*(w) \\ &\le \limsup_{k\to\infty} \frac{1}{t_k^2}\{f(x + t_k u + t_k^2 w) - f(x) - t_k x^*(u)\} - x^*(w) \\ &\le f''_+(x; x^*, u). \end{aligned}$$

Since, by Lemma 2.4(iv), $x^*(u)$ may be replaced by $f'(x; u)$, (4.1) is seen to hold. As (4.2) follows immediately from (4.1), the proof is complete. □

COROLLARY 4.2. *Let $x, u \in X$ and $\partial_u f(x) = \{x^*\}$. If $f''(x; x^*, u)$ exists, then so does $D^2 f(x; u, w)$ and $f''(x; x^*, u) = D^2 f(x; u, w) - x^*(w) = D^2 f(x; u, 0)$ for each $w \in \mathbb{R}^n$.*

To prepare for the proof of our main result (Theorem 4.5), we need the following technical proposition. Part (i) is due to Ioffe [15, Prop. 1]; Part (ii) may be regarded as a second-order version of (i). For convenience, we include both proofs here.

PROPOSITION 4.3. *Let $x, u \in \mathbb{R}^n$, and suppose that $D^2_- f(x; u, v) \ge 0$ for all v in $\mathbb{R}^n$. Then for any $\varepsilon, M > 0$, there exists $T > 0$ such that*

(i) *if the given u is zero, then*

$$f(x) \le f(x + tv) + \varepsilon t\|v\|$$

for any $t \in [0, T]$ and $v \in B[0, M]$;

(ii) *if the given u is nonzero, then*

$$f(x) \le f(x + tu + t^2 v) - tD_+ f(x; u) + \varepsilon t^2 \|u + tv\|^2 \tag{4.3}$$

for any $t \in [0, T]$ and $v \in B[0, M]$.

Proof. If the desired conclusion in (ii) is false, then there exist $\varepsilon_0, M_0 > 0$, a sequence $(t_k) \downarrow 0$, and a sequence (v_k) in $B[0, M_0]$ such that

$$f(x) > f(x + t_k u + t_k^2 v_k) - t_k D_+ f(x; u) + \varepsilon_0 t_k^2 \|u + t_k v_k\|^2.$$

By considering a subsequence if necessary, we can assume that $v_k \to v_0$. Hence, we have

$$\begin{aligned} -\varepsilon_0\|u\|^2 &\ge \liminf_{k\to\infty} \frac{1}{t_k^2}\{f(x + t_k u + t_k^2 v_k) - f(x) - t_k D_+ f(x; u)\} \\ &= \liminf_{k\to\infty} \frac{1}{t_k^2}\{f(x + t_k u + t_k^2 v_0) - f(x) - t_k D_+ f(x; u)\} \\ &\ge D^2_- f(x; u, v_0), \end{aligned}$$

where the equality holds because f is Lipschitzian near x. But this contradicts the assumption that $D^2_- f(x; u, v_0) \ge 0$.

For (i), one notes that $D_-^2 f(x; 0, v) = D_- f(x; v) \geq 0$ for all v by the definitions and the assumption. Further, if (i) is not true, then one has $\varepsilon_0 > M_0 > 0$, a sequence $(t_k) \downarrow 0$, and a sequence (v_k) in $B[0, M_0]$ such that

$$f(x) > f(x + t_k v_k) + \varepsilon_0 t_k \|v_k\|.$$

Then, $v_k \neq 0$, so without loss of generality one can suppose that $w_k := v_k/\|v_k\| \to w$. Note that $\tau_k := t_k\|v_k\| \to 0$ and

$$-\varepsilon_0 > \frac{f(x + \tau_k w_k) - f(x)}{\tau_k}.$$

Therefore, one has $-\varepsilon_0 \geq D_- f(x; w)$, contradicting the given assumption. □

PROPOSITION 4.4. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function; and suppose $u \neq 0$, x are vectors in $\mathbb{R}^n$ such that*

(I) $D_+ F(x; u) = 0$,

(II) $\inf\{D_-^2 F(x; u, w) : w \in \mathbb{R}^n\} \geq 0$.

If $D_-^2 F(x; u, w) < \varepsilon$ for some $w \in \mathbb{R}^n$ and $0 < \varepsilon < 1/2(2 + \|u\|)^2$, then there exist sequences $(t_k) \downarrow 0$ and (ξ_k) in $\mathbb{R}^n$ with the following properties for each k:

(1) $\xi_k \in B(x + t_k u + t_k^2 w, t_k^2)$, *and so* $(\xi_k) \to x$;

(2) $\|(x + t_k u + t_k^2 w - \xi_k)/t_k^2\| \leq (2\varepsilon)^{1/2}(2 + \|u\|)$;

(3) $(x + t_k u + t_k^2 w - \xi_k)/t_k^2 \in \partial F(\xi_k)$;

(4) $(F(\xi_k) - F(x))/t_k^2 < \varepsilon$;

(5) $(F(\xi_k) - F(x))/t_k^2 \geq -\varepsilon$.

Proof. By assumptions (I), (II), applying Proposition 4.3(ii) with ε and $M := \|w\| + 1$ one can find $T > 0$ such that

$$-\varepsilon t^2 \|u + tv\|^2 \leq F(x + tu + t^2 v) - F(x) \tag{4.4}$$

for all $t \in [0, T]$ and $v \in B[0, \|w\| + 1]$. On the other hand, because of (I) and the assumption $D_-^2 F(x; u, w) < \varepsilon$ one can find a sequence $(t_k) \downarrow 0$ such that

$$\frac{1}{t_k^2}\{F(x + t_k u + t_k^2 w) - F(x)\} < \varepsilon \tag{4.5}$$

for all k. In addition, we can of course assume that

$$t_k \in [0, T]. \tag{4.6}$$

Let $B_k := B[x + t_k u + t_k^2 w, t_k^2]$. Motivated by the Moreau-Yosida approximation [1], let ξ_k be a minimum point of the function

$$y \mapsto F(y) + \frac{1}{2t_k^2}\|y - (x + t_k u + t_k^2 w)\|^2 \quad \text{on } B_k.$$

In particular,

$$F(\xi_k) + \frac{1}{2t_k^2}\|\xi_k - (x + t_k u + t_k^2 w)\|^2 \leq F(x + t_k u + t_k^2 w). \tag{4.7}$$

By definition of B_k, one can express ξ_k in the form

$$\xi_k = x + t_k u + t_k^2(w + v_k)$$

with some $v_k \in B[0, 1]$; it follows from (4.4) and (4.7) that

$$\begin{aligned} -\varepsilon t_k^2 \|u + t_k(w + v_k)\|^2 &\le F(\xi_k) - F(x) \\ &\le F(x + t_k u + t_k^2 w) - F(x) - \frac{1}{2t_k^2}\|\xi_k - (x + t_k u + t_k^2 w)\|^2. \end{aligned} \tag{4.8}$$

Dividing by t_k^2 it follows from (4.5) that

$$-\varepsilon\|u + t_k(w + v_k)\|^2 \le \varepsilon - \frac{1}{2t_k^4}\|\xi_k - (x + t_k u + t_k^2 w)\|^2$$

and so

$$\begin{aligned} \frac{1}{t_k^2}\|\xi_k - (x + t_k u + t_k^2 w)\| &\le (2\varepsilon)^{1/2}(1 + \|u + t_k(w + v_k)\|^2)^{1/2} \\ &\le (2\varepsilon)^{1/2}(2 + \|u\|) < 1 \end{aligned}$$

(provided that k is sufficiently large), because $\varepsilon < 1/2(2 + \|u\|)^2$. Thus by deleting finitely many terms if necessary, one can suppose that each ξ_k is in the interior of B_k. Therefore, (2) and (1) hold. It follows from the minimality of ξ_k and results of Clarke [9, Props. 2.3.2 and 2.3.3] that

$$0 \in \partial F(\xi_k) + \frac{\xi_k - (x + t_k u + t_k^2 w)}{t_k^2},$$

so (3) holds. Part (4) follows clearly from (4.7) and (4.5). Finally, since the sequence (v_k) is bounded, we can assume without loss of generality that it converges, say to some v_0. Then, by (I) and the Lipschitz property of F, we have

$$\begin{aligned} D_-^2 F(x; u, w + v_0) &\le \liminf_{k\to\infty} \frac{F(x + t_k u + t_k^2(w + v_0)) - F(x)}{t_k^2} \\ &= \liminf_{k\to\infty} \frac{F(x + t_k u + t_k^2(w + v_k)) - F(x)}{t_k^2} \\ &= \liminf_{k\to\infty} \frac{F(\xi_k) - F(x)}{t_k^2}. \end{aligned}$$

By (II), it follows that $-\varepsilon < D_-^2 F(x; u, w + v_0) \le \liminf_{k\to\infty}[(F(\xi_k) - F(x))/t_k^2]$; thus, by deleting finitely many terms if necessary, (5) holds. □

Remark. In view of assumption (II), w may be regarded as an "approximate minimum point" of $D_-^2 f(x; u, \cdot)$ on $\mathbb{R}^n$. If $u = 0$, then as $D_-^2 f(x; u, \cdot) = D_- f(x; \cdot)$ the true minimum point (namely, zero) always exists; in this case, the natural choice for w is zero. Explicitly we have the following counterpart of Proposition 4.4.

PROPOSITION 4.4*. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function and x in $\mathbb{R}^n$ such that*

(II)* $D_- F(x; v) \ge 0$ *for all* $v \in \mathbb{R}^n$.

Let $0 < \varepsilon$ with $(2\varepsilon)^{1/2} < 1$. Then there exist sequences $(t_k) \downarrow 0$ and (ξ_k) in $\mathbb{R}^n$ with the following properties for each k:

(1)* $\xi_k \in B(x, t_k^2)$, so $(\xi_k) \to x$;

(2)* $\|(x - \xi_k)/t_k^2\| \le (2\varepsilon)^{1/2}$;

(3)* $(x - \xi_k)/t_k^2 \in \partial F(\xi_k)$;

(4)* $(F(\xi_k) - F(x))/t_k^2 \leq 0$;

(5)* $(F(\xi_k) - F(x))/t_k^2 \geq -\varepsilon$.

Proof. Let $M = 1$. By assumption (II)*, applying Proposition 4.3(i) one can find $T > 0$ such that

$$-\varepsilon t^2\|v\| \leq F(x + t^2 v) - F(x) \quad \forall t^2 \in [0, T],\ v \in B[0, 1]. \tag{4.4)*}$$

Take a sequence (t_k) in $[0, T^{1/2}]$ such that $(t_k) \downarrow 0$. Let $\xi_k := x + t_k^2 v_k$ be a minimum point of the function

$$y \mapsto F(y) + \|y - x\|^2/2t_k^2 \quad \text{on } B[x, t_k^2],$$

where each $v_k \in B[0, 1]$. Then

$$F(\xi_k) \leq F(\xi_k) + \|\xi_k - x\|^2/2t_k^2 \leq F(x) \leq F(\xi_k) + \varepsilon t_k^2\|v_k\| \tag{4.9}$$

and in particular $\|\xi_k - x\|/t_k^2 \leq (2\varepsilon)^{1/2}\|v_k\|^{1/2} \leq (2\varepsilon)^{1/2} < 1$, showing that ξ_k is in the interior of $B[x, t_k^2]$ so, as before, $0 \in \partial F(\xi_k) + (\xi_k - x)/t_k^2$. Thus, (1)*, (2)*, and (3)* are established. Parts (4)* and (5)* also follow immediately from (4.9). □

THEOREM 4.5. *Let $x, u, x^* \in \mathbb{R}^n$ satisfy the following properties:*

(I) $x^* \in \partial_u f(x)$ *and* $x^*(u) = D_+ f(x; u)$.

(II) $\alpha := \inf\{D_-^2 f(x; u, w) - x^*(w) : w \in \mathbb{R}^n\}$ *is a finite number.*

Then

$$f''_-(x; x^*, u) \leq \inf\{D_-^2 f(x; u, w) - x^*(w) : w \in \mathbb{R}^n\} \leq f''_+(x; x^*, u).$$

Consequently, whenever $f''(x; x^*, u)$ exists and (I), (II) *are satisfied, we have*

$$f''(x; x^*, u) = \inf\{D_-^2 f(x; u, w) - x^*(w) : w \in \mathbb{R}^n\}.$$

Remark 1. (I) is automatically satisfied by any $x^* \in \partial_u f(x)$ if either (a) $\partial_u f(x)$ is a singleton (see Lemma 2.4(iv)) or (b) f is semismooth (for definitions and basic properties see [17]).

Remark 2. In the terminology of convex analysis reviewed at the beginning of this section, the number α is the value at x^* of the conjugate function of $D_-^2 f(x; u, \cdot)$. Thus, this theorem implies that this value coincides with $f''(x; x^*, u)$ if the latter exists, and (I), (II) are satisfied.

Remark 3. Theorems 4.1 and 4.5 are independent; neither implies the other (see §6).

Proof of Theorem 4.5. We first consider the case when $u \neq 0$. Then we define

$$F := f - x^* - h, \tag{4.10}$$

where $h(\cdot) = (\alpha\|\cdot - x\|^2)/\|u\|^2$. A list of some properties of F are as follows.

(i) $D_+ F(x; u) = 0$.

(ii) $\inf\{D_-^2 F(x; u, w) : w \in \mathbb{R}^n\} = 0$.

(iii) $D_-^2 F(x; u, w) = D_-^2 f(x; u, w) - x^*(w) - \alpha$.

(iv) $0 \in \partial_u F(x)$.

(v) $F''_-(x; 0, u) = f''_-(x; x^*, u) - \alpha$ and $F''_+(x; 0, u) = f''_+(x; x^*, u) - \alpha$.

In fact, since $\partial h(x) = \{0\}$, (i) follows from (I), (iii) follows from

$$\begin{aligned} D_-^2 F(x; u, w) &= D_-^2 f(x; u, w) - D^2 x^*(x; u, w) - D^2 h(x; u, w) \\ &= D_-^2 f(x; u, w) - x^*(w) - \alpha, \end{aligned}$$

and (iv) follows from the assumption $x^* \in \partial_u f(x)$ and Lemma 2.4(ii). Clearly (ii) follows from (iii) and the definition of α. By (iv) the second-order derivatives in (v) are well defined. Moreover, the sequences (x_k), (x_k^*), (t_k) satisfy (a), (b), and (c) in Definition 2.3 for $f''_-(x; x^*, u)$ if and only if (x_k), $(x_k^* - x^*)$, (t_k) satisfy the same for $F''_-(x; 0, u)$ (note that, by Lemma 2.4 again, $x_k^* \in \partial f(x_k)$ if and only if $x_k^* - x^* \in \partial F(x_k)$). Passing to the appropriate limits in

$$\frac{1}{t_k^2}\{F(x_k) - F(x) - 0\} = \frac{1}{t_k^2}\{f(x_k) - f(x) - x^*(x_k - x) - [h(x_k) - h(x)]\},$$

(v) is seen to hold.

Take $(\varepsilon_m) \downarrow 0$ such that $\varepsilon_m < 1/2(2 + \|u\|)^2$ for all m. By (ii), there exists $w_m \in \mathbb{R}^n$ such that $D_-^2 F(x; u, w_m) < \varepsilon_m$. With any fixed m, one may apply Proposition 4.4 to the pair (ε_m, w_m) to obtain sequences $(t_{mk})_k \downarrow 0$ and $(\xi_{mk})_k$ satisfying the corresponding properties (1)–(5) of Proposition 4.4 (with ε_m, w_m in place of ε, w); for example, (1) reads

$$\|\xi_{mk} - (x + t_{mk}u + t_{mk}^2 w_m)\| < t_{mk}^2 \quad \forall k,$$

which implies that

$$\left\| \frac{\xi_{mk} - x}{t_{mk}} - u \right\| < t_{mk} + t_{mk}\|w_m\|.$$

Do this for each m and select some k_m such that $t_{mk_m}\|w_m\| < \varepsilon_m$; in general one can ensure further that $(t_{mk_m})_m \downarrow 0$. For convenience we denote the sequences $(t_{mk_m})_m$ and $(\xi_{mk_m})_m$ by (τ_m), (η_m), respectively. Then we have, for all m, that

(1) $\eta_m \in B(x + \tau_m u + \tau_m^2 w_m, \tau_m^2)$, and so $\eta_m \to x$;
(2) $\|(x + \tau_m u + \tau_m^2 w_m - \eta_m)/\tau_m^2\| \le (2\varepsilon_m)^{1/2}(2 + \|u\|)$;
(3) $(x + \tau_m u + \tau_m^2 w_m - \eta_m)/\tau_m^2 \in \partial F(\eta_m)$;
(4) $(F(\eta_m) - F(x))/\tau_m^2 < \varepsilon_m$;
(5) $(F(\eta_m) - F(x))/\tau_m^2 > -\varepsilon_m$.

Since $\tau_m w_m \to 0$, it follows from (1), (2), and (3) that the three sequences

$$(\eta_m), \left(\frac{x + \tau_m u + \tau_m^2 w_m - \eta_m}{\tau_m^2}\right), \text{ and } (\tau_m)$$

have the properties (a), (b), and (c) in Definition 2.3 for $F''_-(x; 0, u)$; hence it follows from (4) that

$$F''_-(x, 0, u) \le \liminf_{m\to\infty} \frac{F(\eta_m) - F(x)}{\tau_m^2} \le 0,$$

which implies, by (v), that $f''_-(x; x^*, u) \le \alpha$. Similarly, by (5) and (v), we have

$$0 \le \limsup_{m\to\infty} \frac{F(\eta_m) - F(x)}{\tau_m^2} \le F''_+(x; 0, u),$$

so $\alpha \le f''_+(x; x^*, u)$. This proves the theorem for the case when $u \neq 0$. Henceforth, we let $u = 0$. Then note that

$$D_- f(x; 0) = 0 \quad \text{and} \quad D_-^2 f(x; 0, w) = D_- f(x; w)$$

for any $w \in \mathbb{R}^n$. Hence, by (II), we have

$$\alpha = \inf\{D_- f(x; w) - x^*(w); w \in \mathbb{R}^n\} \le 0.$$

Since the function $w \mapsto D_-f(x; w) - x^*(w)$ is positively homogeneous, the finiteness assumption of α implies that α must be zero. Thus, in place of (4.10) if one defines $F := f - x^*$, then it still satisfies (i)–(v) listed at the beginning of our proof (with $\alpha = 0$, $u = 0$). Take $(\varepsilon_m) \downarrow 0$ such that $(2\varepsilon_m)^{1/2} < 1$ for all m. By (ii), with any fixed m, one may apply Proposition 4.4* to obtain sequences $(t_{mk})_k \downarrow 0$ and $(\xi_{mk})_k$ satisfying (1)*–(5)* (with ε_m in place of ε); then the triple of sequences (η_m), $((x - \eta_m)/\tau_m^2)$, (τ_m) (where η_m and τ_m are defined as in the first part of the proof) satisfies properties (a), (b), and (c) of Definition 2.3 for $F''_+(x; 0, 0)$. Hence, the proof is completed as above for the case when $u \neq 0$. □

Dually, we have the following result.

THEOREM 4.6. *Let $x, u, x^* \in \mathbb{R}^n$ satisfy the properties:*

(I) $x^* \in \partial_u f(x)$ *and* $x^*(u) = D_-f(x; u)$;

(II) $\gamma := \sup\{D^2_+ f(x; u, w) - x^*(w) : w \in \mathbb{R}^n\}$ *is a finite number.*

Then

$$f''_+(x; x^*, u) \geq \sup\{D^2_+ f(x; u, w) - x^*(w) : w \in \mathbb{R}^n\} \geq f''_-(x; x^*, u).$$

Consequently, whenever $f''(x; x^*, u)$ exists and (I), (II) *are satisfied, we have*

$$f''(x; x^*, u) = \sup\{D^2_+ f(x; u, w) - x^*(w) : w \in \mathbb{R}^n\}.$$

Proof. It follows from Lemma 2.4(i) that $x^* \in \partial_u f(x)$ if and only if

$$-x^* \in \partial_u(-f)(x) \quad \text{and} \quad (-f)''_-(x; -x^*, u) = -f''_+(x; x^*, u).$$

Note also that $D^2_-(-f)(x; u, w) = -D^2_+ f(x; u, w)$. Thus, replacing f and x^* in Theorem 4.5 by $-f$ and $-x^*$, respectively, we obtain the required results. □

Combining Theorem 4.5 and Theorem 4.6, we have the following result.

THEOREM 4.7. *Let $x, u, x^* \in \mathbb{R}^n$ satisfy the properties:*

(I) $x^* \in \partial_u f(x)$ *and* $x^*(u) = f'(x; u)$;

(II) *both constants α and γ appearing in Theorems* 4.5 *and* 4.6 *are finite*;

(III) $f''(x; x^*, u)$ *exists.*

Then $D^2 f(x; u, w)$ exists, and

$$f''(x; x^*, u) = D^2 f(x; u, w) - x^*(w) = D^2 f(x; u, 0)$$

for all $w \in \mathbb{R}^n$.

Proof. Combining Theorems 4.6 and 4.7 we have

$$\begin{aligned} f''(x; x^*, u) &= \inf\{D^2_- f(x; u, w) - x^*(w) : w \in \mathbb{R}^n\} \\ &= \sup\{D^2_+ f(x; u, w) - x^*(w) : w \in \mathbb{R}^n\}. \end{aligned} \tag{4.11}$$

It follows that $D^2_- f(x; u, w) \geq D^2_+ f(x; u, w)$ and consequently $D^2 f(x; u, w)$ exists. Thus, (4.11) implies that

$$f''(x; x^*, u) = D^2 f(x; u, w) - x^*(w),$$

for all $w \in \mathbb{R}^n$. □

5. On a special class of functions. In this section we apply the results in the preceding one to real-valued functions f of the form $f(x) = \sum_{i=1}^m g_i(h_i(x))$, studied in [3] and [5]. Here the real-valued functions g_i and h_i are as follows. It is assumed that

$$h_i(x) = \max\{h_{ij}(x) : j = 1, 2, \ldots, p_i\}$$

and each h_{ij} is of class C^2 on $\mathbb{R}^n$. Moreover, it is assumed that each g_i is of the class C^2 on a neighborhood of the set $h_i(\mathbb{R}^n)$. Finally, we assume that for some x_0 in $\mathbb{R}^n$, $g_i'(h_i(x_0)) \geq 0$ for each $i = 1, 2, \ldots, m$. We shall henceforth keep x_0 fixed.

In [3], Ben-Tal and Zowe showed that $D^2 f(x_0; u, w)$ exists for any $u, w \in \mathbb{R}^n$ and gave the explicit formula

(5.1)

$$D^2 f(x_0; u, v) = \frac{1}{2} \sum_{i=1}^{m} g_i''(h_i(x_0))[h_i'(x_0; u)]^2 + \sum_{i=1}^{m} g_i'(h_i(x_0)) \max\{\nabla h_{ij}(x_0) \cdot v + \frac{1}{2} u \cdot \nabla^2 h_{ij}(x_0) u : j \in I_i(x_0, u)\},$$

where for each i

$$I_i(x_0, u) = \{j \in I_i(x_0) : \nabla h_{ij}(x_0) \cdot u = \max_{j \in I_i(x_0)} \nabla h_{ij}(x_0) \cdot u\}$$

and

$$I_i(x_0) = \{j : h_{ij}(x_0) = \max_{1 \leq j \leq p_i} h_{ij}(x_0)\}$$

(we emphasize that here $D^2 f(x_0; u, v)$ is finite thanks to (5.1)). Likewise, Chaney showed in [5, Thm. 4.2] that for functions of this form, his second-order directional derivative $f''(x; x^*, u)$ exists for each $x^* \in \partial_u f(x)$.

Let $z \in \mathbb{R}^n$ and $z^* \in \partial f(z)$. We recall from [9, Props. 2.3.12 and 2.3.9] that z^* has the form

$$z^* = \sum_{i=1}^{m} g_i'(h_i(z)) \sum_{j=1}^{p_i} w_{ij} \nabla h_{ij}(z), \tag{5.2}$$

where for each i, $\sum_{j=1}^{p_i} w_{ij} = 1$, $w_{ij} \geq 0$, and $w_{ij} = 0$ if $j \notin I_i(z)$. Similarly, we have the following description for elements of $\partial_u f(x)$.

LEMMA 5.1. *Let $x_0^* \in \partial_u f(x_0)$. Then there exist $\lambda_{ij} \geq 0$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, p_i$, such that*

(1) *for each i, $\sum_{j=1}^{p_i} \lambda_{ij} = 1$, $\lambda_{ij} \geq 0$, and $\lambda_{ij} = 0$ if $j \notin I_i(x_0, u)$;*

(2) $x_0^* = \sum_{i=1}^{m} g_i'(h_i(x_0)) \sum_{j=1}^{p_i} \lambda_{ij} \nabla h_{ij}(x_0)$.

Consequently, we have

(3) $x_0^*(u) = f'(x_0; u)$.

Proof. By Definition 2.2 there exist sequences $(x_k) \to_u x_0$ and $x_k^* \in \partial f(x_k)$ with $(x_k^*) \to x_0^*$. By (5.2) each x_k^* has the form

$$x_k^* = \sum_{i=1}^{m} g_i'(h_i(x_k)) \sum_{j=1}^{p_i} \lambda_{ij}^{(k)} \nabla h_{ij}(x_k)$$

where for each i, $\sum_{j=1}^{p_i} \lambda_{ij}^{(k)} = 1$, $\lambda_{ij}^{(k)} \geq 0$, and $\lambda_{ij}^{(k)} = 0$ if $j \notin I_i(x_k)$. By considering a subsequence if necessary, we can assume that $(\lambda_{ij}^{(k)})_k \to \lambda_{ij}$; hence,

$$x_0^* = \sum_{i=1}^{m} g_i'(h_i(x_0)) \sum_{j=1}^{p_i} \lambda_{ij} \nabla h_{ij}(x_0),$$

verifying (2). For (1), suppose $\lambda_{ij} > 0$. Then $\lambda_{ij}^{(k)} > 0$ for large k and so $j \in I_i(x_k)$ for large k. Thus, $j \in I_i(x_0)$ since $I_i(x_k) \subseteq I_i(x_0)$ for large k. Hence

$$\begin{aligned} h_i'(x_0; u) &= \lim_{k\to\infty} \frac{\|u\|}{\|x_k - x_0\|}\{h_i(x_k) - h_i(x_0)\} \\ &= \lim_{k\to\infty} \frac{\|u\|}{\|x_k - x_0\|}\{h_{ij}(x_k) - h_{ij}(x_0)\} = \nabla h_{ij}(x_0)u, \end{aligned} \tag{5.3}$$

showing that $j \in I_i(x_0; u)$. Therefore (1) holds. Finally, (5.3) also implies that $\sum_{j=1}^{p_i} \lambda_{ij} \times \nabla h_{ij}(x_0)u = h_i'(x_0; u)$ and hence (3) follows from (2) and the chain rule [3, Lem. 3.1]. □

We are now ready to provide a short proof of the following result of Chaney [5].

THEOREM 5.2. *Let $x_0^* \in \partial_u f(x_0)$. Then*

$$f''(x_0; x_0^*, u) = \inf\{D^2 f(x_0; u, v) - x_0^*(v) : v \in \mathbb{R}^n\} \tag{5.4}$$

and the common value is finite.

Proof. By (5.1) the right-hand side is not $+\infty$. We will show it is not $-\infty$. To do this, we express x_0^* in the form stated in the preceding lemma. Since $g_i'(h_i(x_0)) \geq 0$ by assumption and since the maximum of any finite set of real numbers majorizes their convex combinations, the formula (5.1) implies that

$$D^2 f(x_0; u, v) \geq Q + \sum_{i=1}^{m} g_i'(h_i(x_0)) \sum_{j=1}^{p_i} \lambda_{ij}\{\nabla h_{ij}(x_0)v + \frac{1}{2}u \cdot \nabla^2 h_{ij}(x_0)u\}, \tag{5.5}$$

where $Q := \frac{1}{2}\sum_{i=1}^{m} g_i''(h_i(x_0))[h_i'(x_0; u)]^2$. Taking the difference of (5.5) with (2) in Lemma 5.1, we have for all $v \in \mathbb{R}^n$ that

$$D^2 f(x_0; u, v) - x_0^*(v) \geq Q + \frac{1}{2}\sum_{i=1}^{m} g_i'(h_i(x_0)) \sum_{j=1}^{p_i} \lambda_{ij} u \cdot \nabla^2 h_{ij}(x_0)u,$$

where the right-hand side is independent of v. Therefore condition (II) of Theorem 4.5 is satisfied. Since condition (I) is also satisfied by assumption and part (3) of Lemma 5.1, we see that Theorem 5.2 follows from Theorem 4.5.

6. Examples. In this section we give examples to show that Theorems 4.1 and 4.5 are independent; that is, neither implies the other. Moreover, the relations established in these theorems between Chaney's generalized second-order directional derivative and that of Ben-Tal and Zowe do not necessarily degenerate to the simpler relation (1.1), valid for C^2-functions. Even in the case when $\partial_u f(x_0)$ is a singleton, Examples 6.1 and 6.2 show that both the infimum and supremum in (4.2) may still be $+\infty$ or $-\infty$. Thus, the assumption $\partial_u f(x) = \{x^*\}$ in Theorem 4.1 does not imply the finiteness of α in Theorem 4.5, and therefore Theorem 4.1 is not a corollary of Theorem 4.5. On the other hand, it may happen that α is finite but $\partial_u f(x)$ is not a singleton, as Example 6.3 shows; therefore, Theorem 4.5 is not a corollary of Theorem 4.1.

Example 6.1. Consider the C^1-function $f(x) := x^{3/2}$ on $\mathbb{R}$. Let $x_0 = 0$, $u = 1$. Then $\partial_u f(x_0) = \{\nabla f(x_0)\} = \{0\}$. Since for any $u_t \to u$,

$$\lim_{t\downarrow 0} \frac{f(x_0 + tu_t) - f(x_0)}{t^2} = \lim_{t\downarrow 0} \frac{(tu_t)^{3/2}}{t^2} = +\infty,$$

it is seen that $f''(x_0; 0, u) = +\infty$. One can compute directly or apply (4.1) to conclude that $D^2 f(x_0; u, w) = +\infty$ for each $w \in \mathbb{R}$.

$-f$ provides an example showing that this phenomenon can also happen for $-\infty$ in place of $+\infty$.

Example 6.2. Let $f(x) = x^{3/2}\sin(x^{-1/3})$ for $x \neq 0$, and let $f(0) = 0$. Then f is a C^1-function and $\nabla f(0) = 0$. Let $x_0 = 0$, $u = 1$, $x_0^* = 0$. Then $\partial_u f(x_0) = \{x_0^*\}$. Now we take a sequence $(t_k) \downarrow 0$ with $(t_k + t_k^2)^{-1/3} = 2k\pi - \frac{\pi}{2}$. Then $\sin(t_k + t_k^2)^{-1/3} = -1$ for each k, so

$$D_-^2 f(x_0; u, u) = \liminf_{t \downarrow 0} \frac{f(t + t^2)}{t^2} \leq \liminf_{k \to \infty} \left(\frac{t_k + t_k^2}{t_k^{4/3}} \right)^{3/2} (-1) = -\infty.$$

By Theorem 4.1, it follows that $f''_-(x_0; x_0^*, u) = -\infty$. Similarly, we also have

$$D_+^2 f(x_0; u, u) \geq +\infty \quad \text{and} \quad f''_+(x_0; x_0^*, u) = +\infty.$$

Thus, the infimum appearing in (4.2) (for $x = x_0$, $u = 1$) is $-\infty$ while the supremum is $+\infty$, even though $\partial_u f(x_0)$ is a singleton.

Example 6.3. Let $h\colon \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$h(x) = \max\{f_1(x), f_2(x)\} \qquad (x \in \mathbb{R}^2),$$

where $f_1(x_1, x_2) = (x_1)^2 + x_2$ and $f_2(x_1, x_2) = x_1 + (x_2)^2$ for all $x = (x_1, x_2) \in \mathbb{R}^2$. Let $x_0 = (a, a) \in \mathbb{R}^2$. By symmetry, it is clear that

$$h(\cdot) = f_1(\cdot) = f_2(\cdot) \quad \text{and} \quad \nabla f_1(x_0)(\cdot) = \nabla f_2(x_0)(\cdot) = h'(x_0; \cdot)$$

on the subset G of $\mathbb{R}^2$ consisting of vectors of the form (c, c) with $c \in \mathbb{R}$. Note also that

$$\nabla^2 f_1(x_0)(u, u) = \nabla^2 f_2(x_0)(u, u), \quad \forall u \in G.$$

Hence, by the formula of Ben-Tal and Zowe (see (5.1)), one has for each $v \in \mathbb{R}^2$, $u \in G$ that

$$\begin{aligned} D^2 h(x_0; u, v) &= \max\left\{ \nabla f_i(x_0)v + \frac{1}{2}\nabla^2 f_i(x_0)(u, u)\colon\ i = 1, 2 \right\} \\ &= \frac{1}{2}\nabla^2 f_1(x_0)(u, u) + \max\{\nabla f_i(x_0)v\colon\ i = 1, 2\}. \end{aligned}$$

Note that the right-hand side is clearly convex but not affine in v since $\nabla f_1(x_0) \neq \nabla f_2(x_0)$. Therefore, the function

$$D^2 h(x_0; u, \cdot) - x_0^*(\cdot) \tag{6.1}$$

is also convex but not affine on $\mathbb{R}^2$, where $x_0^* \in \partial_u h(x_0)$. In particular, if we denote the infimum and the supremum of the function (6.1) on $\mathbb{R}^2$ by α and γ, respectively, then α must be strictly smaller than γ. By Theorem 5.2, this α must be finite and must equal $h''(x_0; x_0^*, u)$. It follows from Theorem 4.7 that γ must be $+\infty$ (this can also be seen directly from the fact that the function is convex and nonaffine). Moreover, consider any sequence (x_k) of vectors in $\mathbb{R}^2$ convergent to x_0 in the direction u. Then, since

$$\partial h(x_k) = \operatorname{co}\{\nabla f_1(x_k), \nabla f_2(x_k)\}$$

thanks to [9, Prop. 2.3.12], it is easily seen that

$$\partial_u h(x_0) = \operatorname{co}\{\nabla f_1(x_0), \nabla f_2(x_0)\} = \operatorname{co}\{(2a, 1), (1, 2a)\},$$

which is a nondegenerate line segment. Thus, Theorems 5.2 and 4.5 are not corollaries of Theorem 4.1.

Acknowledgments. The authors would like to thank the referees and the associate editor for some valuable suggestions that, in particular, prompted the authors to produce the examples given in §6.

REFERENCES

[1] J.-P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, John Wiley, New York, 1984.
[2] A. BEN-TAL, *Second-order and related extremality conditions in nonlinear programming*, J. Optim. Theory Appl., 31 (1980), pp. 143–165.
[3] A. BEN-TAL AND J. ZOWE, *Necessary and sufficient optimality conditions for a class of nonsmooth minimization problems*, Math. Programming, 24 (1982), pp. 70–91.
[4] R. W. CHANEY, *On second derivatives for nonsmooth functions*, Nonlinear Anal. Theory Methods Appl., 9 (1985), pp. 1189–1209.
[5] ———, *Second-order directional derivatives for nonsmooth functions*, J. Math. Anal. Appl., 128 (1987), pp. 495–511.
[6] ———, *Second-order necessary conditions in constrained semismooth optimization*, SIAM J. Control Optim., 25 (1987), pp. 1072–1081.
[7] ———, *Second-order necessary conditions in semismooth optimization*, Math. Programming, 40 (1988), pp. 95–109.
[8] ———, *Second-order sufficient conditions in nonsmooth optimization*, Math. Oper. Res., 13 (1988), pp. 660–673.
[9] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
[10] R. COMINETTI AND R. CORREA, *A generalized second-order derivative in nonsmooth optimization*, SIAM J. Control Optim., 28 (1990), pp. 789–809.
[11] C. N. DO, *Generalized second-order derivatives of convex functions in reflexive Banach spaces*, Trans. Amer. Math. Soc., 334 (1992), pp. 281–301.
[12] W. L. CHAN, L. R. HUANG, AND K. F. NG, *On generalized second-order derivatives and Taylor expansions in nonsmooth optimization*, SIAM J. Control Optim., 32 (1994), pp. 591–611.
[13] L. R. HUANG AND K. F. NG, *Second-order necessary and sufficient conditions in nonsmooth optimization*, Math. Programming, 66 (1994), pp. 379–402.
[14] ———, *On Lower Bounds of the Second-Order Directional Derivatives of Ben-Tal-Zowe and Chaney*, Math. Oper. Res., to appear.
[15] A. D. IOFFE, *Calculus of Dini subdifferentials of functions and contingent coderivatives*, Nonlinear Anal. Theory Methods Appl., 8 (1984), pp. 517–539.
[16] H. KAWASAKI, *Second-order necessary and sufficient optimality conditions for minimizing a sup-type function*, Appl. Math. Optim., 26 (1992), pp. 195–220.
[17] R. MIFFLIN, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.
[18] J.-P. PENOT, *Generalized Higher Order Derivatives and Higher Order Optimality Conditions*, preprint, Université de Pau, 1985.
[19] ———, *Second-order generalized derivatives: comparisons of two types of epi-derivatives*, in Advances in Optimization (Lambrecht, 1991), Lecture Notes in Econom. and Math. Systems 382, Springer, Berlin, 1992, pp. 52–76.
[20] ———, *Second-order generalized derivatives: relationships with convergence notions*, in Nonsmooth Optimization Methods and Applications, F. Giannessi, ed., Gordon and Breach, New York, 1992, pp. 303–322.
[21] R. T. ROCKAFELLAR, *First- and second-order epi-differentiability in nonlinear programming*, Trans. Amer. Math. Soc., 307 (1988), pp. 75–108.
[22] ———, *Second-order optimality conditions in nonlinear programming obtained by way of epi-derivatives*, Math. Oper. Res., 14 (1989), pp. 462–484.
[23] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[24] M. STUDNIARSKI, *Second-order necessary conditions for optimality in nonsmooth nonlinear programming*, J. Math. Anal. Appl., 154 (1991), pp. 303–317.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1235–1269, July 1996

007

CONSISTENT APPROXIMATIONS FOR OPTIMAL CONTROL PROBLEMS BASED ON RUNGE–KUTTA INTEGRATION*

A. SCHWARTZ† AND E. POLAK†

Abstract. This paper explores the use of Runge–Kutta integration methods in the construction of families of finite-dimensional, consistent approximations to nonsmooth, control and state constrained optimal control problems. Consistency is defined in terms of epiconvergence of the approximating problems and hypoconvergence of their optimality functions. A significant consequence of this concept of consistency is that stationary points and global solutions of the approximating discrete-time optimal control problems can only converge to stationary points and global solutions of the original optimal control problem. The construction of consistent approximations requires the introduction of appropriate finite-dimensional subspaces of the space of controls and the extension of the standard Runge–Kutta methods to piecewise-continuous functions.

It is shown that in solving discrete-time optimal control problems that result from Runge–Kutta integration, a non-Euclidean inner product and norm must be used on the control space to avoid potentially serious ill-conditioning effects.

Key words. optimal control, discretization theory, consistent approximations, Runge–Kutta integration

AMS subject classifications. 49J15, 49M25, 49J45, 65L06

1. Introduction.

We consider approximations to constrained optimal control problems resulting from the replacement of the differential equations that describe the system dynamics with difference equations that arise from Runge–Kutta integration of those differential equations. In particular, we show that there is a class of higher order, explicit *Runge–Kutta* (RK) methods that provide *consistent approximations* to the original problem, with consistency defined according to [24]. Consequently, we are assured that stationary points of the approximating problems converge to stationary points of the original problem and that global solutions (or strict local solutions with nonvanishing radii of attraction) of the approximating problems converge to global (or local) solutions of the original problem, as the step-size of the RK method is decreased.

The theory in [24] requires that the approximating problems be defined on finite-dimensional subspaces of the control space to which RK methods can be extended. The selection of the control subspaces affects both the accuracy of numerical integration and the accuracy with which solutions of the original problem are approximated. Once the approximating problems are defined, their numerical solution is carried out by means of standard mathematical programming algorithms in the space of coefficients associated with the bases that define the control subspaces. We construct two such families of control subspaces. The "natural" basis functions for one family are piecewise polynomial functions and, for the other, piecewise constant functions. Neither of these sets of basis functions is orthonormal. Hence, to preserve the L_2 inner product and norm used in the control subspace, a nonstandard inner product and norm must be used in the associated space of coefficients. Failing to do so introduces a "changed metric" effect that can adversely affect the performance of algorithms. The possible severity of this phenomenon is demonstrated by our computational results in §6. To remove the need to modify nonlinear programming software written for problems defined on a Euclidean space, we introduce coordinate transformations that change our original bases in the control space to an orthonormal set and change the associated coefficient space to a Euclidean space.

Daniel [13] presents one of the first attempts at characterizing, in a general framework, consistency of approximations to an optimization problem as well as an application of this

*Received by the editors May 9, 1994; accepted for publication (in revised form) February 25, 1995. This research was sponsored by National Science Foundation grant ECS-93-02926.

†Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720.

framework to approximations of optimal control problems obtained using the Euler integration formula. It can be shown that Daniel's conditions for consistency imply epiconvergence [2, 14], i.e., the convergence, in the Kuratowski sense [3], of the constrained epigraphs of the approximating problems to the constrained epigraph of the original problem. Epiconvergence ensures convergence of the global minimizers (or strict local minimizers with nonvanishing radii of attraction) of the approximating problems to global minimizers (or local minimizers) of the original problem.

Polak, in [24], characterizes first-order optimality conditions in terms of zeros of *optimality functions*. To define consistency of approximations, he augments the requirement of epiconvergence of the approximating problems with a related requirement for their optimality functions. As a result, consistency, in the Polak sense, ensures convergence of global (local) solutions, and stationary points, of the approximating problems to global (local) solutions, and stationary points, of the original problem. Furthermore, the Polak definition of consistency indirectly imposes the requirement that the mathematical characterization of the constraints of the approximating problems satisfy certain congruence conditions and that derivatives of the approximating problem functions converge to those of the original problem. In addition to a definition of consistency, we find in [24] diagonalization strategies, in the form of master algorithms, that call nonlinear programming algorithms as subroutines. These algorithms enable one to obtain efficiently an approximate, numerical "solution" to an original infinite-dimensional problem.

With the exception of [13] and [24], the analysis of the approximating properties of numerical integration techniques (see, e.g., [7, 10, 11, 19, 20, 30]) in optimal control is not carried in the framework of a general theory.[1] Convergence of global solutions, or in some cases, of stationary points, of approximating problems obtained using Euler integration to those of the original problem was established in [7, 10, 11, 13, 19, 20, 24]. Of these, perhaps the most extensive treatment can be found in [20]. The rate of convergence of stationary points of approximating problems, obtained from discretization of unconstrained optimal control problems using a class of RK methods, to those of the original problem was explored in [15].

Organization. This paper is organized as follows. Section 2 summarizes the theory of consistent approximations. Section 3 defines the optimal control problem and develops an optimality function for it. In §4 the approximating problems are constructed and epiconvergence of the approximating problems is proved. In §5 optimality functions for the approximating problems are derived and are shown to hypoconverge to the optimality function for the original problem. This completes the proof that the approximating problems are consistent approximations to the original problem. Section 6 introduces a transformation that defines orthonormal bases for the control subspaces and presents a rate of convergence result for the most commonly used RK method, that is, RK4. Some numerical results are also included.

Because of the complexity of the notation in this paper, to assist the reader, we begin with a glossary of notation.

Notation.

Spaces and elements.

$\mathbb{R}^n$	Euclidean n-space
$\underset{r}{\times} \mathbb{R}^m$	Cartesian product of r copies of $\mathbb{R}^m$
$L^m_{\infty,2}[0,1]$	$(L^m_\infty[0,1], \langle\cdot,\cdot\rangle_{L^m_2[0,1]}, \|\cdot\|_{L^m_2[0,1]})$
L^i_N	finite-dimensional subspace of $L^m_{\infty,2}[0,1]$, $i = 1, 2$

[1] This is also true for collocation and other techniques (see, e.g., [12, 21, 26, 29, 31]).

$\bar{L}_N^i$	time samples of elements in L_N^i, $i = 1, 2$
$H_{\infty,2}$	$\mathbb{R}^n \times L_{\infty,2}^M[0, 1]$
H_N	$\mathbb{R}^n \times L_N^1$ or $\mathbb{R}^n \times L_N^2$, $H_N \subset H_{\infty,2}$
$\bar{H}_N$	$\mathbb{R}^n \times \bar{L}_N^1$ or $\mathbb{R}^n \times \bar{L}_N^2$
$\bar{u}_k$	$(\bar{u}_k^1, \ldots, \bar{u}_k^r) \in \mathbb{R}^m \times \cdots \times \mathbb{R}^m$
$\bar{u}$	$(\bar{u}_0, \ldots, \bar{u}_{N-1}) \in \bar{L}_N$
η	$\eta = (\xi, u) \in H_{\infty,2}$
η_N	$\eta_N = (\xi, u_N) \in H_N$
$\bar{\eta}$	$\bar{\eta} = (\xi, \bar{u}) \in \bar{H}_N$

Functions.

$\langle \cdot, \cdot \rangle_{\mathcal{H}}$	inner product in Hilbert space $\mathcal{H}$
$\| \cdot \|_{\mathcal{H}}$	norm in Hilbert space $\mathcal{H}$
$V_{\mathbf{A},N}$	$V_{\mathbf{A},N} : L_N \to \bar{L}_N$
$W_{\mathbf{A},N}$	$W_{\mathbf{A},N} : H_N \to \bar{H}_N$, $W_{\mathbf{A},N}((\xi, u)) = (\xi, V_{\mathbf{A},N}(u))$
t_k	$k\Delta$, $\Delta = 1/N$
$\tau_{k,i}$	$t_k + c_i \Delta$
$u[\tau_{k,i}]$	control sample at time $\tau_{k,i}$
$\nabla g(\eta)$	$g_\eta(\eta)^T$
$d\psi(\eta; h)$	directional derivative
$F(x, w)$	right-hand side of difference equation produced by RK discretization

Sets.

$\mathbb{N}$	$\{0, 1, 2, \ldots\}$
$\mathbf{N}$	$\{d^n\}_{n=1}^{\infty}$, d an integer
$\mathcal{N}$	$\{0, 1, 2, \ldots, N-1\}$
$\mathbf{q}$	$\{1, \ldots, q\}$
$\mathbf{r}$	$\{1, \ldots, r\}$
$\mathbf{A}$	Runge–Kutta parameters
$B(x, \rho)$	$\{x' \in \mathcal{H} \mid \|x' - x\|_{\mathcal{H}} \le \rho\}$

Constraint sets.

$U \subset \mathbb{R}^m$	pointwise control constraint set
$\mathbf{U} \subset L_{\infty,2}$	set of feasible controls
$\bar{\mathbf{U}}_N \subset \bar{\mathbf{L}}_N$	$u \in \bar{\mathbf{U}}_N \Rightarrow \bar{u}_k^j \in U$
$\mathbf{U}_N$	$\mathbf{U}_N = V_{\mathbf{A},N}^{-1}(\bar{\mathbf{U}}_N) \subset L_{\infty,2}$
$\mathbf{H}$	$\mathbf{H} = \mathbb{R}^n \times \mathbf{U} \subset H_{\infty,2}$
$\bar{\mathbf{H}}_N$	$\bar{\mathbf{H}}_N = \mathbb{R}^n \times \bar{\mathbf{U}}_N \subset \bar{H}_N$
$\mathbf{H}_N$	$\mathbf{H}_N = \mathbb{R}^n \times V_{\mathbf{A},N}^{-1}(\bar{\mathbf{U}}_N)$

Differential and difference equations.

$x^\eta(t)$	solution at time t of differential equation given $\eta = (\xi, u)$: initial condition ξ and control input u
$\bar{x}_k^{\bar{\eta}}$	solution at time step k of difference equation, resulting from RK discretization, for $\bar{\eta} = (\xi, \bar{u})$: initial condition ξ and control samples $\bar{u}$
$\bar{x}_k^{\eta_N}$	$\bar{x}_k^{\eta_N} = \bar{x}_k^{\bar{\eta}_N}$ with $\bar{\eta} = W_{\mathbf{A},N}(\eta_N)$

Sequences.

$\underline{\lim}_{i \to \infty} x_i$	limit inferior
$\overline{\lim}_{i \to \infty} x_i$	limit superior
$x_i \to^K x$	$\{x_i\}_{i \in K}$ converges to x

2. Consistent approximations. Let $\mathcal{H}$ be a normed linear space and $\mathbf{B} \subset \mathcal{H}$ a convex set, and consider the problem

$$\text{(2.1a)}\quad \mathbf{P} \qquad \min_{\eta \in \mathbf{F}} \psi(\eta)$$

where $\psi : \mathbf{B} \to \mathbb{R}$ is (at least) lower semicontinuous and $\mathbf{F} \subset \mathbf{B}$ is the feasible set. Next, let $\mathbb{N} \triangleq \{1, 2, 3, \ldots\}$, let $\mathbf{N}$ be an infinite subset of $\mathbb{N}$, and let $\{\mathcal{H}_n\}_{N \in \mathbf{N}}$ be a family of finite-dimensional subspaces of $\mathcal{H}$ such that $\mathcal{H}_{N_1} \subset \mathcal{H}_{N_2}$ for all $N_1, N_2 \in \mathbf{N}$ such that $N_1 < N_2$. Now consider a family of approximating problems

$$\text{(2.1b)}\quad \mathbf{P}_N \qquad \min_{\eta_N \in \mathbf{F}_N} \psi_N(\eta_N), \qquad N \in \mathbf{N},$$

where $\psi_N : \mathcal{H}_N \to \mathbb{R}$ is (at least) lower semicontinuous and $\mathbf{F}_N \subset \mathcal{H}_N \cap \mathbf{B}$.

In [24] we find a characterization of the consistency of the approximating problems $\mathbf{P}_N$ in terms of two concepts. The first is epiconvergence of the $\mathbf{P}_N$ to $\mathbf{P}$ [2], which can be shown to be equivalent to Kuratowski convergence [3] of the restricted epigraphs of the cost functions of the approximating problems to the restricted epigraph of the original problem. Epiconvergence does not involve derivatives of the cost function nor the specific description of the constraint sets, hence it is a kind of "zero-order" property. The second concept consists of the characterization of stationary points as zeros of an "optimality function" and a kind of upper semicontinuity property of the optimality functions of the approximating problems. Optimality functions do depend on derivatives and the specific description of the constraint set, hence they add important first-order and structural information.

DEFINITION 2.1. *We will say that the problems in the family* $\{\mathbf{P}_N\}_{N \in \mathbf{N}}$ converge epigraphically *(or* epiconverge*) to* $\mathbf{P}$ ($\mathbf{P}_N \to^{\text{Epi}} \mathbf{P}$) *if*

(a) *for every* $\eta \in \mathbf{F}$, *there exists a sequence* $\{\eta_N\}_{N \in \mathbf{N}}$, *with* $\eta_N \in \mathbf{F}_N$, *such that* $\eta_N \to \eta$ *and* $\overline{\lim} \psi_N(\eta_N) \leq \psi(\eta)$;

(b) *for every infinite sequence* $\{\eta_N\}_{N \in K}$, $K \subset \mathbf{N}$, *satisfying* $\eta_N \in \mathbf{F}_N$ *for all* $N \in K$ *and* $\eta_N \to^K \eta$, *we have that* $\eta \in \mathbf{F}$ *and* $\underline{\lim}_{N \in K} \psi_N(\eta_N) \geq \psi(\eta)$.

There are two subsets involved in our formulation of this definition. The subset $\mathbf{N}$ is used to provide nesting of the finite-dimensional subspaces $\mathcal{H}_N$. The subset $K \subset \mathbf{N}$ is required so that Definition 2.1 is equivalent to Kuratowski convergence.

In [2, 14, 24] we find the following result.

THEOREM 2.2. *Suppose that* $\mathbf{P}_N \to^{\text{Epi}} \mathbf{P}$.

(a) *If, for* $N \in \mathbf{N}$, $\hat{\eta}_N$ *is a global minimizer of* $\mathbf{P}_N$ *and* $\hat{\eta}$ *is any accumulation point of the sequence* $\{\hat{\eta}_N\}_{N \in \mathbf{N}}$, *then* $\hat{\eta}$ *is a global minimizer of* $\mathbf{P}$.

(b) *If, for* $N \in \mathbf{N}$, $\hat{\eta}_N$ *is a strict local minimizer of* $\mathbf{P}_N$ *whose radius of attraction is bounded away from zero and* $\hat{\eta}$ *is any accumulation point of the sequence* $\{\hat{\eta}_N\}_{N \in \mathbf{N}}$, *then* $\hat{\eta}$ *is a local minimizer of* $\mathbf{P}$. □

Epigraphical convergence does not eliminate the possibility of stationary points of $\mathbf{P}_N$ converging to a nonstationary point of $\mathbf{P}$—a most inconvenient outcome from a numerical optimization point of view. For example, let $\mathcal{H} = \mathbb{R}^2$ with $\eta = (x, y)$, and let $f(\eta) = f_N(\eta) = (x-2)^2$, $N \in \mathbb{N}$. Choose

$$\text{(2.2a)}\qquad \mathbf{F} \triangleq \{(x, y) \in \mathbb{R}^2 | x^2 + y^2 - 2 \leq 0\},$$

$$\text{(2.2b)}\quad \mathbf{F}_N \triangleq \{(x, y) \in \mathbb{R}^2 | (x-y)^2(x^2+y^2-2) \leq 0,\ x^2 + y^2 \leq 2 + 1/N\}, \qquad N \in \mathbb{N}.$$

Then we see that $\mathbf{P}_N \to^{\text{Epi}} \mathbf{P}$. Nevertheless, the point (1,1) is feasible and satisfies the F. John optimality condition for all $\mathbf{P}_N$ but is not a stationary point for the problem $\mathbf{P}$. The reason for

this is an incompatibility of the constraint sets $\mathbf{F}_N$ with the constraint set $\mathbf{F}$ which shows up only at the level of optimality conditions. Hypotheses precluding this pathology, at least for first-order nonstationary points, were introduced in [24] using optimality functions as a tool for ensuring a kind of "first-order" approximation result that implicitly enforces convergence of derivatives and restricts the forms chosen for the description of the sets $\mathbf{F}$ and $\mathbf{F}_N$.

DEFINITION 2.3. *We will say that a function* $\theta : \mathbf{B} \to \mathbb{R}$ *is an* optimality function *for* $\mathbf{P}$ *if* (i) $\theta(\cdot)$ *is (at least) upper semicontinuous*; (ii) $\theta(\eta) \leq 0$ *for all* $\eta \in \mathbf{B}$; *and* (iii) *for* $\hat{\eta} = \mathbf{F}$, $\theta(\hat{\eta}) = 0$ *if* $\hat{\eta}$ *is a local minimizer for* $\mathbf{P}$. *Similarly, we will say that function* $\theta_N : H_N \to \mathbb{R}$ *is an* optimality function *for* $\mathbf{P}_N$ *if* (i) $\theta_N(\cdot)$ *is (at least) upper semicontinuous*; (ii) $\theta_N(\eta_N) \leq 0$ *for all* $\eta_N \in H_N$; *and* (iii) *if* $\hat{\eta}_N \in \mathbf{F}_N$ *is a local minimizer for* $\mathbf{P}_N$, *then* $\theta_N(\hat{\eta}_N) = 0$.

DEFINITION 2.4. *Consider the problems* $\mathbf{P}$, $\mathbf{P}_N$, *defined in* (2.1a,b). *Let* $\theta(\cdot)$, $\theta_N(\cdot)$, $N \in \mathbf{N}$, *be optimality functions for* $\mathbf{P}$, $\mathbf{P}_N$, *respectively. We will say that the pairs* $(\mathbf{P}_N, \theta_N)$ *in the sequence* $\{(\mathbf{P}_N, \theta_N)\}_{N \in \mathbf{N}}$ *are* consistent approximations *to the pair* $(\mathbf{P}, \theta)$, *if* (i) $\mathbf{P}_N \to^{\text{Epi}} \mathbf{P}$ *and* (ii) *for any sequence* $\{\eta_N\}_{N \in K}$, $K \subset \mathbf{N}$, *with* $\eta_N \in \mathbf{F}_N$ *for all* $N \in K$, *such that* $\eta_N \to^K \eta$, $\overline{\lim}\, \theta_N(\eta_N) \leq \theta(\eta)$.

Note that part (ii) of Definition 2.4 rules out the possibility of stationary points (points such that $\theta_N(\eta_N) = 0$) for the approximating problems converging to nonstationary points of the original problem. In the sequel, we will prove a stronger condition than is required by Definition 2.4, namely, Kuratowski convergence of the hypographs of $\theta_N(\cdot)$ to the hypograph of $\theta(\cdot)$ (that is, $-\theta_N \to^{\text{Epi}} -\theta$).

In addition to the characterization of consistency, the theory of consistent approximations in [24] includes various master algorithm models for efficiently solving problems such as $\mathbf{P}$. Given a level of discretization defined by N, the master algorithms construct an approximating problem $\mathbf{P}_N$, execute a nonlinear programming or discrete-time optimal control algorithm as a subroutine for a certain number of iterations on $\mathbf{P}_N$, and then increase N. Then the process is repeated. For specific examples, see [16, 25].

3. Problem definition. We will consider optimal control problems with dynamics described by ordinary differential equations of the form

$$\dot{x}(t) = h(x(t), u(t)) \quad \text{a.e. for } t \in [0, 1], \ x(0) = \xi, \tag{3.1}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and hence $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$.

To establish continuity and differentiability of solutions of (3.1) with respect to controls, one must assume that the controls are bounded in $L_\infty^m[0, 1]$. However, the finite-dimensional approximating control subspaces that we will introduce must be treated as Hilbert spaces. This can cause complications in establishing the required approximation properties of the optimality functions for the approximating problems that we will construct. To circumvent this difficulty, we will, as in [24], assume that the controls are elements of the pre-Hilbert space

$$L_{\infty,2}^m[0, 1] \triangleq (L_\infty^m[0, 1], \langle \cdot, \cdot \rangle_2, \| \cdot \|_2), \tag{3.2a}$$

which consists of the elements of $L_\infty^m[0, 1]$ but is endowed with the $L_2^m[0, 1]$ inner product and norm. Note that $L_{\infty,2}^m[0, 1]$ is dense in $L_2^m[0, 1]$.

We will define our optimal control problems on the pre-Hilbert space

$$H_{\infty,2} \triangleq \mathbb{R}^n \times L_{\infty,2}^m[0, 1] \triangleq (\mathbb{R}^n \times L_\infty^m[0, 1], \langle \cdot, \cdot \rangle_H, \| \cdot \|_H), \tag{3.2b}$$

whose elements η consist of pairs of initial states and controls, i.e., $\eta = (\xi, u)$. Note that $H_{\infty,2}$ is a dense subspace of the Hilbert space

$$H_2 = \mathbb{R}^n \times L_2^m[0, 1]. \tag{3.2c}$$

The inner product $\langle \cdot, \cdot \rangle_H$ and norm $\| \cdot \|_H$ on H_2, and hence also on $H_{\infty,2}$, are defined as follows. For any $\eta = (\xi, u) \in H_2$ and $\eta' = (\xi', u') \in H_2$,

$$\langle \eta, \eta' \rangle_H \triangleq \langle \xi, \xi' \rangle + \langle u, u' \rangle_2, \tag{3.2d}$$

where $\langle \xi, \xi' \rangle$ denotes the Euclidean inner product, and the L_2 inner product $\langle u, u' \rangle_2$ is defined by $\langle u, u' \rangle_2 \triangleq \int_0^1 \langle u(t), u'(t) \rangle \, dt$. Consequently, for any $\eta = (\xi, u) \in H_2$,

$$\|\eta\|_H^2 \triangleq \langle \eta, \eta \rangle_H = \|\xi\|^2 + \|u\|_2^2. \tag{3.2e}$$

Next, we introduce a compact, convex control constraint set $U \subset B(0, \rho_{\max}) \triangleq \{u \in \mathbb{R}^m | \|u\| \leq \rho_{\max}\}$, where $\rho_{\max}$ is assumed to be sufficiently large to ensure that all the controls $u(\cdot)$ with which we expect to deal take values in the interior of $B(0, \rho_{\max})$. We then define the set of admissible controls by

$$\mathbf{U} \triangleq \{u \in L_{\infty,2}^m[0, 1] | u(t) \in U, \text{ a.e. on } [0,1]\} \tag{3.3a}$$

and the set of admissible initial state-control pairs by

$$\mathbf{H} \triangleq \mathbb{R}^n \times \mathbf{U} \subset H_{\infty,2}. \tag{3.3b}$$

The set $\mathbf{H}$ is contained in the larger set

$$\mathbf{B} \triangleq \mathbb{R}^n \times \{u \in L_{\infty,2}^m[0, 1] | u(t) \in B(0, \rho_{\max}), \text{ a.e. on } [0,1]\} \subset H_{\infty,2} \tag{3.3c}$$

inside which all of our results concerning differential equations are valid. Finally, solutions of (3.1) corresponding to a particular $\eta \in \mathbf{B}$ will be denoted by $x^\eta(\cdot)$.

We will consider the canonical constrained minimax optimal control problem

$$\mathbf{CP} \qquad \min_{\eta \in \mathbf{H}} \{\psi_o(\eta) | \psi_c(\eta) \leq 0\}, \tag{3.4a}$$

where the objective function $\psi_o : \mathbf{B} \to \mathbb{R}$ and the state endpoint constraint function $\psi_c : \mathbf{B} \to \mathbb{R}$ are defined by

$$\psi_o(\eta) \triangleq \max_{\nu \in \mathbf{q}_o} f^\nu(\eta), \qquad \psi_c(\eta) \triangleq \max_{\nu \in \mathbf{q}_c + q_o} f^\nu(\eta), \tag{3.4b}$$

where the νth function $f^\nu : \mathbf{H} \to \mathbb{R}$ is defined by

$$f^\nu(\eta) \triangleq \zeta^\nu(\xi, x^\eta(1)), \tag{3.4c}$$

with $\zeta^\nu : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, and $\mathbf{q}_o \triangleq \{1, 2, \ldots, q_o\}$, $\mathbf{q}_c \triangleq \{1, 2, \ldots, q_c\}$ (with q_o and q_c positive integers). The set $\mathbf{q}_c + q_o \triangleq \{1 + q_o, \ldots, q_c + q_o\}$. In what follows, we will let $\mathbf{q} \triangleq \{1, 2, \ldots q\}$ with $q = q_o + q_c$. By defining the feasible set $\mathbf{F} \triangleq \{\eta \in \mathbf{H} | \psi_c(\eta) \leq 0\}$, we can write **CP** in the equivalent form of problem **P** in (2.1a).

Various optimal control problems, such as nonautonomous, integral cost, and free-time problems, can be transcribed into this canonical form. Also, the endpoint constraint in (3.4a) can be discarded by setting $\psi_c(\eta) \equiv -\infty$, and control unconstrained problems can be included by setting $U = \mathbf{B}(0, \rho_{\max})$ and choosing $\rho_{\max}$ sufficiently large to ensure that the solutions $u^*(\cdot)$ of **CP** take values in the interior of U.

Properties of the defining functions. We will require the following assumptions.

Assumption 3.1. (a) The function $h(\cdot, \cdot)$ in (3.1) is continuously differentiable, and there exists a Lipschitz constant $\kappa < \infty$ such that for all $x', x'' \in \mathbb{R}^n$ and $v', v'' \in B(0, \rho_{\max})$ the following relations hold:

$$\|h(x', v') - h(x'', v'')\| \leq \kappa[\|x' - x''\| + \|v' - v''\|], \tag{3.5a}$$

$$\|h_x(x', v') - h_x(x'', v'')\| \leq \kappa[\|x' - x''\| + \|v' - v''\|], \tag{3.5b}$$

$$\|h_u(x', v') - h_u(x'', v'')\| \leq \kappa[\|x' - x''\| + \|v' - v''\|], \tag{3.5c}$$

(b) The functions $\zeta^\nu(\cdot, \cdot)$, $\zeta^\nu_\xi(\cdot, \cdot)$, and $\zeta^\nu_x(\cdot, \cdot)$, with $\nu \in \mathbf{q}$, are Lipschitz continuous on bounded sets.

The following results can be found in [4].

THEOREM 3.2. *If Assumption* 3.1 *is satisfied, then*

(i) *there exists a* $\kappa < \infty$ *such that for all* $\eta', \eta'' \in \mathbf{B}$ *and for all* $t \in [0, 1]$

$$\|x^{\eta'}(t) - x^{\eta''}(t)\| \leq \kappa \|\eta' - \eta''\|_H;$$

(ii) *there exists an* $L < \infty$ *such that for all* $\eta \in \mathbf{B}$ *and all* $t \in [0, 1]$

$$\|x^\eta(t)\| \leq L(1 + \|\xi\|);$$

(iii) *the functions* $\psi_0 : \mathbf{B} \to \mathbb{R}$ *and* $\psi_c : \mathbf{B} \to \mathbb{R}$ *are Lipschitz continuous on bounded sets*;

(iv) *the functions* $f^\nu(\cdot)$, $\nu \in \mathbf{q}$, *have continuous Gateaux differentials* $Df^\nu : \mathbf{B} \times H_{\infty,2} \to \mathbb{R}$ *that have the form* $Df^\nu(\eta; \delta\eta) = \langle \nabla f^\nu(\eta), \delta\eta \rangle_H$;

(v) *the gradients* $\nabla f^\nu : \mathbf{B} \to H_{\infty,2}$, $\nabla f^\nu(\eta) = (\nabla_\xi f^\nu(\eta), \nabla_u f^\nu(\eta))$, $\nu \in \mathbf{q}$, *are given by*

$$\nabla_\xi f^\nu(\eta) = \nabla_\xi \zeta^\nu(\xi, x^\eta(1)) + p^{\nu,\eta}(0), \tag{3.6a}$$

$$\nabla_u f^\nu(\eta)(t) = h_u(x(t), u(t))^T p^{\nu,\eta}(t), \quad \forall t \in [0, 1], \tag{3.6b}$$

where $p^{\nu,\eta}(t) \in \mathbb{R}^n$ *is the solution to the adjoint equation*

$$\dot{p}^\nu = -h_x(x^\eta, u)^T p^\nu, \qquad p^\nu(1) = \nabla_x \zeta^\nu(\xi, x^\eta(1)), \quad t \in [0, 1], \tag{3.6c}$$

and are Lipschitz continuous on bounded sets in $\mathbf{B}$.

An optimality function. Referring to [9], the following result holds because of Theorem 3.2.

THEOREM 3.3. *For any* $\eta \in \mathbf{B}$ *let*

$$\psi_c(\eta)_+ \triangleq \max\{0, \psi_c(\eta)\}, \tag{3.7a}$$

and for any $\eta, \eta' \in \mathbf{B}$ *and* $\sigma > 0$ *let*

$$\Psi(\eta, \eta') \triangleq \max\{\psi_0(\eta) - \psi_0(\eta') - \sigma\psi_c(\eta')_+, \psi_c(\eta) - \psi_c(\eta')_+\}. \tag{3.7b}$$

If Assumption 3.1 *is satisfied and* $\hat{\eta} \in \mathbf{H}$ *is a local minimizer of the problem* **CP**, *then*

$$d_2\Psi(\hat{\eta}, \hat{\eta}; \eta - \hat{\eta}) \geq 0, \quad \forall \eta \in \mathbf{H}, \tag{3.8}$$

where $d_2\Psi$ *indicates the directional derivative of* $\Psi(\cdot, \cdot)$ *with respect to its second argument.*

Next we define an optimality function $\theta : \mathbf{B} \to \mathbb{R}$ for **CP**. For any $\eta, \eta' \in \mathbf{B}$ and $\nu \in \mathbf{q}$, we define a first-order quadratic approximation to $f^\nu(\cdot)$ at η by

$$\tilde{f}^\nu(\eta, \eta') \triangleq f^\nu(\eta) + \langle \nabla f^\nu(\eta), \eta' - \eta \rangle_H + \frac{1}{2}\|\eta' - \eta\|_H^2. \tag{3.9a}$$

We define the optimality function, with the same fixed $\sigma > 0$ used in (3.7b), by

$$\theta(\eta) \triangleq \min_{\eta' \in \mathbf{H}} \max \left\{ \max_{\nu \in \mathbf{q}_o} \tilde{f}^\nu(\eta, \eta') - \psi_o(\eta) - \sigma\psi_c(\eta)_+, \max_{\nu \in \mathbf{q}_c + q_o} \tilde{f}^\nu(\eta, \eta') - \psi_c(\eta)_+ \right\}. \tag{3.9b}$$

The existence of the minimum in (3.9b) follows from the convexity of the constraint set **H** and of the max functions in (3.9b) with respect to η' and the fact that $\tilde{f}^\nu(\eta, \eta') \to \infty$ as $\|\eta'\| \to \infty$ [6, Cor. III.20, p. 46]. Note that if $f^\nu(\eta) \equiv -\infty$ for all $\nu \in \mathbf{q}_c + q_o$, so that $\psi_c(\eta) \equiv -\infty$, then (3.9b) reduces to

$$\theta(\eta) \triangleq \min_{\eta' \in \mathbf{H}} \max_{\nu \in \mathbf{q}_o} f^\nu(\eta) + \langle \nabla f^\nu(\eta), \eta' - \eta \rangle_H + \frac{1}{2}\|\eta' - \eta\|_H^2 - \psi_o(\eta). \tag{3.9c}$$

Referring once again to [4], we find the following result.

THEOREM 3.4. *Let* $\theta : \mathbf{B} \to \mathbb{R}$ *be defined by* (3.9b). *If Assumption* 3.1 *holds, then* (i) $\theta(\cdot)$ *is negative valued and continuous and* (ii) *the relation* (3.8) *holds if and only if* $\theta(\hat{\eta}) = 0$.

4. Approximating problems. The construction of a family of approximating problems for our problem **CP**, in (3.4a), satisfying the axioms of the theory of consistent approximations, requires the construction of nested families of finite-dimensional subspaces of the initial state-control space $H_{\infty,2}$, approximating cost functions, and approximating constraint sets. Our selection of these approximations is largely determined by our intention to use explicit, fixed step-size Runge–Kutta (RK) methods [8,17] for integrating the dynamic equations (3.1).

Finite-dimensional initial-state-control subspaces. We begin by defining families of finite-dimensional subspaces H_N, with $H_N = \mathbb{R}^n \times L_N \subset H_{\infty,2}$, where the L_N are finite-dimensional subspaces of $L^m_{\infty,2}[0, 1]$, spanned by piecewise-continuous functions to which RK methods can be extended. Hence, given an explicit, fixed step-size RK integration method, using step-size $\Delta = 1/N$, we impose the following conditions on the subspaces L_N.

(i) For any bounded subset S of **B**, there exists a $\kappa < \infty$ such that for any $\eta \in S \cap H_N$ the RK method results in an integration error no greater than κ/N in solving the differential equation (3.1).

(ii) The data used by the RK integration method are an initial state and a set of control samples.[2] We will require that each set of control samples correspond to a unique element $u \in L_N$.

Condition (i) will be needed to prove that our approximating problems epiconverge to the original problem. For the subspaces L_N that we will present, we will actually be able to prove more than first-order accuracy. Condition (ii) facilitates the definition of the approximating problems and makes it possible to define gradients for the approximating cost and constraint functions.

We will now show how the choice of an RK integration method affects the selection of the subspaces L_N. The generic, explicit fixed step-size, s-stage RK method computes an approximate solution to a differential equation of the form

$$\dot{x}(t) = \tilde{h}(t, x(t)), \quad x(0) = \xi, \quad t \in [0, 1], \tag{4.1a}$$

[2]The term *control samples* will be clarified shortly.

where $\tilde{h} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous in t and Lipschitz continuous in x. It does so by solving the difference equation

$$\bar{x}_{k+1} = \bar{x}_k + \Delta \sum_{i=1}^{s} b_i K_{k,i}, \quad \bar{x}_0 = x(0) = \xi, \quad k \in \mathcal{N} \triangleq \{0, 1, \ldots, N-1\}, \tag{4.1b}$$

with $\Delta = 1/N$, $t_k \triangleq k\Delta$, and $K_{k,i}$ defined by the recursion

$$K_{k,1} = \tilde{h}(t_k + c_1\Delta, \bar{x}_k), \quad K_{k,i} = \tilde{h}\left(t_k + c_i\Delta, \bar{x}_k + \Delta \sum_{j=1}^{i-1} a_{i,j} K_{k,j}\right), \quad i = 2, \ldots, s. \tag{4.1c}$$

The variable $\bar{x}_k$ is the computed estimate of $x(t_k)$.

The parameters $a_{i,j}$, c_i, and b_i in (4.1b) and (4.1c) determine the RK method. These parameters are collected in the Butcher array $\mathbf{A} = [c, A, b]$. The following assumption on the b parameters will hold throughout this paper (conditions on the c parameters will be added later).

Assumption 4.1. For all $i \in \mathbf{s}$, $b_i > 0$ and $\sum_{i=1}^{s} b_i = 1$.

Remark 4.2. The condition $\sum_{i=1}^{s} b_i = 1$ is satisfied by all convergent RK methods. Other conditions must be satisfied to achieve higher order convergence for multistage RK methods.

Now, in our case, $\tilde{h}(t, x) = h(x, u(t))$ and the elements $u(\cdot)$ of the subspaces L_N will be allowed to be discontinuous from the left at the points $t_k + c_i\Delta$ (i.e., $\lim_{t \uparrow \tau} u(t) \neq u(\tau)$, for $\tau = t_k + c_i\Delta$). To obtain an accurate integration method for such functions, the values $u(t_k + c_i\Delta)$ must sometimes be replaced by left limits, as appropriate for the particular choice of the subspace L_N. We will refer to these values as "control samples" and denote them by $u[\tau_{k,i}]$, where we have introduced the notation

$$\tau_{k,i} \triangleq t_k + c_i\Delta. \tag{4.2}$$

The times $\tau_{k,i}$ at which $u[\tau_{k,i}]$ is a left-limit are dictated by the definition of L_N. Clearly if $u(\cdot)$ is continuous at $t_k + c_i\Delta$, then $u[t_k + c_i\Delta] = u(t_k + c_i\Delta)$.

The recursion (4.1c) evaluates $\tilde{h}(\cdot, \cdot)$ s times for each time step $k \in \mathcal{N}$. If we collect the corresponding s control samples into a matrix $\omega_k \triangleq (u[\tau_{k,1}] \ldots u[\tau_{k,s}])$, we can replace (4.1b) and (4.1c) with

$$\bar{x}_{k+1} = \bar{x}_k + \Delta \sum_{i=1}^{s} b_i K_{k,i}, \quad \bar{x}_0 = x(0) = \xi, \quad k \in \mathcal{N}, \tag{4.3a}$$

where $K_{k,i} \triangleq K_i(\bar{x}_k, \omega_k)$, which is defined by the recursion

$$K_1(x, \omega) = h(x, \omega^1), \qquad K_i(x, \omega) = h\left(x + \Delta \sum_{j=1}^{i-1} a_{i,j} K_j(x, \omega), \omega^i\right), \quad i = 2, \ldots, s, \tag{4.3b}$$

where ω^i is the ith column of ω.

We will define the control subspace L_N in such a way that there is a one-to-one correspondence between elements $u \in L_N$ and the samples of $u[t_k + c_i\Delta]$ used by the RK method. The definition of L_N is somewhat complicated by the fact that some of the c_i elements of the Butcher array may have the same value. This causes the RK method to use samples at times $t_k + c_i\Delta$ more than once and hence leads to a reduction of the dimension in the associated subspace L_N. To keep track of the distinct values of the c_i elements of the Butcher array, we define the ordered set of indices

$$I \triangleq \{i_1, i_2, \ldots, i_r\} = \{i \in \mathbf{s} | c_j \neq c_i, \ \forall j \in \mathbf{s}, \ j < i\} \tag{4.4a}$$

and let

$$I_j \triangleq \{i \in \mathbf{s} | c_i = c_{i_j},\ i_j \in I\}, \qquad j \in \mathbf{r}. \tag{4.4b}$$

Thus, the total number of distinct values taken by the elements c_i in the Butcher array is r. For example, if $c = \{0, 1/2, 1/2, 1\}$ (as in the most commonly used fourth-order RK method), then $r = 3$, $I = \{i_1 = 1, i_2 = 2, i_3 = 4\}$, $I_1 = \{1\}$, $I_2 = \{2, 3\}$, and $I_3 = \{4\}$. If each c_i is distinct, then $r = s$, $i_j = j$, and I_j is the singleton $\{j\}$. Otherwise, $r < s$ and $i_j \geq j$.

Clearly, the r distinct sampling times in the interval $[t_k, t_{k+1}]$, $k \in \mathcal{N}$, are given by τ_{k,i_j}, $j \in \mathbf{r}$, $i_j \in I$. Corresponding to each sampling time there is a control sample $u[\tau_{k,i_j}]$. The collection of these control samples can be viewed as a vector $\bar{u} \in \times_N \times_r \mathbb{R}^m$, where the symbol $\times_N$ indicates the Cartesian product of N spaces. We will partition vectors $\bar{u} \in \times_N \times_r \mathbb{R}^m$ into N blocks as

$$\bar{u} = (\bar{u}_0, \bar{u}_1, \ldots, \bar{u}_{N-1}), \tag{4.5a}$$

where each block $\bar{u}_k \in \times_r \mathbb{R}^m$, $k \in \mathcal{N}$, is of the form

$$\bar{u}_k = (\bar{u}_k^1, \ldots, \bar{u}_k^r), \tag{4.5b}$$

and $\bar{u}_k^j \in \mathbb{R}^m$, $j \in \mathbf{r}$, corresponds to the samples $u[\tau_{k,i_j}]$, $i_j \in I$, used by the RK integration during the kth time interval. Our algebraic expressions are simplified if we treat $\bar{u}$ as the $m \times Nr$ matrix $[\bar{u}_0^1 \ldots \bar{u}_0^r \ldots \bar{u}_{N-1}^1 \ldots \bar{u}_{N-1}^r]$; i.e., we will identify $\times_N \times_r \mathbb{R}^m$ with the space $\mathbb{R}^{m \times Nr}$ of $m \times Nr$ matrices. Similarly, in algebraic expressions, we will treat $\bar{u}_k$ as the $m \times r$ matrix $[\bar{u}_k^1 \ldots \bar{u}_k^r]$. The standard inner product on $\times_N \times_r \mathbb{R}^m$ is the l_2 inner product given by

$$\langle \bar{u}, \bar{v} \rangle_{l_2} = \sum_{N=0}^{N-1} \sum_{j=1}^{r} \langle \bar{u}_k^j, \bar{v}_k^j \rangle. \tag{4.5c}$$

Let G be the $r \times s$ matrix defined by

$$G = \begin{bmatrix} \mathbf{1}_1 & & & \\ & \mathbf{1}_2 & & \\ & & \ddots & \\ & & & \mathbf{1}_r \end{bmatrix}, \tag{4.5d}$$

where, for each $j \in \mathbf{r}$, $\mathbf{1}_j = (1, 1, \ldots, 1)$ is a row vector of dimension $|I_j|$ ($|I_j|$ is the number of elements in I_j). Then we can associate the components $\bar{u}_k$, $k \in \mathcal{N}$, of a vector $\bar{u} \in \times_N \times_r \mathbb{R}^m$, with the matrices ω_k used by the RK method (4.3a,b) by setting $\omega_k = \bar{u}_k G = [\bar{u}_k^1 \ldots \bar{u}_k^r] G$, $k \in \mathcal{N}$.

We now present two control representations that define subspaces $L_N^i \subset L_{\infty,2}^m[0, 1]$, $i = 1, 2$, $N \in \mathbb{N}$, of dimension Nrm, such that $\bigcup_{N=1}^{\infty} L_N^1$ and $\bigcup_{N=1}^{\infty} L_N^2$ are dense in $L_{\infty,2}^m[0, 1]$. Both representations reduce to simple square pulses for Euler's method ($r = 1$). The basis functions $\{e_l \Phi_{N,k,j}^i\}_{j=1,k=1,l=1}^{N,r,m}$, $i = 1, 2$, with e_l the lth unit vector in $\mathbb{R}^m$ and $\Phi_{N,l,k}^i : [0, 1] \to \mathbb{R}$, that we use to construct the spaces L_N^i are not orthonormal. Hence, for numerical calculations, we associate with these spaces Nrm-dimensional spaces of real coefficients of the form

$$\bar{L}_N^i \triangleq \left(\underset{N}{\times} \underset{r}{\times} \mathbb{R}^m, \langle \cdot, \cdot \rangle_{\bar{L}_N^i}, \| \cdot \|_{\bar{L}_N^i} \right), \qquad i = 1, 2,\ N \in \mathbb{N}, \tag{4.5e}$$

where the inner products and norms are chosen so that for any $u, u' \in L_N^i$, with $u(t) = \sum_{j=1,k=1}^{N,r} \bar{u}_k^j \Phi_{N,j,k}^i(t)$ and $u'(t) = \sum_{j=1,k=1}^{N,r} \bar{u}_k'^j \Phi_{N,j,k}^i(t)$, $t \in [0, 1]$,

$$\langle u, u' \rangle_2 = \langle \bar{u}, \bar{u}' \rangle_{\bar{L}_N^i}, \qquad \|u\|_2 = \|\bar{u}\|_{\bar{L}_N^1}, \tag{4.5f}$$

where $\bar{u} \in \bar{L}_N^i$ is defined as in (4.5a, b). The spaces $\bar{L}_N^i$ will be needed in defining gradients for the cost and constraint functions of the approximating problems as well as in setting up numerical implementations of optimal control algorithms. The reason that we choose an L_2 norm preserving, nonstandard inner product on $\bar{L}_N^i$ is that if we had elected to use the standard l_2 inner product and norm on $\bar{L}_N^i$ (as is commonly done), we might have, unwittingly, caused serious deterioration in the performance of numerical algorithms that solve the approximating problems in the coefficient spaces $\bar{L}_N^i$. The extent of this ill-conditioning effect is illustrated in §6. Of course, if our basis for L_N had been orthonormal, then a standard l_2 inner product would be the appropriate choice.

Representation (R1). Piecewise rth-order polynomials.

Assumption 4.3. For all $i \in \mathbf{s}$, $c_i \in [0, 1]$.

For each $k \in \mathcal{N}$, define the sub-intervals $T_k^1 \triangleq [t_k, t_{k+1})$ and define pulse functions

$$\Pi_{N,k}^1(t) \triangleq \begin{cases} 1 & \text{if } t \in T_k^1, \\ 0 & \text{elsewhere;} \end{cases} \tag{4.6a}$$

and let

$$L_N^1 \triangleq \{u \in L_2^m[0, 1] | u(t) = \sum_{k=0}^{N-1} \sum_{j=1}^{r} \bar{u}_k^j \Phi_{N,k,j}^1(t),\ \bar{u}_{k,j} \in \mathbb{R}^m,\ \forall t \in [0, 1]\}, \tag{4.6b}$$

where

$$\Phi_{N,k,j}^1(t) \triangleq \phi_{N,k,j} \Pi_{N,k}^1(t), \qquad k \in \mathcal{N}, \tag{4.6c}$$

and $\phi_{N,k,j}(t)$ is the jth Lagrange polynomial for the points $\{\tau_{k,i_j}\}_{j=1}^r$, $i_j \in I$, defined by

$$\phi_{N,k,j}(t) \triangleq \prod_{\substack{l=1 \\ (l \neq j)}}^{r} \frac{(t - \tau_{k,i_l})}{(\tau_{k,i_j} - \tau_{k,i_l})}, \qquad k \in \mathcal{N},\ j \in \mathbf{r}, \tag{4.6d}$$

with the property that $\phi_{N,k,j}(\tau_{k,i_l}) = 1$ if $l = j$ and $\phi_{N,k,j}(\tau_{k,i_l}) = 0$ if $l \neq j$. By construction of the set I, $i_l, i_j \in I$ implies that $\tau_{k,i_j} \neq \tau_{k,i_l}$ if $l \neq j$. Hence, the functions $\phi_{N,k,j}(\cdot)$ are well defined and the functions $\Phi_{N,k,j}^1(\cdot)$ are linearly independent. For L_N^1 we define the control samples as

$$u[\tau_{k,i}] \triangleq \begin{cases} u(\tau_{k,i}) & \text{if } \tau_{k,i} \in T_k^i, \\ \lim\limits_{t \uparrow \tau_{k,i}} u(t) & \text{if } \tau_{k,i} = t_{k+1}, \end{cases} \quad k \in \mathcal{N}, i \in I. \tag{4.6e}$$

PROPOSITION 4.4. *Let L_N^1 be defined as in* (4.6b), *and let $V_{\mathbf{A},N}^1 : L_N^1 \to \times_N \times_r \mathbb{R}^m$ be defined by $V_{\mathbf{A},N}^1(u) = \bar{u}$, with $\bar{u}_k^j = u[\tau_{k,i_j}]$, $i_j \in I$, $j \in \mathbf{r}$, $k \in \mathcal{N}$. Suppose Assumption* 4.3 *holds. Then $V_{\mathbf{A},N}^1$ is invertible.*

Proof. Let $u(t) = \sum_{j=0}^{N-1} \sum_{k=1}^{r} \bar{u}_k^j \Phi_{N,k,j}^1(t)$ be an arbitrary element of L_N^1. Assumption 4.3 implies that $\tau_{k,i_j} \in [t_k, t_{k+1}]$. Next, it follows from (4.6e) that $u[\tau_{k,i_j}] = \sum_{j=1}^{r} \bar{u}_k^j \phi_{N,k,j}(\tau_{k,i_j}) = \bar{u}_k^j$ because of the interpolation property of Lagrange polynomials. Hence $V_{\mathbf{A},N}^1$ is invertible. □

The polynomial pulse functions $\{\Phi^1_{N,k,j}(t)\}_{k=0,j=0}^{N-1,r-1}$ are linearly independent but are neither orthogonal nor normal with respect to the L_2 inner product and norm. To complete the definition of the spaces $\bar{L}^1_N$ in (4.5e), we now define the required inner product, which, in turn, defines the norm. First, let $u \in L^1_N$ and note that we can write each rth-order polynomial piece $\sum_{j=1}^r \bar{u}^j_k \phi_{N,k,j}(t)$ in (4.6b) as a power series $\alpha_k P(t - t_k)$, where α_k is an $m \times r$ matrix of coefficients and the function $P : \mathbb{R} \to \mathbb{R}^r$ is defined by

$$P(t) \triangleq [1 \; t/\Delta \ldots (t/\Delta)^{r-1}]^T. \tag{4.7}$$

If $\bar{u} = V^1_{\mathbf{A},N}(u)$, then from Proposition 4.4, $\bar{u}^j_k = \alpha_k P(c_{i_j}\Delta)$, $j \in \mathbf{r}$, $i_j \in I$. Hence, $\bar{u}_k = [\bar{u}^1_k \ldots \bar{u}^r_k] = \alpha_k T^{-1}$, where

$$T^{-1} \triangleq \left[P(c_{i_1}\Delta)P(c_{i_2}\Delta)\ldots P(c_{i_r}\Delta)\right] = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ c_{i_1} & c_{i_2} & & c_{i_r} \\ \vdots & & \ddots & \\ c_{i_1}^{r-1} & c_{i_2}^{r-1} & & c_{i_r}^{r-1} \end{bmatrix}_{r\times r}. \tag{4.8}$$

The matrix T^{-1} is a Vandermonde matrix and the r values c_{i_j}, $i_j \in I$, are distinct. Therefore, T^{-1} is nonsingular and $\alpha_k = \bar{u}_k T$. Hence, for each $k \in \mathcal{N}$, $u(t) = \bar{u}_k T P(t - t_k)$ for $t \in [t_k, t_{k+1})$.

We now define the inner product between two vectors $\bar{u}, \bar{v} \in \bar{L}^1_N$, with $u = (V^1_{\mathbf{A},N})^{-1}(\bar{u})$ and $v = (V^1_{\mathbf{A},N})^{-1}(\bar{v})$, by

$$\begin{aligned} \langle \bar{u}, \bar{v}\rangle_{\bar{L}^1_N} = \langle u, v\rangle_2 &= \sum_{k=0}^{N-1} \int_0^\Delta \langle u(t_k + t), v(t_k + t)\rangle \, dt \\ &= \sum_{k=0}^{N-1} \int_0^\Delta \langle \bar{u}_k T P(t), \bar{v}_k T P(t)\rangle \, dt \\ &= \Delta \sum_{k=0}^{N-1} \operatorname{trace}\left(\bar{u}_k T \frac{1}{\Delta}\int_0^\Delta P(t)P(t)^T \, dt T^T \bar{v}_k^T\right) \\ &= \Delta \sum_{k=0}^{N-1} \operatorname{trace}(\bar{u}_k M_1 \bar{v}_k^T), \end{aligned} \tag{4.9a}$$

where T was defined by (4.8), $P(\cdot)$ was defined in (4.6d), and

$$M_1 \triangleq T\left[\frac{1}{\Delta}\int_0^\Delta P(t)P(t)^T \, dt\right]T^T = T \operatorname{Hilb}(r) T^T \tag{4.9b}$$

is an $r \times r$ symmetric, positive definite matrix with

$$\operatorname{Hilb}(r) = \begin{bmatrix} 1 & 1/2 & 1/3 & & 1/r \\ 1/2 & 1/3 & 1/4 & \cdots & 1/(r+1) \\ 1/3 & 1/4 & 1/5 & & \\ & \vdots & & \ddots & \\ 1/r & 1/(r+1) & & & 1/(2r-1) \end{bmatrix}_{r\times r} \tag{4.9c}$$

the Hilbert matrix whose i, jth entry is $1/(i + j - 1)$. Note that both $\operatorname{Hilb}(r)$ and T are ill-conditioned matrices. However, the product in (4.9b) is well-conditioned (the product corresponds to switching from the power-series polynomial representation back to the Lagrange

expansion). The matrix M_1 is positive definite because Hilb(r) is positive definite and T is nonsingular. Given $\bar{u} \in \bar{L}_N^1$, its norm is $\|\bar{u}\|_{\bar{L}_N^1}^2 = \langle \bar{u}, \bar{u} \rangle_{\bar{L}_N^1}$.

Remark 4.5. A special class of functions with representation R1 is the subspace of rth-order, m-dimensional splines [5]. The dimension of the spline subspace is only a fraction of the dimension of L_N^1. Our results for R1 can be extended to splines; this extension is presented in [28].

Representation (R2). Piecewise constant functions.

For $j \in \mathbf{r}$, I_j defined in (4.4b), let

$$\tilde{b}_j \triangleq \sum_{i \in I_j} b_i, \tag{4.10a}$$

$$d_j \triangleq \Delta \sum_{i=1}^{j} \tilde{b}_i, \qquad d_0 \triangleq 0. \tag{4.10b}$$

If all the c_i elements of the Butcher array have distinct values, then $d_j = \Delta \sum_{i=1}^{j} b_i$. At this point, we can replace Assumption 4.1 with the following weaker assumption.

Assumption 4.1′. For all $j \in \mathbf{r}$, $\tilde{b}_j > 0$ and $d_r = \Delta$.

Note that Assumption 4.1′ implies that for all $j \in \mathbf{r}$, $d_j > d_{j-1}$ and that $t_k + d_j \in [t_k, t_{k+1}]$, $k \in \mathcal{N}$.

Next, we introduce an additional assumption which is stronger than Assumption 4.3.

Assumption 4.6. For $j \in \mathbf{r}$ and $i_j \in I$, $d_{j-1} \le c_{i_j} \Delta \le d_j$, so that $\tau_{k,i_j} \in [t_k + d_{j-1}, t_k + d_j]$.

With $T_{k,j}^2 \triangleq [t_k + d_{j-1}, t_k + d_j)$ define the basis functions $\Phi_{N,k,j}^2 : \mathbb{R} \to \mathbb{R}$, $k \in \mathcal{N}$, $j \in \mathbf{r}$, by

$$\Phi_{N,k,j}^2(t) \triangleq \begin{cases} 1 & \text{if } t \in T_{k,j}^2, \\ 0 & \text{elsewhere} \end{cases} \tag{4.11a}$$

and let

$$L_N^2 \triangleq \{u \in L_2^m[0,1] | u(t) = \sum_{k=0}^{N-1} \sum_{j=1}^{r} \bar{u}_k^j \Phi_{N,k,j}^2(t),\ \bar{u}_k^j \in \mathbb{R}^m,\ \forall t \in [0,1]\}. \tag{4.11b}$$

For L_N^2, we define the control samples as

$$u[\tau_{k,i_j}] \triangleq \begin{cases} u(\tau_{k,i_j}) & \text{if } \tau_{k,i_j} \in T_{k,j}^2, \\ \lim\limits_{t \uparrow \tau_{k,i_j}} u(t) & \text{if } \tau_{k,i_j} = t_k + d_j, \end{cases} \qquad k \in \mathcal{N},\ i_j \in I,\ j \in \mathbf{r}. \tag{4.11c}$$

PROPOSITION 4.7. *Let L_N^2 be defined as in* (4.11b); *and let $V_{\mathbf{A},N}^2 : L_N^2 \to \times_N \times_r \mathbb{R}^m$ be defined by $V_{\mathbf{A},N}^2(u) = \bar{u}$, with $\bar{u}_k^j = u[\tau_{k,i_j}]$, $j \in \mathbf{r}$, $i_j \in I$, $k \in \mathcal{N}$. Suppose Assumptions* 4.1′ *and* 4.6 *hold. Then $V_{\mathbf{A},N}^2$ is invertible.*

Proof. Assumption 4.1′ ensures that the support for each $\Phi_{N,k,j}^2(\cdot)$ is of nonzero length. This ensures a one-to-one correspondence between the elements of L_N^2 and the vector coefficients $\bar{u}_k^j$ in (4.11b). Next, Assumption 4.6 together with the definition (4.11c) of $u[\cdot]$ implies that for any $u \in L_N^2$, with $u(t) = \sum_{k=0}^{N-1} \sum_{j=1}^{r} \bar{u}_k^j \Phi_{N,k,j}^2(t)$, $u[\tau_{k,j}] = \bar{u}_k^j$ for all $k \in \mathcal{N}$ and $j \in \mathbf{r}$. Hence, $V_{\mathbf{A},N}^2$ is invertible. □

To complete the definition, in (4.5e), of the spaces $\bar{L}_N^2$ we will now define the required inner product and norm. We define the inner product between two vectors $\bar{u}, \bar{v} \in \bar{L}_N^2$, with

$u = (V^2_{\mathbf{A},N})^{-1}(\bar{u})$ and $v = (V^2_{\mathbf{A},N})^{-1}(\bar{v})$, by

$$\begin{aligned}\langle \bar{u}, \bar{v}\rangle_{\bar{L}^2_N} = \langle u, v\rangle_2 &= \sum_{k=0}^{N-1}\sum_{j=1}^{r}\int_{d_{j-1}}^{d_j}\langle u(t_k+t), v(t_k+t)\rangle\, dt \\ &= \Delta\sum_{k=0}^{N-1}\sum_{j=1}^{r}\tilde{b}_j\langle \bar{u}_k^j, \bar{v}_k^j\rangle\, dt \\ &= \Delta\sum_{k=0}^{N-1}\operatorname{trace}(\bar{u}_k M_2 \bar{v}_k), \qquad (4.12a)\end{aligned}$$

where

$$M_2 = \begin{bmatrix} \tilde{b}_1 & & 0 \\ & \ddots & \\ 0 & & \tilde{b}_r \end{bmatrix}. \qquad (4.12b)$$

Since all $\tilde{b}_j > 0$, M_2 is diagonal, positive definite. Given $\bar{u} \in \bar{L}^2_N$, its norm is $\|\bar{u}\|^2_{\bar{L}^2_N} = \langle \bar{u}, \bar{u}\rangle_{\bar{L}^2_N}$.

Remark 4.8. In place of (4.10b), we could have used the alternate definition $d_j \triangleq \Delta\sum_{i=1}^{j} b_i$ and set $\bar{u}_k^j = u[\tau_{k,j}]$ for all $j \in \mathbf{s}$, $k \in \mathcal{N}$. In this way, samples corresponding to repeated values of c_j in the Butcher array would be treated as independent values and the space L_N would have to be correspondingly enlarged. However, Proposition 6.1 in §6 indicates that (4.10b) is the preferable definition.

Definition of approximating problems. For $N \in \mathbb{N}$ let

$$H_N \triangleq \mathbb{R}^n \times L_N, \qquad (4.13a)$$

where $L_N = L^1_N$ for representation R1 or $L_N = L^2_N$ for representation R2. Since $H_N \subset H_{\infty,2}$, it inherits the inner product from $H_{\infty,2}$ which, for $\eta', \eta'' \in H_N$, with $\eta' = (\xi', u')$ and $\eta = (\xi'', u'')$, is given by

$$\langle \eta', \eta''\rangle_H \triangleq \langle \xi', \xi''\rangle + \langle u', u''\rangle_2. \qquad (4.13b)$$

Also, for any $\eta \in H_N$, $\|\eta\|^2_H = \langle \eta, \eta\rangle_H$. Similarly, for $N \in \mathbb{N}$, we define the coefficient spaces $\overline{H}_N$ by

$$\overline{H}_N \triangleq \mathbb{R}^n \times \bar{L}_N, \qquad (4.14a)$$

where $\bar{L}_N = \bar{L}^1_N$ or $\bar{L}_N = \bar{L}^2_N$. The inner product on $\overline{H}_N$ is defined by

$$\langle \bar{\eta}', \bar{\eta}''\rangle_{\bar{H}_N} \triangleq \langle \xi', \xi''\rangle + \langle \bar{u}', \bar{u}''\rangle_{\bar{L}_N}, \qquad (4.14b)$$

and the norm correspondingly. Let $W_{\mathbf{A},N} : H_N \to \overline{H}_N$ be defined by $W_{\mathbf{A},N}(\eta) = (\xi, V^1_{\mathbf{A},N}(u))$ for representation R1 and $W_{\mathbf{A},N}(\eta) = (\xi, V^2_{\mathbf{A},N}(u))$ for representation R2, where $\eta = (\xi, u)$. Then we see that $W_{\mathbf{A},N}$ is a nonsingular map and, with our definition of the norms on $\overline{H}_N$, provides an isometric isomorphism between H_N and $\overline{H}_N$. Thus, we can use the spaces H_N and $\overline{H}_N$ interchangeably.

Next, we define control constraint sets for the approximating problems as follows. Let U be the convex, compact set used to define $\mathbf{U}$ in (3.3a). Then, with $\kappa_U < \infty$, we define

$$\overline{\mathbf{U}}^1_N \triangleq \left\{\bar{u} \in \bar{L}^1_N \,|\, \bar{u}^j_k \in U,\ j \in \mathbf{r},\ \|\bar{u}_k T_j\|_\infty \le \frac{\Delta}{(j-1)(r-1)}\kappa_U,\ j = 2, \ldots, r,\ \forall k \in \mathcal{N}\right\}, \qquad (4.15a)$$

$$\overline{\mathbf{U}}^2_N \triangleq \left\{\bar{u} \in \bar{L}^2_N \,|\, \bar{u}^j_k \in U,\ \forall j \in \mathbf{r},\ k \in \mathcal{N}\right\}, \qquad (4.15b)$$

where T_j is the jth column of the matrix T, defined by its inverse in (4.8), and $\Delta = 1/N$, as before. Finally, we define the constraint sets for the approximating problems by

$$\mathbf{H}_N \triangleq \mathbb{R}^n \times V_{\mathbf{A},N}^{-1}(\overline{U}_N) \subset H_N, \tag{4.15c}$$

and their reflections in coefficient space by

$$\overline{H}_N \triangleq \mathbb{R}^n \times \overline{\mathbf{U}}_N \subset \overline{H}_N, \tag{4.15d}$$

with $\overline{\mathbf{U}}_N = \overline{\mathbf{U}}_N^1$ and $V_{\mathbf{A},N} = V_{\mathbf{A},N}^1$ for representation R1 and $\overline{\mathbf{U}}_N = \overline{\mathbf{U}}_N^2$ and $V_{\mathbf{A},N} = V_{\mathbf{A},N}^2$ for representation R2. We assume that $\rho_{\max}$ was chosen large enough in (3.3c) to ensure that $\mathbf{H}_N \subset \mathbf{B}$.

Remark 4.9. The constraints on $\|\bar{u}_k T_j\|_\infty$ appearing in the definition of $\mathbf{U}_N^1$ were introduced to ensure that each polynomial piece, $\sum_{j=1}^r \bar{u}_k^j \Phi_{N,k,j}^1(\cdot)$, of $u = V_{\mathbf{A},N}^{-1}(\bar{u})$ is Lipschitz continuous on $[t_k, t_{k+1}]$ with Lipschitz constant κ_U, independent of N. This is needed to establish that the accuracy of the RK integration increases at least linearly with decreasing step-size (Lemmas A.1 and 4.10(i), but see Remark A.2) and for the proofs of Theorems 4.2 and 5.6. When the system dynamics are linear and time-invariant with respect to u and the RK method is of order r, Lemma 4.10(i) is valid without this Lipschitz constant, and hence, the constraints in the definition of $\mathbf{U}_N^1$ are not needed. It is not clear if the constraints on $\|\bar{u}_k T_j\|_\infty$ are needed in practice because if a sequence of controls converges to a piecewise Lipschitz continuous function, then the members of that sequence will all be piecewise Lipschitz continuous (see Remark 4.13 and Conjecture 5.11 added to the end of this paper).

Next, with $\eta = (\xi, u) \in H_N$ and $\bar{\eta} = (\xi, \bar{u}) = W_{\mathbf{A},N}(\eta)$, we will denote the solutions of (4.3a,b), with $\omega_k = \bar{u}_k G$, $k \in \mathcal{N}$, by $\{\bar{x}_k^\eta\}_{k=0}^N$ or, equivalently, $\{\bar{x}_k^{\bar{\eta}}\}_{k=0}^N$. The variable $\bar{x}_k^{\bar{\eta}}$ is thus the computed estimate of $x^\eta(t_k)$. Finally, for $\nu \in \mathbf{q}$, let $f_N^\nu : H_N \to \mathbb{R}$ and $\bar{f}_N^\nu : \overline{H}_N \to \mathbb{R}$ be defined by

$$f_N^\nu(\eta) \triangleq \zeta^\nu(\xi, \bar{x}_N^\eta) \equiv \bar{f}_N^\nu(\bar{\eta}) \triangleq \zeta^\nu(\xi, \bar{x}_N^{\bar{\eta}}), \qquad \nu \in \mathbf{q}, \tag{4.16}$$

where $\zeta^\nu(\cdot,\cdot)$ was used to define $f^\nu(\cdot)$ in (3.4c). We can now state the approximating problems as

$$\mathbf{CP}_N \qquad \min_{\eta \in \mathbf{H}_N} \{\psi_{o,N}(\eta) | \psi_{c,N}(\eta) \le 0\}, \tag{4.17a}$$

where $\psi_{o,N}(\eta) \triangleq \max_{\nu \in \mathbf{q}_o} f_N^\nu(\eta)$ and $\psi_{c,N}(\eta) \triangleq \max_{\nu \in \mathbf{q}_c + q_o} f_N^\nu(\eta)$, or equivalently, in the form in which they must be solved numerically as

$$\overline{\mathbf{CP}}_N \qquad \min_{\bar{\eta} \in \overline{\mathbf{H}}_N} \{\bar{\psi}_{o,N}(\bar{\eta}) | \bar{\psi}_{c,N}(\bar{\eta}) \le 0\}, \tag{4.17b}$$

where $\bar{\psi}_{o,N}(\bar{\eta}) \triangleq \max_{\nu \in \mathbf{q}_o} \bar{f}_N^\nu(\bar{\eta})$ and $\bar{\psi}_{c,N}(\bar{\eta}) \triangleq \max_{\nu \in \mathbf{q}_c + q_o} \bar{f}_N^\nu(\bar{\eta})$. By defining the feasible set as $\mathbf{F}_N \triangleq \{\eta \in \mathbf{H}_N | \psi_{c,N}(\eta) \le 0\}$, we can write $\mathbf{CP}_N$ in the equivalent form of problem $\mathbf{P}_N$ in (2.1b).

Note that for any $u \in \mathbf{U} \cap L_N^i$, $i = 1, 2$, where $\mathbf{U}$ was defined in (3.3a), $\bar{u} = V_{\mathbf{A},N}^1(u)$ satisfies $\bar{u}_k^j \in U$, for $k \in \mathcal{N}$, $j \in \mathbf{r}$, because $u(t) \in U$ for all $t \in [0, 1]$. Hence, for representation R2, (4.15b,c) imply that $\mathbf{H} \cap H_N \subset \mathbf{H}_N$. Conversely, $\bar{u} \in \overline{\mathbf{U}}_N^2 \Leftrightarrow (V_{\mathbf{A},N}^2)^{-1}(\bar{u}) \in \mathbf{U}$, and therefore $\mathbf{H}_N \subset \mathbf{H} \cap H_N$. Consequently, for representation R2, $\mathbf{H}_N = \mathbf{H} \cap H_N$. Unfortunately, for representation R1 $\mathbf{H}_N \ne \mathbf{H} \cap H_N$. First, $\mathbf{H} \cap H_N \not\subset \mathbf{H}_N$ because elements $u \in \mathbf{U} \cap L_N^1$ do not necessarily satisfy the Lipschitz continuity constraint imposed by (4.15a). Second, if $r \ge 2$ (except for the case $r = 2$ and the Butcher array elements $c = (0, 1)$), $\mathbf{H}_N \not\subset \mathbf{H} \cap H_N$ because, given $\bar{u} \in \bar{L}_N^1$, generally $\|V_{\mathbf{A},N}^1(\bar{u})\|_\infty > \|\bar{u}\|_\infty$, [5, p. 24]. Hence, if $\{\eta_N = (\xi, u_N)\}_{N \in \mathbf{N}}$, $\mathbf{N} \subset \mathbb{N}$, is a sequence of approximate solutions to the problems $\mathbf{CP}_N$ using representation R1,

it is possible for η_N to violate the control constraints in **CP**. However, as we will see, the limit points of such a sequence do satisfy the control constraints in **CP**. This problem of constraint violations for representation R1 could have been avoided by choosing $\mathbf{H}_N \triangleq \mathbf{H} \cap H_N$ (as in [24]) and letting $\overline{\mathbf{H}}_N \triangleq W_{A,N}(\mathbf{H}_N)$. But the set $\overline{\mathbf{H}}_N$ would then be difficult to characterize and we would have to impose a Lipschitz continuity constraint directly on the set **H** which would be unacceptable.

Nesting. The theory of consistent approximations is stated in terms of nested subspaces H_N. This allows the approximate solution of an approximating problem $\mathbf{CP}_{N_1}$ to be used as a "warm-start" for an algorithm solving an approximating problem $\mathbf{CP}_{N_2}$ with a higher discretization level ($N_2 > N_1$) (see [16, 25]).

For representation R1, for any $N \in \mathbb{N}$, $N \geq 1$, $L^1_N \subset L^1_{2N}$, and therefore doubling the discretization level nests the subspaces. If $u \in L^1_N$, then $\bar{v} = V^1_{\mathbf{A},2N}(u)$ can be determined from $\bar{u} = V^1_{\mathbf{A},N}(u)$ using (4.7) and (4.8), as follows. For $k \in \mathcal{N}$ and $j \in \mathbf{r}$, $\bar{v}^j_{2k} = \bar{u}_k T P(c_j/2N)$ and $\bar{v}^j_{2k+1} = \bar{u}_k T P((c_j + 1)/2N)$. For representation R2, $L^2_N \subset L^2_{dN}$, where d is the smallest common denominator of the parameters b_j, $j \in \mathbf{s}$, in the Butcher array, which is finite assuming, as is typically the case, that the b_j are rational. Thus, the discretization level must be increased by factors of d to achieve nesting. If $u \in L^2_N$ and $\bar{u} = V^2_{\mathbf{A},N}(u)$, then $\bar{v} = V^2_{\mathbf{A},dN}(u)$ is given, for $k \in \mathcal{N}$, $i, j \in \mathbf{r}$, and $l = 1, \ldots, d$, by $\bar{v}^i_{dk+l} = \bar{u}^j_k$ for $d_{j-1} \leq l/d < d_j$, where d_j is defined in (4.10b).

Epiconvergence. We are now ready to establish the epiconvergence of the approximating problems. First we present convergence properties for the solutions computed by Runge–Kutta integration on H_N. The proof of the following lemma, given in the appendix, differs from standard Runge–Kutta results because of the presence of (possibly discontinuous) controls in the differential equations.

LEMMA 4.10. *For representation* R1, *suppose that Assumptions* 3.1(a), 4.1′, *and* 4.3 *hold. For representation* R2, *suppose that Assumptions* 3.1(a), 4.1′, *and* 4.6 *hold.*

(i) *Convergence. For any bounded subset $S \subset \mathbf{B}$, there exist $\kappa < \infty$ and $N^* < \infty$, such that for any $\eta \in S \cap \mathbf{H}_N$ and $N \geq N^*$,*

$$\|x^\eta(t_k) - \bar{x}^\eta_k\| \leq \frac{\kappa}{N}, \qquad k \in \{0, 1, \ldots, N\}. \tag{4.18a}$$

(ii) *Order of Convergence. Additionally, suppose the Runge–Kutta method is order ρ (see* [8, 17]*), and $h(\cdot, \cdot)$ is $\rho - 1$ times Lipschitz continuously differentiable. Let*

$$\mathbf{H}^{(\rho)}_N \triangleq \left\{ \eta = (\xi, u) \in \mathbf{H}_N \,\middle|\, \left\| \frac{d^{\rho-1}}{dt^{\rho-1}}(u(t_1) - u(t_2)) \right\| \leq \kappa', \ \forall t_1, t_2 \in [t_k, t_{k+1}), \ k \in \mathcal{N} \right\}, \tag{4.18b}$$

where $\kappa' < \infty$ is independent of N. Then for representation R1, *there exist $\kappa < \infty$ and $N^* < \infty$ such that, if $\eta \in S \cap \mathbf{H}^{(\rho)}_N$, or if $\eta \in S \cap \mathbf{H}_N$ and $h(x, u) = \tilde{h}(x) + Bu$, where B is an $n \times m$ constant matrix, then for any $N \geq N^*$*

$$\|x^\eta(t_k) - \bar{x}^\eta_k\| \leq \frac{\kappa}{N^\rho}, \qquad k \in \{0, 1, \ldots, N\}. \tag{4.18c}$$

The same result holds for representation R2 *for any $\eta \in S \cap \mathbf{H}_N$ if $h(x, u) = \tilde{h}(x) + Bu$.*

In proving consistency, we will need to add a version of Slater's constraint qualification on the problem **CP**.

Assumption 4.11. For every $\eta \in \mathbf{H}$ such that $\Psi_c(\eta) \leq 0$, there exists a sequence $\{\eta_i\}_{i=1}^\infty$ such that $\eta_i \in \mathbf{H}$, $\Psi_c(\eta_i) < 0$, and $\eta_i \to \eta$ as $i \to \infty$.

THEOREM 4.12 (Epiconvergence). *For representation* R1, *suppose that Assumptions* 3.1, 4.1′, 4.3, *and* 4.11 *hold and let $d = 2$. For representation* R2, *suppose that Assumptions*

3.1, 4.1′, 4.6, *and* 4.11 *hold and let* d *be the least common denominator for the elements* b_j, $j \in \mathbf{s}$, *of the Butcher array. Let* $\mathbf{N} = \{d^l\}_{l=1}^{\infty}$. *Then the problems* $\{\mathbf{CP}_N\}_{N\in\mathbf{N}}$ *converge epigraphically to the problem* $\mathbf{CP}$ *as* $N \to \infty$.

Proof. Let $S \subset \mathbf{B}$ be bounded. Then, by Assumption 3.1(b) and Lemma 4.10(i), there exist $\kappa', \kappa < \infty$ such that for any $v \in \mathbf{q}$ and for any $\eta_N \in S \cap \mathbf{H}_N$

$$|f^v(\eta_N) - f_N^v(\eta_N)| = |\zeta^v(\xi_N, x^{\eta_N}(1)) - \zeta^v(\xi_N, \bar{x}_N^{\eta_N})| \le \kappa' \|x^{\eta_N}(1) - \bar{x}_N^{\eta_N}\| \le \frac{\kappa}{N}. \tag{4.19a}$$

Now, let $v' \in \mathbf{q}_o$ be such that $\Psi_o(\eta_N) = f^{v'}(\eta_N)$. Then,

$$\Psi_o(\eta_N) - \Psi_{o,N}(\eta_N) = f^{v'}(\eta_N) - \Psi_{o,N}(\eta_N) \le f^{v'}(\eta_N) - f_N^{v'}(\eta_N) \le \frac{\kappa}{N}. \tag{4.19b}$$

By reversing the roles of $\Psi_o(\eta_N)$ and $\Psi_{o,N}(\eta_N)$ we can conclude that

$$|\Psi_o(\eta_N) - \Psi_{o,N}(\eta_N)| \le \frac{\kappa}{N}. \tag{4.20a}$$

Similarly,

$$|\Psi_c(\eta_N) - \Psi_{c,N}(\eta_N)| \le \frac{\kappa}{N}. \tag{4.20b}$$

Now, given $\eta \in \mathbf{H}$ such that $\Psi_c(\eta) \le 0$, there exists, by Assumption 4.11, a sequence $S = \{\eta_i\}_{i\in\mathbf{N}}$, with $\eta_i \in \mathbf{H}$, such that $\eta_i \to \eta$ as $i \to \infty$ (hence S is a bounded set) and $\Psi_c(\eta_i) < 0$ for all i. Now, clearly for each i, there exist $N_i \in \mathbf{N}$ and $\eta'_{N_i} \in \mathbf{H}_{N_i}$ such that (a) $\kappa/N_i \le -1/2\Psi_c(\eta_i)$; (b) $\|\eta'_{N_i} - \eta_i\| \le 1/N_i$, since, for both control representations the union of the subspaces H_N is dense in H_2, which contains $H_{\infty,2}$ and $\mathbf{H} \cap H_N \subset \mathbf{H}_N$; (c) $\Psi_c(\eta'_{N_i}) \le 1/2\Psi_c(\eta_i)$ due to Theorem 3.2(iii); and (d) $N_i < N_{i+1}$. It follows from (4.20b) that $\Psi_{c,N_{i+k}}(\eta'_{N_i}) \le \Psi_c(\eta'_{N_i}) + \kappa/N_i \le 1/2\Psi_c(\eta_i) + \kappa/N_i \le 0$ for any $i, k \in \mathbb{N}$. Now consider the sequence $S'' = \{\eta''_M\}_{M\in\mathbf{N}}$ defined as follows. If $M = N_i$ for some $i \in \mathbf{N}$, then $\eta''_M = \eta'_{N_i}$ for $M = N_i, N_i + d, N_i + 2d, \ldots, N_{i+1} - d$. Then we see that $\Psi_{c,M}(\eta''_M) \le 0$ for all $M \in \mathbf{N}$, $\eta''_M \to \eta$ as $M \to \infty$ (hence S'' is bounded), and by (4.20a) and Theorem 3.2(iii) that $\lim_{M\in\mathbf{N}} \Psi_{o,M}(\eta''_M) = \Psi_o(\eta)$. Thus, part (a) of Definition 2.1 is satisfied.

Now let $S = \{\eta_N\}_{N\in K}$, $K \subset \mathbf{N}$, be a sequence with $\eta_N = (\xi_N, u_N) \in \mathbf{H}_N$ and $\Psi_{c,N}(\eta_N) \le 0$ for all $N \in K$, and suppose that $\eta_N \to^K \eta = (\xi, u)$. First, we want to show that $\eta \in \mathbf{H}$. For any $v \in \mathbb{R}^m$, let $d(v, U) \triangleq \min_{v'\in U} \|v - v'\|$. Since $\bar{u} = V^i_{\mathbf{A},N}(u_N) \in \overline{\mathbf{U}}^i_N$, $i = 1, 2$, for each N, $\bar{u}^j_k \in U$ for all $k \in \mathcal{N}$, $j \in \mathbf{r}$. For representation R1, $\overline{\lim}_{t\in[0,1],N\in K} d(u_N(t), U) = 0$ since elements $u_N \in \mathbf{U}^1_N$ are piecewise Lipschitz continuous polynomials, with Lipschitz constant independent of N, defined over progressively smaller intervals.[3] For representation R2, $d(u_N(t), U) = 0$ for all $N \in \mathbf{N}$ and $t \in [0, 1]$ since $u_N \in \mathbf{U}^2_N$ is piecewise constant. This implies that $u \in \mathbf{U}$; hence $\eta \in \mathbf{H}$. Furthermore, $\Psi_c(\eta) \le 0$ by (4.20b) and the continuity of $\Psi_c(\cdot)$. Finally, by (4.20a) and Theorem 3.2(iii), $\lim_{N\in K} \Psi_{o,N}(\eta_N) = \Psi_o(\eta)$. Thus, part (b) of Definition 2.1 holds. □

Remark 4.13. In [15], Hager empirically observes that methods with $b_j = 0$ for some j, such as the modified Euler method, cannot be used to discretize optimal control problems. This requirement, formalized in Assumption 4.1, is used in our proof of epiconvergence. However, for representation R1, epiconvergence of $\mathbf{P}_N$ to $\mathbf{P}$ can be established even if, for some j, $\tilde{b}_j \le 0$. This is because of the Lipschitz continuity constraint imposed on the set $\mathbf{U}^1_N$ in (4.15a); see Conjecture 5.11 added to the end of this paper.

[3] It can also be shown by contradiction that $d(u_N(\cdot), U) \to 0$ a.e. on [0,1] without requiring, in (4.15a), elements of $\mathbf{U}^1_N$ to have a uniform piecewise Lipschitz constant.

Nonetheless, our experimental evidence suggests that using an RK method with $\tilde{b}_j \leq 0$ is unwise. For example, the third-order method with Butcher array $b = (-1/6, 8/9, 5/18)$, $c = (0, 1/4, 1)$ and nonzero entries of A given by $a_{2,1} = 1/4, a_{3,1} = -7/5, a_{3,2} = 12/5$ was used to discretize the problem (6.3) with discretization level $N = 10$. The solutions u_N^* for different values of Lipschitz constant κ_U are plotted in Figure 1(a). For comparison, the solutions of the approximating problems produced with the third-order RK method with Butcher array $b = (1/6, 2/3, 1/6)$, $c = (0, 1/2, 1)$ and nonzero entries of A given by $a_{2,1} = 1/2, a_{3,1} = -1$, $a_{3,2} = 2$ are presented in Figure 1(b). For both, with κ_U small, the quadratic polynomial pieces in each time interval are forced to be fairly flat. But, as κ_U is increased, the solutions for the "bad" method become increasingly worse and the control solutions remain pushed against the Lipschitz continuity constraints. On the other hand, the solutions for the "good" method become better as κ_U is increased. In fact, when κ_U is bigger than the Lipschitz constant of the true solution u^*, the Lipschitz continuity constraints are inactive for the "good" method (see Remark 4.9). This is seen in Figure 1(b) since the solutions for $\kappa_U = 1$ and $\kappa_U = 10$ are identical. As κ_U is increased from 0.1 to 10, the error $\max_{k,j} |u_N^*[\tau_{k,j}] - u^*(\tau_{k,j})|$ goes from 0.0332 to 7.9992e-4 for the "good" method and goes from 0.0332 to 1.9119 for the "bad" method.

The conditions imposed by Assumptions 4.3 and 4.6 on the c parameters of the Butcher array are needed because of the discontinuites in the controls $u \in L_N^i$, $i = 1, 2$.

Factors in selecting the control representation. The choice of selecting $L_N = L_N^1$ versus $L_N = L_N^2$ depends on the relative importance of approximation error versus constraint satisfaction. It follows from the proof of epiconvergence that irrespective of which representation is used, if $\{\eta_N\}_{N\in\mathbf{N}}$ is a sequence such that $\eta_N \in \mathbf{H}_N$ and $\eta_N \to \eta$, then $\eta \in \mathbf{H}$. Thus η satisfies the control constraints. However, as mentioned earlier, if representation R1 is used, then η_N may not satisfy the control constraints for any finite N (except for the case $r = 2$ and $c = (0, 1)$). Since a numerical solution must be obtained after a finite number of iterations, representation R2 should be used if absolute satisfaction of control constraints is required.

If some violation of control constraints is permissible, then representation R1 may be preferable to representation R2 (although, see comment about transformation of simple control bounds in §6) because a tighter bound for the error of the approximate solution can be established for R1 than for R2. To see this, let $\eta_N^* = (\xi_N^*, u_N^*)$, $N \in \mathbb{N}$, be a local minimizer of the finite-dimensional problem $\mathbf{CP}_N$. This solution is computed by setting $\eta_N^* = W_{\mathbf{A},N}^{-1}(\bar{\eta}_N^*)$, where $\bar{\eta}_N^*$ is the result of a numerical algorithm implemented on a computer using the formulae to be presented in the following sections. The error, $\|u^* - u_N^*\|_2$, of the approximate control solutions u_N^* can be determined as follows. Assume that $u_N^* \to u^*$ as $N \to \infty$ and that u^* is a local minimizer of **CP** (if the u_N^* solutions are uniformly strict minimizers, then u^* must be a local minimizer by Theorem 2.2). Let $\bar{u}^* \in \times_N \times_r \mathbb{R}^m$ be such that $\bar{u}_k^{*j} = u^*(\tau_{k,j})$ for $k \in \mathcal{N}$, $j \in \mathbf{r}$ (assuming $u^*(\tau_{k,j})$ exists). Then, with $\bar{u}_N^* = V_{\mathbf{A},N}(u_N^*)$,

$$\|u^* - u_N^*\|_2 \leq \|u^* - V_{\mathbf{A},N}^{-1}(\bar{u}^*)\|_2 + \|V_{\mathbf{A},N}^{-1}(\bar{u}^*) - u_N^*\|_2 = \|u^* - V_{\mathbf{A},N}^{-1}(\bar{u}^*)\|_2 + \|\bar{u}^* - \bar{u}_N^*\|_{\bar{L}_N}. \tag{4.21}$$

By Proposition 5.5, the quantity $\|\bar{u}^* - \bar{u}_N^*\|_{\bar{L}_N}$ is not affected by the choice of control representations. For smooth, unconstrained problems discretized by symmetric RK methods, a bound for $\|\bar{u}^* - \bar{u}_N^*\|_{l_\infty}$ can be found in [16, Thm. 3.1] (see Proposition 6.1 in this paper for an improved bound for RK4). The quantity $\|u^* - V_{\mathbf{A},N}^{-1}(\bar{u}^*)\|_2$ is the error between u^* and the element of L_N^1 or L_N^2 that interpolates $u^*(t)$ at $t = \tau_{k,j}$, $k \in \mathcal{N}$, and $j \in \mathbf{r}$. The piecewise polynomials of representation R1 are generally better interpolators for $u^*(\cdot)$, except for non-

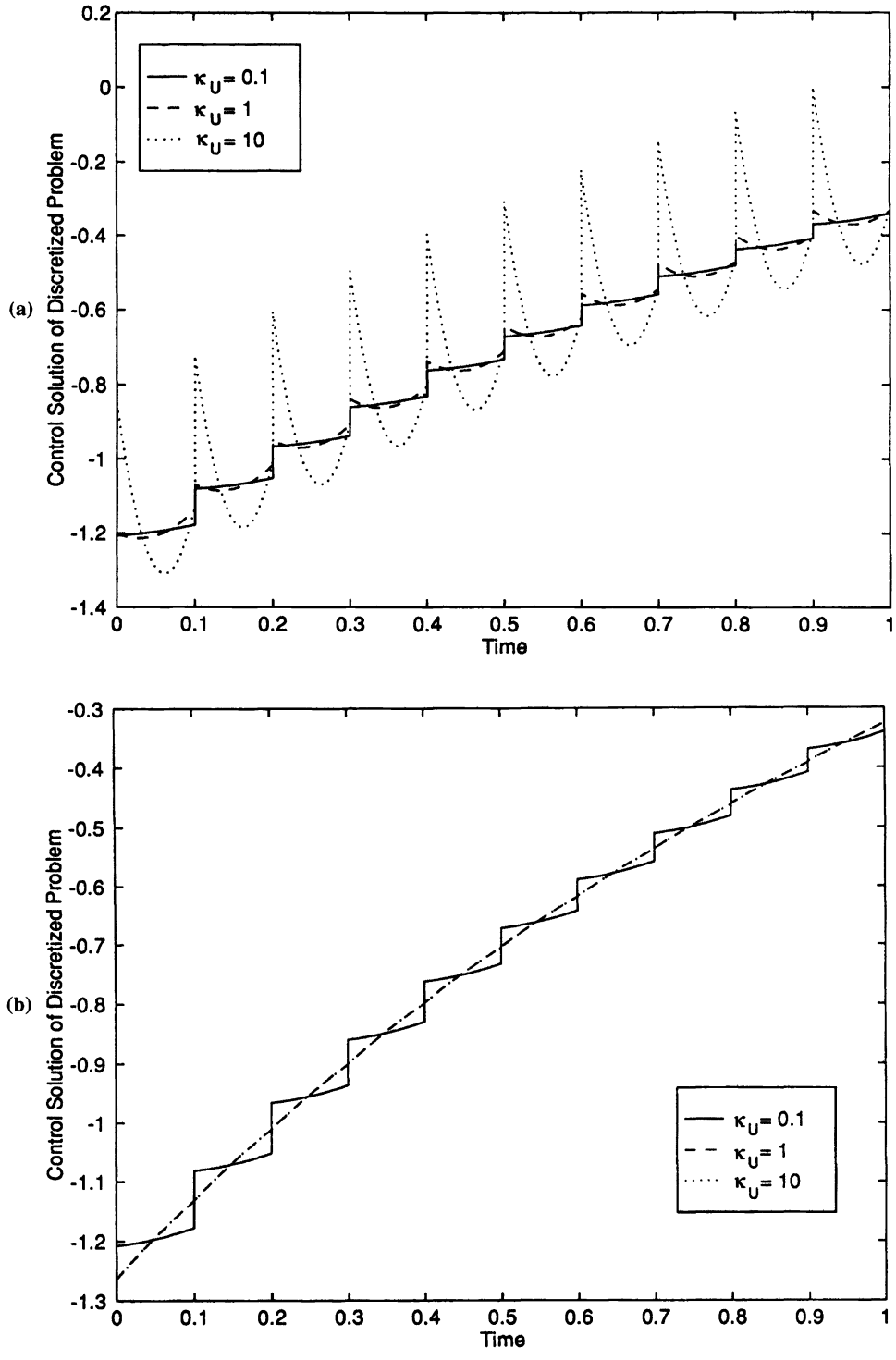


FIG. 1. *Effect of the Lipschitz constant* κ_U *on the solution of problem* (6.3) *discretized with an RK method that has* (a) $b_1 < 0$ *and gets worse as* κ_U *is increased and* (b) *all* $b_j > 0$ *and gets better as* κ_U *is increased until the point where the Lipschitz continuity constraints on* u_N^* *become inactive, as is the case for* $\kappa_U = 1$ *and* $\kappa_U = 10$.

smooth points, than the functions of R2. For $u^*(\cdot)$ sufficiently smooth, $\|u^* - V_{\mathbf{A},N}^{-1}(\bar{u}^*)\|_\infty$ is of order r for representation R1 (see [5]) but only first order of representation R2.

5. Optimality functions for the approximating problems. To develop optimality functions for our approximating problems, we must determine the gradients of the cost and constraint functions for the approximating problems.

At each time step, the RK integration formula is a function of the current state estimate $\bar{x}_k$ and the r control samples $\bar{u}_k = (\bar{u}_k^1, \ldots, \bar{u}_k^r)$. So, let $F : \mathbb{R}^n \times (\times_s \mathbb{R}^m) \to \mathbb{R}^n$ be defined by

$$F(x, w) = x + \Delta \sum_{i=1}^{s} b_i K_i(x, wG), \tag{5.1}$$

where $w = (w^1, \ldots, w^r) \in \times_r \mathbb{R}^m$ is being treated as the $m \times r$ matrix $[w^1 \cdots w^r]$, G was defined in (4.5d), and $K_i(x, \omega)$ was defined in (4.3b) ($\omega = wG \in \mathbb{R}^{m \times s}$). Then, referring to (4.3a,b), we see that for any $\bar{\eta} = (\xi, \bar{u}) \in \bar{H}_N$, with $\bar{H}_N$ defined in (4.14a),

$$\bar{x}_{k+1}^{\bar{\eta}} = F(\bar{x}_k, \bar{u}_k), \quad \bar{x}_0 = \xi, \quad k \in \mathcal{N}. \tag{5.2}$$

The derivative of $F(\cdot, \cdot)$ with respect to the jth component of w is, with I_j defined in (4.4b),

$$\begin{aligned} F_{w^j}(x, w) &= \Delta \frac{\partial}{\partial w^j} \sum_{i=1}^{s} b_i K_i(x, wG) \\ &= \Delta \sum_{l \in I_j} \frac{\partial}{\partial \omega^l} \sum_{i=1}^{s} b_i K_i(x, \omega) \\ &= \Delta \sum_{l \in I_j} \left[b_l h_u(Y_l(x, \omega), w^j) + \Delta \sum_{i=1}^{s} b_i h_x(Y_i(x, \omega), \omega^j) \sum_{p=1}^{i-1} \frac{\partial}{\partial \omega^l} K_p(x, \omega) \right], \end{aligned} \tag{5.3}$$

where $\omega = wG$ and $Y_i(x, \omega) \triangleq x + \Delta \sum_{j=1}^{i-1} a_{i,j} K_j(x, \omega)$.

The next theorem provides an expression for the gradients of the functions $f_N^\nu(\cdot)$, $\nu \in \mathbf{q}$, given by (4.16).

THEOREM 5.1. *Let $N \in \mathbb{N}$, $\eta \in H_N$, and $\bar{\eta} = W_{\mathbf{A},N}(\eta)$. Also, let $\mathbf{M}_N \in \mathbb{R}^{Nr \times Nr}$ be the N-block diagonal matrix defined by*

$$\mathbf{M}_N \triangleq \operatorname{diag}[\Delta M, \Delta M, \ldots, \Delta M], \tag{5.4}$$

where $M = M_1$ for representation R1 *and $M = M_2$ for representation* R2. *Then, for each $\nu \in \mathbf{q}$, the gradient of $f_N^\nu(\cdot)$, $\nabla f_N^\nu : H_N \to H_N$, is given by*

$$\nabla f_N^\nu(\eta) = \left(\nabla_\xi f_N^\nu(\eta), \nabla_u f_N^\nu(\eta)\right) = \left(\bar{\gamma}_\xi^\nu(\bar{\eta}), V_{\mathbf{A},N}^{-1}\left(\bar{\gamma}_u^\nu(\bar{\eta})\mathbf{M}_N^{-1}\right)\right), \tag{5.5a}$$

where $V_{\mathbf{A},N} = V_{\mathbf{A},N}^1$ for representation R1, *$V_{\mathbf{A},N} = V_{\mathbf{A},N}^2$ for representation* R2, *and $\bar{\gamma}^\nu(\bar{\eta}) = (\bar{\gamma}_\xi^\nu(\bar{\eta}), \bar{\gamma}_u^\nu(\bar{\eta})) \in \overline{H}_N$ is defined by*

$$\bar{\gamma}_\xi^\nu(\bar{\eta}) = \nabla_\xi \zeta^\nu(\xi, \bar{x}_N^{\bar{\eta}}) + \bar{p}_0^{\nu,\bar{\eta}}, \tag{5.5b}$$

$$\bar{\gamma}_u^\nu(\bar{\eta})_k^j = F_{w_j}(\bar{x}_k^{\bar{\eta}}, \bar{u}_k)^T \bar{p}_{k+1}^{\nu,\bar{\eta}}, \qquad k \in \mathcal{N}, \ j \in \mathbf{r}, \tag{5.5c}$$

with $\bar{p}_k^{\nu,\bar{\eta}}$ determined by the adjoint equation

$$\bar{p}_k^\nu = F_x(\bar{x}_k^{\bar{\eta}}, \bar{u}_k)^T \bar{p}_{k+1}^\nu, \quad \bar{p}_N^\nu = \zeta_x^\nu(\xi, \bar{x}_N^{\bar{\eta}})^T, \quad k \in \mathcal{N}, \tag{5.5d}$$

and where $F_x(\cdot, \cdot)$ and $F_{w^j}(\cdot, \cdot)$ denote the partial derivatives of $F(x, w)$ with respect to x and the jth component of w.

Proof. First, we note that $V_{\mathbf{A},N}^1$ is invertible by Proposition 4.4 and $V_{\mathbf{A},N}^2$ is invertible by Proposition 4.7. Next, referring to [23, p. 68], we see that $\bar{\gamma}_\xi^\nu(\bar{\eta})$ is the gradient of $\bar{f}_N^\nu(\bar{\eta})$ with

respect to ξ. Similarly, $\bar{\gamma}_u^\nu(\bar{\eta})$ is the gradient of $\bar{f}_N^\nu(\bar{\eta})$ with respect to $\bar{u} \in \times_N \times_r \mathbb{R}^m$ endowed with the standard l_2 inner product. Hence, the Gateaux differential of f_N^ν is given by

$$\begin{aligned} Df_N^\nu(\eta_N; \delta\eta_N) = D\bar{f}_N^\nu(\bar{\eta}; \delta\bar{\eta}) &= \langle \bar{\gamma}_\xi^\nu(\bar{\eta}), \delta\xi\rangle + \langle \bar{\gamma}_u^\nu(\bar{\eta}), \delta\bar{u}\rangle_{l_2} \\ &= \langle \bar{\gamma}_\xi^\nu(\bar{\eta}), \delta\xi\rangle + \langle \bar{\gamma}_u^\nu(\bar{\eta})\mathbf{M}_N^{-1}, \delta\bar{u}\rangle_{\bar{L}_N} \\ &= \langle \gamma_\xi^\nu(\bar{\eta}), \delta\xi\rangle + \langle V_{\mathbf{A},N}^{-1}(\gamma_u^\nu(\bar{\eta})\mathbf{M}_N^{-1}), \delta u\rangle_2, \end{aligned} \tag{5.6}$$

where $\delta\eta_N = (\delta\xi, \delta u_N) \in H_N$ and $\delta\bar{\eta} = (\delta\xi, \delta\bar{u}) = W_{\mathbf{A},N}(\delta\eta_N)$. Since by the definition of $\nabla f_N^\nu(\eta_N)$, $Df_N^\nu(\eta_N; \delta\eta_N) = \langle \nabla f_N^\nu(\eta_N), \delta\eta_N\rangle_H$ for all $\delta\eta_N \in H_N$, the desired result follows from (5.6). □

A simpler expression for $\bar{\gamma}_u(\bar{\eta})$ for a certain class of RK methods can be found in [15].

Note that for $\eta \in H_N$, $\bar{\eta} = W_{\mathbf{A},N}(\eta)$, we have $\bar{\gamma}_u^\nu(\bar{\eta})_k \in \underset{r}{\times} \mathbb{R}^m$ and

$$\left(\nabla_u f_N^\nu(\eta)[\tau_{k,i_1}] \cdots \nabla_u f_N^\nu(\eta)[\tau_{k,i_r}]\right) = \frac{1}{\Delta}\left(\bar{\gamma}_u^\nu(\bar{\eta})_k^1 \cdots \bar{\gamma}_u^\nu(\bar{\eta})_k^r\right) M^{-1}, \tag{5.7}$$

where $i_j \in I$, $j \in \mathbf{r}$, and $\nabla_u f_N^\nu(\eta_N)[\tau_{k,i_j}]$ is computed according to (4.6e) or (4.11c).

Remark 5.2. At this point, we can draw one very important conclusion. For every $\nu \in \mathbf{q}$, the steepest descent direction, in $\overline{H}_N$, for the function $\bar{f}_N^\nu(\cdot)$, at $\bar{\eta}$, is given by $-(\bar{\gamma}_\xi^\nu(\bar{\eta}), \bar{\gamma}_u^\nu(\bar{\eta})\mathbf{M}_N^{-1})$ and not by $-(\bar{\gamma}_\xi^\nu(\bar{\eta}), \bar{\gamma}_u^\nu(\bar{\eta}))$, which is the steepest descent direction that one would obtain using the standard l_2 inner product on $\times_N \times_r \mathbb{R}^m$. The naive approach of solving the discrete-time optimal control problem $\mathbf{CP}_N$ using the latter steepest descent directions amounts to a change of metric that can result in severe ill-conditioning, as we will illustrate in §6.

We can now define optimality functions for the approximating problems using the form of the optimality function presented in (3.9b) for the original problem. For $\mathbf{CP}_N$, we define $\theta_N : H_N \to \mathbb{R}$, with $\sigma > 0$ and the set $\mathbf{H}_N$ defined in (4.15c) by

$$\theta_N(\eta) \triangleq \min_{\eta' \in \mathbf{H}_N} \max\left\{ \max_{\nu \in \mathbf{q}_o} \tilde{f}_N^\nu(\eta, \eta') - \psi_{o,N}(\eta) - \sigma\psi_{c,N}(\eta)_+, \max_{\nu \in \mathbf{q}_c + q_o} \tilde{f}_N^\nu(\eta, \eta') - \psi_{c,N}(\eta)_+ \right\}, \tag{5.8a}$$

where $\psi_{c,N}(\eta)_+ \triangleq \max\{0, \psi_{c,N}(\eta)\}$, and for $\nu \in \mathbf{q}$,

$$\tilde{f}_N^\nu(\eta, \eta') \triangleq f_N^\nu(\eta) + \langle \nabla f_N^\nu(\eta), \eta' - \eta\rangle_H + \frac{1}{2}\|\eta' - \eta\|_H^2. \tag{5.8b}$$

If needed for a particular numerical algorithm (e.g., [22]), $\theta_N(\eta) = \bar{\theta}_N(\bar{\eta})$, where $\bar{\eta} = W_{\mathbf{A},N}(\eta)$ and

$$\bar{\theta}_N(\bar{\eta}) \triangleq \min_{\bar{\eta}' \in \overline{\mathbf{H}}_N} \frac{1}{2}\|\bar{\eta}' - \bar{\eta}\|_{\bar{H}_N}^2 + \Theta_N(\bar{\eta}, \bar{\eta}'), \tag{5.9a}$$

with

$$\begin{aligned} \Theta_N(\bar{\eta}, \bar{\eta}') = \max\{ &\max_{\nu \in \mathbf{q}_o} \bar{f}_N^\nu(\bar{\eta}) + \langle (\bar{\gamma}_\xi^\nu(\bar{\eta}), \bar{\gamma}_u^\nu(\bar{\eta})\mathbf{M}_N^{-1}), \bar{\eta}' - \bar{\eta}\rangle_{\bar{H}_N} - \bar{\psi}_{o,N}(\bar{\eta}) - \sigma\bar{\psi}_{c,N}(\bar{\eta})_+, \\ &\max_{\nu \in \mathbf{q}_c + q_o} \bar{f}_N^\nu(\bar{\eta}) + \langle (\bar{\gamma}_\xi^\nu(\bar{\eta}), \bar{\gamma}_u^\nu(\bar{\eta})\mathbf{M}_N^{-1}), \bar{\eta}' - \bar{\eta}\rangle_{\bar{H}_N} - \bar{\psi}_{c,N}(\bar{\eta})_+\}, \end{aligned} \tag{5.9b}$$

and the set $\overline{\mathbf{H}}_N$ is defined in (4.15d).

It should be obvious that these optimality functions are well defined because of the form of the quadratic term and the fact that the minimum is taken over a set of finite dimension.

The following theorem confirms that (5.8a) satisfies the definition for an optimality function. The proof is essentially the same as the proof in [4, Thms. 3.6 and 3.7].

THEOREM 5.3. (i) $\theta_N(\cdot)$ *is continuous.*

(ii) *For every* $\eta \in H_N$, $\theta_N(\eta) \leq 0$.

(iii) *If* $\hat{\eta} \in \mathbf{H}_N$ *is a local minimizer for* $\mathbf{CP}_N$, *then* $\theta_N(\hat{\eta}) = 0$.

Remark 5.4. It can also be shown that $\theta_N(\hat{\eta}) = 0$ for $\hat{\eta} \in \mathbf{H}_N$ if and only if $d_2\Psi_N(\hat{\eta}, \hat{\eta}; \eta - \hat{\eta}) \geq 0$ for all $\eta \in \mathbf{H}_N$ where $\Psi_N(\eta, \eta') \triangleq \max\{\psi_{o,N}(\eta) - \psi_{o,N}(\eta') - \sigma\psi_{c,N}(\eta')_+, \psi_{c,N}(\eta) - \psi_{c,N}(\eta')_+\}$.

PROPOSITION 5.5. *The stationary points for problem* $\overline{\mathbf{CP}}_N$, *that is, the points* $\bar{\eta} \in \overline{\mathbf{H}}_N$ *such that* $\bar{\theta}_N(\bar{\eta}) = 0$, *do not depend on the control representation.*

Proof. First, $\bar{\eta} \in \overline{\mathbf{H}}_N$ is such that $\bar{\theta}_N(\bar{\eta}) = 0$ if and only if $\Theta_N(\bar{\eta}, \bar{\eta}') = 0$ for all $\bar{\eta}' \in \overline{\mathbf{H}}_N$. The "if" direction is obvious. For the "only if" direction, $\bar{\theta}_N(\bar{\eta}) = \min_{\bar{\eta}' \in \overline{\mathbf{H}}_N}\{1/2\|\bar{\eta}' - \bar{\eta}\|^2_{\bar{H}_N} + \Theta_N(\bar{\eta}, \bar{\eta}')\} = 0$. This implies that $\Theta_N(\bar{\eta}, \bar{\eta}') = 0$ because $\Theta_N(\bar{\eta}, \bar{\eta}')$ is linear in $\bar{\eta}'$ whereas $1/2\|\bar{\eta}' - \bar{\eta}\|^2_{\bar{H}_N}$ is quadratic in $\bar{\eta}'$. Second, let $\delta\bar{\eta} = (\delta\xi, \delta\bar{u}) = \bar{\eta}' - \bar{\eta}$. Then, for each $\nu \in \mathbf{q}$,

$$\langle(\bar{\gamma}^\nu_\xi(\bar{\eta}), \bar{\gamma}^\nu_u(\bar{\eta})\mathbf{M}_N^{-1}), \bar{\eta}' - \bar{\eta}\rangle_{\bar{H}_N} = \langle\bar{\gamma}^\nu_\xi(\bar{\eta}), \delta\bar{\xi}\rangle + \langle\bar{\gamma}^\nu_u(\bar{\eta}), \delta\bar{u}\rangle_{l_2}, \tag{5.9c}$$

since $\mathbf{M}_N$ is nonsingular. Hence, $\Theta_N(\bar{\eta}, \bar{\eta}')$ does not depend on the control representation. Thus, the points $\bar{\eta}$ such that $\bar{\theta}_N(\bar{\eta}) = 0$ do not depend on the control representations. □

Consistency of the approximations. To complete our demonstration of consistency of approximations we will show that the optimality functions of the approximating problems hypoconverge to the optimality function of the original problem. First we will present a simple algebraic condition that implies convergence of the gradients. We will use the column vector $\tilde{b} = (\tilde{b}_1 \ldots \tilde{b}_r)^T \in \mathbb{R}^r$, with components $\tilde{b}_j$ defined in (4.10a), and the values d_j defined in (4.10b).

THEOREM 5.6. *For representation* R1, *suppose that Assumptions* 3.1, 4.1′, *and* 4.3 *hold. For representation* R2, *suppose that Assumptions* 3.1, 4.1′, *and* 4.6 *hold. For* $N \in \mathbb{N}$, *let* H_N *be defined as in* (4.13a), *with* $L_N = L^1_N$ *or* $L_N = L^2_N$, *and let* $f^\nu_N : H_N \to \mathbb{R}$, $\nu \in \mathbf{q}$, *be defined by* (4.16). *Let* $M = M_1$ *if* $L_N = L^1_N$, *and let* $M = M_2$ *if* $L_N = L^2_N$. *Let* S *be a bounded subset of* $\mathbf{B}$. *If*

$$M^{-1}\tilde{b} = \mathbf{1}, \tag{5.10a}$$

where $\mathbf{1}$ *is a column vector of* r *ones, then there exist a* $\kappa < \infty$ *and an* $N^* < \infty$ *such that for all* $\eta = (\xi, u) \in S \cap \mathbf{H}_N$ *and* $N \geq N^*$,

$$\|\nabla f^\nu(\eta) - \nabla f^\nu_N(\eta)\|_H \leq \frac{\kappa}{N}. \tag{5.10b}$$

Proof. To simplify notation, we replace $\bar{x}^{\bar{\eta}}_k$ by $\bar{x}_k$ and $\bar{p}^{\nu,\bar{\eta}}_k$ by $\bar{p}^\nu_k$. Let $S \subset \mathbf{B}$ be bounded, and let $\eta = (\xi, u) \in S \cap \mathbf{H}_N$. Let $\bar{u} = V_{\mathbf{A},N}(u)$ and $\bar{\eta} = (\xi, \bar{u})$ where $V_{\mathbf{A},N} = V^1_{\mathbf{A},N}$ for representation R1 and $V_{\mathbf{A},N} = V^2_{\mathbf{A},N}$ for representation R2. For each $j \in \mathbf{r}$ and $k \in \mathcal{N}$, $F_{w^j}(\bar{x}_k, \bar{u}_k)$ is given by (5.3). So, with $Y_{k,i} \triangleq \bar{x}_k + \Delta\sum_{j=1}^{i-1} a_{i,j}K_j(\bar{x}_k, \omega_k)$ and $\omega_k = \bar{u}_k G$, there exists $\kappa_1 < \infty$ such that

$$\begin{aligned}
&\|F_{w^j}(\bar{x}_k, \bar{u}_k) - \Delta\tilde{b}_j h_u(\bar{x}_k, \bar{u}^j_k)\| \\
&\leq \left\|F_{w^j}(\bar{x}_k, \bar{u}_k) - \Delta\sum_{l\in I_j} b_l h_u(Y_{k,l}, \bar{u}^j_k)\right\| + \left\|\Delta\sum_{l\in I_j} b_l h_u(Y_{k,l}, \bar{u}^j_k) - \Delta\tilde{b}_j h_u(\bar{x}_k, \bar{u}^j_k)\right\| \\
&\leq \Delta^2\left\|\sum_{l\in I_j}\sum_{i=1}^{s} b_i h_x(Y_{k,i}, \omega^i_k)\sum_{p=1}^{i-1}\frac{\partial}{\partial\omega^l}K_p(\bar{x}_k, \omega_k)\right\| + \Delta\sum_{l\in I_j} b_l\|h_u(Y_{k,l}, \bar{u}^j_k) - h_u(\bar{x}_k, \bar{u}^j_k)\| \\
&\leq \kappa_1\Delta^2,
\end{aligned} \tag{5.11a}$$

where we used the Lipschitz continuity of $h_u(\cdot,\cdot)$ and the fact that S bounded implies that $\bar{x}_k$ and $\bar{u}_k^j$ are bounded, which implies that for all $j \in \mathbf{r}$, $\|h_u(\bar{x}_k, \bar{u}_k^j)\|$ and $\|h_x(\bar{x}_k, \bar{u}_k^j)\|$ are bounded. Therefore, it follows from (5.5c) that

$$\bar{\gamma}_u^\nu(\bar{\eta})_k = [F_{w^1}(\bar{x}_k, \bar{u}_k)^T \bar{p}_{k+1}^\nu \cdots F_{w^r}(\bar{x}_k, \bar{u}_k)^T \bar{p}_{k+1}^\nu]$$

$$= \Delta[\tilde{b}_1 h_u^T(\bar{x}_k, \bar{u}_k^1)\bar{p}_{k+1}^\nu \cdots \tilde{b}_r h_u^T(\bar{x}_k, \bar{u}_k^r)\bar{p}_{k+1}^\nu] + O(\Delta^2), \tag{5.11b}$$

where $\lim_{\Delta\to 0} |O(\Delta)/\Delta| < \infty$. From (5.5a), $V_{\mathbf{A},N}(\nabla_u f_N^\nu(\eta)) = \bar{\gamma}_u^\nu(\bar{\eta})\mathbf{M}_N^{-1}$. Therefore, from (5.11b) we obtain

$$V_{\mathbf{A},N}(\nabla_u f_N^\nu(\eta))_k = \frac{\Delta}{\Delta}(\tilde{b}_1 h_u(\bar{x}_k, \bar{u}_k^1)^T \bar{p}_{k+1}^\nu \cdots \tilde{b}_r h_u(\bar{x}_k, \bar{u}_k^r)^T \bar{p}_{k+1}^\nu)M^{-1} + \frac{O(\Delta^2)}{\Delta}, \quad k \in \mathcal{N}. \tag{5.11c}$$

At this point we must deal with our two control representations separately. For representation R1, $u(\cdot) \in \mathbf{U}_N^1$ is a Lipschitz continuous polynomial on each interval $[t_k, t_{k+1})$, with Lipschitz constant κ_U given in (4.15a). Thus, for any $i_j, i_l \in I$, with $j, l \in \mathbf{r}$ and I defined in (4.4a),

$$\|\bar{u}_k^j - \bar{u}_k^l\| = \|u[\tau_{k,i_j}] - u[\tau_{k,i_l}]\| \le \kappa_U \|\Delta(c_{i_j} - c_{i_l})\| \le \kappa_U \Delta, \tag{5.12}$$

where Assumption 4.3 was used to justify the last inequality. Now, let

$$D \triangleq [\tilde{b}_1 h_u^T(\bar{x}_k, \bar{u}_k^1)\bar{p}_{k+1}^\nu \cdots \tilde{b}_r h_u^T(\bar{x}_k, \bar{u}_k^r)\bar{p}_{k+1}^\nu]M^{-1}, \tag{5.13a}$$

and let D^j, $j \in \mathbf{r}$, denote the jth column of D, so that, from (5.11c),

$$\nabla_u f_N^\nu(\eta)[\tau_{k,i_j}] = V_{\mathbf{A},N}(\nabla_u f_N^\nu(\eta))_k^j = D^j + O(\Delta). \tag{5.13b}$$

It follows from Assumptions 3.1(a) and 4.1′, (5.12), and the fact that $\bar{p}_{k+1}^\nu$ is bounded for any $\eta \in S$ that there exists $\kappa_2, \kappa_3 < \infty$, such that for any $j \in \mathbf{r}$ and $i_j \in I$, and with $M_{i,j}^{-1}$ denoting the i, jth entry of M^{-1},

$$\|D^j - h_u(\bar{x}_k, \bar{u}_k^j)^T p_{k+1}^\nu \sum_{i=1}^r \tilde{b}_i M_{i,j}^{-1}\| \le \left\| \sum_{i=1}^r \tilde{b}_i [h_u(\bar{x}_k, \bar{u}_k^i) - h_u(\bar{x}_k, \bar{u}_k^j)]^T p_{k+1}^\nu M_{i,j}^{-1} \right\|$$

$$\le \sum_{i=1}^r \kappa_2 \|\bar{u}_k^i - \bar{u}_k^j\| \|p_{k+1}^\nu M_{i,j}^{-1}\| \le \kappa_3 \Delta. \tag{5.13c}$$

Also, if $M^{-1}\tilde{b} = \mathbf{1}$, then $\sum_{i=1}^r M_{i,j}^{-1}\tilde{b}_i = 1$ since M is symmetric. Hence for any $j \in \mathbf{r}$,

$$\|D^j - h_u(x_k, u_k^j)^T \bar{p}_{k+1}^\nu\| \le \kappa_3 \Delta. \tag{5.13d}$$

Therefore, from (5.13b),

$$\nabla_u f_N^\nu(\eta)[\tau_{k,i_j}] = h_u(x_k, \bar{u}_k^j)^T \bar{p}_{k+1}^\nu + O(\Delta). \tag{5.13e}$$

For representation R2, $\bar{u}(\cdot)$ is not Lipschitz continuous on $[t_k, t_{k+1})$, so (5.12) does not hold. However, since $M = M_2$ is diagonal, (5.13e) is seen to be true directly from (5.11c) if $M^{-1}\tilde{b} = \mathbf{1}$.

Next, since S is bounded, (i) by Lemmas 4.10(i) and A.4 there exists $\kappa_4 < \infty$ such that $\|\bar{x}_k - x^\eta(t_k)\| \le \kappa_4 \Delta$ and $\|\bar{p}_{k+1}^\nu - p^{\eta,\nu}(t_{k+1})\| \le \kappa_4 \Delta$ and (ii) $\bar{p}_{k+1}^\nu$ and $h_u(\bar{x}^k, u[\tau_{k,i_j}])$ are bounded. Thus, making use of Theorem 3.2(v), (5.13e), the fact that both $x^\eta(\cdot)$ and $p^{\nu,\eta}(\cdot)$ are Lipschitz continuous, and $u[\tau_{k,i_j}] = \bar{u}_k^j$, we conclude that there exists $\kappa_5 < \infty$ such that

$$\|\nabla_u f^\nu(\eta)[\tau_{k,i_j}] - \nabla_u f_N^\nu(\eta)[\tau_{k,i_j}]\|$$

$$= \|h_u(x^\eta(\tau_{k,i_j}), u[\tau_{k,i_j}])^T p^{\nu,\eta}(\tau_{k,i_j}) - h_u(\bar{x}_k, u[\tau_{k,i_j}]^T \bar{p}_{k+1}^\nu)\| + O(\Delta) \le \kappa_5 \Delta. \tag{5.14}$$

Next, for $j \in \mathbf{r}$, $i_j \in I$, $k \in \mathcal{N}$, and $t \in [0, 1]$ we have that

$$\begin{aligned}\|\nabla_u f^\nu(\eta)(t) - \nabla_u f_N^\nu(\eta)(t)\| \leq\ & \|\nabla_u f^\nu(\eta)(t) - \nabla_u f^\nu(\eta)[\tau_{k,i_j}]\| \\ & + \|\nabla_u f^\nu(\eta)[\tau_{k,i_j}] - \nabla_u f_N^\nu(\eta)[\tau_{k,i_j}]\| \\ & + \|\nabla_u f_N^\nu(\eta)[\tau_{k,i_j}] - \nabla_u f_N^\nu(\eta)(t)\|. \end{aligned} \tag{5.15a}$$

The second term in (5.15a) is of order $O(\Delta)$ by (5.14). We will show that the first and third terms in (5.15a) are also of order $O(\Delta)$. First consider representation R1. It follows by inspection of (3.6b) in Theorem 3.2(v) that $\nabla_u f^\nu(\eta)(\cdot)$ is Lipschitz continuous on $t \in [t_k, t_{k+1})$, $k \in \mathcal{N}$, because $u \in L_N^1$ is Lipschitz continuous on these intervals. Since $\nabla_u f_N^\nu(\eta)(\cdot) \in L_N$, it is also Lipschitz continuous on these intervals. Finally, by Assumption 4.3, $\tau_{k,i_j} \in [t_k, t_{k+1}]$ for all $k \in \mathcal{N}$. Thus, the first and third terms are of order $O(\Delta)$ for all $t \in [0, 1]$. For representation R2, $\nabla_u f_N^\nu(\eta)(\cdot) \in H_N$ is constant on $t \in [t_k + d_{j-1}, t_k + d_j)$, $j \in \mathbf{r}$, and $k \in \mathcal{N}$. Since $u \in L_N^2$ is constant on these intervals, it again follows by inspection of (3.6b) in Theorem 3.2(v) that $\nabla_u f^\nu(\eta)(\cdot)$ is Lipschitz continuous on these intervals. Finally, by Assumption 4.6, $\tau_{k,i_j} \in [t_k + d_{j-1}, t_k + d_j]$, for all $k \in \mathcal{N}$ and $j \in \mathbf{r}$. Since $d_0 = 0$ and $d_r = \Delta$, the first and third terms are of order $O(\Delta)$ for all $t \in [0, 1]$. We conclude that there exist $\kappa_6 < \infty$ such that

$$\|\nabla_u f^\nu(\eta)(t) - \nabla_u f_N^\nu(\eta)(t)\| \leq \kappa_6 \Delta, \qquad t \in [0, 1], \tag{5.15b}$$

which implies that

$$\|\nabla_u f^\nu(\eta) - \nabla_u f_N^\nu(\eta)\|_2 \leq \kappa_6 \Delta. \tag{5.15c}$$

Next we consider the gradient with respect to initial conditions ξ. From Theorem 3.2(v) and (5.5b), $\|\nabla_\xi f^\nu(\eta) - \bar{\gamma}_\xi^\nu(\eta)\| \leq \|\nabla_\xi \zeta^\nu(\xi, x^\eta(1)) - \nabla_\xi \zeta^\nu(\xi, \bar{x}_N)\| + \|p^{\nu,\eta}(0) - \bar{p}_0^\nu\|$. Thus, since S is bounded, it follows from Assumption 3.1(b) and Lemmas 4.10 and A.4 that there exists $\kappa_7 < \infty$ such that

$$\|\nabla_\xi f^\nu(\eta) - \bar{\gamma}_\xi^\nu(\eta)\| \leq \kappa_7(\|x^\eta(1) - \bar{x}_N\| + \|p^{\nu,\eta}(0) - \bar{p}_0^\nu\|) \leq \kappa_7 \Delta. \tag{5.16}$$

Combining (5.15c) and (5.16), we see that there exists $\kappa < \infty$ such that for any $\eta_N \in S \cap H_N$

$$\|\nabla f^\nu(\eta_N) - \nabla f_N^\nu(\eta_N)\|_H \leq \frac{\kappa}{N}. \qquad \square \tag{5.17}$$

The following proposition states conditions for (5.10a) to hold.

PROPOSITION 5.7. (a) *Suppose* $M = M_1$. *Then* (5.10a) *holds if and only if the coefficients of the Butcher array satisfy*

$$\sum_{j=1}^{s} b_j c_j^{l-1} = \frac{1}{l}, \qquad l = 1, \ldots, r. \tag{5.18}$$

(b) *Suppose* $M = M_2$. *Then* (5.10a) *holds if and only if for all* $j \in \mathbf{r}$, $\tilde{b}_j \neq 0$.

Proof. (a) For $M = M_1$, it follows from (4.9b) that $M^{-1}\tilde{b} = \mathbf{1}$ if and only if

$$T^{-T} \operatorname{Hilb}(s)^{-1} T^{-1} \tilde{b} = \mathbf{1}. \tag{5.19a}$$

Now, it is easy to see that

$$T^{-1}\tilde{b} = \begin{bmatrix} \sum_{j=1}^{r} \tilde{b}_j \\ \sum_{j=1}^{r} \tilde{b}_j c_{i_j} \\ \vdots \\ \sum_{j=1}^{r} \tilde{b}_j c_{i_j}^{r-1} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{s} b_j \\ \sum_{j=1}^{s} b_j c_j \\ \vdots \\ \sum_{j=1}^{s} b_j c_j^{r-1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \\ \vdots \\ 1/r \end{bmatrix}, \tag{5.19b}$$

where the last equality holds if and only if (5.18) holds. Note that $T^{-1}\tilde{b}$ is then the first column of Hilb(r). Consequently,

$$\text{Hilb}(r)^{-1}T^{-1}\tilde{b} = \text{Hilb}(r)^{-1}\begin{bmatrix} 1 \\ 1/2 \\ \vdots \\ 1/r \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{5.19c}$$

which leads us to conclude that

$$M^{-1}\tilde{b} = T^{-T}\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & c_{i_1} & \cdots & c_{i_1}^{r-1} \\ 1 & c_{i_2} & & c_{i_2}^{r-1} \\ & & \ddots & \\ 1 & c_{i_r} & \cdots & c_{i_r}^{r-1} \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{5.19d}$$

(b) For $M = M_2$, given by (4.12b), M^{-1} is nonsingular if and only if $\tilde{b}_j \neq 0$. Also, (5.10a) holds if and only if $M\mathbf{1} = \tilde{b}$. Clearly then, if $\tilde{b}_j \neq 0$, (5.10a) holds because

$$M\mathbf{1} = \begin{bmatrix} \tilde{b}_1 & & \\ & \ddots & \\ & & \tilde{b}_r \end{bmatrix}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \tilde{b}. \qquad \square \tag{5.20}$$

Remark 5.8. The conditions (5.18) on the coefficients of the Butcher array for representation R1 are necessary conditions for the RK methods to be rth-order accurate [8, 17]. The condition with $l = 1$ in (5.18) is the same as the second part of Assumption 4.1′.

THEOREM 5.9. *For representation* R1, *suppose that Assumptions* 3.1, 4.1′, *and* 4.3 *and* (5.18) *hold and let* $d = 2$. *For representation* R2, *suppose that Assumptions* 3.1, 4.1′, *and* 4.6 *hold and let* d *be the least common denominator for the elements* b_j, $j \in \mathbf{s}$, *of the Butcher array. Let* $\mathbf{N} \triangleq \{d^l\}_{l=1}^{\infty}$ *and suppose that* $\{\eta_N\}_{N\in K}$, $K \subset \mathbf{N}$, *is such that* $\eta_N \in \mathbf{H}_N$ *for all* $N \in K$ *and* $\eta_N \to \eta$ *as* $N \to \infty$. *Then* $\theta_N(\eta_N) \to^K \theta(\eta)$ *as* $N \to \infty$.

Proof. Let $\tilde{\Psi} : \mathbf{H} \times \mathbf{H} \to \mathbb{R}$ be defined by

$$\tilde{\Psi}(\eta, \eta') \triangleq \max\left\{\max_{\nu\in\mathbf{q}_o} \tilde{f}^{\nu}(\eta, \eta') - \psi_o(\eta) - \sigma\psi_c(\eta)_+, \max_{\nu\in\mathbf{q}_c+q_o} \tilde{f}^{\nu}(\eta, \eta') - \psi_c(\eta)_+\right\}, \tag{5.21a}$$

and let $\tilde{\Psi}_N : \mathbf{H}_N \times \mathbf{H}_N \to \mathbb{R}$ be defined by

$$\tilde{\Psi}_N(\eta, \eta') \triangleq \max\left\{\max_{\nu\in\mathbf{q}_o} \tilde{f}_N^{\nu}(\eta, \eta') - \psi_{o,N}(\eta) - \sigma\psi_{c,N}(\eta)_+, \max_{\nu\in\mathbf{q}_c+q_o} \tilde{f}_N^{\nu}(\eta, \eta') - \psi_{c,N}(\eta)\right\}, \tag{5.21b}$$

so that $\theta(\eta) = \min_{\eta' \in \mathbf{H}} \tilde{\Psi}(\eta, \eta')$ and $\theta_N(\eta) = \min_{\eta' \in \mathbf{H}_N} \tilde{\Psi}_N(\eta, \eta')$. Now, suppose that $\{\eta_N\}_{N \in K}$ is a sequence such that, for all N, $\eta_N \in \mathbf{H}_N$ and $\eta_N \to^K \eta$. From the proof of Theorem 4.12, $\eta \in \mathbf{H}$. Let $\hat{\eta} \in \mathbf{H}$ be such that $\theta(\eta) = \tilde{\Psi}(\eta, \hat{\eta})$, and let $\{\eta'_N\}_{N \in K}$ be any sequence such that, for all N, $\eta'_N \in \mathbf{H}_N$ and $\eta'_N \to^K \hat{\eta}$. Then,

$$\begin{aligned}\theta_N(\eta_N) &\leq \tilde{\Psi}_N(\eta_N, \eta'_N)\\ &\leq \tilde{\Psi}(\eta_N, \eta'_N) + \max\Big\{\max_{\nu \in \mathbf{q}_o}\{\tilde{f}^\nu_N(\eta_N, \eta'_N) - \tilde{f}^\nu(\eta_N, \eta'_N)\}\\ &\qquad -[\psi_{o,N}(\eta_N) - \psi_o(\eta_N)] - [\sigma\psi_{c,N}(\eta_N)_+ - \sigma\psi_c(\eta_N)_+],\\ &\qquad \max_{\nu \in \mathbf{q}_c + q_o}\{\tilde{f}^\nu_N(\eta_N, \eta'_N) - \tilde{f}^\nu(\eta_N, \eta'_N)\} - [\psi_{c,N}(\eta_N) - \psi_c(\eta_N)]\Big\}.\end{aligned} \tag{5.22}$$

It follows from Theorem 4.12, Theorem 5.6, Proposition 5.7, and the fact that $\{\eta_N\}_{N \in K}$ is a bounded set that each part of the max term on the right-hand side of (5.22) converges to zero as $N \to \infty$. The quantity $\tilde{\Psi}(\eta_N, \eta'_N)$ converges to $\theta(\eta)$ since $\eta_N \to^K \eta$, $\eta'_N \to^K \hat{\eta}$, and $\tilde{\Psi}(\cdot, \cdot)$ is continuous. Thus, taking limits of both sides of (5.22), we obtain that $\overline{\lim}\theta_N(\eta_N) \leq \theta(\eta)$ (this proves that Definition 2.4 holds for the optimality functions of the approximating problems). Now, for all $N \in K$, let $\hat{\eta}_N \in \mathbf{H}_N$ be such that $\theta_N(\eta_N) = \tilde{\Psi}_N(\eta_N, \hat{\eta}_N)$. Then, $\theta(\eta_N) \leq \tilde{\Psi}(\eta_N, \hat{\eta}_N)$ and proceeding in a similar fashion as (5.22) and taking limits, we see that $\theta(\eta) \leq \underline{\lim}\theta_N(\eta_N)$. Hence, together with the previous result, we can conclude that $\theta_N(\eta_N) \to^K \theta(\eta)$ as $N \to \infty$. □

Since the union of the spaces H_N is dense in $H_{\infty,2}$ and Theorem 5.9 holds, it follows that the hypographs of the optimality functions $\theta_N(\cdot)$ converge to the hypograph of the optimality function $\theta(\cdot)$, in the Kuratowski sense, i.e., the $-\theta_N \to^{\text{Epi}} -\theta$.

The following corollary is a direct result of Theorem 4.12 (epiconvergence) and Theorem 5.9.

COROLLARY 5.10 (consistency). *For representation* R1, *suppose that Assumptions* 3.1, 4.1′, 4.3, *and* 4.11 *and* (5.18) *hold. For representation* R2, *suppose that Assumptions* 3.1, 4.1′, 4.6, *and* 4.11 *hold. Let* $\mathbf{N} = \{d^l\}_{l=1}^\infty$ *where* $d = 2$ *for representation* R1 *and* d *is the least common denominator of the* $\tilde{b}_j$, $j \in \mathbf{s}$, *for representation* R2. *Then, the approximating pairs* $(\mathbf{CP}_N, \theta_N)$, $N \in \mathbf{N}$ *are consistent approximations to the pair* $(\mathbf{CP}, \theta)$.

6. Numerical results. The problems $\overline{\mathbf{CP}}_N$ can be solved using existing optimization methods (e.g., [18] and [22]). These methods, however, are defined on a Euclidean space and existing code would have to be modified for use on the coefficient spaces $\tilde{L}^i_N$, $i = 1, 2$. To avoid this difficulty, we will now define a change of coordinates in coefficient space that implicitly defines an orthonormal basis for the subspace L^i_N and, hence, turns the coefficient space into a Euclidean space.

Let $L_N = L^1_N$ or L^2_N and, correspondingly, $V_{\mathbf{A},N} = V^1_{\mathbf{A},N}$ or $V^2_{\mathbf{A},N}$. Recall from (5.5a) that, for $\eta = (\xi, u) \in H_N$ and $\nu \in \mathbf{q}$, $\nabla_u f^\nu_N(\eta) = V^{-1}_{\mathbf{A},N}(\bar{\gamma}^\nu_u(\bar{\eta})\mathbf{M}^{-1}_N)$, where $\bar{\eta} = (\xi, \bar{u}) = W_{\mathbf{A},N}(\eta)$ and $\bar{\gamma}^\nu_u(\bar{\eta})$, defined in (5.5c), is the gradient of $\bar{f}^\nu_N(\cdot)$ with respect to the standard l_2 inner product on $\underset{N}{\times}\,\underset{r}{\times}\,\mathbb{R}^m$. The gradient of $\bar{f}^\nu_N(\cdot)$ with respect to the inner product on $\tilde{L}_N$ is given by $\nabla_u \bar{f}^\nu_N(\bar{\eta}) = V_{\mathbf{A},N}(\nabla_u f^\nu_N(\eta)) = \bar{\gamma}^\nu_u(\bar{\eta})\mathbf{M}^{-1}_N$ and satisfies

$$\langle \nabla_u f^\nu_N(\eta), \delta u\rangle_2 = \langle \nabla_u \bar{f}^\nu_N(\bar{\eta}), \delta\bar{u}\rangle_{\tilde{L}_N} = \langle \bar{\gamma}^\nu_u(\bar{\eta}), \delta\bar{u}\rangle_{l_2} \tag{6.1}$$

for any $\delta u \in H_N$ and $\delta\bar{u} = V_{\mathbf{A},N}(\delta u)$. Introduce a new coefficient space, $\tilde{L}_N = \times_N \times_r \mathbb{R}^m$, endowed with the standard l_2 inner product and norm, and the transformation $Q : \tilde{L}_N \to \tilde{L}_N$

defined by

$$\tilde{u} = Q(\bar{u}) = \bar{u}\mathbf{M}_N^{1/2}, \tag{6.2a}$$

where $\mathbf{M}_N$ is defined in (5.4). Let $\tilde{\eta} = (\xi, \tilde{u})$, and for each $\nu \in \mathbf{q}$ let $\tilde{f}_N^\nu : \mathbb{R}^n \times \tilde{L}_N \to \mathbb{R}$ be defined by

$$\tilde{f}_N^\nu(\tilde{\eta}) \triangleq \bar{f}_N^\nu((\xi, Q^{-1}(\tilde{u}))). \tag{6.2b}$$

Finally, let $\bar{\eta} = (\xi, Q^{-1}(\tilde{u}))$. Then, by the chain rule,

$$\nabla_{\tilde{u}} \tilde{f}_N^\nu(\tilde{\eta}) = Q^{-1}(\nabla_{\bar{u}} \bar{f}_N^\nu(\bar{\eta})) = \bar{\gamma}_u^\nu(\bar{\eta})\mathbf{M}_N^{-1/2}. \tag{6.2c}$$

Thus, $\langle \nabla_{\tilde{u}} \tilde{f}_N^\nu(\tilde{\eta}), \delta\tilde{u}\rangle_{l_2} = \langle \nabla_{\bar{u}} \bar{f}_N^\nu(\bar{\eta}), \delta\bar{u}\rangle_{\bar{L}_N} = \langle \nabla_u f_N^\nu(\eta_N), \delta u_N\rangle_2$, where $\delta\tilde{u} = Q(\delta\bar{u})$.

Implicitly, the transformation Q creates an orthonormal basis for L_N because under this transformation the inner-product and norm on L_N are equal to the l_2 inner-product and norm on the coefficient space. With this transformation, the approximating problems $\overline{\mathbf{CP}}_N$ can be solved using standard nonlinear programming methods without introducing ill-conditioning. It is important to note, however, that control constraints are also transformed. Thus, the constraint $\bar{u} \in \overline{\mathbf{U}}_N$ becomes $\tilde{u}\mathbf{M}_N^{-1/2} \in \overline{\mathbf{U}}_N$. For representation R1, since $\mathbf{M}_N^{-1/2}$ is not diagonal (except if $r = 1$), this means that the transformed control constraints will, for each $k \in \mathcal{N}$, involve linear combinations of the control samples $\tilde{u}_k^j$, $j \in \mathbf{r}$.

We will now present a numerical example that shows, in particular, that this transformation can make a substantial difference in the performance of an algorithm.

Example. Consider the linear-quadratic problem taken from [15]

$$\min_{u \in \mathbf{U}} f(u), \qquad f(u) \triangleq x_2^u(1), \tag{6.3a}$$

where $x(t) = (x_1(t), x_2(t))^T$ and

$$\dot{x} = \begin{bmatrix} 0.5x_1 + u \\ 0.625x_1^2 + 0.5x_1u + 0.5u^2 \end{bmatrix}, \quad x(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad t \in [0, 1]. \tag{6.3b}$$

The solution to this problem is given by

$$u^*(t) = -(\tanh(1-t) + 0.5)\cosh(1-t)/\cosh(1), \qquad t \in [0, 1], \tag{6.4}$$

with optimal cost $x_2^*(1) = e^2 \sinh(2)/(1+e^2)^2 \approx 0.380797$.

The approximating cost functions are $f_N(u) = (0\ 1)\bar{x}_N^u$ where $\{\bar{x}_k^u\}_{k=0}^N$ is the RK solution for a given control $u \in L_N$. We discretized the dynamics using two common RK methods. The first is a third-order method defined by the Butcher array $\mathbf{A}_1 = [c, A, b]$ with $c = (0, 1/2, 1)$, $b = (1/6, 2/3, 1/6)$, and the nonzero entries of A are $a_{2,1} = 1/2$, $a_{3,1} = -1$, and $a_{3,2} = 2$. The matrices $\mathbf{M}_N$ used to define the transformation Q in (6.2a) are given by (5.4) with

$$M = M_1 = \frac{1}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix}, \qquad M = M_2 = \frac{1}{6}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{6.5}$$

We solved the approximating problems using steepest descent with the step-size determined by an Armijo rule augmented with a quadratic fit based on the value of $f_N(\cdot)$ at the last

TABLE 1
Conditioning effect of the transformation Q on approximating problems.

	Number of iterations	
N	$M = M_i, i = 1, 2$	$M = \frac{1}{N} I$
10	4	19
20	4	19
40	5	23
80	5	24

two evaluations in the line search. The stopping criterion[4] was $\|\bar{\gamma}_u\|_2 \leq (3.1e-4)/N$; and the initial guess was $u(t) = 0, t \in [0, 1]$. Table 1 shows the number of iterations required to solve the approximating problems for different discretization levels N with and without the transformation (6.2a,b). We see that solving the discretized problems without the transformation requires about five times the number of iterations required for solving the problem with the transformation. The situation can be even worse for other RK methods. The choice of representation R1 versus representation R2 had no effect on the number of iterations required.

The second RK method is the fourth-order method RK4 defined by the Butcher array $\mathbf{A}_2 = [c, A, b]$ with $c = (0, 1/2, 1/2, 1)$, $b = (1/6, 1/3, 1/3, 1/6)$ and the nonzero entries of A are $a_{2,1} = a_{3,2} = 1/2$ and $a_{4,3} = 1$. The matrices M_1 and M_2, equation (6.5), are the same for this method, since $c_2 = c_3 = 1/2$ implies $r = 3$ and $\tilde{b}_2 = 2/3$. We use this method because it is very common and demonstrates the advantage of treating the samples arising from repeated c_i values in the Butcher array as the same sample (see Remark 4.8).

To see this advantage, let $\{u_N^*\}_{N \in \mathbf{N}}$, $\mathbf{N} \subset \mathbb{N}$, be solutions of $\mathbf{CP}_N$, and suppose $u_N^* \to u^*$ where u^* is a solution of $\mathbf{CP}$. In [16, Thm. 3.1], Hager establishes, for symmetric RK methods [1, 27], a tight upper bound on the error $E_N^u \triangleq \|V_{\mathbf{A},N}(u^*) - V_{\mathbf{A},N}(\bar{u}_N^*)\|_\infty$, of second order in $\Delta = 1/N$ for smooth, unconstrained problems. Note that $V_{\mathbf{A},N}(u^*)_k^j = u^*(\tau_{k,j})$, $k \in \mathcal{N}$ and $j \in \mathbf{r}$, because $u^*(\cdot)$ is smooth for smooth problems [26]. Hager used the problem given in (6.4a) to demonstrate the tightness of this bound. For the particular RK method given by the Butcher array $\mathbf{A}_2$, we can state the following improved result (which, according to Proposition 5.5, does not depend on the control representation).

PROPOSITION 6.1. *Let* $\mathbf{CP} \triangleq \min_{u \in \mathbf{U}} f(x^u(1))$, *u unconstrained. Suppose the approximating problems* $\mathbf{CP}_N$ *are produced by discretizing* $\mathbf{CP}$ *with the fourth-order RK method with Butcher array* $\mathbf{A}_2 : c = (0, 1/2, 1/2, 1)$, $b = (1/6, 1/3, 1/3, 1/6)$ *and the nonzero entries of A are* $a_{2,1} = a_{3,2} = 1/2$ *and* $a_{4,3} = 1$. *Further suppose that the conditions of Lemma* 4.10(ii) *hold with* $\rho = 4$. *Let* $\{u_N^*\}_{N \in \mathbf{N}}$, $\mathbf{N} \subset \mathbb{N}$, *be solutions of* $\mathbf{CP}_N$, *and suppose* $u_N^* \to u^*$ *where* u^* *is a solution of* $\mathbf{CP}$. *Then* $E_N^u \triangleq V_{\mathbf{A},N}(u^*) - V_{\mathbf{A},N}(u_N^*)\|_\infty = O(\Delta^3)$.

Sketch of Proof. In [15], it is shown, using a reasonable nonsingularity assumption on the Hessians of $f_N(\cdot)$, that the accuracy of the solutions of the approximating problems is determined by N times the size of the discrete-time gradient (using the standard l_2 inner-product) of the approximating problem at $\bar{u}^* \triangleq V_{\mathbf{A},N}(u^*)$, that is, $E_N^u \sim \|\bar{\gamma}_u(\bar{u}^*)\|$. This, in turn, is a function of the accuracy of the state and adjoint approximations. For the RK method under consideration, Hager shows that, for $k \in \mathcal{N}$, the variables $\bar{u}_k^{*1} = u^*(t_k)$ and $\bar{u}_k^{*3} = u(t_k + \Delta)$ are third-order approximations to $u^*(t_k)$ and $u^*(t_k + \Delta)$, respectively. Thus, we need only show that $\bar{u}_k^{*2} = u^*(t_K + \Delta/2)$ is a third-order approximation to $u^*(t_k + \Delta/2)$.

Let $Y_{k,2} = \bar{x}_k + \Delta/2h(\bar{x}_k, \bar{u}_k^{*1})$ and $Y_{k,3} = \bar{x}_k + \Delta/2h(Y_{k,2}, \bar{u}_k^{*2})$ represent certain intermediate values used by the RK method at the kth time step. In Hager's notation, $Y_{k,2} = y(1, k)$ and $Y_{k,3} = y(2, k)$. Hager introduces a clever transformation, specific to symmetric

[4]Higher precision was difficult to achieve when the Q transformation was not used.

TABLE 2
Rate of convergence; conditioning effect of the transformation Q.

	Accuracy of solutions				*Number of iterations*	
N	E_N^u	E_N^u/E_{2N}^u	E_N^f	E_N^f/E_{2N}^f	$M = M_i,\ i = 1, 2$	$M = \frac{1}{N} I$
10	1.48e-4	7.91	2.86e-7	16.22	4	21
20	1.87e-5	7.99	1.76e-8	16.13	5	21
40	2.34e-6	7.62	1.09e-9	16.07	5	23
80	3.07e-7		6.80e-11		5	23

RK methods, for the adjoint variables so that they can be viewed as being calculated with the same RK method used to compute the state variables but run backward in time. The intermediate adjoint variables of interest here are denoted by $q(2, k)$ and $q(1, k)$. With this transformation, the discrete-time gradients for the approximating problems have the same form as the continuous-time gradient for the original problem. Since $c_2 = c_3 = 1/2$, $\bar{\gamma}_u(\bar{u}^*)_k^2 = 2\Delta/3[1/2h_u(Y_{k,2}, \bar{u}_k^{*2})^T q(1, k) + 1/2h_u(Y_{k,3}, \bar{u}_k^{*2})^T q(2, k)]$. Further, since $2\Delta/3h_u(x^{u^*}(t_k + \Delta/2), u^*(t_k + \Delta/2))^T p^{u^*}(t_k + \Delta/2) = 0$, $\|\bar{\gamma}_u(\bar{u}^*)_k^2\|$ is bounded by $\frac{2}{3}\Delta$ times the maximum of $\|(Y_{k,2} + Y_{k,3})/2 - x^{u^*}(t_k + \Delta/2)\|$ and $\|(q(2, k) + q(1, k))/2 - p^{u^*}(t_k + \Delta/2)\|$. Let

$$w(k) \triangleq \frac{Y_{k,2} + Y_{k,3}}{2} = \bar{x}_k + \frac{\Delta}{4}\left[h(\bar{x}_k, u_k^{*1}) + h\left(\bar{x}_k + \frac{\Delta}{2}h(\bar{x}_k, \bar{u}_k^{*1}), \bar{u}_k^{*2}\right)\right]$$

$$= \bar{x}_k + \frac{\Delta'}{2}[h(\bar{x}_k, u_k^{*1}) + h(\bar{x}_k + \Delta' h(\bar{x}_k, \bar{u}_k^{*1}), \bar{u}_k^{*2})], \tag{6.6}$$

where $\Delta' = \Delta/2$. Thus, $w(k)$ is produced by the improved Euler rule applied to $\bar{x}_k$. Since the local truncation error for the improved Euler rule is of order $O(\Delta^3)$ and $\bar{x}_k$ is of order $O(\Delta^4)$, $\|w(k) - x^{u^*}(t_k + \Delta/2)\|$ is of order $O(\Delta^3)$. In the same way, it can be shown that $\|q(2, k) + q(1, k))/2 - p^{u^*}(t_k + \Delta/2)\|$ is $O(\Delta^3)$. Thus, we can conclude that $\|\bar{\gamma}_u(\bar{u}^*)_k^2\| = O(\Delta^4)$ for all $k \in \mathcal{N}$. This implies that the solutions of the approximating problems satisfy $\|\bar{u}_{N,k}^{*j} - u^*(\tau_{k,j})\| = O(\Delta^3)$ for all $k \in \mathcal{N}$ and $j \in \mathbf{r}$. □

Table 2 summarizes our numerical results using the RK method with Butcher array $\mathbf{A}_2$. The first column gives the discretization level. Columns 2 and 3 show that doubling the discretization results in an eightfold reduction in the control error. Thus, as predicted by Proposition 6.1, E_N^u is $O(\Delta^3)$. The next two columns, agreeing with Hager's observations that the optimal trajectories of the approximating problem converge to those of the original problem with the same order as the order of the symmetric RK method, show that $E_N^f \triangleq |f(u^*) - f_N(\bar{u}_N^*)|$ is of order $O(\Delta^4)$. The numbers in columns 2 and 4 were obtained by solving the discretized problems to full precision. Finally, we include in the last two columns the number of iterations required to solve the approximate problem with and without the transformation Q. The stopping criterion was the same as used for Table 1. As with the previous method, the effect of the Q transformation is quite significant. The solution of the untransformed problem requires about five times the number of iterations required to solve the transformed problem.

The last table shows the accuracy of the gradients for the approximating problems produced with the second RK method (Butcher array $\mathbf{A}_4$) evaluated at the control $u(t) = -1 + 2t$. The first column shows the discretization level N. The second and third columns confirm that the gradients, $\nabla \bar{f}_N(\bar{u}) = \bar{\gamma}_u \mathbf{M}_N^{-1}$, for the approximating problems converge to the gradients of the original problem. Note that, based on the proof of Theorem 5.6, it is enough to show that the gradients converge at the points τ_{k,i_j}, $k \in \mathcal{N}$, $j \in \mathbf{r}$, and $i_j \in I$. The fourth column of Table 3 shows that the gradients that result if one uses the standard l_2 inner product on $\times_N \times_r \mathbb{R}^m$ do not converge.

TABLE 3
Convergence of gradients.

N	$M = M_1$ $\|V_{\mathbf{A},N}(\nabla f(u)) - \nabla \bar{f}_N(\bar{u})\|_\infty$	$M = M_2$ $\|V_{\mathbf{A},N}(\nabla f(u)) - \nabla \bar{f}_N(\bar{u})\|_\infty$	$M = \frac{1}{N} I$ $\|V_{\mathbf{A},N}(\nabla f(u)) - N\bar{\gamma}_u(\bar{u})\|_\infty$
10	1.67e-3	6.46e-4	1.48
20	3.77e-4	8.31e-5	1.48
40	9.94e-5	1.05e-5	1.48
80	2.55e-5	1.33e-6	1.48

7. Conclusion. We have shown that a large class of Runge–Kutta integration methods can be used to construct consistent approximations to continuous-time optimal control problems. The construction of consistent approximations is not unique: it is determined by the selection of families of finite-dimensional subspaces of the control space. Because the elements of these subspaces are discontinuous functions, appropriate extensions of Runge–Kutta methods must be used. Not all convergent Runge–Kutta methods, however, produce consistent approximations. This was observed both numerically and by failure to prove consistency of approximation with these methods. We have considered two selections of control subspaces in this paper, one defined by piecewise polynomial functions and one by piecewise constant functions. Splines can also be used and are treated in [28]. Each selection has some advantages and some disadvantages. A final selection has to be made on the basis of secondary considerations such as the importance of approximate solutions satisfying the original control constraints, the form that the control constraints take in the discrete-time optimal problems or the accuracy with which the differential equation is integrated.

As in our case, the basis functions that are used implicitly to define the finite-dimensional control subspaces may turn out to be nonorthonormal. In this case care must be taken to introduce a nonstandard inner product and corresponding norm in solving the resulting approximating discrete-time optimal control problems. Neglecting to do so amounts to a change of coordinates that can lead to serious ill-conditioning. This ill-conditioning is demonstrated in §6.

Finally, the use of the framework of consistent approximations opens up the possibility of developing optimal discretization strategies, such as those considered for semi-infinite programming in [16]. Such a strategy provides rules for selecting the number of approximating problems to be used as well as the discretization level, the order of the RK method, and the number of iterations of a particular optimization algorithm to be applied for each such approximating problem, so as to minimize the computing time needed to reach a specified degree of accuracy in solving an optimal control problem.

Appendix A. In this appendix we collect a few results used in the analysis of §§4 and 5. We will continue to use the notation of §4, that is, $\Delta = 1/N$, $t_k = k\Delta$, and $\tau_{k,i} = t_k + c_i\Delta$.

LEMMA A.1. *For the representation* R1, *suppose that Assumptions* 3.1(a), 4.1′, *and* 4.3 *hold. For representation* R2, *suppose that Assumptions* 3.1(a), 4.1′, *and* 4.6 *hold. For any bounded subset* $S \subset \mathbf{B}$, *there exists a* $\kappa < \infty$ *such that for any* $\eta = (\xi, u) \in S \cap \mathbf{H}_N$, $\|\delta_k\| \leq \kappa\Delta^2$ *for all* $k \in \mathcal{N}$, *where*

$$\delta_k \triangleq x^\eta(t_k) - x^\eta(t_{k+1}) + \Delta \sum_{i=1}^{s} b_i h(x^\eta(t_k), u[\tau_{k,i}]), \qquad k \in \mathcal{N}, \tag{A.1}$$

with $x^\eta(\cdot)$ *the solution of the differential equation* (3.1) *and* $u[\tau_{k,i}]$ *defined by* (4.6e) *for representation* R1 *and* (4.11c) *for representation* R2.

Proof. Let $\tilde{b}_j$ and d_j be as defined in (4.10), and for $j \in \mathbf{r}$ let $i_j \in I$ where I is given by (4.4a). Then, writing $x(\cdot) = x^\eta(\cdot)$, since the solution of (3.1) satisfies $x(t_{k+1}) = x(t_k) + \int_{t_k}^{t_{k+1}} h(x(t), u(t))\,dt$, we see that

$$\begin{aligned}\delta_k &= \Delta \sum_{i=1}^{s} b_i h(x(t_k), u[\tau_{k,i}]) - \int_{t_k}^{t_{k+1}} h(x(t), u(t))\,dt \\ &= \sum_{j=1}^{r} \int_{t_k+d_{j-1}}^{t_k+d_j} h(x(t_k), u[\tau_{k,i_j}])\,dt - \sum_{j=1}^{r} \int_{t_k+d_{j-1}}^{t_k+d_j} h(x(t), u(t))\,dt \qquad \text{(A.2a)}\end{aligned}$$

because $d_j - d_{j-1} = \Delta\tilde{b}_j$, $u[\tau_{k,i_j}] = u[\tau_{k,i}]$ for all $i \in I_j$, $d_0 = 0$ and, by Assumption 4.1′, $d_r = \Delta \sum_{j=1}^{r} \tilde{b}_j = \Delta \sum_{j=1}^{s} b_j = \Delta$. Since $d_j - d_{j-1} > 0$ by Assumption 4.1′, we have that

$$\begin{aligned}\|\delta_k\| &\le \sum_{j=1}^{r} \int_{t_k+d_{j-1}}^{t_k+d_j} \|h(x(t_k), u[\tau_{k,i_j}]) - h(x(t), u(t))\|\,dt \\ &\le \sum_{j=1}^{r} \int_{t_k+d_{j-1}}^{t_k+d_j} \kappa_1[\|x(t_k) - x(t)\| + \|u[\tau_{k,i_j}] - u(t)\|]\,dt, \qquad \text{(A.2b)}\end{aligned}$$

where $\kappa_1 < \infty$ is as in Assumption 3.1(a). Now, for $t \in [t_k, t_{k+1}]$, there exists $\kappa_2 < \infty$ such that

$$\|x(t_k) - x(t)\| \le \int_{t_k}^{t} \|h(x(t), u(t))\|\,dt \le \int_{t_k}^{t_{k+1}} \kappa_2[\|x(t)\| + 1]\,dt \qquad \text{(A.3)}$$

by Assumption 3.1(a) and the fact that S is bounded. Also because S is bounded, it follows from Theorem 3.2(ii) that there exists $L < \infty$ such that $\|x(t)\| \le \kappa_3[\|\xi\| + 1] \le L$. Thus, for $t \in [t_k, t_{k+1}]$, $\|x(t_k) - x(t)\| \le \int_{t_k}^{t_{k+1}} \kappa_2[L+1]\,dt = \Delta\kappa_2(L+1)$. Next, for representation R1, for any $k \in \mathcal{N}$, $j \in \mathbf{r}$, and $t \in [t_k + d_{j-1}, t_k + d_j)$, $\|u[\tau_{k,i_j}] - u(t)\| \le \kappa_U \Delta$, where κ_U is used in (4.15a), since by construction, $\bar{u} \in \mathbf{U}_N^1$ is a Lipschitz continuous polynomial on $[t_k, t_{k+1})$ with Lipschitz constant κ_U independent of N, $\tau_{k,i_j} \in [t_k, t_k + \Delta]$ by Assumption 4.3, and $0 \le d_j \le \Delta$ for $j = 0, \ldots, r$ by Assumption 4.1′, which implies that $[t_k + d_{j-1}, t_k + d_j) \subset [t_k, t_k + \Delta)$. The same holds for representation R2, since $u \in L_N^2$ is constant on $t \in [t_k + d_{j-1}, t_k + d_j)$ and $\tau_{k,i_j} \in [t_k + d_{j-1}, t_k + d_j]$ by Assumption 4.6. Therefore,

$$\|\delta_k\| \le \sum_{j=1}^{r} \int_{t_k+d_{j-1}}^{t_k+d_j} \kappa_1(\kappa_2(L+1) + \kappa_U)\Delta\,dt = \kappa\Delta \sum_{j=1}^{r} \int_{t_k+d_{j-1}}^{t_k+d_j} dt = \kappa\Delta^2, \qquad \text{(A.4)}$$

where $\kappa = \kappa_1(\kappa_2(L+1) + \kappa_U)$. This completes our proof. □

Remark A.2. The result in Lemma A.1 can be shown to hold even if the constraints on $\|u_K T_j\|$ in the definition (4.15a) of $\overline{\mathbf{U}}_N^1$ were removed if $h(x, u) = \tilde{h}(x) + Bu$ and the RK method is of order r. Starting from (A.2a), we have

$$\delta_k = \sum_{j=1}^{r} \int_{t_k+d_{j-1}}^{t_k+d_j} \tilde{h}(x(t_k)) - \tilde{h}(x(t))\,dt + \Delta \sum_{j=1}^{r} \tilde{b}_j B u[\tau_{k,j}] - \int_{t_k}^{t_{k+1}} Bu(t)\,dt. \qquad \text{(A.5a)}$$

The first term is $O(\Delta^2)$ by the argument already presented. For the remaining part, we see that

$$B\left[\Delta \sum_{j=1}^{r} \tilde{b}_j u[\tau_{k,j}] - \int_0^{\Delta} u(t + t_k)\,dt\right] = 0, \qquad \text{(A.5b)}$$

since a ρth-order Runge–Kutta method, $\rho \geq r$, integrates the equation $\dot{x} = u(t + t_k)$ exactly for any rth order polynomial u.

The next lemma concerns the functions $K_{k,i} = K_i(\bar{x}_k, \omega_k)$ of the RK method defined by (4.3a,b). The proof of this result is easily obtained from the proof for Lemma 222A in [8, p. 131].

LEMMA A.3. *Suppose Assumption* 3.1(a) *holds. Let* $S \subset \mathbf{B}$ *be bounded. Then there exists* $L < \infty$ *and* $N^* < \infty$ *such that for all* $N \geq N^*$, $\eta \in S \cap H_N$, $k \in \mathcal{N}$, *and* $i \in \mathbf{s}$,

$$\|K_{k,i} - h(\bar{x}_k, u[\tau_{k,i}])\| \leq L\Delta. \tag{A.6}$$

Next, we present a proof of Lemma 4.10.

Proof of Lemma 4.10. (i) Convergence. Let $\eta = (\xi, u) \in S \cap \mathbf{H}_N$, and for $k \in \mathcal{N}$ let $e_k \triangleq \bar{x}_k^\eta - x^\eta(t_k)$. Then $\|e_0\| = 0 \leq \kappa\Delta$ and by adding and subtracting terms,

$$\begin{aligned} e_{k+1} &= \bar{x}_k^\eta + \Delta \sum_{i=1}^{s} b_i K_{k,i} - x^\eta(t_{k+1}) \\ &= e_k + \left[x^\eta(t_k) - x^\eta(t_{k+1}) + \Delta \sum_{i=1}^{s} b_i h(x^\eta(t_k), u[\tau_{k,i}]) \right] \\ &\quad + \Delta \sum_{i=1}^{s} b_i \left[K_{k,i} - h(x^\eta(t_k), u[\tau_{k,i}]) \right]. \end{aligned} \tag{A.7}$$

The norm of the second term in this expression is bounded by $\kappa_1\Delta^2$ by Lemma A.1 where $\kappa_1 < \infty$. Using Lemma A.3, Assumption 3.1(a), and the fact that $|b_i| \leq 1$ by Assumption 4.1′, we conclude for the third term that there exists $\kappa_2 < \infty$ such that

$$\begin{aligned} &\Delta \left\| \sum_{i=1}^{s} b_i \left[K_{k,i} - h(x^\eta(t_k), u[\tau_{k,i}]) \right] \right\| \\ &\quad \leq \Delta \sum_{i=1}^{s} \|K_{k,i} - h(\bar{x}_k^\eta, u[\tau_{k,i}])\| + \Delta \sum_{i=1}^{s} \|h(\bar{x}_k^\eta, u[\tau_{k,i}]) - h(x^\eta(t_k), u[\tau_{k,i}])\| \\ &\quad \leq \Delta^2 Ls + \Delta\kappa_2 s \|e_k\|. \end{aligned} \tag{A.8}$$

Thus, for all $k \in \mathcal{N}$,

$$\|e_{k+1}\| \leq (1 + \kappa_2\Delta s)\|e_k\| + \kappa_3\Delta^2, \tag{A.9}$$

where $\kappa_3 = \kappa_1 + Ls$. Solving (A.9), we see that for all $k \in \mathcal{N}$, $\|e_k\| \leq (1 + \kappa_2\Delta s)^N \|e_0\| + \kappa_3'\Delta \leq \kappa\Delta$. This proves (4.18a).

(ii) Order of Convergence. We prove (4.18c) in two steps. First suppose that $H_N = H_N^1 = \mathbb{R}^n \times L_N^1$ and let $\eta_1 = (\xi, u_1) \in S \cap H_N^1$ be given. The expansion based on higher order derivatives (see [8]) needed to prove (4.18c) requires smoothness of $h(x, u)$ between time steps. The stated assumptions provide enough smoothness if uniform piecewise smoothness of $u_1(\cdot)$ is assumed. Alternatively, the result can also be shown to hold without this assumption on u_1 if the differential equations describing the system dynamics are linear and time-invariant with respect to u (as in Remark A.2). In either case, using the same type of reasoning as in the proof of Lemma A.1, we conclude that there exists $\kappa < \infty$, independent of η, such that (4.18c) holds for representation R1. Next, to prove (4.18c) for representation R2, let $H_N = H_N^2 = \mathbb{R}^n \times L_N^2$. Let $\eta_2 = (\xi, u_2) \in S \cap H_N^2$ be given, and let $\eta_1 = (\xi, u_1) \in H_N^1$ with

$u_1 = (V^1_{\mathbf{A},N})^{-1}(V^2_{\mathbf{A},N}(u_2))$ so that $V^1_{\mathbf{A},N}(u_1) = V^2_{\mathbf{A},N}(u_2)$. Then for any $t \in [0, 1]$,

$$\begin{aligned}\|x^{\eta_1}(t) - x^{\eta_2}(t)\| &= \left\| \int_0^t h(x^{\eta_1}(s), u_1(s)) - h(x^{\eta_2}(s), u_2(s))\, ds \right\| \\ &\le \left\| \int_0^t \tilde{h}(x^{\eta_1}(s)) - \tilde{h}(x^{\eta_2}(s)) + B(u_1(s) - u_2(s))\, ds \right\| \\ &\le \kappa_1 \int_0^t \|x^{\eta_1}(s) - x^{\eta_2}(s)\|\, ds + \left\| \int_0^t B(u_1(s) - u_2(s))\, ds \right\|, \end{aligned} \tag{A.10a}$$

by Assumption 3.1(a). Using the Bellman–Gronwall lemma, we conclude that for any $t \in [0, 1]$

$$\|x^{\eta_1}(t) - x^{\eta_2}(t)\| \le \kappa_1 e^{\kappa_1} \|B\| \left\| \int_0^t (u_1(s) - u_2(s))\, ds \right\|. \tag{A.10b}$$

Now, let $\dot{z}^1(t) \triangleq u_1(t)$, $t \in [0, 1]$, $z^1(0) = \xi$ and $\dot{z}^2(t) \triangleq u_2(t)$, $t \in [0, 1]$, $z^2(0) = \xi$. Let $\bar{z}^1_k$ and $\bar{z}^2_k$, $k \in \mathcal{N}$, be the computed solutions of $z^1(t)$ and $z^2(t)$, respectively, using the RK method under consideration. We note that $\bar{z}^1_k = \bar{z}^2_k$ for all $k \in \mathcal{N}$ since $V^1_{\mathbf{A},N}(u_1) = V^2_{\mathbf{A},N}(u_2)$. Then, since u^1 is an rth-order polynomial, any pth-order RK method, $p \ge r$, integrates $\dot{z}^1(t)$ exactly. Thus $\bar{z}^1_k = z^1(t_k)$ for all $k \in \mathcal{N}$. Also, from (4.3a,b),

$$\bar{z}^2_{k+1} = \bar{z}^2_k + \sum_{i=1}^{s} b_i u_2[\tau_{k,i}] = \bar{z}^2_k + \sum_{j=1}^{r} \int_{t_k + d_{j-1}}^{t_k + d_j} u_2(s)\, ds = z^2(t_{k+1}), \tag{A.10c}$$

since $\tau_{k,j} \in [\tau_k + d_{j-1}, t_k + d_j)$ (by Assumption 4.6) with $u_2(\cdot)$ constant on these intervals and $d_r = \Delta$ by Assumption 4.1′. Since $\bar{z}^1_k = \bar{z}^2_k$, we must have

$$z^1(t_k) - z^2(t_k) = \bar{z}^1_k - \bar{z}^2_k = 0 \quad \forall k \in \mathcal{N}. \tag{A.10d}$$

Hence, we conclude that

$$\left\| \int_0^{t_k} (u_1(s) - u_2(s))\, ds \right\| = \|z^1(t_k) - z^2(t_k)\| = 0. \tag{A.10e}$$

Therefore,

$$\|x^{\eta_2}(t_k) - \bar{x}^{\eta_2}_k\| \le \|x^{\eta_2}(t_k) - x^{\eta_1}(t_k)\| + \|x^{\eta_1}(t_k) - \bar{x}^{\eta_1}_k\| + \|\bar{x}^{\eta_1}_k - \bar{x}^{\eta_2}_k\| \le \kappa'/N^\rho \quad \forall k \in \mathcal{N}, \tag{A.10f}$$

where we have used (A.10b) and (A.10e), the fact that $\|x^{\eta_1}(t_k) - \bar{x}^{\eta_1}_k\| \le \kappa_2/N^\rho$ since (4.18c) holds for $\eta_1 \in S \cap H^1_N$ by the first part of this discussion, and the fact that $\bar{x}^{\eta_1}_k = \bar{x}^{\eta_2}_k$ since $u_1[\tau_{k,i}] = u_2[\tau_{k,i}]$. Thus (4.18c) holds for representation R2 under the stated conditions. □

LEMMA A.4. *Suppose that Assumptions* 3.1, 4.1′, *and* 4.3 *hold for representation* R1 *and that Assumptions* 3.1, 4.1′, *and* 4.6 *hold for representation* R2. *For any* $S \subset \mathbf{B}$ *bounded, there exist* $\kappa < \infty$ *and* $N^* < \infty$ *such that for any* $\eta \in S \cap \mathbf{H}_N$ *and* $N \ge N^*$

$$\|\bar{p}^\nu_k - p^\nu(t_k)\| \le \frac{\kappa}{N}, \qquad k \in \{0, \ldots, N\},\ \nu \in \mathbf{q}, \tag{A.11}$$

where $p^\nu(\cdot)$ *is the solution to the adjoint differential equation* (3.6c) *and* $\{\bar{p}^\nu_k\}_{k=0}^N$ *is the solution to the corresponding adjoint difference equation* (5.5d).

Proof. Proceeding as in the proof of Lemma 4.10(i), if we define $e_{k+1} \triangleq \bar{p}^{\nu}_{k+1} - p^{\nu}(t_{k+1})$ we can show that

$$\|e_k\| \leq L_1\|e_{k+1}\| + L_2\Delta^2, \qquad k \in \mathcal{N}, \tag{A.12}$$

where $L_1, L_2 < \infty$, using (i) the fact that

$$\bar{p}^{\nu}_k = F_x(\bar{x}_k, \bar{u}_k)^T \bar{p}_{k+1} = \bar{p}^{\nu}_{k+1} + \Delta \sum_{i=1}^{s} b_i h_x(\bar{x}_k, u[\tau_{k,i}])^T \bar{p}_{k+1} + O(\Delta^2), \tag{A.13}$$

(ii) Lemma A.1 with $h(x(t_k), u[\tau_{k,i}])$ replaced by $-h_x(x(t_k), u[\tau_{k,i}])^T p^{\nu}(t_{k+1})$, and (iii) the result of Lemma 4.10(i) that $\|x(t_k) - \bar{x}_k\| \leq \kappa\Delta$ for all $k \in \mathcal{N}$. Now, by Assumption 3.1(b) and Lemma 4.10(i), there exists $\kappa_1 < \infty$ such that

$$\|e_N\| = \|\bar{p}^{\nu}_N - p^{\nu}(1)\| \leq \|\zeta_x(\xi, \bar{x}_N)^T - \zeta_x(\xi, x(1))^T\| \leq \kappa_1\|\bar{x}_N - x(1)\| \leq \kappa_2\Delta, \tag{A.14}$$

where $\kappa_2 = \kappa\kappa_1$. Thus, solving (A.15) we conclude that for all $k \in \mathcal{N}$,

$$\|e_k\| \leq (L_1)^N\|e_N\| + L_2'\Delta,$$

which, with (A.17), proves (A.14). □

Acknowledgments. We wish to thank the referees and associate editor for carefully reading this paper, for pointing out an error in one of our proofs, and for their suggestions for improving readability.

Note added in proof. The following conjecture concerns the constraints on $\|\bar{u}_k T_j\|_{\infty}$ used to define $\overline{\mathbf{U}}^1_N$ in (4.15a). Recall from Remark 4.9 that these constraints impose a Lipschitz continuity constraint on the individual polynomial pieces of $u \in \mathbf{U}^1_N = V^{-1}_{\mathbf{A},N}(\overline{\mathbf{U}}^1_N)$ that is needed to ensure accurate RK integration for controls defined by representation R1. Clearly, the addition of these constraints, which do not appear in the original problem **CP**, is a nuisance. Conjecture 5.11 proposes conditions under which these constraints are not needed to define consistent approximating problems $(\mathbf{CP}_N, \theta_N)$ using control representation R1. Assumption 4.6 (needed for control representation R2) is required in place of Assumption 4.3.

CONJECTURE 5.11. *Suppose that the approximating problems* $\mathbf{CP}_N$ *are defined according to* (4.17a) *with* $\mathbf{H}_N \doteq \mathbb{R} \times V^{-1}_{\mathbf{A},N}(\overline{\mathbf{U}}^1_N)$, *where*

$$\overline{\mathbf{U}}^1_N \doteq \{\bar{u} \in \bar{L}^1_N | \bar{u}^j_k \in U \ \forall j \in \mathbf{r}, k \in \mathcal{N}\}. \tag{5.23}$$

Furthermore, assume that Assumptions 3.1, 4.1′, 4.6, *and* 4.11 *and* (5.18) *hold. Let* $\mathbf{N} = \{2^l\}_{l=1}^{\infty}$. *Then the approximating pairs* $(\mathbf{CP}_N, \theta_N)$, $N \in \mathbf{N}$, *are consistent approximations to the pair* $(\mathbf{CP}, \theta)$. □

The basis for this conjecture is the fact that, according to Proposition 5.5, the control samples of the approximating problem solutions do not depend on the control representation and consistency for approximating problems defined with control representation R2 does not require a piecewise Lipschitz continuity contraint on the controls. More details are provided in [28].

REFERENCES

[1] U. ASCHER AND G. BADER, *Stability of collocation at Gaussian points*, SIAM J. Numer. Anal., 23 (1986), pp. 412–422.

[2] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Pitman, London, 1984.

[3] J. P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.

[4] T. E. BAKER AND E. POLAK, *On the optimal control of systems described by evolution equations*, SIAM J. Control Optim., 32 (1994), pp. 224–260.
[5] C. DE BOOR, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.
[6] H. BREZIS, *Analyse Fonctionnelle*, Masson, Paris, 1983.
[7] B. M. BUDAK, E. M. BERKOVICH, AND E. N. SOLOV'EVA, *Difference approximations in optimal control problems*, SIAM J. Control, 7 (1969), pp. 18–31.
[8] J. C. BUTCHER, *The Numerical Analysis of Ordinary Differential Equations*, John Wiley and Sons, Chichester, England, 1987.
[9] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
[10] J. CULLUM, *Discrete approximations to continuous optimal control problems*, SIAM J. Control, 7 (1969), pp. 32–49.
[11] ———, *An explicit procedure for discretizing continuous, optimal control problems*, J. Optim. Theory Appl., 8 (1971), pp. 15–35.
[12] J. E. CUTHRELL AND L. T. BIEGLER, *On the optimization of differential-algebraic process systems*, AIChE J., 33 (1987), pp. 1257–1270.
[13] J. W. DANIEL, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
[14] S. DOLECKI, G. SALINETTI, AND R. J.-B. WETS, *Convergence of functions: Equisemicontinuity*, Trans. Amer. Math. Soc., 276 (1983), pp. 409–429.
[15] W. W. HAGER, *Rates of convergence for discrete approximations to unconstrained control problems*, SIAM J. Numer. Anal., 13 (1976), pp. 449–472.
[16] L. HE AND E. POLAK, *An optimal diagonalization strategy for the solution of a class of optimal design problems*, IEEE Trans. Automat. Control., 35 (1990), pp. 258–267.
[17] J. D. LAMBERT, *Numerical Methods for Ordinary Differential Systems*, John Wiley and Sons, Chichester, England, 1991.
[18] C. LAWRENCE, J. L. ZHOU, AND A. L. TITS, *User's Guide for CFSQP Version* 2.1*: A C Code for Solving (Large Scale) Constrained Nonlinear (Minimax) Optimization Problems*, Institute for Systems Research, Univ. of Maryland, TR-94-16rl.
[19] B. SH. MORDUKHOVICH, *On difference approximations of optimal control systems*, J. Appl. Math. Mech., 42 (1978), pp. 452–461.
[20] ———, *Methods of Approximation in Optimal Control Problems*, Nauka, Moscow, 1988. (In Russian.)
[21] C. P. NEUMAN AND A. SEN, *A suboptimal control algorithm for constrained problems using cubic splines*, Automatica J., 9 (1973), pp. 601–613.
[22] E. POLAK AND L. HE, *Unified steerable phase* I*-phase* II *method of feasible directions for semi-infinite optimization*, J. Optim. Theory Appl., 69 (1991), pp. 83–107.
[23] E. POLAK, *Computational Methods in Optimization*, Academic Press, New York, 1971.
[24] ———, *On the use of consistent approximations in the solution of semi-infinite optimization and optimal control problems*, Math. Programming, 62 (1993), pp. 385–415.
[25] E. POLAK AND L. HE, *Rate-preserving discretization strategies for semi-infinite programming and optimal control*, SIAM J. Control Optim., 30 (1992), pp. 548–572.
[26] G. W. REDDIEN, *Collocation at gauss points as a discretization in optimal control*, SIAM J. Control Optim., 17 (1979), pp. 298–306.
[27] R. SCHERE AND H. TURKE, *Reflected and transposed Runge-Kutta methods*, BIT, 23 (1983), pp. 262–266.
[28] A. SCHWARTZ, *Theory and Implementation of Numerical Methods Based on Runge–Kutta Integration for Solving Optimal Control Problems*, Ph.D. thesis, University of California at Berkeley, 1996.
[29] O. STRYK AND R. BULIRSCH, *Direct and indirect methods for trajectory optimization*, Ann. Oper. Res., 37 (1992), pp. 357–373.
[30] V. VELIOV, *Second-order discrete approximations to linear differential inclusions*, SIAM J. Numer. Anal., 29 (1992), pp. 439–451.
[31] J. VLASSENBROECK AND R. V. DOOREN, *A Chebyshev technique for solving nonlinear optimal control problems*, IEEE Trans. Automat. Control, 33 (1988), pp. 333–340.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1270–1290, July 1996

008

ON L^2 SUFFICIENT CONDITIONS AND THE GRADIENT PROJECTION METHOD FOR OPTIMAL CONTROL PROBLEMS*

J. C. DUNN†

Abstract. L^2-local convergence and active constraint identification theorems are proved for gradient projection iterates in sets of L^∞ functions on [0, 1] with range in a polyhedron $U \subset \mathbb{R}^m$. These theorems extend earlier results for $U = [0, \infty) \subset \mathbb{R}^1$ and are based on an infinite-dimensional variant of the Karush–Kuhn–Tucker second-order sufficient conditions in polyhedral subsets of $\mathbb{R}^n$. The new sufficient conditions and convergence results proved here are directly applicable to continuous-time optimal control problems with smooth nonconvex nonquadratic objective functions and Hamiltonians that are quadratic in the control input vector u. In particular, these theorems apply to nonconvex nonquadratic regulator problems with control-linear state equations and vector-valued inputs $u(t)$ satisfying unqualified affine inequality constraints at almost all t in [0, 1].

Key words. gradient projection, infinite-dimensional programs, affine inequality constraints, nonconvex objectives, L^2-local optimality, second-order sufficient conditions, L^2-local convergence, active constraint identification, optimal control, constrained inputs

AMS subject classifications. 49M07, 49M10, 49K15, 65K10, 90C06

1. Introduction.

Infinite-dimensional extensions of the Karush–Kuhn–Tucker (KKT) second-order sufficient conditions in nonnegative orthants are closely linked to the local convergence properties of gradient projection iterates in the cone of nonnegative L^2 functions [1]–[3]. Similar L^2-local convergence and active constraint identification theorems are based here on second-order sufficient conditions related to those in [4] for infinite-dimensional constrained minimization problems

$$\min_{\Omega} J(u), \tag{1.1a}$$

$$\Omega = \{u \in L^\infty_m[0, 1] : u(t) \overset{\text{a.e.}}{\in} U\}, \tag{1.1b}$$

where U is a nonempty polyhedral convex set in $\mathbb{R}^m$, $L^\infty_m[0, 1]$ is the vector space of essentially bounded measurable functions $u : [0, 1] \to \mathbb{R}^m$, and J is a nonconvex real-valued function with first and second Gâteaux differentials satisfying structure and continuity assumptions that are justifiable in a control-theoretic context, as explained below and in [2].

We assume that for all u in $L^\infty_m[0, 1]$, there exist $\phi(u) \in L^\infty_m[0, 1]$, $S(u) \in L^\infty_{m\times m}[0, 1]$, and $\mathcal{K}(u) \in L^2_{m\times m}([0, 1] \times [0, 1])$, such that

$$d^1J(u\,;\,v) = \langle \nabla J(u), v\rangle_2 = \int_0^1 \langle \nabla J(u)(t), v(t)\rangle dt, \tag{1.2a}$$

$$\nabla J(u)(t) = \phi(u)(t) + S(u)(t)u(t), \tag{1.2b}$$

$$d^2J(u\,;\,v, w) = \langle v, \nabla^2 J(u)w\rangle_2 = \int_0^1 \langle v(t), (\nabla^2 J(u)w)(t)\rangle dt, \tag{1.3a}$$

$$(\nabla^2 J(u)v)(t) = S(u)(t)v(t) + \int_0^1 \mathcal{K}(u)(t, s)v(s)ds, \tag{1.3b}$$

*Received by the editors April 15, 1994; accepted for publication (in revised form) March 9, 1995. This research was supported by NSF research grant DMS-9205240.

†Mathematics Department, Box 8205, North Carolina State University, Raleigh, NC 27695-8205.

for all v and w in $L_m^\infty[0, 1]$, and almost all t in $[0, 1]$ and (s, t) in $[0, 1] \times [0, 1]$, with

$$\lim_{\|v-u\|_2 \to 0} \|\phi(v) - \phi(u)\|_\infty \overset{\text{def}}{=} \lim_{\|v-u\|_2 \to 0} \operatorname{ess\,sup}_{t \in [0,1]} \|\phi(v)(t) - \phi(u)(t)\| = 0, \tag{1.4a}$$

$$\lim_{\|v-u\|_2 \to 0} \|S(v) - S(u)\|_\infty \overset{\text{def}}{=} \lim_{\|v-u\|_2 \to 0} \operatorname{ess\,sup}_{t \in [0,1]} \|S(v)(t) - S(u)(t)\| = 0, \tag{1.4b}$$

(1.4c)

$$\lim_{\|v-u\|_2 \to 0} \|\mathcal{K}(v) - \mathcal{K}(u)\|_2 \overset{\text{def}}{=} \lim_{\|v-u\|_2 \to 0} \left(\int_0^1 \int_0^1 \|\mathcal{K}(v)(t, s) - \mathcal{K}(u)(t, s)\|^2 dt ds \right)^{\frac{1}{2}} = 0.$$

We also assume that the $m \times m$ matrices $S(u)(t)$ and $\mathcal{K}(u)(t, s)$ are symmetric with $\mathcal{K}(u)(t, s) = \mathcal{K}(u)(s, t)$ and that the vector and matrix norms on $\mathbb{R}^m$ and $\mathbb{R}^{m \times m}$ in (1.2)–(1.4) are induced by the standard Euclidean inner product $\langle \xi, \eta \rangle = \sum_{i=1}^m \xi_i \eta_i$ on $\mathbb{R}^m$. Portions of these conditions are invoked in the L^2-local optimality sufficiency analysis of [4] and the convergence analysis of [2]; together, they insure that J is twice continuously Fréchet differentiable on the pre-Hilbert space $\{L_m^\infty[0, 1], \|\cdot\|_2\}$ and permit infinite-dimensional extensions of established proof strategies for sufficient conditions, active constraint identification theorems, and local convergence theorems for gradient projection iterations in polyhedral subsets of $\mathbb{R}^m$.

Section 6 in [2] shows that conditions (1.2)–(1.4) are met by a large class of continuous-time optimal control problems with Bolza objective functions

$$J(u) = P(x(u)(1)) + \int_0^1 f_0(t, x(u)(t), u(t))\, dt, \tag{1.5a}$$

where $x(u) : [0, 1] \to \mathbb{R}^n$ is the unique (absolutely continuous) solution of an initial value problem

$$\frac{dx}{dt}(t) \overset{\text{a.e.}}{=} f(t, x(t), u(t)), \tag{1.5b}$$

$$x(0) = x^0, \tag{1.5c}$$

corresponding to u in $L_m^\infty[0, 1]$. In these equations, x^0 is fixed in $\mathbb{R}^n$, and the functions P, f, and f_0 map $\mathbb{R}^n$ to $\mathbb{R}^1$, $(\mathbb{R}^1 \times \mathbb{R}^n \times \mathbb{R}^m)$ to $\mathbb{R}^n$, and $(\mathbb{R}^1 \times \mathbb{R}^n \times \mathbb{R}^m)$ to $\mathbb{R}^1$, respectively, and satisfy suitable smoothness and growth restrictions. In particular, conditions (1.2)–(1.4) are implied by Assumptions A1–A3 in [2], repeated below (or analogous weaker assumptions of the Carathéodory type [40]).

ASSUMPTION 1.1.

(i) *P is twice continuously differentiable.*

(ii) *For $i = 0, \ldots, n$,*

$$f_i(t, x, u) = q_i(t, x) + \langle r_i(t, x), u \rangle + \frac{1}{2} \langle u, s_i(t, x)u \rangle,$$

where $s_i(t, x) \in \mathbb{R}^{m \times m}$ is symmetric, and the functions $q_i : \mathbb{R}^1 \times \mathbb{R}^n \to \mathbb{R}^1$, $r_i : \mathbb{R}^1 \times \mathbb{R}^n \to \mathbb{R}^m$, and $s_i : \mathbb{R}^1 \times \mathbb{R}^n \to \mathbb{R}^{m \times m}$ and their first and second partial derivatives with respect to x are continuous on $\mathbb{R}^1 \times \mathbb{R}^n$.

(iii) *For $i = 1, \ldots, n$, the first partial derivatives of q_i, r_i, and s_i with respect to x are bounded on $\mathbb{R}^1 \times \mathbb{R}^n$.*

In such cases, conditions (1.2)–(1.4) hold with $\phi(u)(t)$, $S(u)(t)$, and $\mathcal{K}(u)(t, s)$ obtained from partial Hessians of the Hamiltonian

$$H(t, \psi, x, u) = \langle \psi, f(t, x, u) \rangle + f_0(t, x, u),$$

solutions $\psi(u)$ of the costate (adjoint) equations

$$\frac{d\psi}{dt}(t) \overset{\text{a.e.}}{=} -\frac{\partial f}{\partial x}(t, x(t), u(t))^t \psi - \nabla f_0(t, x(t), u(t)),$$
$$\psi(1) = \nabla P(x(u)(1)),$$

and fundamental solution matrices of the equations of variation for (1.5b) [2]. More specifically, Assumption 1.1 implies that conditions (1.2)–(1.4) hold with

$$\begin{aligned} S(u)(t) &= \nabla_{uu} H(t, \psi(u)(t), x(u)(t), u(t)) \\ &= s_0(t, x(u)(t)) + \sum_{i=1}^{n} \psi_i(u)(t) s_i(t, x(u)(t)), \\ \nabla J(u)(t) &= \nabla_u H(t, \psi(u)(t), x(u)(t), u(t)) \\ &= \phi(u)(t) + S(u)(t)u(t), \end{aligned}$$

where

$$\phi(u)(t) = r_0(t, x(u)(t)) + \sum_{i=1}^{n} \psi_i(u)(t) r_i(t, x(u)(t)).$$

In these expressions, the symbols ψ, x, and u refer to vectors in $\mathbb{R}^n$ or $\mathbb{R}^m$ when they appear in $H(t, \psi, x, u)$, $\nabla_u H(t, \psi, x, u)$, and $\nabla_{uu} H(t, \psi, x, u)$ and to vector-valued *functions* on $[0, 1]$ with range in $\mathbb{R}^n$ or $\mathbb{R}^m$ when they appear in $\psi(u)$ and $x(u)$. Note that the structure condition in Assumption 1.1(ii) holds iff H is quadratic in the control input vector $u \in \mathbb{R}^m$; this happens for the important subclass of Bolza problems with nonconvex control-quadratic running costs f_0 and control-linear state equations (where 1.1(ii) holds with $s_i(t, x) = 0$, for $i = 1, \ldots, n$). Note also that Assumption 1.1 can be weakened if conditions (1.2)–(1.4) are needed only in some L^2 neighborhood of a function u_* for which unique solutions $x(u_*)$ and $\psi(u_*)$ of the state and costate equations are already known to exist. Local versions of (1.2)–(1.4) are actually enough for the local optimality conditions in §4 (cf. [4]) and the local convergence analysis in §§5 and 6. We impose (1.2)–(1.4) in their global form to simplify the exposition.

The gradient projection scheme investigated here employs a Goldstein–Levitin–Polyak iteration map [5], [6] and Bertsekas step length rule [7], i.e.,

$$u \to G(u) = g(a(u), u), \tag{1.6a}$$

$$g(a, u) = P_\Omega(u - a\nabla J(u)), \tag{1.6b}$$

with

$$a(u) = \min a \tag{1.6c}$$

subject to

$$a \in \{\overline{a}, \overline{a}\beta, \overline{a}\beta^2, \ldots\} \tag{1.6d}$$

and

$$J(u) - J(g(a, u)) \geq \sigma \langle \nabla J(u), u - g(a, u) \rangle_2, \tag{1.6e}$$

where $\overline{a}$ is fixed in $(0, \infty)$, σ and β are fixed in $(0, 1)$, and P_Ω is the L^2 metric projection map for Ω. The map P_Ω exists for the set Ω in (1.1b), even though $\{\mathsf{L}^\infty_m[0, 1], \|\cdot\|_2\}$ is incomplete. To see this, note that P_U is continuous and nonexpansive and satisfies the Euclidean minimum distance condition

$$\|u - P_U u\| = \min_{\xi \in U} \|u - \xi\|,$$

for all u in $\mathbb{R}^m$. Hence, the rule

$$v(t) \overset{\text{a.e.}}{=} P_U(u(t)) \tag{1.7}$$

defines a mapping from $L_m^\infty[0, 1]$ to Ω that satisfies the L^2 minimum distance condition

$$\|u - v\|_2 = \min_{\xi \in \Omega} \|u - \xi\|_2,$$

for all u in $L_m^\infty[0, 1]$ and v defined by (1.7). Even in a pre-Hilbert space, there is at most one such v for each u; hence $v = P_\Omega u$ and

$$(P_\Omega u)(t) \overset{\text{a.e.}}{=} P_U(u(t)) \tag{1.8}$$

(cf. the projection theorem proof in [8]). We can now see that if $\nabla J(u)$ is in $L_m^\infty[0, 1]$ for all u in $L_m^\infty[0, 1]$, then $g(a, \cdot)$ is defined for all $a > 0$. Finally, if $\nabla J(\cdot)$ is continuous with respect to the L^2 metric, then the Bertsekas step length $a(u)$ also exists (the Hilbert space proof for this assertion in [9] does not depend on completeness and transfers directly to $\{L_m^\infty[0, 1], \|\cdot\|_2\}$).

In view of (1.8), the projection operator P_Ω is readily computed for (1.1b) when U is an orthant, box, simplex, or some such simple polyhedral set, as is often the case for optimal control problems; moreover, the required gradients are also often cheaply computed for the specially structured objective functions J in optimal control problems. Under these circumstances, the efficacy of a gradient projection method turns on the convergence behavior of its approximate or asymptotically exact implementations in finite-dimensional subspaces converging to $L_m^\infty[0, 1]$ from below. Computational experience [23], [29] and some theoretical studies [24]–[29] show that convergence theories for the underlying infinite-dimensional algorithm can play an important role in predicting this behavior, particularly when new *qualitative* differences emerge in the infinite-dimensional limit. In the present context, L^2-local and L^∞-local convergence theorems are qualitatively different in $L_m^\infty[0, 1]$, and this fact has definite *quantitative* computational implications even though all norms are (qualitatively) equivalent on $\mathbb{R}^n$ (see [23] and the opening paragraphs in §5).

The convergence analysis in §§5 and 6 proceeds from certain fundamental geometric properties of the solution sets of finite systems of affine inequalities in $\mathbb{R}^m$, i.e., the polyhedral convex sets

$$U = \{u \in \mathbb{R}^m : \langle a_i, u\rangle + b_i \leq 0;\ i = 1, \ldots, k\}. \tag{1.9}$$

These properties are derived from the algebraic representation (1.9); however, (1.9) itself is not needed and does not appear explicitly in our theorems or proofs. As in [26], [30], [31], [4], key proof ideas are more readily grasped (and discovered) in this geometric framework, and constraint qualifications are not required. In §2 we give the pertinent properties of polyhedral convex sets $U \subset \mathbb{R}^m$, and in §3 we sketch the proofs of second-order optimality conditions and local convergence theorems for gradient projection iterates in polyhedra. The geometric sufficiency proof strategy outlined in §3 is then adapted in §4 to the specially structured infinite-dimensional programs (1.1) satisfying (1.2)–(1.4). The resulting second-order sufficient conditions for L^2-local optimality in Theorem 4.1 consist of an L^2-coercivity restriction on $\nabla^2 J$ formally analogous to the standard coercivity condition in polyhedra, a norm-independent pointwise counterpart of strict complementarity in Cartesian products of polyhedra, and a third requirement that amounts to a stronger version of Pontryagin's necessary condition in the ordinary differential equation (ODE) control problem context.

Theorem 4.1 adds significantly to the L^2-local sufficiency analysis in [4]. In [4], a two-norm L^∞-local optimality growth estimate of the form

$$\exists \delta_\infty, c_\infty > 0\ \forall u \in \Omega\ \left(\|u - u_*\|_\infty < \delta_\infty \Rightarrow J(u) - J(u_*) \geq c_\infty \|u - u_*\|_2^2\right) \tag{1.10}$$

is deduced from the L^2-coercivity condition, a weakened L^∞ strict complementarity condition, and additional technical constraints imposed near point τ in [0, 1], where $u_*(t)$ passes from one open facet in the polyhedron U to another. The growth condition (1.10) then appears along with the strengthened Pontryagin condition as a *hypothesis* in the L^2-local optimality sufficiency theorem in [4]. In contrast, the proof technique employed here permits weaker regularity hypotheses and strict complementarity conditions by fully exploiting the Pontryagin condition. For additional ground-breaking L^∞-local optimality sufficiency theorems and related applications, see [10]–[22].

The convergence theorem in §5 and the active constraint identification theorem in §6 are based directly on a corollary of Theorem 4.1 that applies when the matrices $S(u_*)(t)$ are essentially uniformly positive-definite on [0, 1] at a minimizer u_*, i.e., when there is a positive number c_P such that

$$(1.11) \qquad \left(\forall \xi \in \mathbb{R}^m \quad \langle \xi, S(u_*)(t)\xi \rangle \geq c_P \|\xi\|^2\right) \qquad \text{a.e. in } [0, 1].$$

Condition (1.11) is a natural requirement for nonconvex regulator problems with control-quadratic running costs and control-linear state equations. For such problems,

$$S(u)(t) = \nabla^2_{uu} H(t, \psi(u)(t), x(u)(t), u(t)) = s_0(t, x(u)(t)),$$

and s_0 is often *constant* or dependent on t only. An example in [23] shows that the weaker Pontryagin-like hypothesis imposed in Theorem 4.1 will not support an L^2-local convergence theory for gradient projection methods, hence (1.11) assumes a special importance for these algorithms. In the control problem context, (1.11) is a strong variant of Legendre's necessary condition.

2. Polyhedra. Each nonempty polyhedron U in (1.9) is a union of d polyhedral faces $\mathcal{F}_1, \ldots, \mathcal{F}_d$. Let $aff\ \mathcal{F}_i$ denote the *affine hull* of $\mathcal{F}_i$, i.e., the smallest linear variety (translated subspace) containing $\mathcal{F}_i$. Let $ri\ \mathcal{F}_i$ and $rb\ \mathcal{F}_i$ denote the *relative interior* and *relative boundary* of $\mathcal{F}_i$, i.e., the interior and boundary of $\mathcal{F}_i$ relative to $aff\ \mathcal{F}_i$. It is shown in [32] that the sets $ri\ \mathcal{F}_i$ make a partition for U, i.e.,

$$(2.12a) \qquad U = \bigcup_{i=1}^{d} ri\ \mathcal{F}_i,$$

$$(2.12b) \qquad ri\ \mathcal{F}_i \cap ri\ \mathcal{F}_j = \emptyset, \qquad i \neq j.$$

In [17], the sets $ri\ \mathcal{F}_i$ are called *open facets*.

At each u in U, let $\mathcal{F}(u)$ be the unique face in $\{\mathcal{F}_1, \ldots, \mathcal{F}_d\}$ containing u in its relative interior, and let $\mathcal{N}_U(u)$ be the corresponding *normal cone*:

$$\mathcal{N}_U(u) = \{w \in \mathbb{R}^m : \forall v \in U \quad \langle w, v - u \rangle \leq 0\},$$

with

$$\mathsf{N}(u) = span\ \mathcal{N}_U(u), \qquad \mathsf{T}(u) = \mathsf{N}(u)^\perp.$$

In polyhedra U, the set-valued maps $\mathcal{N}_U(\cdot)$, $\mathsf{N}(\cdot)$, and $\mathsf{T}(\cdot)$ are *constant* on each open facet $ri\ \mathcal{F}_i$ [30], [31], i.e., there exist cones $\mathcal{N}_1, \ldots, \mathcal{N}_d$ and subspace pairs $(\mathsf{N}_1, \mathsf{T}_1), \ldots, (\mathsf{N}_d, \mathsf{T}_d)$ such that

$$(2.13a) \qquad \mathcal{N}_U(u) = \mathcal{N}_i, \qquad \mathsf{N}(u) = \mathsf{N}_i, \qquad \mathsf{T}(u) = \mathsf{T}_i,$$

for all $i = 1, \ldots, d$ and u in $ri\ \mathcal{F}_i$; moreover,

$$aff\ \mathcal{F}_i = u + \mathsf{T}_i \tag{2.13b}$$

and

$$\mathcal{F}_i = (u + \mathsf{T}_i) \cap U. \tag{2.13c}$$

The partition (2.12) and the Euclidean projection map P_U induce a corresponding partition of $\mathbb{R}^m$ that plays a central role in geometric characterizations of the convergence behavior of gradient projection iterates in polyhedra [30], [31] and in the sets (1.1b) considered here. For nonempty polyhedra U, the projection theorem asserts that P_U is defined on $\mathbb{R}^m$ and that

$$u = P_U(x) \quad \Leftrightarrow \quad x - u \in \mathcal{N}_U(u),$$

for all x in $\mathbb{R}^m$ and u in U. Hence (2.12) implies that

$$P_U^{-1}[ri\ \mathcal{F}_i] = ri\ \mathcal{F}_i + \mathcal{N}_i$$

for $i = 1, \ldots, d$, and therefore

$$\mathbb{R}^m = \bigcup_{i=1}^{d} (ri\ \mathcal{F}_i + \mathcal{N}_i),$$

with

$$(ri\ \mathcal{F}_i + \mathcal{N}_i) \cap (ri\ \mathcal{F}_j + \mathcal{N}_j) = \emptyset, \qquad i \neq j.$$

Furthermore, each of the sets $P_U^{-1}[ri\ \mathcal{F}_i]$ has a nonempty interior in $\mathbb{R}^m$ obtained by adding the relative interiors of $\mathcal{F}_i$ and $\mathcal{N}_i$. This essential fact is noted in [33] and restated below in our first lemma; a closely related result is proved in [31].

For $x \in \mathbb{R}^m$, $Y \subset \mathbb{R}^m$, and $r \geq 0$, let $dist\ (x, Y) = \inf_{y \in Y} \|y - x\|$ and $\overline{B(x; r)} = \{y \in \mathbb{R}^m : \|y - x\| \leq r\}$. For all y in Y, it can be seen that

$$dist\ (y, rb\ Y) = \sup\{r \geq 0 : \overline{B(y; r)} \cap aff\ Y \subset Y\},$$

with $dist\ (y, rb\ Y) > 0$ for all y in $ri\ Y$.

LEMMA 2.1. *Let U be a polyhedral convex set in $\mathbb{R}^m$ with faces $\mathcal{F}_1, \ldots, \mathcal{F}_d$. Then for $i = 1, \ldots, d$,*

$$\operatorname{int} P_U^{-1}[ri\ \mathcal{F}_i] = ri\ \mathcal{F}_i + ri\ \mathcal{N}_i.$$

Moreover, suppose that $u \in ri\ \mathcal{F}_i$, $v \in ri\ \mathcal{N}_i$, $w = u + v$, and $\Delta w \in \mathbb{R}^m$, with

$$\|\Delta w\| < s = \min\{dist\ (u, rb\ \mathcal{F}_i), dist\ (v, rb\ \mathcal{N}_i)\}.$$

Then $P_U(w + \Delta w) \in ri\ \mathcal{F}_i$ and $dist\ (P_U(w + \Delta w), rb\ \mathcal{F}_i) \geq s - \|\Delta w\|$.

Proof. Suppose that $\Delta w \in \mathbb{R}^m$ and $\|\Delta w\| < s$. Write $w + \Delta w = x + y$, with $x = u + P_{\mathsf{T}_i} \Delta w$ and $y = v + P_{\mathsf{N}_i} \Delta w$. Note that $\|P_{\mathsf{T}_i} \Delta w\| < \|\Delta w\|$ and $\|P_{\mathsf{N}_i} \Delta w\| < \|\Delta w\|$. Hence $x \in ri\ \mathcal{F}_i$, $y \in ri\ \mathcal{N}_i$, and therefore $x = P_U(w + \Delta w)$. This proves that $dist\ (P_U(w + \Delta w), rb\ \mathcal{F}_i) \geq s - \|\Delta w\|$ and that $ri\ \mathcal{F}_i + ri\ \mathcal{N}_i \subset \operatorname{int} P_U^{-1}[ri\ \mathcal{F}_i]$.

Conversely, suppose that $w \in \operatorname{int} P_U^{-1}[ri\ \mathcal{F}_i]$. Then there exist vectors $u \in ri\ \mathcal{F}_i$ and $v \in \mathcal{N}_i$, and a real number $r > 0$, such that $w = u + v$ and $(w + \Delta v) \in P_U^{-1}[ri\ \mathcal{F}_i]$ for all

$\Delta v \in \mathsf{N}_i$, with $\|\Delta v\| < r$. Now note that if $\Delta v \in \mathsf{N}_i$ and $P_U(u + v + \Delta v) = u' \in ri\ \mathcal{F}_i$, then $(u + v + \Delta v - u') \in \mathcal{N}_U(u') = \mathcal{N}_i \subset \mathsf{N}_i$, with $(u - u') \in \mathsf{T}_i$ and $(v + \Delta v) \in \mathsf{N}_i$. But in this case, $(u - u') = 0$ and $(v + \Delta v) \in \mathcal{N}_i$. This proves that $v \in ri\ \mathcal{N}_i$ and hence $ri\ \mathcal{F}_i + ri\ \mathcal{N}_i \supset \text{int}\ P_U^{-1}[ri\ \mathcal{F}_i]$. □

Note 2.1. Theorem 2.8 in [31] asserts that int $(\mathcal{F}_i + \mathcal{N}_i) = ri\ \mathcal{F}_i + ri\ \mathcal{N}_i$. The first part of Lemma 2.1 is a corollary of this result; however, the sets $P_U^{-1}[ri\ \mathcal{F}_i]$ have a more immediate intuitive significance for our analysis, and the direct proof for Lemma 2.1 given here is shorter than the proof in [31].

3. Optimality conditions and convergence theorems in polyhedra. In polyhedral sets $U \subset \mathbb{R}^m$, the essential contents of the KKT optimality conditions and the local convergence theorems for gradient projection iterates are best expressed in a representation-free, multiplier-free geometric language. We pursue this further below to motivate proof constructions for the set Ω in (1.1), which is in some sense an infinite-dimensional limit of k-fold polyhedral Cartesian products $\Omega_k = U \times \cdots \times U$.

If $J : \mathbb{R}^m \to \mathbb{R}^1$ is twice continuously differentiable and if u_* is a local minimizer for J in the polyhedron $U \subset \mathbb{R}^m$, then

$$-\nabla J(u_*) \in \mathcal{N}_U(u_*), \tag{3.14a}$$

$$\forall v \in \mathsf{T}(u_*) \quad \langle v, \nabla^2 J(u_*)v\rangle \geq 0. \tag{3.14b}$$

Conversely, if

$$-\nabla J(u_*) \in ri\ \mathcal{N}_U(u_*) \tag{3.15a}$$

and

$$\exists c_{\mathsf{T}} > 0\ \forall v \in \mathsf{T}(u_*) \quad \langle v, \nabla^2 J(u_*)v\rangle \geq c_{\mathsf{T}}\|v\|^2, \tag{3.15b}$$

then u_* is a local minimizer for J in U, and

$$\exists c > 0\ \exists \delta > 0\ \forall u \in U \cap B(u_*; \delta) \quad J(u) - J(u_*) \geq c\|u - u_*\|^2. \tag{3.16}$$

Local minimizers that satisfy the sufficient conditions (3.15) are said to be *nonsingular*.

If u_* lies in the interior of U in $\mathbb{R}^m$, then $ri\ \mathcal{N}_U(u_*) = rb\ \mathcal{N}_U(u_*) = \mathcal{N}_U(u_*) = \{0\}$, and (3.14b) and (3.15b) reduce to the standard second-order necessary condition and sufficient condition for unconstrained local minimizers. If u_* is a frontier point of U, then $ri\ \mathcal{N}_U(u_*)$ is a nonempty proper subset of $\mathcal{N}_U(u_*)$ and (3.15a) is stronger than (3.14a); however, $rb\ \mathcal{N}_U(u_*)$ is nowhere dense in the subspace $\mathsf{N}(u_*)$ and is negligibly small compared with $ri\ \mathcal{N}_U(u_*)$ in a measure-theoretic sense. Similarly, in $\mathbb{R}^m$, the coercivity condition (3.15b) is *equivalent* to the positive-definiteness condition

$$\forall v \in \mathsf{T}(u_*) \quad v \neq 0 \Rightarrow \langle v, \nabla^2 J(u_*)v\rangle > 0, \tag{3.17}$$

and positive-definiteness is generic in the class of linear operators that are positive-semidefinite on a subspace $T \subset \mathbb{R}^m$. In this sense, the gap between the standard necessary conditions (3.14) and sufficient conditions (3.15) is not large in a polyhedron $U \subset \mathbb{R}^m$. We note that the geometric nondegeneracy condition (3.15a) proposed in [30] is equivalent to algebraic strict complementarity in the KKT sufficient conditions [31].

The first-order necessary condition (3.14a) merely says that for all u in U, the corresponding directional derivatives $\langle \nabla J(u_*), u - u_*\rangle$ can't be negative if u_* is a local minimizer; this

is clearly true in *any* convex set U. Since $\mathsf{T}(u_*)$ is the orthogonal complement of $\mathcal{N}_U(u_*)$, it follows at once from (3.14a) that $\langle \nabla J(u_*), v\rangle = 0$ for all v in $\mathsf{T}(u_*)$. This result and the second-order necessary condition (3.14b) are also immediately seen by noting that u_* lies in one of the open facets $ri\ \mathcal{F}_i$, and hence u_* is a local minimizer for the restriction of J to the translated subspace $aff\ \mathcal{F}_i = u_* + \mathsf{T}_i = u_* + \mathsf{T}(u_*)$. The sufficiency of conditions (3.15) for local optimality likewise rests on two simple observations. By continuity, the second-order condition (3.15b) implies a similar coercivity condition in sufficiently small *cones* $C_\epsilon(u_*) = \{v \in \mathbb{R}^m : \|P_{\mathsf{N}(u_*)}\, v\| \le \epsilon\|v\|\}$ containing the subspace $\mathsf{T}(u_*)$. Hence, by Taylor's formula and (3.14a), the increments $J(u) - J(u_*)$ must grow like $\|u-u_*\|^2$ in the intersection of U with the translated cone $u_* + C_\epsilon(u_*)$ and sufficiently small balls $B(u_*, \delta) = \{u \in \mathbb{R}^m : \|u - u_*\| < \delta\}$. On the other hand, if $c_{\mathsf{N}} = dist\ (-\nabla J(u_*), rb\ \mathcal{N}_U(u_*))$, then it follows easily from the first-order condition (3.15a) that $\langle \nabla J(u_*), u - u_*\rangle \ge c_{\mathsf{N}}\ \|P_{\mathsf{N}(u_*)}\ (u - u_*)\|$ for all u in U. In particular, for all u_* in the intersection of U, $B(u_*; \delta)$, and $u_* + C_\epsilon(u_*)^c$, we then have $\langle \nabla J(u_*), u-u_*\rangle \ge \epsilon c_{\mathsf{N}}\|u-u_*\| \ge \epsilon c_{\mathsf{N}}\delta^{-1}\|u-u_*\|^2$. Since the second-order term in Taylor's formula is of order $O(\|u - u_*\|^2)$, it now follows that $J(u) - J(u_*)$ grows like $\|u - u_*\|^2$ in the intersection of U with $u_* + C_\epsilon(u_*)^c$ and $B(u_*, \delta)$, for sufficiently small $\delta > 0$. A similar argument is made in the proof of Theorem 4.1.

For polyhedral sets $U \subset \mathbb{R}^m$, the counterpart of the gradient projection iteration map in §1 is

$$u \to G(u) = g(a(u), u) = P_U(u - a(u)\nabla J(u)), \tag{3.18a}$$

with

$$a(u) = \min\ a \tag{3.18b}$$

subject to

$$a \in \{\overline{a}, \overline{a}\beta, \overline{a}\beta^2, \ldots\}, \tag{3.18c}$$

and

$$J(u) - J\ (g(a, u)) \ge \sigma\ \langle \nabla J(u), u - g(a, u)\rangle. \tag{3.18d}$$

The fundamental local convergence theorem for (3.18) asserts that an iterate sequence $\{u_i\}$ passing sufficiently close to a nonsingular local minimizer u_* will eventually enter and remain in the open facet $ri\ \mathcal{F}(u_*)$ and become an Armijo steepest descent sequence that converges to u_* r-linearly in the translated subspace $u_* + \mathsf{T}(u_*)$. The key elements in the proof for this theorem are already present in the analyses of [7], [36], [37], and [35] for iterates of (3.18) in the images of orthants and boxes under invertible linear transformations. Reference [30] provides a self-contained fully geometric representation-free expression of this proof strategy for arbitrary polyhedral sets and more generally for closed convex sets with embedded open facets. The salient points in the geometric proof for nonsingular local minimizers u_* in polyhedral sets U are outlined below, in the language of §2.

Proof Outline 3.1. (i) Iterate sequences $\{u_i\}$ generated by (3.18) remain in any specified neighborhood of u_* provided u_0 lies in a sufficiently small subneighborhood of u_* in U, i.e., u_* is a *stable* fixed point of the map (3.18).

(ii) The step lengths $a(u)$ are bounded away from 0 in U near u_*.

(iii) Since $\mathcal{N}_U(u_*)$ is a cone, the strict complementarity condition (3.15a), Lemma 2.1, and (ii) imply that $u - a(u)\nabla J(u) \in P_U^{-1}[ri\ \mathcal{F}(u_*)]$ and hence $G(u) \in ri\ \mathcal{F}(u_*)$ in U near u_*.

(iv) The coercivity condition (3.15b) implies that the restriction of J to $u_* + \mathsf{T}(u_*)$ is locally convex and grows like $\|u - u_*\|^2$ near u_*.

(v) In $ri\ \mathcal{F}(u_*)$ near u_*, $G(u) = u - a(u)P_{\mathsf{T}(u_*)}\nabla J(u)$ and G reduces to an unconstrained Armijo steepest descent iteration map for the restriction of J to $u_* + \mathsf{T}(u_*) \supset ri\ \mathcal{F}(u_*)$.

(vi) The Armijo steepest descent iteration in $u_* + \mathsf{T}(u_*)$ is locally linearly convergent to u_*.

Note 3.2. References [34] and [31] provide an active facet identification proof for C^1 cost functions J and arbitrary stationary points u_* satisfying the nondegeneracy condition (3.15a); however, this proof *assumes* that the iterate sequence $\{u_i\}$ converges to u_*. In contrast, the active facet identification step in the foregoing proof outline is used to *prove* convergence of iterate sequences confined to sufficiently small neighborhoods of u_*. Other local convergence proof strategies which do not require the active facet identification step are developed in [30] and [33] for C^1 cost functions and proper local minimizers that are isolated stationary points. Extensions of the basic local convergence theorem have been proved for gradient projection iterations in solution sets of finite systems of smooth inequality constraints [38] and in unions of pairwise disjoint class C^p-identifiable surfaces analogous to open facets [39].

4. L^2-Local optimality sufficient conditions. The normal cone $\mathcal{N}_U(u)$, the associated subspaces $\mathsf{N}(u)$ and $\mathsf{T}(u)$, and the optimality conditions in §3 all have formal counterparts for the set Ω in (1.1b). The formal necessary conditions are actually valid even though Ω is not a polyhedron [4]; however, infinite-dimensional formal counterparts of the KKT sufficient conditions need not imply local optimality in any norm [10] . These points are developed further below.

For u in $\mathsf{L}^\infty_m[0, 1]$ and t in $[0, 1]$ such that $u(t) \in U$, let

$$\mathsf{N}(u)(t) = span\ \mathcal{N}_U(u(t)), \qquad \mathsf{T}(u)(t) = \mathsf{N}(u)(t)^\perp$$

and put

$$\mathcal{N}_\Omega(u) = \{w \in \mathsf{L}^\infty_m[0, 1] : w(t) \overset{\text{a.e}}{\in} \mathcal{N}_U(u(t))\},$$

$$\mathsf{N}(u) = \{w \in \mathsf{L}^\infty_m[0, 1] : w(t) \overset{\text{a.e}}{\in} \mathsf{N}(u)(t)\},$$

$$\mathsf{T}(u) = \{w \in \mathsf{L}^\infty_m[0, 1] : w(t) \overset{\text{a.e}}{\in} \mathsf{T}(u)(t)\}.$$

In addition, let

$$\alpha_i(u) = u^{-1}[ri\ \mathcal{F}_i], \quad i = 1, \ldots, d,$$

where $\mathcal{F}_i$ is the ith polyhedral face in U. Note that for u in Ω, the corresponding sets $\alpha_i(u) \subset [0, 1]$ are measurable and pairwise disjoint, and their union differs from $[0, 1]$ by a set of measure zero; moreover, the set-valued maps $\mathcal{N}_U(u(\cdot))$, $\mathsf{N}(u)(\cdot)$, and $\mathsf{T}(u)(\cdot)$ are *constant* on each set $\alpha_i(u)$ (§2). It is therefore not difficult to prove that in the pre-Hilbert space $\{\mathsf{L}^\infty_m[0, 1], \|\cdot\|_2\}$, the set $\mathcal{N}_\Omega(u)$ is the cone of outer normals to Ω at u, the subspace $\mathsf{N}(u)$ is the closure of the span of $\mathcal{N}_\Omega(u)$, and for each v in $\mathsf{L}^\infty_m[0, 1]$, the rule

$$v_{\mathsf{N}}(t) \overset{\text{a.e.}}{=} P_{\mathsf{N}(u)(t)}v(t), \qquad v_{\mathsf{T}}(t) \overset{\text{a.e.}}{=} P_{\mathsf{T}(u)(t)}v(t) \tag{4.19}$$

defines essentially bounded measurable functions $v_{\mathsf{N}} \in \mathsf{N}(u)$ and $v_{\mathsf{T}} \in \mathsf{T}(u)$, with $v = v_{\mathsf{N}} + v_{\mathsf{T}}$ and $\langle v_{\mathsf{N}}, v_{\mathsf{T}}\rangle_2 = 0$ [4]. Thus, $\mathsf{N}(u)$ and $\mathsf{T}(u)$ are complementary orthogonal closed subspaces in $\{\mathsf{L}^\infty_m[0, 1], \|\cdot\|_2\}$, and v_{N} and v_{T} are the L^2 metric projections of v into $\mathsf{N}(u)$ and $\mathsf{T}(u)$. In fact, $\mathsf{N}(u)$ and $\mathsf{T}(u)$ are clearly *pointwise orthogonal* in the sense that

$$\forall v \in \mathsf{N}(u)\ \forall w \in \mathsf{T}(u) \quad \langle v(t), w(t)\rangle \overset{\text{a.e.}}{=} 0. \tag{4.20}$$

Formal counterparts of the KKT necessary conditions in polyhedra may now be stated as follows. Suppose that $J : \mathsf{L}_m^\infty[0, 1] \to \mathbb{R}^1$ has first and second Gâteaux differential representations (1.2a) and (1.3a), with gradient vectors $\nabla J(u)$ in $\mathsf{L}_m^\infty[0, 1]$ and Hessian operators $\nabla^2 J(u) : \mathsf{L}_m^\infty[0, 1] \to \mathsf{L}_m^\infty[0, 1]$. If u_* is a local minimizer for the restriction of J to any line in Ω, then

$$-\nabla J(u_*) \in \mathcal{N}_\Omega(u_*), \tag{4.21a}$$

$$\forall v \in \mathsf{T}(u_*) \quad \langle v, \nabla^2 J(u_*)v\rangle_2 \geq 0. \tag{4.21b}$$

The inclusion (4.21a) is just a special case of the well-known elementary first-order necessary condition for optimality in convex feasible sets; however, the second-order condition (4.21b) is a nontrivial assertion since Ω is not a polyhedron and the subspace $T(u_*)$ is not contained in the cone of feasible directions to Ω at u_*. When the map $w \to \langle w, \nabla^2 J(u_*)w\rangle_2$ is continuous with respect to the norm $\|\cdot\|_2$, condition (4.21b) can be proved by demonstrating that $T(u_*)$ is a subspace in the L^2 *closure* of the cone of feasible directions at u_* [4]. It is also shown in [4] that (4.21b) is generally *false* if the mapping $w \to \langle w, \nabla^2 J(u_*)w\rangle_2$ is merely continuous in the L^∞ norm.

Conditions (4.21) apply if u_* is any local minimizer on lines in Ω; however, more can be said when u_* is an L^2-local minimizer in Ω, and the structure/continuity conditions (1.2)–(1.4) are met. Under these circumstances, it is shown in [4] that for almost all t in [0, 1],

$$\forall \xi \in U \quad H(u_*\,;\, \xi, t) - H(u_*\,;\, u_*(t), t) \geq 0, \tag{4.22a}$$

where

$$H(u_*\,;\, \xi, t) = \langle \nabla J(u_*)(t), \xi - u_*(t)\rangle + \frac{1}{2}\langle \xi - u_*(t),\, S(u_*)(t)\,(\xi - u_*(t))\rangle. \tag{4.22b}$$

In the control problem context, condition (4.22) amounts to the Pontryagin minimum principle. We note that (4.22) has no analogue in $\mathbb{R}^n$, where all norms are equivalent.

The KKT strict complementarity and coercivity conditions in (3.15) have obvious formal counterparts as well in the pre-Hilbert space $\{\mathsf{L}_m^\infty[0, 1], \|\cdot\|_2\}$, namely,

$$-\nabla J(u_*) \in ri\ \mathcal{N}_\Omega(u_*) \tag{4.23}$$

and

$$\exists c_\mathsf{T} > 0\ \forall v \in \mathsf{T}(u_*) \quad \langle v, \nabla^2 J(u_*)v\rangle_2 \geq c_\mathsf{T}\, \|v\|_2^2, \tag{4.24}$$

where *ri* now means interior relative to the subspace $\mathsf{N}(u_*)$ and the L^2 norm. Unfortunately, these conditions are *vacuous* in Ω since the L^2 relative interior of the normal cone $\mathcal{N}(u_*)$ is typically empty. Nonvacuous sufficient conditions for L^∞-local optimality are obtained in [4] by interpreting *ri* as the interior of $\mathcal{N}(u_*)$ relative to $\mathsf{N}(u_*)$ and the L^∞ norm (4.23); however, this L^∞ strict complementarity condition is quite stringent, since it amounts to requiring that the distance from $-\nabla J(u_*)(t)$ to the relative boundary of $\mathcal{N}_U(u_*(t))$ in $\mathsf{N}_U(u_*(t))$ is essentially bounded away from zero on [0, 1]. The latter condition typically can't hold in commonly encountered situations where $\nabla J(u_*)(\cdot)$ and $u_*(\cdot)$ are continuous at points in the set

$$\gamma(u_*) = \{\tau \in [0, 1] : \exists j \in \{1, \ldots, d\}, \quad \tau \in \partial\alpha_j(u_*)\}.$$

(Here $\partial\alpha_j(u_*)$ denotes the boundary of $\alpha_j(u_*)$ *relative to* [0, 1].) Improved sufficient conditions for L^∞-local optimality are based in [4] on the much weaker strict complementarity condition

$$\mu\big[\gamma(u_*)\big] = 0, \tag{4.25a}$$

$$\text{(4.25b)} \qquad \forall \beta \subset \bigcup_{i=1}^{d} \operatorname{int} \alpha_i(u_*) \quad \left(\beta \text{ compact} \Rightarrow \exists c_\beta > 0 \ \Delta(u_*)(t) \geq c_\beta \text{ a.e. in } \beta\right),$$

where

$$\text{(4.25c)} \qquad \Delta(u_*)(t) = dist\ (-\nabla J(u_*)(t), rb\, \mathcal{N}_U(u_*(t)))$$

and

$$\text{(4.25d)} \qquad rb\, \mathcal{N}_U(u_*(t)) = \mathcal{N}_U(u_*(t)) \setminus ri\, \mathcal{N}_U(u_*(t)).$$

We note that vanishingly small values of $\Delta(u_*)(t)$ are likely to occur only near points in the set $\gamma(u_*)$, where $-\nabla J(u_*)(t)$ crosses a common boundary of contiguous normal cones in U. Condition (4.25b) merely requires that $\Delta(u_*)(t)$ is essentially bounded away from zero in the *exterior* of any open neighborhood of $\gamma(u_*)$ in $[0, 1]$.

It is shown in [4] that (4.24) and (4.25) imply the L^∞-local optimality growth estimate (1.10) when (1.2)–(1.3) hold with a weaker two-norm version of (1.4) and when $\nabla J(u_*)(\cdot)$, $S(u_*)(\cdot)$, and the set-valued map $\mathsf{T}(u_*)(\cdot)$ meet certain additional local restrictions near points in the set $\gamma(u_*)$; the latter restrictions compensate for the weakened L^∞ strict complementarity condition (4.25) by insuring that (4.24) lifts from $\mathsf{T}(u_*)$ to a larger subspace $\hat{\mathsf{T}}$. It is also shown in [4] that the L^∞-local optimality growth condition (1.10) implies L^2-local optimality when (1.2)–(1.4) hold and u_* satisfies the following strengthened variant of the Pontryagin condition (4.22) for some $c_P > 0$ and almost all t in $[0, 1]$:

$$\text{(4.26)} \qquad \forall \xi \in U \quad H(u_*\,;\ \xi, t) - H(u_*\,;\ u_*(t), t) \geq \frac{1}{2}\, c_P \|\xi - u_*(t)\|^2.$$

We now make full use of (4.26) to prove sharper L^2-local optimality sufficient conditions directly. In the following theorem, the weakened L^∞ strict complementarity condition (4.25) is replaced by the still weaker norm-independent pointwise strict complementarity condition

$$\text{(4.27)} \qquad -\nabla J(u_*)(t) \in ri\, \mathcal{N}_U(u_*(t)) \quad \text{a.e. in } [0, 1].$$

We note that (4.22) (and a fortiori, (4.26)) need not hold at an L^∞-local minimizer. We also note that (4.27) is the natural norm-independent formal counterpart of the geometric KKT strict complementarity condition in finite-dimensional k-fold Cartesian products $U \times \cdots \times U$ and that (4.27) *cannot* replace (4.25) in the L^∞-local optimality sufficiency analysis of [4]. Condition (4.27) works in the present L^2 analysis only because of the strengthened Pontryagin condition (4.26).

THEOREM 4.1. *Suppose that J satisfies the structure/continuity requirements* (1.2a), (1.3), (1.4b), *and* (1.4c) *with $\nabla J(u) \in \mathsf{L}^\infty_m[0, 1]$. In addition, assume that the L^2-coercivity condition* (4.24), *the strengthened Pontryagin condition* (4.26), *and the pointwise strict complementarity condition* (4.27) *hold at the point u_* in the set Ω in* (1.1b). *Then u_* is an L^2-local minimizer of J in Ω; more specifically, for each c_2 in the interval $0 < c_2 < \min\{c_\mathsf{T}, c_P\}$, there is a corresponding $\delta_2 > 0$ such that*

$$\text{(4.28)} \qquad J(u) - J(u_*) \geq \frac{1}{2} c_2 \|u - u_*\|_2^2,$$

for all u in the intersection of Ω and the ball

$$B_2(u_*; \delta_2) = \{u \in \mathsf{L}^\infty_m[0, 1] : \|u - u_*\|_2 < \delta_2\}.$$

Proof. For each u in Ω, let φ be a corresponding measurable set in $[0, 1]$, and put $v_u(t) = u_*(t)$ for t in φ, $v_u(t) = u(t)$ for t in $\varphi^c = [0, 1] \setminus \varphi$, and $w_u = v_u - u_*$. Now write $(\varphi^c \times \varphi^c)^c = ([0, 1] \times [0, 1]) \setminus (\varphi^c \times \varphi^c)$, and

$$J(u) - J(u_*) = \langle \nabla J(u_*), w_u \rangle_2 + \frac{1}{2} \langle w_u, \nabla^2 J(u_*) w_u \rangle_2$$

$$+ \int_\varphi \left[\langle \nabla J(u_*)(t), u(t) - u_*(t) \rangle + \frac{1}{2} \langle u(t) - u_*(t), S(u_*)(t)\,(u(t) - u_*(t)) \rangle \right] dt$$

$$+ \frac{1}{2} \int \int_{(\varphi^c \times \varphi^c)^c} \langle u(t) - u_*(t), \mathcal{K}(u_*)(t, s)\,(u(t) - u_*(t)) \rangle dt ds + r(u_* ;\ u - u_*).$$

The growth condition (4.28) is obtained by estimating each term in the right side of the foregoing expression with appropriate choices for φ.

Recall that $\mu \left[[0, 1] \setminus \bigcup_{i=1}^d \alpha_i(u_*) \right] = 0$ and $\mathcal{N}_U(u_*(t)) = \mathcal{N}_i$ for $i = 1, \ldots, d$ and $t \in \alpha_i(u_*) = u^{-1}[ri\ \mathcal{F}_i]$. Note that the function $dist\ (\cdot\ , rb\ \mathcal{N}_i) : \mathbb{R}^m \to \mathbb{R}^1$ is continuous, and

$$\forall t \in \alpha_i(u_*) \quad \Delta(u_*)(t) = dist\ (-\nabla J(u_*)(t), rb\ \mathcal{N}_i),$$

for $i = 1, \ldots, d$. Since $\nabla J(u_*)(\cdot)$ is measurable and u_* satisfies the pointwise strict complementarity condition (4.27), it follows that $\Delta(u_*)(\cdot)$ is measurable, with

$$\Delta(u_*)(t) > 0 \quad \text{a.e. in } [0,1],$$

and, therefore,

$$\lim_{n \to \infty} \mu \left[\left\{ t \in [0, 1] : \Delta(u_*)(t) < \frac{1}{n} \right\} \right] = 0.$$

Consequently, for each $\rho > 0$ there is a measurable set $\beta \subset [0, 1]$ and a positive number c_β such that

$$\mu\ [\beta^c] < \rho, \tag{4.29a}$$

$$\forall t \in \beta \quad \Delta(u_*)(t) \geq c_\beta, \tag{4.29b}$$

where $\beta^c = [0, 1] \setminus \beta$.

Now fix c_2 in the interval $0 < c_2 < \min\{c_T, c_P\}$. Choose $\nu > 0$ so that for all measurable sets $\varphi \subset [0, 1]$,

$$\mu[\varphi] \leq \nu \Rightarrow \left(\int \int_{(\varphi^c \times \varphi^c)^c} \|\mathcal{K}(u_*)(t, s)\|^2 dt ds \right)^{\frac{1}{2}} \leq \frac{1}{4} (\min\{c_T, c_P\} - c_2).$$

By (4.27) and (4.29), there is a measurable set $\beta \subset [0, 1]$ and a $c_\beta > 0$ such that

$$\mu\ [\beta^c] < \frac{\nu}{2}$$

and

$$\forall t \in \beta\ \forall \eta \in \mathsf{N}(u_*)(t) \quad \|\eta\| \leq c_\beta \Rightarrow -\nabla J(u_*)(t) + \eta \in \mathcal{N}_U(u_*(t)). \tag{4.30}$$

Fix $\epsilon \in (0, 1]$, with

$$\left(1 - \epsilon^2\right) c_{\mathsf{T}} - 3\epsilon \|\nabla^2 J(u_*)\|_2 \geq \frac{1}{2} \left(\min\{c_{\mathsf{T}}, c_P\} + c_2\right).$$

By (1.2a), (1.3), (1.4b), (1.4c), and Taylor's theorem in $\mathbb{R}^1$, there is a sufficiently small $\delta \in (0, c_\beta]$ such that

$$\frac{\epsilon^2 c_\beta}{\delta} - \frac{1}{2}\|\nabla^2 J(u_*)\|_2 \geq \frac{1}{4} \left(\min\{c_{\mathsf{T}}, c_P\} + c_2\right)$$

and

$$\forall u \in B_2(u_*; \delta), \quad |r(u_* ; u - u_*)| \leq \frac{1}{8} \left(\min\{c_{\mathsf{T}}, c_P\} - c_2\right) \|u - u_*\|_2^2.$$

Put $\delta_2 = \delta\sqrt{\nu/2}$, $\theta = \{t \in [0, 1] : \|u(t) - u_*(t)\| \leq \delta\}$, and $\varphi = \theta^c \cup \beta^c = (\theta \cap \beta)^c$. Note that for all u in $\Omega \cap B_2(u_*; \delta_2)$, $\mu[\theta^c] \leq \frac{\nu}{2}$, $\mu[\varphi] \leq \nu$ and $\|w_u\|_\infty \leq \delta \leq c_\beta$. Hence, (4.26) and the preceding estimates imply that

$$J(u) - J(u_*) \geq \langle \nabla J(u_*), w_u \rangle_2 + \frac{1}{2} \langle w_u, \nabla^2 J(u_*) w_u \rangle_2$$

$$+ \frac{1}{2} c_P \int_\varphi \|u(t) - u_*(t)\|^2 dt - \frac{1}{4} \left(\min\{c_{\mathsf{T}}, c_P\} - c_2\right) \|u - u_*\|_2^2,$$

for all u in $\Omega \cap B_2(u_*; \delta_2)$. We now complete the proof by estimating the sum of the derivatives on the right side of this inequality.

Let $(w)_{\mathsf{N}}$ and $(w)_{\mathsf{T}}$ be the L^2 metric projections of w into $\mathsf{N}(u_*)$ and $\mathsf{T}(u_*)$, and consider the cone $C_\epsilon = \{w \in \mathsf{L}^\infty_m[0, 1] : \|(w)_{\mathsf{N}}\|_2 \leq \epsilon \|w\|_2\}$. Suppose that $u \in \Omega$ and $w_u = v_u - u_* \in C_\epsilon$. Then conditions (4.27) and (4.24) imply that $\langle \nabla J(u_*), w_u \rangle_2 \geq 0$ and

$$\begin{aligned}
\langle w_u, \nabla^2 J(u_*) w_u \rangle_2 &\geq \langle (w_u)_{\mathsf{T}}, \nabla J(u_*)(w_u)_{\mathsf{T}} \rangle_2 \\
&\quad - \|\nabla^2 J(u_*)\|_2 \left(2\|(w_u)_{\mathsf{N}}\|_2 \|(w_u)_{\mathsf{T}}\|_2 + \|(w_u)_{\mathsf{N}}\|_2^2\right) \\
&\geq \left(\left(1 - \epsilon^2\right) c_{\mathsf{T}} - 3\epsilon \|\nabla^2 J(u_*)\|_2\right) \|w_u\|_2^2 \\
&\geq \frac{1}{2} \left(\min\{c_{\mathsf{T}}, c_P\} + c_2\right) \|w_u\|_2^2,
\end{aligned}$$

and, therefore,

$$\langle \nabla J(u_*), v_u - u_* \rangle_2 + \frac{1}{2} \langle v_u - u_*, \nabla^2 J(u_*)(v_u - u_*) \rangle_2 \tag{4.31}$$

$$\geq \frac{1}{4} \left(\min\{c_{\mathsf{T}}, c_P\} + c_2\right) \int_{\varphi^c} \|u(t) - u_*(t)\|^2 dt.$$

On the other hand, suppose that $u \in \Omega \cap B_2(u_*; \delta_2)$ and $w_u \in C_\epsilon^c$. Put $\zeta = c_\beta \delta^{-1} (w_u)_{\mathsf{N}}$ and, note that $\zeta \in \mathsf{N}(u_*)$ and $\|\zeta\|_\infty \leq c_\beta$ since $\mathsf{N}(u_*)$ and $\mathsf{T}(u_*)$ are pointwise orthogonal; therefore, $\|(w_u)_{\mathsf{N}}(t)\| = \|P_{\mathsf{N}(u_*)(t)} w_u(t)\| \leq \|w_u(t)\|$ almost everywhere in $[0, 1]$. According to (4.30), we then have $-\nabla J(u_*) + \zeta \in \mathcal{N}_\Omega(u_*)$, in which case $\langle -\nabla J(u_*) + \zeta, w_u \rangle_2 \leq 0$ and, therefore,

$$\langle \nabla J(u_*), w_u \rangle_2 \geq \frac{c_\beta}{\delta} \|(w_u)_{\mathsf{N}}\|_2^2 \geq \frac{\epsilon^2 c_\beta}{\delta} \|w_u\|_2^2.$$

Hence, (4.31) holds once again, and thus

$$J(u) - J(u_*) \geq \frac{1}{4}(\min\{c_T, c_P\} + c_2) \int_{\varphi^c} \|u(t) - u_*(t)\|^2 dt + \frac{1}{2} c_P \int_{\varphi} \|u(t) - u_*(t)\|^2 dt$$

$$-\frac{1}{4}(\min\{c_T, c_P\} - c_2) \|u - u_*\|_2^2$$

$$\geq \frac{1}{2} c_2 \|u - u_*\|_2^2,$$

for all u in $\Omega \cap B_2(u_*; \delta_2)$. □

COROLLARY 4.2. *Assume that the hypotheses of Theorem* 4.1 *hold, with the Pontryagin condition* (4.26) *replaced by the stronger Legendre coercivity condition* (1.11). *Then* u_* *is an* L^2*-local minimizer of* J *in* Ω, *and the growth condition* (4.28) *holds near* u_*.

Note 4.1. For Bolza cost functions (1.5) satisfying (1.2a), (1.3), (1.4b), and (1.4c), the function H in the necessary condition (4.22) coincides with the Hamiltonian H in §1, and condition (4.22) reduces to the Pontryagin minimum principle.

Note 4.2. For the special case $U = [0, \infty)$ treated in [1], conditions (1.11) and (4.26) are *equivalent*. In general, (1.11) is considerably stronger than (4.26), but is still a natural hypothesis for nonconvex nonquadratic regulator optimal control problems with control-quadratic running costs and control-linear dynamic equations in (1.5).

5. L^2-Local convergence. It can happen that iterate sequences $\{u_i\}$ generated by (1.6) converge to a minimizer u_* in the norm $\|\cdot\|_2$ from all u_0 in some L^∞ neighborhood of u_*, and yet fail to converge to u_* in the norm $\|\cdot\|_2$ from starting points u_0 that are *arbitrarily close* to u_* in the L^2 sense [2]. In such cases, refined approximate finite-dimensional implementations of (1.6) are sure to converge (nearly) to u_* only from starting points u_0 that are exceedingly close to u_* in quadrature approximations of the norm $\|\cdot\|_2$. For practical purposes, this means that the location of jump discontinuities in u_* or thin boundary layers, where $u_*(t)$ is changing rapidly, must be known very precisely in advance and built into u_0, or (1.6) may not improve u_0 in any useful sense. As in [2], we therefore seek conditions on u_* that insure L^2 convergence of the iterates of (1.6) from all starting points u_0 in some L^2 neighborhood of u_*.

Note that every L^∞-local minimizer is a fixed point of the gradient projection map (1.6) since all such minimizers satisfy the necessary condition (4.21a). Moreover, the sufficient conditions in Theorem 4.1 can hold at u_* and yet every L^2 neighborhood of u_* can contain a *continuum* of L^∞-local minimizers that are actually L^∞-local *stable attractors* for (1.6) [23, Ex. 4]. Something more than the hypotheses in Theorem 4.1 is therefore needed to support an L^2-local convergence theory for gradient projection methods, or any other standard nonlinear programming algorithms that have a fixed point at each u_* satisfying (4.21a). In the following analysis, hypothesis (4.26) is replaced by the stronger Legendre coercivity condition (1.11) on $S(u_*)$, and Corollary 4.2 and (1.11) are then used much as the proof outline of §3 uses the KKT strict complementarity and coercivity conditions in $\mathbb{R}^m$. More specifically, the hypotheses in Corollary 4.2, the structure/continuity conditions (1.2b)–(1.4a), and local bounds for the step lengths $a(u)$ in (1.6) imply that for each u in small L^2 neighborhoods of u_* in Ω, there is a corresponding subspace $\hat{T}(u) \supset T(u^*)$ such that $G(u)$ falls in $(u_* + \hat{T}(u)) \cap \Omega$ and $\nabla^2 J(u_*)$ is coercive on $\hat{T}(u)$ (condition (1.11) is the key to the extension of coercivity from $T(u_*)$ to $\hat{T}(u)$). Furthermore, since u_* is an L^2-locally stable fixed point of G when the uniform growth condition (4.28) is satisfied [30], it follows that for $\|u_0 - u_*\|_2$ sufficiently small, the subsequent G-iterates u_i are confined to an L^2 neighborhood where the preceding assertions are true and where J is *convex* near u_* on the translated subspaces $u_* + \hat{T}(u_{i-1})$. Although G

no longer reduces to an unconstrained Armijo steepest descent map in the infinite-dimensional setting of (1.1), it is now possible to prove local convergence for $\{u_i\}$ *directly*, with estimates from [9] for gradient projection iteration maps and convex cost functions.

LEMMA 5.1. *Let J satisfy the structure/continuity requirements* (1.2)–(1.4) *and suppose that the Legendre coercivity condition* (1.11), *the* L^2*-coercivity condition* (4.24), *and the pointwise strict complementarity condition* (4.27) *in Corollary* 4.2 *hold at a point u_* in the set Ω in* (1.1b). *Then for some $\epsilon > 0$ and each u in $\Omega \cap B_2(u_*; \epsilon)$, there is a corresponding subspace $\hat{\mathsf{T}}(u) \supset \mathsf{T}(u_*)$ in $\mathsf{L}^\infty_m[0, 1]$ such that*

$$G(u) \in (u_* + \hat{\mathsf{T}}(u)) \cap \Omega \tag{5.32a}$$

and

$$\forall v \in B_2(u_*; \epsilon)\ \forall w \in \hat{\mathsf{T}}(u) \quad \langle w, \nabla^2 J(v)w\rangle_2 \geq 0, \tag{5.32b}$$

where G is the gradient projection map (1.6) *on Ω.*

Proof. Recall that for each u in U, $\mathcal{F}(u)$ is the unique face $\mathcal{F}_i \subset U$ such that $u \in ri\ \mathcal{F}_i$ (see §2). Thus, for all u in Ω and almost all t in $[0, 1]$, $\mathcal{F}(u(t))$ is defined and $u(t) \in ri\,\mathcal{F}(u(t))$. We will prove that for some $\epsilon > 0$ and $\rho > 0$, and for each u in $\Omega \cap B_2(u_*; \epsilon)$, there is a corresponding set $\varphi \subset [0, 1]$ such that

$$\min\{dist\ (u_*(t), rb\ \mathcal{F}(u_*(t))), dist\ (-a(u)\nabla J(u_*)(t), rb\ \mathcal{N}_U(u_*)(t))\} \geq \rho, \tag{5.33a}$$

$$\|u(t) - a(u)\nabla J(u)(t) - (u_*(t) - a(u)\nabla J(u_*)(t))\| < \rho, \tag{5.33b}$$

for all t in $\varphi^c = [0, 1] \setminus \varphi$. Hence, by Lemma 2.1,

$$G(u)(t) = P_U(u(t) - a(u)\nabla J(u)(t)) \in ri\ \mathcal{F}(u_*(t)) \subset u_*(t) + \mathsf{T}(u_*)(t),$$

for all t in φ^c; thus (5.32a) holds with

$$\hat{\mathsf{T}}(u) = \{v \in \mathsf{L}^\infty_m[0, 1] : v(t) \overset{\text{a.e.}}{\in} \hat{\mathsf{T}}(u)(t)\} \tag{5.34a}$$

and

$$\hat{\mathsf{T}}(u)(t) = \begin{cases} \mathbb{R}^m, & t \in \varphi, \\ \mathsf{T}(u_*)(t), & t \in \varphi^c. \end{cases} \tag{5.34b}$$

Moreover, $\mu[\varphi]$ and ϵ are small enough to insure that the coercivity condition (4.24) extends from $\mathsf{T}(u_*)$ to the larger subspace $\hat{\mathsf{T}}(u)$ and also that (5.32b) holds in the ball $B_2(u_*; \epsilon)$.

Conditions (1.2a), (1.3), (1.4b), and (1.4c) imply that J is twice continuously Fréchet differentiable relative to the norm $\|\cdot\|_2$, hence ∇J is locally Lipschitz continuous near u_* in this norm. A minor modification of the proof for Lemma A.2 in [38] shows that there are numbers $\delta > 0$ and $\underline{a} > 0$ such that for all u in $\Omega \cap B_2(u_*; \delta)$,

$$a(u) \geq \underline{a} \quad \text{and} \quad \|\nabla^2 J(u) - \nabla^2 J(u_*)\|_2 \leq \frac{1}{2} \min\{c_\mathsf{T}, c_P\},$$

where c_T and c_P are the positive numbers in the coercivity conditions (4.24) and (1.11), and $a(u)$ is the step length in (1.6). Now choose $\nu \in (0, 1]$ so that for all measurable sets $\varphi \subset [0, 1]$,

$$\mu[\varphi] \leq \nu \Rightarrow \left(\int\int_{(\varphi^c \times \varphi^c)^c} \|\mathcal{K}(u)(t, s)\|^2 dt ds\right)^{\frac{1}{2}} \leq \frac{1}{2} \min\{c_\mathsf{T}, c_P\}.$$

With reference to (4.27) and the resulting condition (4.29) in the proof of Theorem 4.1, there is a measurable set β in $[0, 1]$ and a number $c_\beta > 0$ such that $\mu[\beta^c] \leq \nu/3$, $-\nabla J(u_*)(t) \in ri\ \mathcal{N}_U(u_*(t))$, and $dist\ (-\nabla J(u_*)(t), rb\ \mathcal{N}_U(u_*)(t)) \geq c_\beta$, for all t in β. Since $\mathcal{N}_U(u_*(t))$ is a cone, it follows that for all $a > 0$,

$$-a\nabla J(u_*)(t) \in ri\ \mathcal{N}_U(u_*(t))$$

and

$$dist\ (-a\nabla J(u_*)(t), rb\ \mathcal{N}_U(u_*)(t)) \geq ac_\beta,$$

for all t in β. Furthermore, by construction, the set-valued map $\mathcal{F}(u_*(\cdot))$ is *constant* on each of the measurable sets $\alpha_i(u_*)$, with $u_*(t) \in ri\ \mathcal{F}(u_*)(t)$ almost everywhere in $[0, 1]$. Hence $dist\ (u_*(\cdot), rb\ \mathcal{F}(u_*(\cdot))$ is measurable and positive almost everywhere in $[0, 1]$, and there is a number $c_\omega > 0$ and a corresponding set

$$\omega = \{t \in [0, 1] : dist\ (u_*(t), rb\ \mathcal{F}(u_*(t))) \geq c_\omega\}$$

such that $\mu[\omega^c] \leq \nu/3$. The estimate (5.33a) is now seen to hold for all t in $(\omega^c \cup \beta^c)^c = \omega \cap \beta$ with $\rho = \min\{c_\omega, \underline{a}c_\beta\}$. By (1.2b) and (1.4), δ can be reduced further if necessary, so that for some $L > 0$ and all u in $B_2(u_*; \delta)$,

$$\|\nabla J(u)(t) - \nabla J(u_*)(t)\| \leq \frac{\rho}{2\overline{a}} + L\|u(t) - u_*(t)\|$$

almost everywhere in $[0, 1]$, and $(1 + \overline{a}L)\,\delta < \frac{\rho}{2}$, where $\overline{a}$ is the step length *upper* bound in (1.6). For u in Ω, let $\theta = \{t \in [0, 1] : \|u(t) - u_*(t)\| \leq \delta\}$ and $\varphi = \theta^c \cup \omega^c \cup \beta^c$. Put $\epsilon = \delta\sqrt{\nu/3}$ and note that for all u in $\Omega \cap B_2(u_*; \epsilon)$, conditions (5.33) hold almost everywhere in φ^c with $\rho = \min\{c_\omega, \underline{a}c_\beta\}$. Finally, note that $\mu[\theta^c] \leq \frac{\nu}{3}$ and $\mu[\varphi] \leq \nu$, and construct $\hat{\mathsf{T}}(u)$ with (5.34) for u in $\Omega \cap B_2(u_*; \epsilon)$. With (4.24) and (1.11), we now find that for all u in $\Omega \cap B_2(u_*; \epsilon)$ and w in $\hat{\mathsf{T}}(u)$,

$$\begin{aligned}
\langle w, \nabla^2 J(u_*)w\rangle_2 &= \int_{\varphi^c} \langle w(t), S(u_*)(t)w(t)\rangle dt + \int_{\varphi} \langle w(t), S(u_*)(t)w(t)\rangle dt \\
&\quad + \int\int_{\varphi^c \times \varphi^c} \langle w(t), \mathcal{K}(u_*)(t, s)w(t)\rangle dtds \\
&\quad + \int\int_{(\varphi^c \times \varphi^c)^c} \langle w(t), \mathcal{K}(u_*)(t, s)w(t)\rangle dtds \\
&\geq c_{\mathsf{T}} \int_{\varphi^c} \|w(t)\|^2 dt + c_P \int_{\varphi} \|w(t)\|^2 dt - \frac{1}{2} \min\{c_{\mathsf{T}}, c_P\}\|w\|_2^2 \\
&\geq \frac{1}{2} \min\{c_{\mathsf{T}}, c_P\}\|w\|_2^2.
\end{aligned}$$

Hence, $\langle w, \nabla^2 J(v)w\rangle_2 \geq \langle w, \nabla^2 J(u_*)w\rangle_2 - \|\nabla^2 J(v) - \nabla^2 J(u_*)\|_2\|w\|_2^2 \geq 0$, for all v in $B_2(u_*; \epsilon)$ and w in $\hat{\mathsf{T}}(u)$. □

Note 5.1. Reference [36] establishes a positive lower bound for $a(u)$ on bounded subsets of simple polyhedra in $\mathbb{R}^m$ when ∇J is Lipschitz continuous on bounded sets. In infinite-dimensional spaces, this Lipschitz continuity hypothesis does not automatically follow from continuity of the second derivative of J, since closed bounded sets are not compact in such spaces. On the other hand, the step length bound in [38] applies when ∇J is merely Lipschitz continuous near stationary points, and this local property *is* implied by continuity of the second derivative.

THEOREM 5.2. *Let J satisfy the structure/continuity requirements* (1.2)–(1.4) *and suppose that the Legendre coercivity condition* (1.11), *the* L^2*-coercivity condition* (4.24), *and the pointwise strict complementarity condition* (4.27) *in Corollary* 4.2 *hold at a point u_* in the set Ω in* (1.1b). *Then there are numbers $\delta > 0$ and $\lambda \in [0, 1)$ such that for each sequence $\{u_i\}$ generated by the gradient projection iteration* (1.6),

$$u_0 \in \Omega \cap B_2(u_*; \delta) \Rightarrow \forall i \geq 1 \quad J(u_{i+1}) - J(u_*) \leq \lambda \ (J(u_i) - J(u_*)) .$$

Furthermore, the corresponding norms $\|u_i - u_*\|_2$ are bounded above by a real sequence that converges to zero geometrically, with ratio $\sqrt{\lambda}$.

Proof. By Corollary 4.2 and Lemma 5.1, there are positive numbers ϵ and $\underline{a}$ such that $a(u)$ is bounded away from zero by $\underline{a}$ and conditions (4.28) and (5.32) hold at each u in $\Omega \cap B_2(u_*; \epsilon)$. Condition (4.28) and the basic continuity and descent properties of G insure that u_* is an L^2 stable fixed point of G [30]. Hence, there is a $\delta \in (0, \epsilon]$ such that for all sequences $\{u_i\}$ generated by (1.6),

$$u_0 \in \Omega \cap B_2(u_*, \delta) \Rightarrow \forall i \geq 0, \quad u_i \in \Omega \cap B_2(u_*; \epsilon).$$

The remainder of the proof employs local versions of estimates developed in [9] for convex J.

Assume that the G-iterate sequence $\{u_i\}$ is confined to $\Omega \cap B_2(u_*; \epsilon)$. By (5.32a), we then have

$$u_i \in (u_* + \hat{\mathrm{T}}(u_{i-1})) \cap \Omega \cap B_2(u_*; \epsilon),$$

for $i \geq 1$. By (5.32b), this implies that

$$J(u_i) - J(u_*) \leq \langle \nabla J(u_i), u_i - u_* \rangle_2,$$

for $i \geq 1$. Since $u_{i+1} = P_\Omega (u_i - a(u_i)\nabla J(u_i))$, the orthogonality condition in the projection theorem gives

$$\langle u_i - a(u_i)\nabla J(u_i) - u_{i+1}, u_* - u_{i+1} \rangle_2 \leq 0$$

and

$$\langle u_i - a(u_i)\nabla J(u_i) - u_{i+1}, u_i - u_{i+1} \rangle_2 \leq 0.$$

Hence,

$$\begin{aligned} \langle \nabla J(u_i), u_i - u_* \rangle_2 &= \langle \nabla J(u_i), u_i - u_{i+1} \rangle_2 + \langle \nabla J(u_i), u_{i+1} - u_* \rangle_2 \\ &\leq \langle \nabla J(u_i), u_i - u_{i+1} \rangle_2 + \frac{1}{\sqrt{a(u_i)}} \|u_{i+1} - u_*\|_2 \ \langle \nabla J(u_i), u_i - u_{i+1} \rangle_2^{\frac{1}{2}}, \end{aligned}$$

for all $i \geq 1$. Now put $r_i = J(u_i) - J(u_*)$ and note that $a(u_i)$ is bounded below by $\underline{a}$ and that $\sigma \langle \nabla J(u_i), u_i - u_{i+1} \rangle_2$ is bounded above by $r_i - r_{i+1}$, where σ is the step length rule parameter in (1.6). Note also that $c_2 \|u_{i+1} - u_*\|_2^2$ is bounded above by $2r_{i+1}$, and hence by $2r_i$. Consequently, for all $i \geq 1$,

$$0 \leq r_i \leq \frac{r_i - r_{i+1}}{\sigma} + \left(\frac{2r_i}{c_2 \underline{a}}\right)^{\frac{1}{2}} \left(\frac{r_i - r_{i+1}}{\sigma}\right)^{\frac{1}{2}}.$$

By completing the square on the right side of this estimate, we find that for $i \geq 1$,

$$0 \leq r_{i+1} \leq \lambda r_i$$

and

$$\|u_i - u_*\|_2 \leq \sqrt{\frac{2r_i}{c_2}},$$

with

$$\lambda = 1 - \frac{2c_2\underline{a}\sigma}{(1+\sqrt{1+2c_2\underline{a}})^2}. \qquad \square$$

6. L^2-Local active facet identification. In [31], it is shown that if ∇J is continuous and u_* satisfies the geometric strict complementarity condition (3.15a) in a polyhedral set $U \subset \mathbb{R}^m$, then iterates of the projection map (3.18) that converge to u_* are eventually confined to the *active facet* $ri\ \mathcal{F}(u_*)$ at u_*. A similar conclusion can be reached by the analysis in [30] under the stronger local Lipschitz continuity hypothesis needed for the local convergence theory (see Note 3.2). We now prove an *asymptotic* counterpart of the latter result for the gradient projection map (1.6). This theorem subsumes the active constraint identification result in [2] for $U = [0, \infty)$. Results of a similar nature are established in [41] and [42] for constrained compact fixed point problems and discrete-time approximations to optimal control problems. Reference [43] formulates an active constraint identification theorem *and* a local convergence theorem for an infinite-dimensional variant of the projected Newton scheme in [36] applied to optimal control problems with bounded scalar inputs.

Fix u_* in Ω and for each u in Ω put

$$\mathcal{I}(u) = \{t \in [0,1] : u(t) \in U, u_*(t) \in U \text{ and } \mathcal{F}(u(t)) = \mathcal{F}(u_*(t))\}. \tag{6.35}$$

By construction, u *identifies* the active facets $ri\ \mathcal{F}(u_*(t))$ for t in the index set $\mathcal{I}(u)$.

THEOREM 6.1. *Suppose that J is twice continuously Fréchet differentiable relative to the norm $\|\cdot\|_2$ and satisfies conditions* (1.2), (1.4a), *and* (1.4b). *Assume that the pointwise strict complementarity condition* (4.27) *holds at point u^* in the set Ω in* (1.1b). *Then for all gradient projection iterate sequences $\{u_i\}$ generated by* (1.6),

$$\lim_{i\to\infty} \|u_i - u_*\|_2 = 0 \Rightarrow \lim_{i\to\infty} \mu[\mathcal{I}(u_i)^c] = 0,$$

where $\mathcal{I}(u_i)$ is defined by (6.35) *and $\mathcal{I}(u_i)^c = [0,1] \setminus \mathcal{I}(u_i)$.*

Proof. For $u \in \Omega$ and $j = 1, \ldots, d$, let $\alpha_j(u) = u^{-1}\ [ri\ \mathcal{F}_j]$ as before. By construction, $\alpha_j(u) \cap \alpha_k(u) = \emptyset$ for $j \neq k$, with

$$\mu\left[[0,1] \setminus \bigcup_{j=1}^{d} \alpha_j(u)\right] = 0,$$

and thus

$$\mu\left[[0,1] \setminus \bigcup_{j,k=1}^{d} \left(\alpha_j(u) \cap \alpha_k(u_*)\right)\right] = 0.$$

Note that

$$\mathcal{I}(u) = \bigcup_{j=1}^{d} \left(\alpha_j(u) \cap \alpha_j(u_*)\right).$$

Hence $[0,1] \setminus \mathcal{I}(u)$ differs from $\bigcup_{j\neq k} \left(\alpha_j(u) \cap \alpha_k(u_*)\right)$ by a set of measure zero, and it therefore suffices to show that for all $j, k = 1, \ldots, d$,

$$j \neq k \Rightarrow \lim_{i\to\infty} \mu[\alpha_j(u_i) \cap \alpha_k(u_*)] = 0. \tag{6.36}$$

We now use Lemma 2.1 to prove (6.36).

Suppose that $\{u_i\}$ is generated by (1.6) and that $\lim_{i\to\infty} \|u_i - u_*\|_2 = 0$. Since J is twice continuously differentiable relative to $\|\cdot\|_2$, the map $\nabla J(\cdot)$ is locally Lipschitz continuous relative to $\|\cdot\|_2$; hence, for some $\underline{a} > 0$, the step lengths $a(u_i)$ are bounded away from zero by $\underline{a}$ for all i. Fix $\epsilon > 0$, k and $j \neq k$. By (4.27) and the resulting condition (4.29), there is a measurable set $\beta \subset \alpha_k(u_*)$ and a $c_\beta > 0$ such that $\mu[\alpha_k(u_*) \setminus \beta] \leq \frac{\epsilon}{3}$, $-\nabla J(u_*)(t) \in ri\ \mathcal{N}_U(u_*(t))$, and $dist\ (-\nabla J(u_*)(t), rb\ \mathcal{N}_U(u_*(t))) \geq c_\beta$, for all t in β. Since $dist\ (u_*(\cdot), rb\ \mathcal{F}_k)$ is measurable and positive on $\alpha_k(u_*)$, there is a measurable set $\omega \subset \alpha_k(u_*)$ and a $c_\omega > 0$ such that $\mu[\alpha_k(u_*)\setminus\omega] \leq \frac{\epsilon}{3}$, $u_*(t) \in ri\ \mathcal{F}_k$, and $dist\ (u_*(t), rb\ \mathcal{F}_k) \geq c_\omega$, for all t in ω. Put $\rho = \min\{c_\omega, \underline{a}c_\beta\}$, and note that $\mathcal{N}_U(u_*(t))$ is a cone. Hence, for all i and all t in $\beta \cap \omega$,

$$\min\{dist\ (-a(u_i)\nabla J(u_*)(t), rb\ \mathcal{N}_U(u_*(t))), dist\ (u_*(t), rb\ \mathcal{F}(u_*(t)))\} \geq \rho.$$

Since $\lim_{i\to\infty} \|u_i - u_*\|_2 = 0$, conditions (1.4a), (1.4b) imply that for some $L \geq 0$ and all sufficiently large i,

$$\|\nabla J(u_i)(t) - \nabla J(u_*)(t)\| \leq \frac{\rho}{2\bar{a}} + L\|u_i(t) - u_*(t)\|$$

almost everywhere in $[0, 1]$, where $\bar{a}$ is the step length upper bound in (1.6). Now let $\eta = \frac{\rho}{2(1+\bar{a}L)}$ and $\theta_i = \{t \in \alpha_k(u_*) : \|u_i(t) - u_*(t)\| < \eta\}$, and note that for sufficiently large i and almost all t in θ_i,

$$\|u_i(t) - a(u_i)\nabla J(u_i)(t) - (u_*(t) - a(u_i)\nabla J(u_*)(t))\| < \rho.$$

Hence, Lemma 2.1 insures that for sufficiently large i,

$$u_{i+1}(t) = P_U(u_i(t) - a(u_i)\nabla J(u_i)(t)) \in ri\ \mathcal{F}(u_*(t)) = ri\ \mathcal{F}_k$$

almost everywhere in $\beta \cap \omega \cap \theta_i$; thus

$$\mu[\alpha_j(u_{i+1}) \cap \alpha_k(u_*) \cap \beta \cap \omega \cap \theta_i] = 0,$$

since $ri\ \mathcal{F}_j \cap ri\ \mathcal{F}_k = \emptyset$ for $j \neq k$. In addition, since convergence in the norm $\|\cdot\|_2$ implies convergence in measure, it follows that $\mu[\alpha_k(u_*) \setminus \theta_i] \leq \frac{\epsilon}{3}$ and, therefore, $\mu[\alpha_k(u_*) \setminus (\beta \cap \omega \cap \theta_i)] \leq \epsilon$, for sufficiently large i. In the limit as $i \to \infty$, we now obtain

$$\begin{aligned} 0 &\leq \limsup_{i\to\infty}\ \mu[\alpha_j(u_{i+1}) \cap \alpha_k(u_*)] \\ &\leq \limsup_{i\to\infty}\ \mu[\alpha_j(u_{i+1}) \cap \alpha_k(u_*) \cap \beta \cap \omega \cap \theta_i] \\ &\quad + \limsup_{i\to\infty}\ \mu[\alpha_j(u_{i+1}) \cap \alpha_k(u_*) \cap (\alpha_k(u_*) \setminus (\beta \cap \omega \cap \theta_i))] \\ &\leq \epsilon. \end{aligned}$$

Since ϵ is an arbitrarily small positive number, this proves (6.36). □

Note 6.1. Theorem 6.1 and its proof remain valid for objective functions J with gradients that are locally Lipschitz continuous relative to the norm $\|\cdot\|_2$; this weaker hypothesis automatically holds for the twice continuously Fréchet differentiable functions of interest here.

Acknowledgment. The author is grateful to the referees for several helpful comments on exposition and a valuable observation pertaining to the proof of Theorem 4.1. In the original version of this paper, Theorem 4.1 invoked the strict complementarity condition (4.25), and one of the referees noticed that the weaker condition (4.29) seemed to serve equally well in

place of (4.25). The proof originally designed for (4.25) does indeed work for (4.29), and appears here virtually unchanged, apart from notational adjustments and new material in the second paragraph. The new material establishes measurability of $\Delta(u_*)(\cdot)$, proves that the proposed condition (4.29) is actually equivalent to the weak pointwise strict complementarity condition (4.27), and links (4.29) to KKT strict complementarity in Cartesian products of polyhedra.

REFERENCES

[1] J. C. DUNN AND T. TIAN, *Variants of the Kuhn-Tucker sufficient conditions in cones of non-negative functions*, SIAM J. Control Optim., 30 (1992), pp. 1361–1384.

[2] T. TIAN AND J. C. DUNN, *On the gradient projection method for optimal control problems with nonnegative L^2 inputs*, SIAM J. Control Optim., 32 (1994), pp. 516–537.

[3] T. TIAN, *Convergence Analysis of a Projected Gradient Method for a Class of Optimal Control Problems*, Ph.D. Dissertation, North Carolina State University, Raleigh, NC, 1992.

[4] J. C. DUNN, *Second order optimality conditions in sets of L^∞ functions with range in a polyhedron*, SIAM J. Control Optim., 33 (1995), pp. 1603–1635.

[5] A. A. GOLDSTEIN, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.

[6] E. S. LEVITIN and B. T. POLJAK, *Constrained minimization problems*, USSR Comp. Math. Phys., 6 (1966), pp. 1–50.

[7] D. P. BERTSEKAS, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Trans. Automat. Control, AC-10 (1976), pp. 174–184.

[8] D. G. LUENBERGER, *Optimization by Vector Space Methods*, Wiley, New York, 1969.

[9] J. C. DUNN, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, SIAM J. Control. Optim., 19 (1981), pp. 368–400.

[10] H. MAURER AND J. ZOWE, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.

[11] H. MAURER, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Study, 14 (1981), pp. 163–177.

[12] ———, *The two-norm approach for second order sufficiency conditions in mathematical programming and optimal control*, Tech. Rept. 6/92 - N, Inst. f. Angew. Math. Inform., Universität Münster, 1992.

[13] ———, *Solution differentiability for parametric nonlinear control problems with control-state constraints*, Control Cybernet., 23 (1994), pp. 201–227.

[14] A. S. DONTCHEV, W. W. HAGER, A. B. POORE, AND B. YANG, *Optimality, stability and convergence in nonlinear control*, Appl. Math. Optim., 31 (1995), pp. 297–326.

[15] V. ZEIDAN, *Sufficient conditions for the generalized problem of Bolza*, Trans. Amer. Math. Soc., 275 (1983), pp. 561–586.

[16] ———, *Sufficiency criteria via focal points and via coupled points*, SIAM J. Control Optim., 30 (1992), pp. 82–98.

[17] D. ORRELL AND V. ZEIDAN, *Another Jacobi sufficiency criterion for optimal control with smooth constraints*, J. Optim. Theory Appl., 58 (1988), pp. 283–300.

[18] W. ALT, *The Lagrange-Newton method for infinite-dimensional optimization problems*, Numer. Funct. Anal. Optim., 11 (1990), pp. 201–224.

[19] W. ALT AND K. MALANOWSKI, *The Lagrange-Newton method for nonlinear optimal control problems*, Comp. Optim. Appl., 2 (1993), pp. 77–100.

[20] K. MALANOWSKI, *Sensitivity analysis of optimization problems in Hilbert space, with application to optimal control*, Appl. Math. Optim., 21 (1990), pp. 1–20.

[21] ———, *Second order conditions and constraint qualifications in stability and sensitivity analysis of solutions to optimization problems in Hilbert space*, Appl. Math. Optim., 25 (1992), pp. 51–79.

[22] ———, *Two-norm approach in stability and sensitivity analysis of optimization and optimal control problems*, Adv. Math. Sci. Appl., 2 (1993), pp. 397–443.

[23] J. C. DUNN, *Gradient-related constrained minimization algorithms in function spaces: Convergence properties and computational implications*, in Large Scale Optimization: State of the Art, Kluwer Academic Publishers, Dordrecht, 1994.

[24] ———, *Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals*, SIAM J. Control Optim., 17 (1979), pp. 187–211.

[25] G. C. HUGHES AND J. C. DUNN, *Newton-Goldstein convergence rates for convex constrained minimization problems with singular solutions*, Appl. Math. Optim., 12 (1984), pp. 203–230.

[26] J. C. DUNN, *Extremal types for certain L^p-minimization problems and associated large scale nonlinear programs*, Appl. Math. Optim., 10 (1983), pp. 303–335.

[27] J. C. DUNN AND E. W. SACHS, *The effects of perturbations on the convergence rates of optimization algorithms*, Appl. Math. Optim., 10 (1983), pp. 143–157.
[28] E. W. SACHS, *Rates of convergence for adaptive Newton methods*, J. Optim. Theory Appl., 48 (1986), pp. 175–190.
[29] J. C. DUNN, *Diagonally modified conditional gradient methods for input constrained optimal control problems*, SIAM J. Control Optim., 24 (1986), pp. 1177–1191.
[30] ———, *On the convergence of projected gradient processes to singular critical points*, J. Optim. Theory Appl., 55 (1987), pp. 203–215.
[31] J. V. BURKE AND J. J. MORÉ, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197–1211.
[32] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[33] J. C. DUNN, *A subspace decomposition principle for scaled gradient projection methods: Local theory*, SIAM J. Control Optim., 31 (1993),pp. 219–246.
[34] P. H. CALAMAI AND J. J. MORÉ, *Projected gradient methods for linearly constrained problems*, Math. Programming, 39 (1987), pp. 93–116.
[35] E. M. GAFNI AND D. P. BERTSEKAS, *Two-metric projection methods for constrained minimization*, SIAM J. Control Optim., 22 (1984), pp. 936–964.
[36] D. P. BERTSEKAS, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221–246.
[37] ———, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.
[38] M. GAWANDE AND J. C. DUNN, *Variable metric gradient projection processes in convex feasible sets defined by nonlinear inequalities*, Appl. Math. Optim., 17 (1988), pp. 103–119.
[39] S. J. WRIGHT, *Identifiable surfaces in constrained optimization*, SIAM J. Optim., 31 (1993), pp. 1063–1079.
[40] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1974.
[41] C. T. KELLEY AND E. W. SACHS, *Multi-level algorithms for constrained compact fixed point problems*, SIAM J. Sci. Comput., 15 (1994), pp. 645–667.
[42] ———, *Mesh independence of the gradient projection method for optimal control problems*, SIAM J. Control Optim., 30 (1992), pp. 477–493.
[43] ———, *Solution of optimal control problems by a pointwise projected Newton method*, Rept. CRSC-TR93-13, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, 1993.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1291–1299, July 1996

009

LARGE-TIME LOCAL CONTROLLABILITY VIA HOMOGENEOUS APPROXIMATIONS*

HENRY HERMES†

Abstract. If all points in a neighborhood of a rest solution of an n-dimensional, affine control system can be attained in some sufficiently large time $t_1 > 0$, we say that the system is large-time locally controllable at the rest solution. Sufficient conditions for large-time local controllability are given in terms of small-time local controllability of homogeneous approximating systems. The major result, Theorem 3, is a geometric test for large-time local controllability in terms of the (coordinate-free) structure of Lie products of the vector fields which define the system, evaluated at the rest solution. Large-time local controllability has implications for the problem of the existence of an asymptotically stabilizing feedback control.

Key words. homogeneous approximations, local controllability, Lie algebras

AMS subject classifications. 93B29, 93B05, 57R27, 17B70

Introduction. We study the problem of when, for sufficiently large time, can one reach all points in a neighborhood of the origin by solutions of the nonlinear, affine, single input control system on $\mathbb{R}^n$

$$\dot{x} = X(x) + uY(x), \quad X(0) = 0,\ Y(0) \neq 0 \tag{1}$$

with initial data $x(0) = 0$. Here X and Y are assumed real analytic vector fields on $\mathbb{R}^n$, while the admissible control set, denoted Ω^α, consists of Lebesgue-measurable functions $u : [0, \infty) \to [-\alpha, \alpha]$, $\alpha > 0$. The attainable set at time t, i.e., the set of all points that can be reached in time t by solutions of (1) using controls $u \in \Omega^\alpha$, will be denoted $A_1^\alpha(t)$. System (1) is *small-time locally controllable* (STLC) at zero if given any $t_1 > 0, \alpha > 0, 0 \in \text{int} A_1^\alpha(t_1)$. We define system (1) to be *large-time locally controllable* (LTLC) at zero if given any $\alpha > 0$ there exists a $t_1 > 0$ such that $0 \in \text{int} A_1^\alpha(t_1)$. STLC corresponds to the ability to correct small deviations from the reference solution, corresponding to $u = 0$, in arbitrarily small time. It has been extensively studied (see [4]–[6], [9]) with a major computable, sufficient condition given by Sussmann in [11]. Large-time local controllability is more basic to the question of asymptotic stabilization of the zero solution of (1) via feedback control. Indeed, if the time reversal system of system (1), i.e., $\dot{x} = -X(x) + vY$, is LTLC at zero, it follows that for any point p in some neighborhood of zero there exists a control $u \in \Omega^\alpha$ such that the corresponding solution $x(t, u)$ of system (1) for initial data $x(0) = p$ satisfies $\lim_{t\to\infty} x(t, u) = 0$. This is an obvious necessary condition for the existence of an asymptotically stabilizing feedback control [1]. Our sufficient condition for large-time local controllability will be given in terms of Lie brackets of X and Y evaluated at zero; hence it suffices to study system (1) since the bracket structure of the time-reversed system is the same (up to signs) as that of system (1).

The methods, here, depend on high-order homogeneous approximations of system (1). A *dilation*, δ_ε^r, is a map from $\mathbb{R}^n$ to $\mathbb{R}^n$ of the form $\delta_\varepsilon^r x = (\varepsilon^{r_1} x_1, \ldots, \varepsilon^{r_n} x_n)$, $\varepsilon > 0$, $1 \leq r_1 \leq r_2 \leq \cdots \leq r_n$ integers. A polynomial $h : \mathbb{R}^n \to \mathbb{R}^1$ is homogeneous of degree m with respect to δ_ε^r if $h(\delta_\varepsilon^r x) = \varepsilon^m h(x)$. The set of polynomials homogeneous of degree m will be denoted P_m. P_0 consists of constant functions, and we set $P_m = \{0\}$ if $m < 0$. A vector field $X(x) = \sum_{j=1}^n a_j(x)\partial/\partial x_j$ on $\mathbb{R}^n$ is homogeneous of degree m with respect to δ_ε^r if $a_j \in P_{r_j+m-1}$. This definition (while not universally used; e.g., see [8], [4]) agrees with the classical definition; i.e., a vector field $X(x) = Ax$ linear in the local coordinates is homogeneous of degree 1 with respect to the *standard dilation* δ_ε^1 having $r_1 = r_2 = \cdots = r_n = 1$.

*Received by the editors May 20, 1994; accepted for publication (in revised form) March 10, 1995. This research was supported by NSF grant DMS–9301039.

†Department of Mathematics, University of Colorado, Boulder, CO 80309.

A vector field X homogeneous of degree m will usually be denoted $X^{(m)}$. In §1, we will not make local coordinate changes, and it is convenient for later results of §2 to *not* necessarily assume $Y(x) = \partial/\partial x_1$ in the given coordinates. Expand X and Y in terms of homogeneous vector fields as

$$\begin{aligned} X(x) &= X^{(m)}(x) + X^{(m+1)}(x) + \cdots, \\ Y(x) &= Y^{(1-\ell)}(x) + Y^{(2-\ell)}(x) + \cdots, \end{aligned} \tag{2}$$

where $X(0) = 0$ implies $X^{(m)}(0) = 0$ and $Y(0) \neq 0$ implies $\ell \geq 1$ and we assume that the leading terms $X^{(m)}$, $Y^{(1-\ell)}$ are not zero vector fields. The homogeneous approximation of (1) relative to the dilation δ_ε^r used in (2) is

$$\dot{x} = X^{(m)}(x) + vY^{(1-\ell)}(x). \tag{3}$$

A basic result in the study of STLC is the following theorem.

THEOREM 1 (see [9], [2]). *If there exists a dilation δ_ε^r for which $m = 1$ in the expansion of X (or $m \leq 1$ and $m + \ell - 1 \geq 0$) and the approximating system* (3) *is STLC at zero, then system* (1) *is STLC at zero. If system* (3) *is STLC at zero and has $m > 1$, then system* (1) *is LTLC.*

The first statement in the above theorem can be found in [9, Thm. 3.2]. The second statement is a consequence of [2, Rem. 2.5]. The basic idea in the proof is to make the variable changes $\tau = e^{-(m-1)s}t$ or $t = t(\tau) = e^{(m-1)s}\tau$ and $x^s(\tau) = \delta_{e^s}^r x(t(\tau), u)$. Let $A_1^\alpha(t)$ denote the attainable set, at time t, of system (1) with controls $u \in \Omega^\alpha$; let $A_3^\alpha(\tau)$ denote the attainable set, at time τ, of the approximating system (3) with controls in Ω^α and $A_s^\alpha(\tau)$ denote the attainable set at time τ for the system x^s satisfies. Then one can show

$$A_s^\alpha(\tau) = \delta_{e^s}^r A_1^{\alpha e^{-s(m+\ell-1)}}(e^{(m-1)s}\tau)$$

and $0 \in \operatorname{int} A_3^\alpha(\tau)$ implies $0 \in \operatorname{int} A_s^\alpha(\tau)$ for s sufficiently large. Thus if $m = 1$, the first statement of Theorem 1 follows, but for $m > 1$, the time $e^{(m-1)s}\tau$ may be very large for large s. Indeed $s = s(\tau)$ and hence $e^{(m-1)s(\tau)}\tau$ need not tend to zero as $\tau \to 0$ (Example 1.1 has such behavior), so for $m > 1$ we cannot conclude that STLC of (3) implies STLC of (1).

Examples are given of systems for which the only STLC homogeneous approximating systems have $m \geq 2$; i.e., the original system is LTLC and can be shown to not be STLC at zero. These examples involve guessing a dilation; the results that can be concluded depend on the dilation chosen. Furthermore, no changes in the local coordinates used to initially describe the vector fields in (1) are made.

In §2 the main result is Theorem 3, which gives a "geometric" sufficient condition for large-time local controllability in terms of the (coordinate-free) structure of Lie products of X and Y evaluated at zero. Example 2.1 illustrates the nature of Theorem 3 and introduces some notation. The proof of Theorem 3 uses weight-induced filtrations of the Lie algebra, $L(X, Y)$, generated by X and Y and is constructive in that the preferred local coordinates, dilation, and STLC approximating system can be computed, although this is not necessary for applications of the theorem. This construction is illustrated in Example 2.2.

Our notation will be $(\operatorname{ad} X, Y) = [X, Y]$, the Lie product of vector fields X, Y, and inductively $(\operatorname{ad}^{k+1} X, Y) = [X, (\operatorname{ad}^k X, Y)]$. For S a subset of $L(X, Y)$, $S(0)$ denotes the elements of S evaluated at zero. We assume, throughout, that $\dim L(X, Y)(0) = n$, and all examples will satisfy this. For X a smooth vector field on $\mathbb{R}^n$ and $p \in \mathbb{R}^n$, we will often use $\exp(X)(p)$ to denote the solution, at time t, of $\dot{x} = X(x)$, $x(0) = p$.

1. Examples.

Example 1.1. We consider the system

$$\dot{x}_1 = u, \quad \dot{x}_2 = x_1^3, \quad \dot{x}_3 = x_2^3 + x_1^4. \tag{4}$$

Here $X(x) = x_1^3 \partial/\partial x_2 + (x_2^3 + x_1^4)\partial/\partial x_3$, $Y = \partial/\partial x_1$. If one chooses the standard dilation δ_ε^1, then $X(x) = X^{(3)}(x) + X^{(4)}(x)$ with $X^{(3)}(x) = x_1^3\ \partial/\partial x_2 + x_2^3\ \partial/\partial x_3$, $X^{(4)}(x) = x_1^4\ \partial/\partial x_3$, and $Y = Y^{(0)} = \partial/\partial x_1$. The homogeneous approximating system (3) becomes $\dot{x} = X^{(3)}(x) + u(t)Y$ and is STLC. Indeed, this is an "odd" system; i.e., the brackets which are linearly independent at zero are Y, $(\mathrm{ad}^3 Y, X)$ and $(\mathrm{ad}^3(\mathrm{ad}^3 Y, X), X)$, which all have an odd number of factors Y. This approximating system is a cubic integrator and hence admits a smooth asymptotic stabilizing feedback control (see [3]), which is therefore a local asymptotically stabilizing feedback control for (4). By Theorem 1, system (4) is LTLC. However, in the original system, the bracket $(\mathrm{ad}^4 Y, X)(0)$ is linearly independent of Y, $(\mathrm{ad}^3 Y, X)(0)$ and is an obstruction to small-time local controllability (see [10]); i.e., the original system is not STLC. Had one chosen the dilation as δ_ε^r with $r = (1, 3, 4)$, the expansion of X becomes $X(x) = X^{(1)}(x) + X^{(6)}(x)$ with $X^{(1)}(x) = x_1^3\ \partial/\partial x_2 + x_1^4\ \partial/\partial x_3$, $X^{(6)}(x) = x_2^3\ \partial/\partial x_3$, $Y = Y^{(0)}$, and the approximating system $\dot{x} = X^{(1)}(x) + u(t)Y$ is *not* STLC. □

This example illustrates the importance of the choice of dilation. This choice, and the proper choice of local coordinates, will be dealt with in §2.

Example 1.2. On $\mathbb{R}^3$, consider the system

$$\dot{x}_1 = u, \quad \dot{x}_2 = x_1^3 + x_1^2 x_3, \quad \dot{x}_3 = x_3^2 + x_2^5, \tag{5}$$

i.e., $X(x) = (x_1^3 + x_1^2 x_3)\ \partial/\partial x_2 + (x_3^2 + x_2^5)\ \partial/\partial x_3$, $Y = \partial/\partial x_1$. If we choose the standard dilation δ_ε^1, then the homogeneous approximating system is $\dot{x} = X^{(2)}(x) + uY^{(0)}$ with $X^{(2)}(x) = x_3^2\ \partial/\partial x_3$, $Y^{(0)} = Y$. This system is not STLC, and we gain no information.

If we choose δ_ε^r with $r = (1, 1, 3)$, then $x_1^3 \in P_3 = P_{r_2+3-1}$, $x_2^5 \in P_5 = P_{r_3+3-1}$, and the homogeneous approximating system is $\dot{x} = X^{(3)}(x) + uY^{(0)}$ with $X^{(3)}(x) = x_1^3\ \partial/\partial x_2 + x_2^5\ \partial/\partial x_3$. This approximating system is an "odd" system; the relevant brackets which are linearly independent at zero are Y, $(\mathrm{ad}^3 Y, X)$ and $(\mathrm{ad}^5\ (\mathrm{ad}^3 Y, X), X)$. Theorem 1 applies with $m > 1$, showing that the original system is LTLC.

However, had we chosen δ_ε^r with $r = (1, 3, 15)$, then $x_1^3 \in P_3 = P_{r_2+1-1}$, $x_2^5 \in P_{15} = P_{r_3+1-1}$, while the remaining terms are of higher homogeneous order. Relative to this dilation, the approximating system is $\dot{x} = X^{(1)}(x) + uY^{(0)}$ with $X^{(1)}(x) = x_1^3\ \partial/\partial x_2 + x_2^5\ \partial/\partial x_3$. This approximation is STLC; since the approximating vector field of X is homogeneous of degree one, Theorem 1 applies with $m = 1$, and system (5) is actually STLC. □

In §1 we have not changed local coordinates; i.e., in the examples the dilation was changed but the local coordinates relative to which the vector fields were described were not changed. Linearity of a vector field, in chosen local coordinates, is homogeneity of degree one with respect to the standard dilation δ_ε^1. But linearity is not a coordinate-free notion, and neither is homogeneity relative to a given dilation. Example 1.2 illustrates the role of choosing the "correct" dilation, but the approach was by guessing, i.e., not constructive. Section 2 will deal with a coordinate-free construction based on the bracket structure, at zero, of the original vector fields X, Y. If the construction can be accomplished, system (1) will be LTLC. Behind the scenes (as seen in the proof of Theorem 3) the construction leads to the proper local coordinates and dilation relative to which the approximating system should be considered.

2. A constructive test for large-time local controllability. Before stating and proving the main theorem of this section, we give an example to illustrate notation and ideas which will be involved in this theorem.

(1, 0)	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)
×	×	•	•	•	•
	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)
	×	•	•	•	•
	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)
	•	•	•	×	•
	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)
	•	•	•	•	•

FIG. 1.

Example 2.1. On $\mathbb{R}^3$, consider the system

$$\dot{x}_1 = u, \quad \dot{x}_2 = x_1, \quad \dot{x}_3 = x_1^2 + x_2^3, \tag{6}$$

i.e., $X(x) = x_1\,\partial/\partial x_2 + (x_1^2 + x_2^3)\partial/\partial x_3$, $Y = \partial/\partial x_1$. The relevant brackets are $[X, Y](x) = \partial/\partial x_2 + 2x_1\,\partial/\partial x_3$, $(\mathrm{ad}^2 Y, X) = 2\partial/\partial x_3$, and $(\mathrm{ad}^3[X, Y], X) = -6\,\partial/\partial x_3$. We denote a bracket with k factors Y and ℓ factors X to be a bracket of type (k, ℓ) and graph them on integer lattice points of the plane as in Figure 1.

Since small-time local controllability at zero implies large-time local controllability, one should certainly first check the constructive sufficient conditions for the former as given in [11]; see also [6], [7]. In this example, one does not have small-time local controllability; the (2,1) bracket gives an obstruction (see [10]).

The brackets in Figure 1 which do not vanish at zero are marked with an ×. In general, we begin by examining the first row (i.e., brackets of the form $(\mathrm{ad}^j X, Y)$, $j = 0, 1, 2, \ldots$) and let m be such that the brackets $(1, \ell)$, $\ell \geq (m + 1)$ are zero when evaluated at zero. Note that we may have brackets of type $(1, \ell)$, $\ell \leq m$, equal to zero when evaluated at zero. If a $(1, j)$ bracket is ever zero at zero, since $X(0) = 0$, we will have the $(1, \ell)$ brackets zero at zero for $\ell \geq j$. In the above example choose $m = 1$. Now for any integer $k \geq 1$, we consider a line through the point $(1, m) = (1, 1)$ with slope $-k/(mk + 1)$. In the above example, choose $k = 2$. If we can find such a line so that all brackets to its right vanish at zero, while if we slide the line to the left parallel to itself, the increases in rank of the set of brackets (evaluated at zero) encountered occur at brackets of type (odd, anything) or (even, even), i.e., "good" brackets in the sense of Sussmann [11], then the original system is LTLC. In this example, the original line encounters the brackets (1,1) and (3,4). As we slide the line to the left, the next bracket which is not zero at zero is (1,0), and these three, at zero, span $\mathbb{R}^3$. Thus system (6) is LTLC. □

This illustrates the idea, and an example, of the application of Theorem 3. The proof will proceed by showing that the existence of such a line implies that we can choose local coordinates and a dilation relative to which the approximating system of (1) has the form $\dot{x} = X^{(k+1)}(x) + uY^{(-mk)}(x)$, or for system (6) in the above example, $\dot{x} = X^{(3)}(x) + uY^{(-2)}(x)$, which is STLC at zero. Then large-time local controllability follows by Theorem 1.

Before stating and proving Theorem 3, some preliminary results on homogeneous vector fields are needed.

PROPOSITION 2.1. *If $V^{(m)}$ is a vector field homogeneous of degree m with respect to a dilation δ_ε^r and $V^{(m)}(x) = \sum_{j=1}^n a_j(x)\,\partial/\partial x_j$, then $a_j(0) = 0$ if $r_j \neq (1 - m)$.*

Proof. $V^{(m)}$ homogeneous of degree m implies $a_j \in P_{r_j+m-1}$. Recall that $P_\ell = \{0\}$ if $\ell < 0$; P_0 consists of constant polynomials, and polynomials in P_ℓ vanish at zero if $\ell > 0$. Thus $a_j(0) = 0$ except when $r_j = 1 - m$. □

PROPOSITION 2.2. *Let* $W(x) = \sum_{j=1}^{n} b_j(x)\,\partial/\partial x_j$ *and have expansion in terms of homogeneous vector fields relative to a dilation* δ_ε^r, $W(x) = W^{(m)}(x) + W^{(m+1)}(x) + \cdots$. *Let* $W^{(k)}(x) = \sum_{j=1}^{n} b_j^{(k)}(x)\,\partial/\partial x_j$. *Then* $b_j(0) = b_j^{(m)}(0)$ *for* $r_j = 1 - m$, $b_j(0) = b_j^{(m+1)}(0)$ *for* $r_j = -m$, $b_j(0) = b_j^{(m+2)}(0)$ *for* $r_j = -m - 1$, *etc.*

Proof. This follows immediately from Proposition 2.1.

PROPOSITION 2.3. X *is homogeneous of degree* m *if and only if* $Xh \in P_{k+m-1}$ *whenever* $h \in P_k$.

Proof. If $X(x) = \sum a_j(x)\,\partial/\partial x_j$ is homogeneous of degree m, $a_j \in P_{r_j+m-1}$, while $h \in P_k$ implies $\partial h/\partial x_j \in P_{k-r_j}$. Thus $Xh = \sum a_j\,\partial h/\partial x_j \in P_{k+m-1}$. Conversely, if $\sum a_j\,\partial h/\partial x_j \in P_{k+m-1}$ for $h \in P_k$, this implies $a_j \in P_{r_j+m-1}$ or X is homogeneous of degree m. □

PROPOSITION 2.4. *If* $X^{(m)}$ *is homogeneous of degree* m *and* $Y^{(\ell)}$ *is homogeneous of degree* ℓ, *then* $[X^{(m)}, Y^{(\ell)}]$ *is homogeneous of degree* $(m + \ell - 1)$.

Proof. By Proposition 2.3, for $h \in P_k$, $[X^{(m)}, Y^{(\ell)}]h = X^{(m)}(Y^{(\ell)}h) - Y^{(\ell)}(X^{(m)}h) \in P_{k+m+\ell-2} = P_{k+(m+\ell-1)-1}$, which implies that $[X^{(m)}, Y^{(\ell)}]$ is homogeneous of degree $(m + \ell - 1)$. □

DEFINITION. *An extended filtration,* $\mathcal{F}$, *of* $L(X, Y)$ *at zero is a sequence of subspaces* $\{F_j : -\infty < j < \infty\}$ *of* $L(X, Y)$ *such that for all integers* i, j

(i) $F_j \subset F_{j+1}$,

(ii) $[F_i, F_j] \subset F_{i+j}$,

(iii) $\bigcup_j F_j = L(X, Y)$,

(iv) $X \in F_j$ *with* $j \leq 0$ *implies* $X(0) = 0$.

Such filtrations of $L(X, Y)$ will be constructed as follows. Assign integer weights to X, Y, denoted wt X, wt Y (negative integers are permitted), and let the weight of a Lie product be the sum of the weights of its factors. The weight-induced filtration $\mathcal{F}$ then has subspace F_j consisting of all elements in $L(X, Y)$ having weight $i \leq j$. Note that condition (iv) puts a severe restriction on the admissible weights that one can assign to X and Y.

Filtration-induced local coordinates and dilations. Let $\mathcal{F} = \{F_j : -\infty < j < \infty\}$ be an extended filtration of $L(X, Y)$ at zero and $n_k = \dim F_k(0)$, $-\infty < k < \infty$. Property (iv) shows that $n_k = 0$ if $k \leq 0$, while $\dim L(X, Y)(0) = n$ means $\dim F_N(0) = n$ for some integer N. Choose $X_{\pi_1}, \ldots, X_{\pi_{n_1}} \in F_1$ such that they are linearly independent at zero. Adjoin $X_{\pi_{n_1+1}}, \ldots, X_{\pi_{n_2}} \in F_2$ such that $X_{\pi_1}(0), \ldots, X_{\pi_{n_2}}(0)$ are linearly independent, and continue in this fashion to get $X_{\pi_1}, \ldots, X_{\pi_n}$ with

$$X_{\pi_i} \in F_j, \qquad n_{j-1} + 1 \leq i \leq n_j. \tag{7}$$

Let $j \geq 1$ be the smallest integer such that $n_j \neq 0$. Choose $r_i = j$ for $1 \leq i \leq n_j$, $r_i = j + 1$ for $n_j \leq i \leq n_{j+1}$, and so on. The dilation δ_ε^r with $r = (r_1, \ldots, r_n)$ chosen as above is called the *filtration-induced dilation.* In a specific problem, vector fields are initially given relative to some local coordinates—say, $x = (x_1, \ldots, x_n)$—for a neighborhood of zero. Define a local coordinate change $y = \varphi^{-1}(x)$, where

$$x = \varphi(y) = (\exp y_1 X_{\pi_1}) \circ \cdots \circ (\exp y_n X_{\pi_n})(0). \tag{8}$$

Then φ is a local diffeomorphism with $\varphi(0) = 0$, and the coordinates $y = (y_1, \ldots, y_n)$ are *local coordinates induced* by the filtration.

As mentioned in the introduction, in some recent papers a vector field $X(x) = \sum_{j=1}^{n} a_j(x)\,\partial/\partial x_j$ was defined to be homogeneous of degree m if $a_j \in P_{r_j-m}$, $j = 1, \ldots, n$. We have instead changed the definition to $a_j \in P_{r_j+m-1}$ in order to be in keeping with classical terminology.

The next theorem is merely a translation of [4, Thm. 2.1] from the previous definition of vector field homogeneity to the one being used here.

THEOREM 2 (see [4, Thm. 2.1]). *Let $\mathcal{F} = \{F_j : -\infty < j < \infty\}$ be an extended filtration at zero for $L(X, Y)$ with $y = (y_1, \ldots, y_n)$ induced local coordinates and δ_ε^r the induced dilation. Then if $X \in F_{1-m}$,*

$$X(y) = X^{(m)}(y) + X^{(m+1)}(y) + \cdots, \tag{9}$$

where $X^{(j)}$ is homogeneous of degree j with respect to δ_ε^r.

Now assume that a lattice of bracket types (k, ℓ) as in Figure 1 has been drawn. The main result of this section is the following theorem.

THEOREM 3. *Assume* $\dim L(X, Y)(0) = n$ *for system* (1) *and that there exists an m such that the bracket $(1, m+1)$, evaluated at zero, is zero (and hence this will be true for brackets of type $(1, \ell)$, $\ell \geq m+1$, since $X(0) = 0$). If for some integer $k \geq 1$ one can draw a line through the $(1, m)$ point in the lattice with slope $-k/(mk+1)$ such that*

(a) *all brackets to the right of the line vanish at zero. All (even, odd) brackets on the line vanish at zero.*

(b) *as you slide the line to the left parallel to itself, each time that you reach a bracket of type (even, odd) which when evaluated at zero is not zero, its value at zero is a linear combination of brackets (at zero) to the right of the line.*

Then system (1) *is LTLC.*

Remark 1. If we were to allow $k = 0$, the line would have slope zero. Condition (a) is vacuously satisfied with "to the right" replaced as above. If condition (b) is satisfied as you slide the line down (rather than to the left), the original system would be STLC at zero.

Remark 2. If $m = 0$, i.e., $(\text{ad } X, Y)(0) = 0$, one can consider a line through the (1,0) point of slope $-k$ with $k \geq 1$ arbitrary. However, if $m \geq 1$, the limiting (but not attainable) possible slope is $-1/m$. Indeed, the possible slope values lie in the interval $(-1/m, -1/(m+1))$. Choosing m large may be advantageous.

Remark 3. There is no loss of generality in assuming that the local coordinates are such that $Y = \partial/\partial x_1$ and, using some feedback if necessary, $X(x) = \sum_{j=2}^n a_j(x)\partial/\partial x_j$. This ensures that $\partial/\partial x_1$ cannot also occur as some nontrivial product of the vector fields X, Y evaluated at zero. Thus if we expand Y as a sum of homogeneous vector fields relative to a (weight) filtration-induced dilation as in Theorem 2, the leading term will not vanish at zero.

Proof. Assign $\text{wt } X = -k$, $\text{wt } Y = mk + 1$. This makes the weight of the $(1, m)$ bracket one; brackets of types (i, j) which lie on the line will have weight one, and those to the right of the line will have weights less than or equal to zero and hence belong to F_j with $j \leq 0$, where $\mathcal{F} = \{F_j : -\infty < j < \infty\}$ is our candidate for a weight-induced filtration. But by (a), such brackets vanish at zero; hence condition (iv) of an extended filtration at zero is satisfied, and our weight assignment does give such a filtration. Brackets of type (i, j) which lie on or to the left of the line will have positive weights, and their values at zero could be nonzero.

Let δ_ε^r be the filtration-induced dilation and $y = (y_1, \ldots, y_n)$ be the induced local coordinates. By Theorem 3, in these coordinates and relative to this dilation, we can write

$$\begin{aligned} X(y) &= X^{(k+1)}(y) + X^{(k+2)}(y) + \cdots, \\ Y(y) &= Y^{(-mk)}(y) + Y^{(1-mk)}(y) + \cdots. \end{aligned} \tag{10}$$

The homogeneous approximation of system (1) is then

$$\dot{y} = X^{(k+1)}(y) + uY^{(-mk)}(y) \tag{11}$$

with $Y^{(-mk)}(0) \neq 0$ by Remark 3. The proof proceeds by showing that condition (b) implies that system (11) is STLC at zero. In particular, we first show that if W is an ordered product of X, Y of (even, odd) type which does not vanish at zero but is a linear combination of brackets

(evaluated at zero) already encountered, then this same ordered product of $X^{(k+1)}$, $Y^{(-mk)}$ vanishes at zero.

LEMMA 1. *Let $W \in F_k$; hence in the induced local coordinates, and relative to the induced dilation, by Theorem* 2

$$W(y) = W^{(1-k)}(y) + W^{(2-k)}(y) + \cdots.$$

If $W(0) \neq 0$ but is a linear combination of elements in $F_\ell(0)$ for $\ell < k$, then $W^{(1-k)}(0) = 0$.

Proof. Suppose $W(y) = \sum_{j=1}^n b_j(y)\, \partial/\partial y_j$. Then $W(0)$, a linear combination of elements of $F_\ell(0)$ for $\ell < k$, implies $b_j(0) = 0$ for $r_j = k$ by Proposition 2.2. Now let $W^{(m-k)}(y) = \sum_{j=1}^n b_j^{(m)}(y)\, \partial/\partial y_j$, $m = 1, 2, \ldots$. From Proposition 2.1 we see $b_j^{(m)}(0) = 0$ except if $r_j = 1 + k - m$. In other words, $b_j(0) = b_j^{(m)}(0)$ for $r_j = 1 + k - m$. Then with $m = 1$ we have $0 = b_j(0) = b_j^{(1)}(0)$ for $r_j = k$ and hence $W^{(1-k)}(0) = 0$. □

We continue with the proof of Theorem 3. Suppose that W is an ordered product of X, Y of type (s, ℓ) and hence of weight $(smk + s - \ell k)$ which we call m_1, i.e., $W \in F_{m_1}$ and

$$W(y) = W^{(1-m_1)}(y) + W^{(2-m_1)}(y) + \cdots. \tag{12}$$

Then the term of lowest homogeneous order, i.e., $W^{(1-m_1)}$, is the same ordered product of $X^{(k+1)}$ and $Y^{(-mk)}$. Hypothesis (b), together with Lemma 1, shows that any product of the vector fields $X^{(k+1)}$, $Y^{(-mk)}$ of type (even, odd) vanishes at zero.

LEMMA 2. $\dim L(X, Y)(0) = n$ *implies* $\dim L(X^{(k+1)}, Y^{(-mk)})(0) = n$.

Proof. Let $W \in F_\ell$ be a product of factors X, Y such that $W(0) \notin F_{\ell-1}(0)$. As in (12), expand W as $W(y) = W^{(1-\ell)}(y) + W^{(2-\ell)}(y) + \cdots$. Then $W^{(1-\ell)}$ will be the same ordered product of $X^{(k+1)}$, $Y^{(-mk)}$ as W was of X, Y. From Proposition 2.2, if we write $W(0) = \xi + \eta$ with ξ in the quotient space $F_\ell(0)/F_{\ell-1}(0)$, $\eta \in F_{\ell-1}(0)$, then $W^{(1-\ell)}(0) = \xi$. If we assign $\operatorname{wt} X^{(k+1)} = -k$, $\operatorname{wt} Y^{(-mk)} = mk + 1$ and let $\mathcal{G} = \{G_j : -\infty < j < \infty\}$ be the weight-induced filtration of $L(X^{(k+1)}, Y^{(-mk)})$, this shows $\dim G_j(0) = \dim F_j(0)$ for all j, from which the result follows. □

To complete the proof of Theorem 3, we again note that hypothesis (b) and Lemma 1 imply that for the approximating system (11), the (even, odd) brackets evaluated at zero are zero. Since $\dim L(X^{(k+1)}, Y^{(-mk)})(0) = n$, Sussmann's theorem [11] applied to the approximating system shows that it is STLC at zero. Theorem 1 now applies to make system (1) LTLC. □

We next give another example to illustrate the use of Theorem 3 to show large-time local controllability. Furthermore, to see how the details of the proof of this theorem work in a specific example, we compute the filtration-induced local coordinates and dilation and the approximating system relative to them. Then Lemma 1 can be illustrated as we check small-time local controllability of the approximating system.

Example 2.2. On $\mathbb{R}^5$ we consider a system of the form (1) with

$$X(x) = x_1 \frac{\partial}{\partial x_2} + (x_1^4 + x_1^2 x_4)\frac{\partial}{\partial x_3} + x_5 \frac{\partial}{\partial x_4} + x_1^3 \frac{\partial}{\partial x_5}, \qquad Y(x) = \frac{\partial}{\partial x_1}. \tag{13}$$

The relevant brackets, which do not vanish at zero, are

$$Y(0) = \frac{\partial}{\partial x_1}, \quad (\operatorname{ad} X, Y)(0) = \frac{\partial}{\partial x_2}, \quad (\operatorname{ad}^3 Y, X)(0) = -6\frac{\partial}{\partial x_5},$$

$$(\operatorname{ad}^4 Y, X)(0) = \frac{\partial}{\partial x_3}, \quad [(\operatorname{ad}^3 Y, X), X](0) = 6\frac{\partial}{\partial x_4}, \quad \text{and}$$

$$V(0) = [(\operatorname{ad}^3 Y, X),\ [(\operatorname{ad}^2 Y, X),\ X]](0) = 12\frac{\partial}{\partial x_3}.$$

These are, respectively, of types (1,0), (1,1), (3,1), (4,1), (3,2), and (5,3).

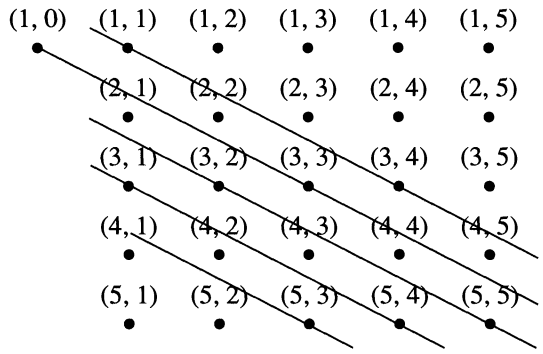


FIG. 2.

Here the (1,1) bracket is linearly independent of the (1,0) bracket, while brackets of the form $(1, \ell)$, $\ell \geq 2$, vanish at zero. We choose $m = 1$. The "bad" bracket is that of type (4,1), and we see that we need a line having a slope so that the (5,3) bracket is encountered before the (4,1). Choose $k = 2$, giving a line of slope $-2/3$ through the point (1,1). As seen in Figure 2, this line meets the requirements of Theorem 3, and the system is LTLC.

It is instructive to essentially follow the steps of the proof of Theorem 3 for this example. For $m = 1$ and $k = 2$, the weight assignments are $\operatorname{wt} X = -2$ and $\operatorname{wt} Y = 3$. Then $\operatorname{wt}(\operatorname{ad} X, Y) = 1$, $\operatorname{wt}(\operatorname{ad}^3 Y, X) = 7$, $\operatorname{wt}(\operatorname{ad}^4 Y, X) = 10$, $\operatorname{wt}[(\operatorname{ad}^3 Y, X), X] = 5$, and $\operatorname{wt}[(\operatorname{ad}^3 Y, X), [(\operatorname{ad}^2 Y, X), X]] = 9$. The weight-induced filtration is δ_ε^r with $r = (1, 3, 5, 7, 9)$. Also, $X_{\pi_1} = (\operatorname{ad} X, Y) \in F_1$, $X_{\pi_2} = Y \in F_3$, $X_{\pi_3} = [(\operatorname{ad}^3 Y, X), X] \in F_5$, $X_{\pi_4} = (\operatorname{ad}^3 Y, X) \in F_7$, and $X_{\pi_5} = [(\operatorname{ad}^3 Y, X), [(\operatorname{ad}^2 Y, X), X]] \in F_9$. The filtration-induced local coordinates are obtained from the diffeomorphism

$$x = \varphi(y) = (\exp y_1 X_{\pi_1}) \circ \cdots \circ (\exp y_5 X_{\pi_5})(0) = \begin{pmatrix} y_2 \\ y_1 \\ 4y_2^3 y_1 + 12 y_1 y_2 y_3 + 12 y_5 \\ 6y_3 \\ 3y_2^2 y_1 - 6y_4 \end{pmatrix}.$$

Abusing notation by letting $X(y)$, $Y(y)$ again denote the original vector fields X, Y in the new y-coordinates, the system $\dot{x} = X(x) + uY$ with $X(x)$, $Y(x)$ given by (13) transforms into

$$\dot{y} = \begin{pmatrix} y_2 \\ 0 \\ (1/2) y_1 y_2^2 - y_4 \\ y_2^3/3 \\ (y_2^4/12) - (y_2^2 y_3/2) - (y_1^2 y_2^3/2) + y_1 y_2 y_4 \end{pmatrix} + u \begin{pmatrix} 0 \\ 1 \\ 0 \\ y_1 y_2 \\ -y_1 y_2^2 - y_1 y_3 \end{pmatrix} \tag{14}$$
$$= X(y) + uY(y).$$

Then relative to the filtration-induced dilation, $X(y) = X^{(3)}(y) + X^{(4)}(y)$, $Y(y) = Y^{(-2)}(y) + Y^{(-1)}(y)$, where

$$X^{(3)}(y) = y_2 \frac{\partial}{\partial y_1} + ((y_1 y_2^2/2) - y_4) \frac{\partial}{\partial y_3} + (y_2^3/3) \frac{\partial}{\partial y_4},$$
$$+ (y_1 y_2 y_4 - (y_2^2 y_3/2) - (y_1^2 y_2^3/2)) \frac{\partial}{\partial y_5},$$

$$X^{(4)}(y) = (y_2^4/12)\frac{\partial}{\partial y_5},$$

$$Y^{(-2)}(y) = \frac{\partial}{\partial y_2} + y_1 y_2 \frac{\partial}{\partial y_4} - y_1 y_3 \frac{\partial}{\partial y_5},$$

$$Y^{(-1)}(y) = -y_1 y_2^2 \frac{\partial}{\partial y_5}.$$

The homogeneous approximating system is $\dot{y} = X^{(3)}(y) + uY^{(-2)}(y)$, which is STLC (it is an "odd" system). One should note that $(\mathrm{ad}^4 Y^{(-2)}, X^{(3)})(0) = 0$ as promised by Lemma 1; i.e., the approximation is such that the "bad brackets" of the original system vanish at zero for the approximating system.

Example 2.3. On $\mathbb{R}^5$ consider system (1) with

$$X(x) = x_1 \frac{\partial}{\partial x_2} + (x_1^4 + x_1^2 x_5)\frac{\partial}{\partial x_3} + x_5 \frac{\partial}{\partial x_4} + x_1^3 \frac{\partial}{\partial x_5}, \qquad Y = \frac{\partial}{\partial x_1}.$$

The relevant brackets not vanishing at zero are $Y(0) = \partial/\partial x_1$, $(\mathrm{ad}\, X, Y)(0) = \partial/\partial x_2$, $(\mathrm{ad}^3 Y, X)(0) = -6\ \partial/\partial x_5$, $[(\mathrm{ad}^3 Y, X), X](0) = 6\ \partial/\partial x_4$, $[(\mathrm{ad}^3 Y, X), (\mathrm{ad}^2 Y, X)](0) = -12\ \partial/\partial x_3$, and $(\mathrm{ad}^4 Y, X)(0) = 24\ \partial/\partial x_3$. These are, respectively, of types (1,0), (1,1), (3,1), (3,2), (5,2), and (4,1). Plotting them on Figure 2, one sees that no line through a $(1, m)$, $m \geq 1$ point, with slope $-1/m < -k/(km+1) \leq -1/(m+1)$, can encounter the (5,2) bracket before the (4,1). Thus Theorem 3 will not apply. □

REFERENCES

[1] R. W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, Progr. Math. 27, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 181–191.

[2] R. Bianchini and G. Stephani, *Graded approximations and controllability along a trajectory*, SIAM J. Control Optim., 28 (1990), pp. 903–924.

[3] W. Dayawansa and C. Martin, *Asymptotic stabilization of low dimensional systems*, in Progress in Systems and Control Theory, C. I. Byrnes and A. Kurzhansky, eds., Birkhäuser, Boston, 1991, pp. 53–67.

[4] H. Hermes, *Nilpotent and high-order approximations of vector field systems*, SIAM Rev., 33 (1991), pp. 238–264.

[5] ———, *Control systems which generate decomposable Lie algebras*, J. Differential Equations, 44 (1982), pp. 166–187.

[6] H. Hermes and M. Kawski, *Local controllability of a single-input affine system*, in Nonlinear Analysis and Applications, Lecture Notes in Pure and Appl. Math. 109, V. Lakshmikantham, ed., Marcel Dekker, New York, 1987, pp. 235–248.

[7] M. Kawski, *High-order small-time local controllability*, in Nonlinear Controllability and Optimal Control, Lecture Notes in Pure and Appl. Math. 133, H. J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 431–467.

[8] L. P. Rothschild and E. Stein, *Hypoelliptic differential operators and nilpotent groups*, Acta Math., 137 (1976), pp. 247–320.

[9] G. Stefani, *Polynomial approximations to control systems and local controllability*, Proc. 24th IEEE Conf. on Decision and Control, I (1985), pp. 33–38.

[10] ———, *Local properties of nonlinear control systems*, in Geometric Theory of Nonlinear Control Systems, B. Jakubczyk, W. Respondek, and K. Tchon, eds., Tech. Univ. Wroclaw, 1984, pp. 219–226.

[11] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1300–1314, July 1996

010

EQUIVALENT SUBGRADIENT VERSIONS OF HAMILTONIAN AND EULER–LAGRANGE EQUATIONS IN VARIATIONAL ANALYSIS*

R. TYRRELL ROCKAFELLAR†

Abstract. Much effort in recent years has gone into generalizing the classical Hamiltonian and Euler–Lagrange equations of the calculus of variations so as to encompass problems in optimal control and a greater variety of integrands and constraints. These generalizations, in which nonsmoothness abounds and gradients are systematically replaced by subgradients, have succeeded in furnishing necessary conditions for optimality that reduce to the classical ones in the classical setting, but important issues have remained unsettled, especially concerning the exact relationship of the subgradient versions of the Hamiltonian equations versus those of the Euler–Lagrange equations. Here it is shown that new, tighter subgradient versions of these equations are actually equivalent to each other. The theory of epi-convergence of convex functions provides the technical basis for this development.

Key words. Euler–Lagrange equations, Hamiltonian equations, variational analysis, nonsmooth analysis, subgradients, optimality

AMS subject classifications. 49B10, 49B34

1. Introduction.

In the classical theory of minimization problems involving an integral functional $\int_{t_0}^{t_1} L\big(t, x(t), \dot{x}(t)\big)dt$ with Lagrangian expression $L(t, x, v)$ on $[t_0, t_1] \times \mathbb{R}^n \times \mathbb{R}^n$, a key role in analyzing the optimality of an arc $x(\cdot) : [t_0, t_1] \to \mathbb{R}^n$ is played by the Euler–Lagrange equation

$$\dot{p}(t) = \nabla_x L\big(t, x(t), \dot{x}(t)\big) \quad \text{for} \quad p(t) = \nabla_v L\big(t, x(t), \dot{x}(t)\big). \tag{1.1}$$

When $L(t, x, v)$ is twice differentiable and the Hessian matrix in v is positive definite, the Legendre transform can be applied in the v argument to get a Hamiltonian $H(t, x, p)$ in terms of which the Euler–Lagrange equation can be expressed equivalently as the Hamiltonian system

$$\dot{x}(t) = \nabla_p H\big(t, x(t), p(t)\big), \qquad -\dot{p}(t) = \nabla_x H\big(t, x(t), p(t)\big). \tag{1.2}$$

The differentiability assumptions in this scheme have long posed difficulties, however.

Many problems of interest fail to meet all the criteria for utilizing the Legendre transform; in such cases (1.2) may only be a consequence of (1.1), not equivalent to it. Then the arcs $x(\cdot)$ and $p(\cdot)$ can have "corners" where their derivatives are discontinuous. Tonelli's theory for the existence of optimal arcs demands an even broader setting; that is, problems must be studied with $x(\cdot)$ merely assumed to be absolutely continuous, so that (1.1) and (1.2), to the degree that they are valid, have to be interpreted in an almost everywhere sense.

Questions of existence have also challenged the suitability of classical assumptions in other ways. Tonelli showed that the convexity of $L(t, x, v)$ in v is a crucial property. If this is lacking, a convexification process can be introduced to achieve it as a justifiable sort of regularization (or relaxation) of a given problem, but convexification can disrupt differentiability. Thus, Lagrangians L need to be admitted for which certain derivatives may be absent. The theory of optimal control has pushed this direction of generalization much further through the recognition that a vast range of applications can be covered "neoclassically" in terms of Lagrangians that are not even continuous everywhere and can take on the value ∞, as a device for representing constraints on $x(t)$ and $\dot{x}(t)$ through infinite penalization when they are violated.

*Received by the editors August 26, 1994; accepted for publication (in revised form) March 13, 1995. This work was supported in part by National Science Foundation grant DMS–9200303 at the University of Washington and by U.S.–Israel Science Foundation grant 90-00455.

†Department of Mathematics, University of Washington, 354350, Seattle, WA 98115-4350.

As far as possible in the face of this far-reaching extension of the classical framework, one would nonetheless like to make sense of the Euler–Lagrange and Hamiltonian equations as necessary conditions for optimality. The Hamiltonian can always be defined by appealing to the Legendre–Fenchel transform of convex analysis [1] instead of the Legendre transform as

$$H(t, x, p) = \sup_{v \in \mathbb{R}^n} \{\langle p, v\rangle - L(t, x, v)\}, \tag{1.3}$$

where $\langle p, v\rangle$ denotes the inner product of two vectors p and v in $\mathbb{R}^n$. Provided that $L(t, x, v)$ as a function of v is convex and lower semicontinuous, one has

$$L(t, x, v) = \sup_{p \in \mathbb{R}^n} \{\langle p, v\rangle - H(t, x, p)\}, \tag{1.4}$$

so that a one-to-one correspondence is set up between Lagrangians and Hamiltonians without calling for their differentiability. In the possible absence of gradients of L and H, the idea is to try to rewrite (1.1) and (1.2) in terms of some kind of "subgradients."

This program was first carried out in the fully convex case, where $L(t, x, v)$ is convex as a function of (x, v) (rather than just v), which corresponds to $H(t, x, p)$ being concave in x and convex in p. Subgradients of convex analysis were used by Rockafellar [2], [3], [4] to establish an Euler–Lagrange condition

$$\big(\dot{p}(t), p(t)\big) \in \partial L\big(t, x(t), \dot{x}(t)\big) \quad \text{a.e. } t \tag{1.5}$$

and a Hamiltonian condition

$$\big(-\dot{p}(t), \dot{x}(t)\big) \in \partial H\big(t, x(t), p(t)\big) \quad \text{a.e. } t, \tag{1.6}$$

where in (1.5) the subgradients are those of $L(t, \cdot, \cdot)$ as a convex function while in (1.6) they are those of $H(t, \cdot, \cdot)$ in the special sense employed for concave-convex functions. The equivalence of these Euler–Lagrange and Hamiltonian conditions was shown through the dualization rules for subgradient relations in convex analysis.

In a major advance, Clarke [5], [6] developed a robust concept of subgradient that could serve for nonconvex functions and be used in pushing the Euler–Lagrange and Hamiltonian conditions further. This concept has evolved considerably since its introduction, both in the pattern of definition and the role of the convex hull operation. The subgradients in question can now be described in several ways, but for purposes here it is easiest to start with proximal subgradients and then take limits.

Consider a function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ (where $\overline{\mathbb{R}}$ denotes the extended reals). A vector z is a *proximal subgradient* of f at $\bar{y}$ if $f(\bar{y})$ is finite, and for some $\rho \geq 0$ and $\delta > 0$ one has

$$f(y) \geq f(\bar{y}) + \langle z, y - \bar{y}\rangle - \tfrac{1}{2}\rho|y - \bar{y}|^2 \quad \text{when} \quad |y - \bar{y}| \leq \delta.$$

It is a *subgradient in the general sense*, expressed by $z \in \partial f(\bar{y})$, if there are sequences $y^\nu \to \bar{y}$ and $z^\nu \to z$ such that z^ν is a proximal subgradient of f at y^ν and $f(y^\nu) \to f(\bar{y})$. It is a *subgradient in the horizon sense*, expressed by $z \in \partial^\infty f(\bar{y})$, if this condition holds with the modification that, instead of $z^\nu \to z$, one has $\lambda^\nu z^\nu \to z$ for some sequence of scalars $\lambda^\nu \searrow 0$. (In these expressions and below, we use the superscript ν as the generic index for sequences.)

When f is continuously differentiable, $\partial f(\bar{y})$ consists of just the gradient $\nabla f(\bar{y})$, while $\partial^\infty f(\bar{y})$ has just the zero vector. When f is convex, $\partial f(\bar{y})$ is a closed, convex set, the same as the subgradient set of convex analysis, which if nonempty has $\partial^\infty f(\bar{y})$ as its recession cone. In general, however, $\partial f(\bar{y})$ and $\partial^\infty f(\bar{y})$ are not convex, although they are always closed. Seeking

a subgradient set that always would be both closed and convex, Clarke, although his notation was different and his definition followed an alternate route, ended up with the set

$$\bar{\partial} f(\bar{y}) \;=\; \text{cl con}\big[\partial f(\bar{y}) + \partial^{\infty} f(\bar{y})\big],$$

where "cl" stands for closure and "con" for convex hull. He especially emphasized the case where f is Lipschitz continuous around $\bar{y}$; then $\partial f(\bar{y})$ is a nonempty compact set, whereas $\partial^{\infty} f(\bar{y}) = \{0\}$, so the formula simplifies to $\bar{\partial} f(\bar{y}) = \text{con}\,\partial f(\bar{y})$. See Loewen [7] for a recent exposition furnishing the details.

Nowadays the convexification in this definition is no longer seen as essential for most applications, thanks to improvements in subgradient calculus achieved by Mordukhovich, Ioffe, and others. In the treatment of the class of problems under discussion here, which was Clarke's chief concern, it has a natural genesis in taking weak limits, however, and the question of the extent to which it is needed has been harder to answer.

With full recourse to such convexification, Clarke was able to demonstrate in some situations where $L(t, \cdot, \cdot)$ is locally Lipschitz continuous [8] the necessity of the Euler–Lagrange condition in the form

$$\big(\dot{p}(t), p(t)\big) \in \bar{\partial} L\big(t, x(t), \dot{x}(t)\big) \quad \text{a.e. } t, \tag{1.7}$$

where the subgradient set $\bar{\partial} L\big(t, x(t), \dot{x}(t)\big)$ refers to the function $L(t, \cdot, \cdot)$ at $\big(x(t), \dot{x}(t)\big)$ and, because of the Lipschitz continuity, equals $\text{con}\,\partial L\big(t, x(t), \dot{x}(t)\big)$. On the other hand, he established in some other situations [9] where $H(t, \cdot, \cdot)$ is locally Lipschitz continuous the necessity of the Hamiltonian condition in the form

$$\big(-\dot{p}(t), \dot{x}(t)\big) \in \bar{\partial} H\big(t, x(t), p(t)\big) \quad \text{a.e. } t, \tag{1.8}$$

where the subgradient set $\bar{\partial} H\big(t, x(t), \dot{x}(t)\big)$ is that of $H(t, \cdot, \cdot)$ at $\big(x(t), p(t)\big)$ and, again because of the Lipschitz continuity, is the same as $\text{con}\,\partial H\big(t, x(t), p(t)\big)$. (See [6], [10] for an overview of this development.)

Although Clarke's conditions (1.7) and (1.8) reduce to (1.1) and (1.2) in the classical case and to (1.5) and (1.6) in the convex case, and then are equivalent, neither necessarily implies the other in general, even when both $L(t, \cdot, \cdot)$ and $H(t, \cdot, \cdot)$ are locally Lipschitz continuous. Their precise relationship has therefore been a mystery.

Loewen and Rockafellar [11] showed, in building on Clarke's results, that for a major class of problems the Euler–Lagrange condition (1.7) and Hamiltonian condition (1.8) do at least have to hold simultaneously for some arc $p(\cdot)$ when $x(\cdot)$ is optimal. Rockafellar proved in [12, Thm. 5.1] that when $H(t, \cdot, \cdot)$ is locally Lipschitz continuous the Hamiltonian condition implies

$$\big(\dot{p}(t), \dot{x}(t)\big) \in \text{con}\,\big\{(-w, v) \;\big|\; \big(w, p(t)\big) \in \partial L(t, x(t), v),\ \ p(t) \in \partial_v L(t, x(t), v)\big\}, \tag{1.9}$$

which is a form of the Euler–Lagrange condition suggested by Mordukhovich [13], [14], [15]. For Hamiltonians arising from bounded differential inclusions, Ioffe [16] established that this implication is an equivalence. Also identified in Rockafellar [12, Thm. 3.4] is a broadly applicable case, beyond the known classical and convex ones, where (1.7) and (1.8) are equivalent even with $\bar{\partial} L$ replaced by ∂L.

More recent work of Loewen and Rockafellar in [17] raised the possibility of establishing the Euler–Lagrange condition in the form

$$\dot{p}(t) \in \text{con}\,\big\{w \;\big|\; \big(w, p(t)\big) \in \partial L\big(t, x(t), \dot{x}(t)\big)\big\} \quad \text{a.e. } t \tag{1.10}$$

with the companion property that

$$p(t) \in \partial_v L\big(t, x(t), \dot{x}(t)\big) \quad \text{a.e. } t. \tag{1.11}$$

They were able to do this in a case where L is the indicator of a possibly unbounded differential inclusion, which should allow extension to other Lagrangians L through consideration of epigraphical mappings. This case also covers, for instance, the case where L is the indicator of a Lipschitz continuous differential inclusion of the kind underlying Clarke's Hamiltonian results. Such an Euler–Lagrange condition has also been obtained by Mordukhovich [18] for a class of nonconvex differential inclusions and by Ioffe and Rockafellar [19] for certain finite functions L. In the special case where $L(t, x, v)$ is essentially strictly convex in v, which corresponds in the theory of the Legendre–Fenchel transform to $H(t, x, p)$ being smooth in p, a case used as a technical stepping stone in [17], (1.9) comes out as saying the same thing as (1.10) and (1.11). In general, though, the combination of (1.10) with (1.11) is distinctly sharper than the versions of Euler–Lagrange in (1.7) and (1.9) because the process of convexification is much more limited.

Here we sidestep the exploration of the full range of situations in which the Euler–Lagrange condition in the form of (1.10) might be necessary for the optimality of an arc $x(\cdot)$. Instead we focus on the relationship between (1.10) and a corresponding version of the Hamiltonian condition, namely,

$$\dot{p}(t) \in \operatorname{con}\big\{w \,\big|\, \big(-w, \dot{x}(t)\big) \in \partial H\big(t, x(t), p(t)\big)\big\} \quad \text{a.e. } t \tag{1.12}$$

along with

$$\dot{x}(t) \in \partial_p H\big(t, x(t), p(t)\big) \quad \text{a.e. } t. \tag{1.13}$$

This is sharper than the Hamiltonian condition (1.8) and has not previously been considered. We'll show it is in fact equivalent to (1.10) in the kinds of circumstances that are typically present in derivations of necessary conditions for the optimality of an arc $x(\cdot)$. Efforts aimed at enlarging the range of cases in which the Euler–Lagrange condition holds in the version (1.10) can thus count on the side benefit of improving Clarke's Hamiltonian condition in a hitherto unsuspected way.

The following theorem is our main result. Through [17] it brings to light, among other things, that (1.8) can be strengthened to (1.12) in Clarke's context [9].

In stating this theorem, we say that the Lagrangian L has the *epi-continuity property along* $x(\cdot)$ if, for almost every t, there is an open set $O(t)$ containing $x(t)$ such that:

(a) $L(t, \cdot, \cdot)$ is lower semicontinuous on $O(t) \times \mathbb{R}^n$;

(b) for every point $(\bar{x}, \bar{v}) \in O(t) \times \mathbb{R}^n$ with $L(t, \bar{x}, \bar{v}) < \infty$, and every sequence $x^\nu \to \bar{x}$, there is a sequence $v^\nu \to \bar{v}$ with $L(t, x^\nu, v^\nu) \to L(t, \bar{x}, \bar{v})$.

Clearly (a) and (b) are satisfied in particular when $L(t, \cdot, \cdot)$ is continuous on $O(t) \times \mathbb{R}^n$.

THEOREM 1.1. *Let $L(t, x, v)$ be convex in v (possibly with the value ∞), and let $H(t, x, p)$ be defined by* (1.3). *Let $x(\cdot)$ be an arc along which L has the epi-continuity property. Suppose for almost every t that*

$$(w, 0) \in \partial^\infty L\big(t, x(t), \dot{x}(t)\big) \implies w = 0, \tag{1.14}$$

this being true in particular if $L(t, \cdot, \cdot)$ is Lipschitz continuous around $\big(x(t), \dot{x}(t)\big)$. Then version (1.10) *of the Euler–Lagrange condition is equivalent to version* (1.12) *of the Hamiltonian condition and automatically entails* (1.11) *and* (1.13). *The same holds when* (1.14) *is replaced by*

$$(w, 0) \in \partial^\infty H\big(t, x(t), p(t)\big) \implies w = 0, \tag{1.15}$$

this being true in particular if $H(t, \cdot, \cdot)$ is Lipschitz continuous around $\big(x(t), p(t)\big)$.

The epi-continuity property invoked in Theorem 1.1 concerns the continuity of the set-valued mapping that associates with each x the epigraph of the function $L(t, x, \cdot)$, as will become clearer in the next section. Assumption (1.14) concerns a kind of localized Lipschitz continuity of this mapping. Such properties of epigraphical mappings have long been implicit in most developments of the subject, in consequence for instance of Lipschitz assumptions placed on L or H, or on some underlying differential inclusion mapping, but their effects on subgradients have not been explored directly. Here they emerge finally in the foreground. Also coming on stage for the first time in such a setting, through the technique we will use to prove Theorem 1.1, will be a number of tools of convex analysis. These include Fenchel's duality theorem in convex optimization, Moreau's theory of proximal regularizations of convex functions, Wijsman's epi-continuity theorem for the Legendre-Fenchel transform, and Attouch's theorem on convergence of subgradients.

2. Dualization framework and epi-continuity. For the task to be accomplished, the t argument doesn't matter; all questions revolve around properties that hold for a fixed t. We therefore suppress t. We consider an open subset O of $\mathbb{R}^m$ and take $L(x, v)$ to be an expression defined for $(x, v) \in O \times \mathbb{R}^n$ such that, for each $x \in O$, $L(x, \cdot)$ is a convex, lower semicontinuous (lsc) function on $\mathbb{R}^n$ that is *proper*, i.e., although possibly extended real-valued does not take on $-\infty$ and is not identically ∞. In the targeted applications to Euler–Lagrange and Hamiltonian conditions we'll have $m = n$, but for the sake of other potential uses of the results to be obtained we allow the dimensions m and n to differ. We define $H(x, p)$ for $(x, p) \in O \times \mathbb{R}^n$ by

$$H(x, p) = \sup_{v \in \mathbb{R}^n} \{\langle p, v\rangle - L(x, v)\}. \tag{2.1}$$

The Legendre–Fenchel transformation, on which this formula is based, has the property that for each $x \in O$, $H(x, \cdot)$ is, like $L(x, \cdot)$, a proper, convex, lsc function on $\mathbb{R}^n$, moreover with

$$L(x, v) = \sup_{p \in \mathbb{R}^n} \{\langle p, v\rangle - H(x, p)\}. \tag{2.2}$$

This symmetric relationship between L and H will enable us to apply any result proved for either function to the other function as well. We can later interpret O as a neighborhood of some particular point of $\mathbb{R}^m$ that happens to be under scrutiny.

The lower semicontinuity of $L(x, v)$ in v and of $H(x, p)$ in p has already been incorporated into our framework, but nothing has been said yet about continuity properties relative to x. At the very least we'll need $L(x, v)$ to be lsc in $(x, v) \in O \times \mathbb{R}^n$ and similarly for $H(x, p)$ in (x, p); but we're going to go further, clarifying along the way the property that provides the simplest dualization scheme and best supports our subsequent analysis. We'll be working with the concept of epi-continuity in the dependence of the functions $L(x, \cdot)$ and $H(x, \cdot)$ on x.

Recall that the *epigraph* of a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is the set

$$\operatorname{epi} f = \{(y, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid \alpha \geq f(y)\}.$$

In general, a sequence of functions $f^\nu : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to *epi-converge* to a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ if the corresponding epigraphs converge, meaning that

$$\operatorname{epi} f = \limsup_{\nu \to \infty} \operatorname{epi} f^\nu = \liminf_{\nu \to \infty} \operatorname{epi} f^\nu$$

in the Painlevé–Kuratowski sense as subsets of $\mathbb{R}^n \times \mathbb{R}$. This is true if and only if

$$\text{(2.3)} \qquad \begin{cases} \liminf_\nu f^\nu(v^\nu) \geq f(v) & \text{for every sequence } v^\nu \to v, \\ \limsup_\nu f^\nu(v^\nu) \leq f(v) & \text{for some sequence } v^\nu \to v. \end{cases}$$

Epi-convergence was introduced for proper, lsc, convex functions by Wijsman [20], who proved that the Legendre–Fenchel transformation was continuous with respect to it. For more background on this topic, see Wets [21] and Salinetti and Wets [22].

PROPOSITION 2.1. *The following six properties are equivalent and imply in particular that L and H are lsc in both arguments jointly, as functions on $O \times \mathbb{R}^n$.*

(a) *The set* epi $L(x, \cdot)$ *in* $\mathbb{R}^n \times \mathbb{R}$ *depends continuously on* $x \in O$.

(b) *Whenever* $x^\nu \to \bar{x}$ *in* O, *the function* $L(x^\nu, \cdot)$ *epi-converges to* $L(\bar{x}, \cdot)$.

(c) *For any* $(\bar{x}, \bar{v}) \in O \times \mathbb{R}^n$ *and sequence* $x^\nu \to \bar{x}$, *one has*

$$\begin{cases} \liminf_\nu L(x^\nu, v^\nu) \geq L(\bar{x}, \bar{v}) & \text{for every sequence } v^\nu \to \bar{v}, \\ \limsup_\nu L(x^\nu, v^\nu) \leq L(\bar{x}, \bar{v}) & \text{for some sequence } v^\nu \to \bar{v}. \end{cases}$$

(d) *The set* epi $H(x, \cdot)$ *in* $\mathbb{R}^n \times \mathbb{R}$ *depends continuously on* $x \in O$.

(e) *Whenever* $x^\nu \to \bar{x}$ *in* O, *the function* $H(x^\nu, \cdot)$ *epi-converges to* $H(\bar{x}, \cdot)$.

(f) *For any* $(\bar{x}, \bar{p}) \in O \times \mathbb{R}^n$ *and sequence* $x^\nu \to \bar{x}$, *one has*

$$\begin{cases} \liminf_\nu H(x^\nu, p^\nu) \geq H(\bar{x}, \bar{p}) & \text{for every sequence } p^\nu \to \bar{p}, \\ \limsup_\nu H(x^\nu, p^\nu) \leq H(\bar{x}, \bar{p}) & \text{for some sequence } p^\nu \to \bar{p}. \end{cases}$$

Proof. Conditions (a) and (b) mean the same, by the definition of epi-convergence, and (c) characterizes this property in accordance with the facts just cited. This pattern holds for (d), (e), and (f) as well. But because $L(x^\nu, \cdot)$ and $H(x^\nu, \cdot)$ are proper convex functions conjugate to each other under the Legendre–Fenchel transformation, which preserves epi-convergence according to Wijsman's theorem, (b) is equivalent to (e). Hence, all the conditions are equivalent to each other. □

For short, we'll say that the *epi-continuity assumption* is satisfied when the six equivalent properties in Proposition 2.1 are present. Obviously this is true in particular when L is continuous on $O \times \mathbb{R}^n$ (through (c)), or when H is continuous on $O \times \mathbb{R}^n$ (through (f)). In typical applications the epi-continuity assumption merely means (through property (c)) that, in addition to taking $L(x, v)$ to be lsc in (x, v), rather than just in x, we suppose that whenever $(\bar{x}, \bar{v})$ is a point of $O \times \mathbb{R}^n$ where L is finite and $\{x^\nu\}$ is a sequence in O converging to $\bar{x}$, there must be a sequence $\{v^\nu\}$ converging to $\bar{v}$ for which $L(x^\nu, v^\nu)$ converges to $L(\bar{x}, \bar{v})$.

Note that the epi-continuity condition used in the hypothesis of Theorem 1.1 merely requires for almost every t that this should hold relative to some neighborhood $O(t)$ of $x(t)$. Proposition 2.1 shows that the condition in question could be expressed in terms of the Hamiltonian just as well as the Lagrangian. It's actually symmetric between the two functions (as long as the Lagrangian is lsc and convex with respect to v).

The study of subgradients of L and H with respect to both of their arguments in $O \times \mathbb{R}^n$ requires working with the definition in §1 in terms of limits of proximal subgradients. But subgradients of L in the v argument and of H in the p argument enjoy the benefits of convexity. Convex analysis informs us that

$$\text{(2.4)} \qquad p \in \partial_v L(x, v) \iff v \in \partial_p H(x, p) \iff L(x, v) + H(x, p) = \langle p, v \rangle$$

(cf. [1, Thm. 23.5]), where from (2.1) and (2.2) we know that $L(x, v) + H(x, p) \geq \langle v, p \rangle$ for all choices of $(x, v, p) \in O \times \mathbb{R}^n \times \mathbb{R}^n$.

PROPOSITION 2.2. *Under the epi-continuity assumption,*

$$\begin{aligned} (w, p) \in \partial L(x, v) &\implies p \in \partial_v L(x, v), \\ (w, v) \in \partial H(x, p) &\implies v \in \partial_p H(x, p). \end{aligned} \tag{2.5}$$

Proof. Due to symmetry, it suffices to deal with the first of these implications. Suppose $(\bar{w}, \bar{p}) \in \partial L(\bar{x}, \bar{v})$. By definition there exist $(x^\nu, v^\nu) \to (\bar{x}, \bar{v})$ and $(w^\nu, p^\nu) \to (\bar{w}, \bar{p})$ such that (w^ν, p^ν) is a proximal subgradient of L at (x^ν, v^ν), and $L(x^\nu, v^\nu) \to L(\bar{x}, \bar{v})$. The proximal subgradient condition refers to the existence of $\rho^\nu \geq 0$ and $\delta^\nu > 0$ such that

$$L(x, v) \geq L(x^\nu, v^\nu) + \langle (w^\nu, p^\nu), (x, v) - (x^\nu, v^\nu) \rangle - \tfrac{1}{2}\rho^\nu \big(|x - x^\nu|^2 + |v - v^\nu|^2\big)$$

when $|(x - x^\nu, v - v^\nu)| \leq \delta^\nu$. In taking $x = x^\nu$ we see that the convex function

$$f^\nu(v) := L(x^\nu, v) - \langle p^\nu, v - v^\nu \rangle + \tfrac{1}{2}\rho^\nu |v - v^\nu|^2$$

must have a local minimum at v^ν. This implies that $0 \in \partial f^\nu(v^\nu) = \partial_v L(x^\nu, v^\nu) - p^\nu$ or, in other words, $p^\nu \in \partial_v L(x^\nu, v^\nu)$, a subgradient condition that, because of the convexity of $L(x, v)$ in v, can be written as the inequality

$$L(x^\nu, v) \geq L(x^\nu, v^\nu) + \langle p^\nu, v - v^\nu \rangle \quad \text{for all } v \in \mathbb{R}^n.$$

Consider now an arbitrary $v \in \mathbb{R}^n$ for which $L(\bar{x}, v) < \infty$. Our epi-continuity assumption ensures the existence of a sequence $\hat{v}^\nu \to v$ with $L(x^\nu, \hat{v}^\nu) \to L(x^\nu, v)$. For each index ν we have

$$L(x^\nu, \hat{v}^\nu) \geq L(x^\nu, v^\nu) + \langle p^\nu, \hat{v}^\nu - v^\nu \rangle.$$

In passing to the limit as $\nu \to \infty$ and using the fact that $L(x^\nu, v^\nu) \to L(\bar{x}, v)$ in particular, we obtain

$$L(\bar{x}, v) \geq L(\bar{x}, \bar{v}) + \langle \bar{p}, v - \bar{v} \rangle.$$

We have shown this inequality to hold for any v with $L(\bar{x}, v)$ finite, but it holds trivially when $L(\bar{x}, v) = \infty$. Hence it holds for all $v \in \mathbb{R}^n$, confirming that $\bar{p} \in \partial_v L(\bar{x}, \bar{v})$. □

Proposition 2.2 suggests approaching the subgradients of L and H in general by looking at the set

$$M := \{(x, v, p) \in O \times \mathbb{R}^n \times \mathbb{R}^n \mid \text{properties (2.4) hold}\} \tag{2.6}$$

and the set-valued mappings

$$\begin{aligned} S_L &: (x, v, p) \mapsto \{w \mid (w, p) \in \partial L(x, v)\}, \\ S_H &: (x, v, p) \mapsto \{w \mid (w, v) \in \partial H(x, p)\}. \end{aligned} \tag{2.7}$$

The graph of S_L, consisting by definition of all (x, v, p, w) such that $w \in S_L(x, v, p)$, is the same then as the graph of ∂L, except for a permutation of arguments; likewise for the graph of S_H in comparison with the graph of ∂H.

PROPOSITION 2.3. *Under the epi-continuity assumption, M is closed in $O \times \mathbb{R}^n \times \mathbb{R}^n$ and the functions $(x, v, p) \mapsto L(x, v)$ and $(x, v, p) \mapsto H(x, p)$ are finite and continuous on M. Moreover, the effective domains of the set-valued mappings S_L and S_H on $O \times \mathbb{R}^n \times \mathbb{R}^n$ (the effective domains being the sets on which the mappings are nonempty-valued) lie in M, and the graphs of these mappings are closed as subsets of $O \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$.*

Proof. We can view M as the graph of the mapping G that associates with each $x \in O$ the set of all $(v, p) \in \mathbb{R}^n \times \mathbb{R}^n$ such that p is a subgradient of the convex function $L(x, \cdot)$ at v. According to Attouch's theorem on subgradient convergence (see [23]), the epi-convergence of $L(x^\nu, \cdot)$ to $L(x, \cdot)$ implies the set convergence of $G(x^\nu)$ to $G(x)$. In particular, this entails the closedness of the graph of G in $O \times \mathbb{R}^n \times \mathbb{R}^n$.

As functions of (x, v, p), both $L(x, v)$ and $H(x, p)$ are lower semicontinuous by Proposition 2.1 and never take on $-\infty$. But on M they are related by $H(x, p) = \langle v, p\rangle - L(x, v)$ and $L(x, v) = \langle v, p\rangle - H(x, p)$, so they cannot take on ∞ either and must be upper semicontinuous as well. Hence they are finite and continuous on M.

The assertion about the effective domains of S_L and S_H just restates Proposition 2.2. Verifying the closedness of the graphs of S_L and S_H comes down to verifying the closedness of the graphs of ∂L and ∂H. The graph of ∂L consists by definition of the closure, in a special way relative to $O \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$, of the set of all (x, v, w, p) such that (w, p) is a proximal subgradient of L at (x, v). The closure consists of all limits of sequences $(x^\nu, v^\nu, w^\nu, p^\nu)$ that not only converge themselves but have the additional property that the values $L(x^\nu, v^\nu)$ converge. But whenever (w^ν, p^ν) is a proximal subgradient of L at (x^ν, v^ν) we have in particular that p^ν is a proximal subgradient of the convex function $L(x^\nu, \cdot)$ at v^ν. This implies $p^\nu \in \partial_v L(x^\nu, v^\nu)$, hence $(x^\nu, v^\nu, p^\nu) \in M$. The convergence of the values $L(x^\nu, v^\nu)$ is then automatic because L is continuous as a function on M. Thus, the special feature of the closure process falls away, and the graph of ∂L is seen to be a closed set in the ordinary sense relative to $O \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$. For ∂H the argument is parallel. □

Our strategy for proving Theorem 1.1 is now perhaps becoming clear. Under the epi-continuity assumption, we need only come up with additional conditions on a point $(\bar{x}, \bar{v}, \bar{p}) \in M$ that guarantee the sets $S_L(\bar{x}, \bar{v}, \bar{p})$ and $-S_H(\bar{x}, \bar{v}, \bar{p})$ have the same convex hull. The fact that these sets don't necessarily agree in advance of taking convex hulls is evident from simple examples. For instance, if $L(x, v) = c(x) + l(v)$ for a finite, continuous function c on $\mathbb{R}^m$ and a proper, lsc, convex function l on $\mathbb{R}^n$, we have $H(x, p) = -c(x) + h(p)$ with h the proper, lsc, convex function on $\mathbb{R}^n$ conjugate to l. Then M is the product of $\mathbb{R}^m$ with the graph of ∂l, and for $(\bar{x}, \bar{v}, \bar{p}) \in M$ we see that

$$S_L(\bar{x}, \bar{v}, \bar{p}) = \partial c(\bar{x}), \qquad S_H(\bar{x}, \bar{v}, \bar{p}) = -\partial(-c)(\bar{x}).$$

While $\partial c(\bar{x})$ is the set of subgradients of c at $\bar{x}$ as defined in the manner explained, from limits of proximal subgradients introduced "from below," $-\partial(-c)(\bar{x})$ has the analogous interpretation with proximal subgradients introduced instead "from above." These two sets are known often to differ for a nonsmooth function c, although they have the same convex hull when c is Lipschitz continuous on a neighborhood of $\bar{x}$.

The key to further progress will be the following *regularized functions* associated with L and H:

$$\begin{aligned} R_L(x, u) &:= \inf_{v \in \mathbb{R}^n} \left\{ L(x, v) + \tfrac{1}{2}|v - u|^2 \right\}, \\ R_H(x, u) &:= \inf_{p \in \mathbb{R}^n} \left\{ H(x, p) + \tfrac{1}{2}|p - u|^2 \right\}, \end{aligned} \tag{2.8}$$

where $|\cdot|$ is the Euclidean norm. Also important will be the corresponding *proximal mappings*:

$$\begin{aligned} P_L(x, u) &:= \arg\min_{v \in \mathbb{R}^n} \left\{ L(x, v) + \tfrac{1}{2}|v - u|^2 \right\}, \\ P_H(x, u) &:= \arg\min_{p \in \mathbb{R}^n} \left\{ H(x, p) + \tfrac{1}{2}|p - u|^2 \right\}. \end{aligned} \tag{2.9}$$

In dealing with these regularized functions and proximal mappings for a fixed u, we draw heavily on the theory of Moreau [24]; see also Rockafellar [1, Thm. 31.5]. For any fixed $x \in O$, the functions $u \mapsto R_L(x, u)$ and $u \mapsto R_H(x, u)$ are finite, convex, and continuously differentiable on $\mathbb{R}^n$. They satisfy the identity

$$R_L(x, u) + R_H(x, u) = \tfrac{1}{2}|u|^2. \tag{2.10}$$

The mappings $u \mapsto P_L(x, u)$ and $u \mapsto P_H(x, u)$ are single-valued from $\mathbb{R}^n$ into $\mathbb{R}^n$ and nonexpansive—globally Lipschitz continuous with constant 1—and related by

$$P_L(x, u) + P_H(x, u) = u, \tag{2.11}$$

$$P_L(x, u) = \nabla_u R_H(x, u), \qquad P_H(x, u) = \nabla_u R_L(x, u). \tag{2.12}$$

The proximal mappings P_L and P_H are especially of interest in providing a convenient parameterization of M in terms of (x, u).

PROPOSITION 2.4. *Under the epi-continuity assumption, $R_L(x, u)$ and $R_H(x, u)$ are not just continuous in u but with respect to $(x, u) \in O \times \mathbb{R}^n$. This holds also for $P_L(x, u)$ and $P_H(x, u)$. The one-to-one correspondence between points $(x, v, p) \in M$ and points $(x, u) \in O \times \mathbb{R}^n$ that is set up by the relations*

$$u = v + p, \qquad (v, p) = \big(P_L(x, u), P_H(x, u)\big)$$

is then a homeomorphism.

Proof. From the theory of epi-convergence [21], [23], the convex functions $L(x^\nu, \cdot)$ epi-converge to $L(x, \cdot)$ if and only if the regularized functions $R_L(x^\nu, \cdot)$ converge pointwise on $\mathbb{R}^n$ to $R_L(x, \cdot)$. These regularized functions being not only convex but finite, their pointwise convergence implies uniform convergence on all bounded subsets of $\mathbb{R}^n$ (cf. [1, Thm. 10.8]), and then their gradient mappings $\nabla R_L(x^\nu, \cdot)$ converge in such a manner to $\nabla R_H(x, \cdot)$ as well [1, Thm. 24.5]. The assertions about R_L and P_L, and similarly those about R_H and P_H, thus follow from the meaning of the epi-continuity assumption as designating the six conditions in Proposition 2.1. (The indicated correspondence is one-to-one because the elements $v = P_L(x, u)$ and $p = P_H(x, u)$ satisfy (2.11) and are characterized by the relations $0 \in \partial_v L(x, v) + (v - u)$ and $0 \in \partial_p H(x, p) + (v - p)$.) □

The gradients of $R_L(x, u)$ and $R_H(x, u)$ with respect to u are pinpointed by (2.12); but we can also determine, or at least estimate, subgradients with respect to (x, u).

PROPOSITION 2.5. *Under the epi-continuity assumption, consider any $(\bar{x}, \bar{u}) \in O \times \mathbb{R}^n$ and let $\bar{v} = P_L(\bar{x}, \bar{u})$ and $\bar{p} = P_H(\bar{x}, \bar{u})$. Then*

$$\begin{aligned}
\partial R_L(\bar{x}, \bar{u}) &\subset \big\{(w, p) \,\big|\, p = \bar{p},\ (w, \bar{p}) \in \partial L(\bar{x}, \bar{v})\big\},\\
\partial^\infty R_L(\bar{x}, \bar{u}) &\subset \big\{(w, p) \,\big|\, p = 0,\ (w, 0) \in \partial^\infty L(\bar{x}, \bar{v})\big\},\\
\partial R_H(\bar{x}, \bar{u}) &\subset \big\{(w, v) \,\big|\, v = \bar{v},\ (w, \bar{v}) \in \partial H(\bar{x}, \bar{p})\big\},\\
\partial^\infty R_H(\bar{x}, \bar{u}) &\subset \big\{(w, v) \,\big|\, v = 0,\ (w, 0) \in \partial^\infty H(\bar{x}, \bar{p})\big\}.
\end{aligned}$$

Proof. Let $f(x, v, u) = L(x, v) + \frac{1}{2}|v - u|^2$, so that $R_L(x, u) = \min_v f(x, v, u)$. For $(\bar{x}, \bar{u})$ the minimum is attained uniquely at $\bar{v}$. A general calculus rule in [25, Thm. 3.1] gives us

$$\begin{aligned}
\partial R_L(\bar{x}, \bar{u}) &\subset \big\{(w, p) \,\big|\, (w, 0, p) \in \partial f(\bar{x}, \bar{v}, \bar{u})\big\},\\
\partial^\infty R_L(\bar{x}, \bar{u}) &\subset \big\{(w, p) \,\big|\, (w, 0, p) \in \partial^\infty f(\bar{x}, \bar{v}, \bar{u})\big\},
\end{aligned}$$

but also in terms of the smooth function $f_0(x, v, u) := \frac{1}{2}|v-u|^2$ we have (cf. [7, Lem. 5A.1]) that

$$\begin{aligned} \partial f(\bar{x}, \bar{v}, \bar{u}) &= \big[\partial L(\bar{x}, \bar{v}) \times \{0\}\big] + \nabla f_0(\bar{x}, \bar{v}, \bar{u}), \\ \partial^\infty f(\bar{x}, \bar{v}, \bar{u}) &= \big[\partial^\infty L(\bar{x}, \bar{v}) \times \{0\}\big], \end{aligned}$$

where $\nabla f_0(\bar{x}, \bar{v}, \bar{u}) = (0, \bar{v}-\bar{u}, \bar{u}-\bar{v})$. The combination of these two sets of relations yields the inclusions claimed in the proposition for $\partial R_L(\bar{x}, \bar{u})$ and $\partial^\infty R_L(\bar{x}, \bar{u})$. Those for $\partial R_H(\bar{x}, \bar{u})$ and $\partial^\infty R_H(\bar{x}, \bar{u})$ follow by symmetry. □

Lipschitz continuity of the regularized functions with respect to x will be critical to us at a certain stage. This property is the subject of the next proposition.

PROPOSITION 2.6. *Under the epi-continuity assumption, the following four properties are equivalent to each other at a point* $(\bar{x}, \bar{u}) \in O \times \mathbb{R}^n$.

(a) R_L *is Lipschitz continuous around* $(\bar{x}, \bar{u})$.

(b) $R_L(x, \bar{u})$ *is Lipschitz continuous in* x *around* $\bar{x}$.

(c) R_H *is Lipschitz continuous around* $(\bar{x}, \bar{u})$.

(d) $R_H(x, \bar{u})$ *is Lipschitz continuous in* x *around* $\bar{x}$.

These properties are present in particular when the unique vectors $\bar{v}$ *and* $\bar{p}$ *satisfying* $\bar{u} = \bar{v} + \bar{p}$ *and* $(\bar{x}, \bar{v}, \bar{p}) \in M$ *are such that*

$$(w, 0) \in \partial^\infty L(\bar{x}, \bar{v}) \implies w = 0 \tag{2.13}$$

or such that

$$(w, 0) \in \partial^\infty H(\bar{x}, \bar{p}) \implies w = 0. \tag{2.14}$$

As a special case, the first of these two conditions is implied by L *being Lipschitz continuous around* $(\bar{x}, \bar{v})$, *whereas the second is implied by* H *being Lipschitz continuous around* $(\bar{x}, \bar{p})$.

Proof. The equivalence is apparent from the identity in (2.10) and the fact that $R_L(x, u)$ and $R_H(x, u)$ are finite, convex functions of u, hence locally Lipschitz continuous in u. (As a matter of fact, they are globally Lipschitz continuous in u with constant 1.)

By a result of Rockafellar [26], the function R_L, because of its lower semicontinuity on $O \times \mathbb{R}^n$ (Proposition 2.1), is Lipschitz continuous around $(\bar{x}, \bar{u})$ if and only if the set $\partial^\infty R_L(\bar{x}, \bar{u})$ contains only $(0, 0)$. This is true under (2.13) by virtue of the second inclusion in Proposition 2.5. In the same way, (2.14) suffices for Lipschitz continuity of R_H.

If L is Lipschitz continuous around $(\bar{x}, \bar{v})$, so that $\partial^\infty L(\bar{x}, \bar{v}) = \{(0, 0)\}$ by the result cited, we have (2.13) trivially. Likewise, if H is Lipschitz continuous around $(\bar{x}, \bar{p})$, so that $\partial^\infty H(\bar{x}, \bar{p}) = \{(0, 0)\}$, we have (2.14) trivially. □

For the record, conditions (2.13) and (2.14) aren't equivalent to each other, and they therefore cannot actually be equivalent to the Lipschitz continuity property in Proposition 2.6 but merely sufficient for it. This is seen through the example of $L(x, v) = x^{4/3}v^2$ on $\mathbb{R}^1 \times \mathbb{R}^1$, which is convex in v and continuously differentiable in (x, v). Since $\nabla L(0, 0) = (0, 0)$, the point $(\bar{x}, \bar{v}, \bar{p}) = (0, 0, 0)$ belongs to M. We have $\partial^\infty L(0, 0) = \{(0, 0)\}$ because L is Lipschitz continuous around $(0, 0)$; thus (2.13) is satisfied. But (2.14) isn't satisfied, which is seen as follows. Our choice of L corresponds to

$$H(x, p) = \begin{cases} p^2/4x^{4/3} & \text{when } x \neq 0, \\ 0 & \text{when } x = 0 \text{ and } p = 0, \\ \infty & \text{when } x = 0 \text{ and } p \neq 0. \end{cases}$$

Away from $x = 0$, H is twice continuously differentiable, so its gradients $\nabla H(x, p) = \big(-p^2/3x^{7/3},\ p/2x^{4/3}\big)$ are proximal subgradients. We aim at constructing a nonzero vector

$(\bar{w}, 0)$ in $\partial^\infty H(0,0)$ in accordance with the definition of that set in terms of limits of proximal subgradients. Consider any sequence $t^\nu \searrow 0$ and let $x^\nu = (t^\nu)^6$, $p^\nu = (t^\nu)^5$. Then $(x^\nu, p^\nu) \to (\bar{x}, \bar{p}) = (0,0)$ and $H(x^\nu, p^\nu) = (t^\nu)^2/4 \to 0$. Let $(w^\nu, v^\nu) = \nabla H(x^\nu, p^\nu) = \big(-(t^\nu)^{-4}, (t^\nu)^{-3}/2\big)$. Then for $\lambda^\nu = (t^\nu)^4$ we have $\lambda^\nu \searrow 0$ and $\lambda^\nu(w^\nu, v^\nu) = (-1, t^\nu/2) \to (-1, 0)$. This limit vector belongs to $\partial^\infty H(0,0)$ and demonstrates that (2.14) fails.

3. The main arguments. With this foundation in place, we can turn to the subgradient arguments that lead to the equivalence relation in Theorem 1.1.

LEMMA 3.1. *Under the epi-continuity assumption, suppose that $(\bar{w}, \bar{p})$ is a proximal subgradient to L at a point $(\bar{x}, \bar{v}) \in O \times \mathbb{R}^n$; in other words, $L(\bar{x}, \bar{v})$ is finite and there exist $\rho > 0$ and $\delta > 0$ such that*

$$L(x,v) \geq L(\bar{x},\bar{v}) + \langle \bar{w}, x-\bar{x}\rangle + \langle \bar{p}, v-\bar{v}\rangle - \tfrac{1}{2}\rho|x-\bar{x}|^2 - \tfrac{1}{2}\rho|v-\bar{v}|^2 \quad \text{when } x \in O,\ |x-\bar{x}| \leq \delta,\ |v-\bar{v}| \leq \delta. \tag{3.1}$$

Then there exists $\varepsilon \in (0, \delta)$ such that

$$L(x,v) \geq L(\bar{x},\bar{v}) + \langle \bar{w}, x-\bar{x}\rangle + \langle \bar{p}, v-\bar{v}\rangle - \tfrac{1}{2}\rho|x-\bar{x}|^2 - \tfrac{1}{2}\rho|v-\bar{v}|^2 \quad \text{for all } v \in \mathbb{R}^n \text{ when } x \in O,\ |x-\bar{x}| \leq \varepsilon. \tag{3.2}$$

Proof. We can write (3.1) equivalently as

$$\rho^{-1}L(\bar{x},\bar{v}) + \langle \rho^{-1}\bar{w}, x-\bar{x}\rangle - \tfrac{1}{2}|x-\bar{x}|^2 \leq f(x) \quad \text{when} \quad x \in O,\ |x-\bar{x}| \leq \delta, \tag{3.3}$$

where

$$\begin{aligned} f(x) :&= \inf_{|v-\bar{v}|\leq\delta} \big\{\rho^{-1}L(x,v) + \tfrac{1}{2}|v-\bar{v}|^2 - \langle \rho^{-1}\bar{p}, v-\bar{v}\rangle\big\} \\ &= \inf_{|v-\bar{v}|\leq\delta} \big\{\rho^{-1}L(x,v) + \tfrac{1}{2}\big|v-(\bar{v}+\rho^{-1}\bar{p})\big|^2\big\} - \tfrac{1}{2}|\rho^{-1}\bar{p}|^2. \end{aligned}$$

The question is whether (3.3) will continue to hold when the constraint $|v-\bar{v}| \leq \delta$ is dropped in the definition of f, at least if δ is replaced by some smaller value in (3.3).

We can answer this by applying the facts about proximal regularization to the function

$$\widehat{L}(x,v) := \begin{cases} \rho^{-1}L(x,v) & \text{when } |v-\bar{v}| \leq \delta, \\ \infty & \text{when } |v-\bar{v}| > \delta, \end{cases}$$

in terms of which we have $f(x) = R_{\widehat{L}}(x, \bar{u})$ for $\bar{u} = \bar{v} + \rho^{-1}\bar{p}$. Here $\widehat{L}(x, \cdot)$ is the sum of two convex functions, namely $\rho^{-1}L(x,\cdot)$ and the indicator of $\{v \mid |v-\bar{v}| \leq \delta\}$. For $x = \bar{x}$, the effective domain of the first function meets the interior of the effective domain of the second, and this is enough to guarantee through the convergence theorem of McLinden and Bergstrom [27] that whenever $x^\nu \to \bar{x}$ and $L(x^\nu, \cdot)$ epi-converges to $L(\bar{x}, \cdot)$, the sum $\widehat{L}(x^\nu, \cdot)$ epi-converges to $\widehat{L}(\bar{x}, \cdot)$. It follows then from our epi-continuity assumption that, for x in some open neighborhood O' of $\bar{x}$ within O, $\widehat{L}(x, \cdot)$ depends epi-continuously on x.

Through Proposition 2.4 we conclude that the associated proximal mapping $P_{\widehat{L}}$, which gives the unique minimizing v in the formula for f, is continuous on $O' \times \mathbb{R}^n$. Moreover, the minimum defining $f(\bar{x})$ is attained at $\bar{v}$ because of the proximal subgradient inequality; thus, $P_{\widehat{L}}(\bar{x}, \bar{u}) = \bar{v}$. There must, then, exist an $\varepsilon \in (0, \delta)$ such that when $x \in O'$ and $|x-\bar{x}| \leq \varepsilon$ we have $\big|P_{\widehat{L}}(x, \bar{u}) - \bar{v}\big| < \delta$. For such x the constraint $|v-\bar{v}| \leq \delta$ in the formula for $f(x)$ is inactive. Since the function of v being minimized in this formula is convex, the constraint can in this case be suppressed without affecting the minimum value that is attained. □

LEMMA 3.2. *Under the epi-continuity assumption, suppose* $(\bar{w}, \bar{p})$ *is a proximal subgradient of* L *at a point* $(\bar{x}, \bar{v}) \in O \times \mathbb{R}^n$, *and let* $\bar{u} = \bar{v} + \bar{p}$. *If* R_H *is Lipschitz continuous around* $(\bar{x}, \bar{u})$, *then* $(-\bar{w}, \bar{v}) \in \bar{\partial} R_H(\bar{x}, \bar{u}) = \operatorname{con} \partial R_H(\bar{x}, \bar{u})$.

Proof. We start by noting that in particular $(\bar{w}, \bar{p}) \in \partial L(\bar{x}, \bar{v})$, hence $\bar{p} \in \partial_v L(\bar{x}, \bar{v})$ by Proposition 2.2. Thus $(\bar{x}, \bar{v}, \bar{p}) \in M$. We have

$$\bar{v} = \nabla_u R_H(\bar{x}, \bar{u}) \tag{3.4}$$

by (2.11); this will be needed later.

Through Lemma 3.1 the proximal subgradient condition can be identified with the existence of $\rho > 0$ and $\varepsilon > 0$ such that (3.2) holds. Replacing the second occurrence of ρ in (3.2) by $\rho + 1$, which certainly maintains the inequality, we get

$$F(\bar{x}) + \langle \bar{v}, x - \bar{x} \rangle - \tfrac{1}{2}\rho |x - \bar{x}|^2 \le F(x) \quad \text{when} \quad |x - \bar{x}| \le \varepsilon \tag{3.5}$$

for the function

$$F(x) := \inf_{v \in \mathbb{R}^n} \left\{ L(x, v) - \langle \bar{p}, v - \bar{v} \rangle + \tfrac{1}{2}(\rho + 1)|v - \bar{v}|^2 \right\},$$

which has $F(\bar{x}) = L(\bar{x}, \bar{v})$. As a consequence of (3.5) we certainly have $\bar{w} \in \partial F(\bar{x})$.

To carry the analysis of $\bar{w}$ further, we'll make use of Fenchel's duality theorem [1, Thm. 31.1] to represent F in a different way. For any fixed x the definition of $F(x)$ can be interpreted as saying that

$$F(x) = \inf_{v \in \mathbb{R}^n} \left\{ \varphi(v) - \psi(v) \right\}$$

for the convex function $\varphi(v) := L(x, v) + \frac{1}{2}|v - \bar{v}|^2$ and the concave function $\psi(v) := \langle \bar{p}, v - \bar{v} \rangle - \frac{1}{2}\rho |v - \bar{v}|^2$. Because ψ is finite everywhere, the effective domains of these functions overlap in the manner required by the duality theorem in question, and we are able to conclude from it that

$$-F(x) = \inf_{p \in \mathbb{R}^n} \left\{ \varphi^*(p) - \psi^*(p) \right\}$$

for the functions φ^* and ψ^* conjugate to φ and ψ. From the definition of the convex conjugate φ^* we calculate that

$$\begin{aligned} \varphi^*(p) &= \sup_{v \in \mathbb{R}^n} \left\{ \langle p, v \rangle - L(x, v) - \tfrac{1}{2}|v - \bar{v}|^2 \right\} \\ &= - \inf_{v \in \mathbb{R}^n} \left\{ L(x, v) + \tfrac{1}{2}|v|^2 - \langle \bar{v} + p, v \rangle + \tfrac{1}{2}|\bar{v} + p|^2 - \tfrac{1}{2}|\bar{v} + p|^2 + \tfrac{1}{2}|\bar{v}|^2 \right\} \\ &= \tfrac{1}{2}|\bar{v} + p|^2 - \tfrac{1}{2}|\bar{v}|^2 - R_L(x, \bar{v} + p) = R_H(x, \bar{v} + p) - \tfrac{1}{2}|\bar{v}|^2, \end{aligned}$$

where the final steps use definition (2.8) and the identity (2.10). The definition of the concave conjugate ψ^* yields

$$\psi^*(p) = \inf_{v \in \mathbb{R}^n} \left\{ \langle p, v \rangle - \langle \bar{p}, v - \bar{v} \rangle + \tfrac{1}{2}\rho |v - \bar{v}|^2 \right\} = \langle p, \bar{v} \rangle - \tfrac{1}{2}\rho^{-1}|p - \bar{p}|^2.$$

Out of these calculations we get

$$-F(x) = \inf_{p \in \mathbb{R}^n} \left\{ R_H(x, \bar{v} + p) - \tfrac{1}{2}|\bar{v}|^2 - \langle p, \bar{v} \rangle + \tfrac{1}{2}\rho^{-1}|p - \bar{p}|^2 \right\},$$

which can be transformed to

$$-\rho F(x) = \inf_{p\in\mathbb{R}^n} \{\rho R_H(x, \bar{v} + p) + \tfrac{1}{2}|p - (\rho\bar{v} + \bar{p})|^2\} - \tfrac{1}{2}\rho(\rho + 1)|\bar{v}|^2. \tag{3.6}$$

In terms of $\widetilde{H}(x, p) := \rho R_H(x, \bar{v} + p)$ this has the interpretation that

$$\rho F(x) = -R_{\widetilde{H}}(x, \tilde{u}) + \tfrac{1}{2}\rho(\rho + 1)|\bar{v}|^2 \quad \text{for} \quad \tilde{u} = \rho\bar{v} + \bar{p}. \tag{3.7}$$

Proposition 2.4 assures us that $\widetilde{H}(x, p)$ is finite and continuous in $(x, p) \in O \times \mathbb{R}^n$. It is convex in p besides. We can therefore apply our proximal regularization results to this function in place of H. By assumption, $\widetilde{H}$ is Lipschitz continuous around $(\bar{x}, \bar{p})$. We have $\rho\bar{v} \in \partial_p\widetilde{H}(\bar{x}, \bar{p})$ by (3.4), hence also $\bar{p} = P_{\widetilde{H}}(\bar{x}, \tilde{u})$ and $\rho\bar{v} = \nabla_u R_{\widetilde{H}}(\bar{x}, \tilde{u})$.

By means of Proposition 2.6 we see that $R_{\widetilde{H}}$ is Lipschitz continuous around $(\bar{x}, \tilde{u})$. The fact that $\bar{w} \in \partial F(\bar{x})$ gives us in (3.7) that $\rho\bar{w} \in \partial_x(-R_{\widetilde{H}})(\bar{x}, \tilde{u})$. But by the Lipschitz continuity of $R_{\widetilde{H}}(\cdot, \tilde{u})$ at $\bar{x}$ the Clarke subgradient relation $\bar{\partial}_x(-R_{\widetilde{H}})(\bar{x}, \tilde{u}) = -\bar{\partial}_x R_{\widetilde{H}}(\bar{x}, \tilde{u})$ holds, i.e., $\operatorname{con}\partial_x(-R_{\widetilde{H}})(\bar{x}, \tilde{u}) = -\operatorname{con}\partial_x R_{\widetilde{H}}(\bar{x}, \tilde{u})$ (cf. [5], [7]). Therefore

$$-\rho\bar{w} \in \operatorname{con}\partial_x R_{\widetilde{H}}(\bar{x}, \tilde{u}). \tag{3.8}$$

Next we analyze the set $\partial_x R_{\widetilde{H}}(\bar{x}, \tilde{u})$. Because of the Lipschitz continuity of $R_{\widetilde{H}}$ around $(\bar{x}, \tilde{u})$, the rule holds that

$$\partial_x R_{\widetilde{H}}(\bar{x}, \tilde{u}) \subset \{w \mid \exists\, v \text{ with } (w, v) \in \partial R_{\widetilde{H}}(\bar{x}, \tilde{u})\}$$

(see [7, Lem. 5A.3]). Now Proposition 2.5, as applied with $\widetilde{H}$ in place of H (using the fact that $\tilde{u} = \rho\bar{v} + \bar{p}$ with $\rho\bar{v} \in \partial_p R_{\widetilde{H}}(\bar{x}, \tilde{u})$), says that

$$\partial R_{\widetilde{H}}(\bar{x}, \tilde{u}) \subset \{(w, v) \mid v = \rho\bar{v},\ (w, \rho\bar{v}) \in \partial\widetilde{H}(\bar{x}, \bar{p})\},$$

where from the choice of $\widetilde{H}$ we have $\partial\widetilde{H}(\bar{x}, \bar{p}) = \rho\partial R_H(\bar{x}, \bar{u})$. In combination with (3.8), this gives us $(-\bar{w}, \bar{v}) \in \operatorname{con}\partial R_H(\bar{x}, \bar{u})$, as claimed. □

THEOREM 3.3. *Under the epi-continuity assumption, consider the mappings S_L and S_H of (2.7) at a point $(\bar{x}, \bar{v}, \bar{p})$ such that either $S_L(\bar{x}, \bar{v}, \bar{p})$ or $S_H(\bar{x}, \bar{v}, \bar{p})$ is nonempty or merely $(\bar{x}, \bar{v}, \bar{p}) \in M$. Suppose for $\bar{u} = \bar{v} + \bar{p}$ that $R_L(\cdot, \bar{u})$ is Lipschitz continuous around $\bar{x}$ or, equivalently, that $R_H(\cdot, \bar{u})$ is Lipschitz continuous around $\bar{x}$. Then both $S_L(\bar{x}, \bar{v}, \bar{p})$ and $S_H(\bar{x}, \bar{v}, \bar{p})$ are nonempty and compact, and*

$$\operatorname{con} S_L(\bar{x}, \bar{v}, \bar{p}) = -\operatorname{con} S_H(\bar{x}, \bar{v}, \bar{p}).$$

Proof. In all cases we have $(\bar{x}, \bar{v}, \bar{p}) \in M$, since otherwise both $S_L(\bar{x}, \bar{v}, \bar{p})$ and $S_H(\bar{x}, \bar{v}, \bar{p})$ are empty by Proposition 2.3. The equivalence of the Lipschitz continuity assumptions is shown by Proposition 2.6, which also reveals that they imply that R_L and R_H are Lipschitz continuous around $(\bar{x}, \bar{u})$.

It will be demonstrated that there is a compact set W such that whenever $\bar{w} \in S_L(\bar{x}, \bar{v}, \bar{p})$ we have $\bar{w} \in W$ and $-\bar{w} \in \operatorname{con} S_H(\bar{x}, \bar{v}, \bar{p})$. The full conclusion of the theorem will follow then by symmetry.

By the definition of subgradients in general, the relation $\bar{w} \in S_L(\bar{x}, \bar{v}, \bar{p})$ implies the existence of proximal subgradients (w^ν, p^ν) to L at points $(x^\nu, v^\nu) \to (\bar{x}, \bar{v})$ such that $(w^\nu, p^\nu) \to (\bar{w}, \bar{p})$. The points $u^\nu = v^\nu + p^\nu$ then converge to $\bar{u}$ so that eventually (x^ν, u^ν) lies in the neighborhood of $(\bar{x}, \bar{u})$ in which R_H is Lipschitz continuous. Once this is true, we can apply Lemma 3.2 at (x^ν, u^ν) to ascertain that $(-w^\nu, v^\nu) \in \bar{\partial} R_H(x^\nu, u^\nu)$. In the limit this

yields $(-\bar{w}, \bar{v}) \in \bar{\partial} R_H(\bar{x}, \bar{u})$. In particular, $\bar{w}$ belongs to the image W of $\bar{\partial} R_H(\bar{x}, \bar{u})$ under the projection $(w, v) \mapsto -w$. This image set W is compact because $\bar{\partial} R_H(\bar{x}, \bar{u})$ is compact (in consequence of the Lipschitz continuity of R_H). Since $\bar{\partial} R_H(x^\nu, u^\nu) = \operatorname{con} \partial R_H(x^\nu, u^\nu)$, the inclusion for $\partial R_H(\bar{x}, \bar{u})$ in Proposition 2.5 gives us

$$(-\bar{w}, \bar{v}) \in \operatorname{con}\big\{(-w, v) \,\big|\, v = \bar{v},\ (-w, \bar{v}) \in \partial H(\bar{x}, \bar{p})\big\}.$$

This implies that $-\bar{w} \in \operatorname{con} S_H(\bar{x}, \bar{v}, \bar{p})$ as required. □

The proof of Theorem 3.3 discloses an additional property of the mappings S_L and S_H, which is worth noting. Recall that a set-valued mapping is *locally bounded* at a given point if some neighborhood of that point has bounded image under the mapping.

PROPOSITION 3.4. *Under the hypothesis of Theorem* 3.3, *the mappings S_L and S_H are locally bounded at $(\bar{x}, \bar{v}, \bar{p})$. The same is true also of the mappings $(x, v, p) \mapsto \operatorname{con} S_L(x, v, p)$ and $(x, v, p) \mapsto \operatorname{con} S_H(x, v, p)$, which in this case must, like S_L and S_H, have closed graphs relative to some neighborhood of $(\bar{x}, \bar{v}, \bar{p})$.*

Thus, whenever $(x^\nu, v^\nu, p^\nu) \to (\bar{x}, \bar{v}, \bar{p})$ and $w^\nu \in \operatorname{con} S_L(x^\nu, v^\nu, p^\nu)$, the sequence $\{w^\nu\}$ must be bounded, and all of its cluster points must belong to $\operatorname{con} S_L(\bar{x}, \bar{v}, \bar{p})$; likewise with S_L replaced by S_H.

Proof. This comes from the observation in the proof of Theorem 3.3 that when $(x, v + p)$ belongs to the neighborhood of $(\bar{x}, \bar{v} + \bar{p})$ on which R_H is Lipschitz continuous, we have $S_L(x, v, p) \subset \big\{w \,\big|\, (-w, v) \in \bar{\partial} R_L(x, v)\big\}$. The mapping $\bar{\partial} R_L$ is known to be locally bounded on such a neighborhood. The local boundedness of S_L along with the closedness of its graph (Proposition 2.3) ensures the same properties of the mapping $\operatorname{con} S_L$. The case of S_H follows by symmetry. □

In conclusion we summarize how the results we have obtained fit together to produce the main result stated in §1.

Proof of Theorem 1.1. Let $L_t = L(t, \cdot, \cdot)$ and $H_t = H(t, \cdot, \cdot)$. The assumption that L has the epi-continuity property along the arc $x(\cdot)$ puts us for almost every t in the picture of L_t and H_t satisfying the epi-continuity assumption of §2 relative to some open set $O(t)$ containing $x(t)$. Then by Proposition 2.2, if either $\big(w, p(t)\big) \in \partial L_t\big(x(t), \dot{x}(t)\big)$ or $\big(-w, \dot{x}(t)\big) \in \partial H_t\big(x(t), p(t)\big)$ we have $p(t) \in \partial_v L_t\big(x(t), \dot{x}(t)\big)$ and $\dot{x}(t) \in \partial_p H_t\big(x(t), p(t)\big)$. Thus, the Euler–Lagrange condition (1.10) and the Hamiltonian condition (1.12) automatically give (1.11) and (1.13). The equivalence of (1.10) and (1.12) follows from Theorem 3.3 whenever the regularized function R_{L_t} happens to be Lipschitz continuous around $\big(x(t), \dot{x}(t)\big)$ for almost every t, or equivalently, the function R_{H_t} is Lipschitz continuous around $\big(x(t), p(t)\big)$ for almost every t. Proposition 2.6 shows that such cases occur under the additional assumptions furnished in Theorem 1.1. □

REFERENCES

[1] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.

[2] ———, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Analysis Appl., 32 (1970), pp. 174–222.

[3] ———, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–40.

[4] ———, *Generalized Hamiltonian equations for convex problems of Lagrange*, Pacific J. Math., 33 (1970), pp. 411–428.

[5] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[6] ———, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983; reprinted as Classics in Applied Mathematics 5, Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[7] P. D. LOEWEN, *Optimal Control via Nonsmooth Analysis*, CRM Proc. Lecture Notes 2, American Mathematical Society, Providence, RI, 1993.

[8] F. H. CLARKE, *The Euler–Lagrange differential inclusion*, J. Differential Equations, 19 (1975), pp. 80–90.
[9] ———, *Extremal arcs and extended Hamiltonian systems*, Trans. Amer. Math. Soc., 64 (1977), pp. 349–367.
[10] ———, *Methods of Dynamic and Nonsmooth Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 57, Society for Industrial and Applied Mathematics, Philadelphia, 1989.
[11] P. D. LOEWEN AND R. T. ROCKAFELLAR, *The adjoint arc in nonsmooth optimization*, Trans. Amer. Math. Soc., 325 (1991), pp. 39–72.
[12] R. T. ROCKAFELLAR, *Dualization of subgradient conditions for optimality*, Nonlinear Analysis Theory Methods Appl., 20 (1993), pp. 627–646.
[13] B. S. MORDUKHOVICH, *Metric approximations and necessary optimality conditions for general classes of nonsmooth extremal problems*, Soviet Math. Dokl., 22 (1980), pp. 526–530.
[14] ———, *Optimization and approximation of differential inclusions*, Cybernetics, 24 (1988), pp. 781–788.
[15] ———, *On variational analysis of differential inclusions*, in Optimization and Nonlinear Analysis, A. D. Ioffe et al., eds., Longman, Harlow, 1992, pp. 199–213.
[16] A. D. IOFFE, *Nonsmooth subdifferentials: their calculus and applications*, in Proc. 1st World Congress of Nonlinear Analysis, de Grueter, Berlin, 1995.
[17] P. D. LOEWEN AND R. T. ROCKAFELLAR, *Optimal control of unbounded differential inclusions*, SIAM J. Control Optim., 32 (1994), pp. 442–470.
[18] B. S. MORDUKHOVICH, *Discrete approximations and refined Euler-Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 33 (1995), pp. 882–915.
[19] A. D. IOFFE AND R. T. ROCKAFELLAR, *The Euler and Weierstrass conditions for nonsmooth variational problems*, Calculus of Variations and Partial Differential Equations, to appear.
[20] R. J. WIJSMAN, *Convergence of sequences of convex sets, cones, and function*, II, Trans. Amer. Math. Soc., 123 (1966), pp. 32–45.
[21] R. J.-B. WETS, *Convergence of convex functions, variational inequalities, and convex optimization problems*, in Variational Inequalities and Complementarity Problems, R. W. Cottle et al., eds., Wiley, New York, 1980, pp. 376–403.
[22] G. SALINETTI AND R. J.-B. WETS, *On the convergence of convex sets in finite dimensions*, SIAM Rev. 21 (1979), pp. 18–33.
[23] H. ATTOUCH, *Variational Convergence of Functions and Operators*, Pitman, New York, 1984.
[24] J.-J. MOREAU, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
[25] R. T. ROCKAFELLAR, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Analysis Theory Methods Appl., 7 (1985), pp. 665–698.
[26] ———, *Clarke's tangent cones and the boundaries of closed sets in* $\mathbb{R}^n$, Nonlinear Analysis Theory Methods Appl., 3 (1979), pp. 145–154.
[27] L. MCLINDEN and R. C. BERGSTROM, *Preservation of convergence of convex sets and functions*, Trans. Amer. Math. Soc., 268 (1981), pp. 127–142.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1315–1328, July 1996

011

AN A PRIORI ESTIMATE FOR DISCRETE APPROXIMATIONS IN NONLINEAR OPTIMAL CONTROL*

ASEN L. DONTCHEV†

Abstract. We examine the convergence of an approximate discretization applied to the first-order optimality conditions for a nonlinear optimal control problem with convex control constraints. Under an assumption of the coercivity type we prove the existence of an optimal control for the original problem such that its L^∞ distance from the approximating sequence is proportional to the error. In a corollary we give conditions for Riemann integrability of the optimal control.

Key words. optimal control, discrete approximations, error estimates

AMS subject classifications. 49M25, 65K10

1. Introduction.

In this paper we obtain an a priori error estimate for a discrete approximation provided by the Euler scheme to a nonlinear optimal control problem with convex control constraints. We show that if the sequence of successive approximations satisfies a coercivity-type condition, then for a sufficiently fine approximation there exists an optimal control, the L^∞ distance from which to the approximating sequence is proportional to the error.

An excellent survey of earlier works on computational optimal control, including discrete approximations, is contained in Polak [48]. Based on an unpublished work by J.-P. Aubin and J.-L. Lions, Daniel [16] (see also his papers [14], [15]) developed an abstract approach for proving convergence of discrete approximations of optimal control problems, parallel to the direct method of the calculus of variations (but formally independent of it; see also [55]). In a series of papers Cullum [11]–[13] showed value and solution convergence of Euler approximations, applying virtually the same idea. In [11] the problem is linear-convex, [12] considers problems of Mayer's type that are linear in control, and [13] is an extension of [12] for a state and control constrained problem. More recent results on convergence analysis of algorithms involving discrete approximations are contained in [26], [37], [46], [49]–[51], [61]. There is also a body of work in Russian focusing mainly on convergence of discrete approximations combined with Tikhonov's regularization; see [2], [5]–[7], [28], [38], [54].

As typical in perturbation analysis of optimal control, the sequence of optimal values of perturbed (approximating) problems converges to the optimal value of the corresponding relaxed problem (e.g., in the sense of Warga [59]); then the value convergence is equivalent to the relaxability of the continuous problem. This was observed first by Mordukhovich [43]; see also [44] and [25, Ch. 6]. In an earlier paper Cullum [13] proved that a sequence of solutions of the discretized problem converges to a solution of the relaxed problem; instead of relaxability she used necessary conditions for optimality of the relaxed problem.

Compared with the extensive literature on computational optimal control, see the monographs [30], [39], [47], [53], [54]; there are relatively few results available providing *error estimates* for discrete approximations in optimal control. Hager [33] considered higher order schemes applied to unconstrained problems, obtaining error estimates under appropriate smoothness assumptions. Related results are given in [52]. Error estimates for the Euler scheme applied to state and control constrained convex optimal control problems were derived in [17]; see also [1], [40]–[42]. As usual in numerical analysis, the estimation of the error provided by a discrete approximation is limited by the regularity of the solution. In a

*Received by the editors June 19, 1994; accepted for publication (in revised form) March 28, 1995. This research was supported by National Science Foundation grant DMS 9404431.

†Mathematical Reviews, 416 Fourth Street, Ann Arbor, MI 48107.

recent paper [21] we consider a nonlinear optimal control problem obtaining an a posteriori error estimate. More specifically, we show in [21] that, if an optimal solution of the continuous problem satisfies a coercivity-type condition, then for a sufficiently small mesh spacing h there exists a local minimizer of the corresponding discrete problem at a distance from the optimal solution that is proportional to the averaged modulus of continuity $\tau(u^*; h)$ of the optimal control u^*. The only assumption for the regularity of the optimal control is Riemann integrability; i.e., the optimal control may have infinitely many points of discontinuity (but must be almost everywhere continuous). If the optimal control is of bounded variation, then we automatically obtain that the error is $O(h)$.

In a series of papers Hager [32], [34]–[36] studied approximations to dual problems in optimal control. Dual problems can be defined in various ways, depending on the assumed spaces for the variables and the constraints. By choosing appropriate spaces one may obtain dual problems that are tractable numerically. The dual approach is exceptionally efficient when one deals with entropy-like minimization problems; see, e.g., [4]. In [19], [20] we applied dual methods to a class of best approximation problems with applications in computer-aided geometric design.

Another approach is based on the dynamic programming formulation related to the Hamilton–Jacobi equation. The Hamilton–Jacobi equation typically has a nonsmooth (viscosity) solution whose computation requires special procedures. An overview of a priori error estimates for the approximation of optimal control problems by discrete models is given in [9]; see also [3], [8], [10], [29], [31].

Related to the subject of the present paper is the work on discrete approximations of differential inclusions; for a survey see [24]. In a recent paper [23] we approximated the Hausdorff distance between the sets of solutions to a controlled boundary value problem and its discrete approximation. For other results in this direction, including higher order schemes and error estimates for the reachable set, see [56]–[58], [60].

Although discrete approximations are primarily intended for computations, they are also a useful tool for obtaining purely theoretical results. The idea of using broken lines to prove existence of solutions to differential equations goes back to Euler. Apparently one can use various approximations, depending on the purpose of the study. For instance, one can derive necessary optimality conditions from necessary conditions for a discretized problem by passing to a limit when the step size goes to zero; for a recent paper using this approach see [45]. Depending on the kind of approximation one may obtain various necessary or sufficient optimality conditions.

In §2 we present our error estimate. A proof of this result is given in §3 and is based on an abstract lemma of inverse mapping type for a sequence of set-valued maps. As a corollary we obtain that if the problem has a unique optimal control and a certain coercivity condition is fulfilled for any regular partition, then the optimal control is Riemann integrable. Section 4 contains a discussion of numerical results.

2. The error estimate. We consider the nonlinear optimal control problem

$$\text{minimize} \int_0^1 g(x(t), u(t))dt \tag{1}$$

subject to

$$\dot{x}(t) = f(x(t), u(t)) \quad \text{for a.e. } t \in [0, 1], x(0) = a;$$

$$u(t) \in U \quad \text{for a.e. } t \in [0, 1], u \in L^\infty, x \in W^{1,\infty},$$

where $x(t) \in \mathbf{R}^n$, U is a closed and convex set in $\mathbf{R}^m$, $g : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$, $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$, and a is a fixed vector in $\mathbf{R}^n$. Throughout L^∞ is the space of essentially bounded vector

functions in [0, 1] with the norm $\| u \|_{L^\infty} = \operatorname{ess\,sup}\{|u(t)| : 0 \le t \le 1\}$ and $W^{1,\infty}$ is the space of functions with values in $\mathbf{R}^n$ that are Lipschitz continuous on [0, 1] equipped with the norm $\| x \|_{W^{1,\infty}} = \| x \|_{L^\infty} + \| \dot{x} \|_{L^\infty}$. We denote by $| \cdot |$ any norm in $\mathbf{R}^n$ and by the superscript ^{T} the transposition.

Assuming that the functions f and g are continuously differentiable, we write the first-order conditions (Pontryagin maximum principle) as a variational inequality of the form

$$\dot{x}(t) = f(x(t), u(t)), \quad x(0) = a; \tag{2}$$

$$\dot{\lambda}(t) = -\nabla_x H(x(t), u(t), \lambda(t)), \quad \lambda(1) = 0; \tag{3}$$

$$\nabla_u H(x(t), u(t), \lambda(t)) \in N(U; u(t)), \tag{4}$$

for a.e. $t \in [0, 1]$, where λ is the adjoint variable, H is the Hamiltonian, $H(x, u, \lambda) = g(x, u) + \lambda^T f(x, u)$, $\nabla_x H$ is the derivative of H with respect to x, and $N(U; u)$ is the normal cone to the set U at the point u; that is,

$$N(U; u) = \begin{cases} \{y \in \mathbf{R}^m : y^T(v - u) \le 0 \text{ for each } v \in U\} & \text{if } u \in U, \\ \emptyset & \text{if } u \notin U. \end{cases}$$

Suppose that to solve problem (2)–(4) we use a discrete approximation provided by the Euler scheme. More specifically, let N be a natural number; let $h = 1/N$ be the mesh spacing; let $t_i = ih$; and let x_i, u_i, λ_i denote approximations of $x(t)$, $u(t)$, and $\lambda(t)$, respectively, at $t = t_i$. The Euler scheme applied to the variational inequality (2)–(4) results in a finite-dimensional variational inequality. There are various numerical techniques for solving such problems; in this paper we shall not discuss this topic. We suppose that, after applying a numerical procedure with certain accuracy, a sequence of vectors $(x^N, u^N, \lambda^N) \in \mathbf{R}^{Nn} \times \mathbf{R}^{Nm} \times \mathbf{R}^{Nn}$ is available that satisfy the following approximation to (2)–(4):

$$x_{i+1} = x_i + hf(x_i, u_i) + h\delta_i^N, \quad x_0 = a; \tag{5}$$

$$\lambda_i = \lambda_{i+1} + h\nabla_x H(x_i, u_i, \lambda_{i+1}) + h\eta_i^N, \quad \lambda_N = \sigma^N; \tag{6}$$

$$\nabla_u H(x_i, u_i, \lambda_{i+1}) + \kappa_i^N \in N(U; u_i), \quad i = 0, 1, \ldots, N - 1. \tag{7}$$

Here the vector $\epsilon_i^N = (\delta_i^N, \eta_i^N, \sigma^N, \kappa_i^N)$, $\delta_i^N \in \mathbf{R}^n$, $\eta_i^N \in \mathbf{R}^n$, $\kappa_i^N \in \mathbf{R}^m$, $i = 0, 1, \ldots, N - 1$, $\sigma^N \in \mathbf{R}^n$, represents the accuracy of the algorithm used for solving the discretized problem. We denote by $x^N(\cdot)$ and $\lambda^N(\cdot)$ the piecewise linear and continuous extensions over [0, 1] of x^N and λ^N, respectively, and by $u_N(\cdot)$ the piecewise constant extension of u^N over [0, 1], which is continuous from the right across the grid points t_i.

In our previous paper with W. Hager [21] we showed that if a given optimal control u^* and the corresponding state x^* and adjoint variable λ^* for the continuous problem (1) satisfy certain conditions, then there exists a solution of the discretized problem that is "close" to (u^*, x^*, λ^*). In this paper we are concerned with the converse estimate; namely, we consider the distance from a given sequence of solutions of the approximating problem (5)–(7) to the set of solutions of the continuous problem (1). More precisely, in Theorem 1 we give an answer to the question of under what conditions on a sequence of successive approximations x^N, u^N, λ^N there exists an optimal control u^* of the original problem (1) such that its L^∞ distance from $u^N(\cdot)$ is proportional to the step size h and to the error ϵ^N.

THEOREM 1. *Assume that the sequence $(x_i^N, u_i^N, \lambda_i^N)$ is contained in a compact set $\mathcal{X} \subset \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^n$ for all $i = 0, 1, \ldots, N$ and for all N and that the functions f and g are twice*

continuously differentiable in an open set containing $\mathcal{X}$. *Let there exist a constant* $\alpha > 0$ *such that for every sufficiently large* N

$$\int_0^1 [x^T(t)Q_N(t)x(t) + 2x^T(t)S_N(t)u(t) + u^T(t)R_N(t)u(t)]dt \geq \alpha \int_0^1 |u(t)|^2 dt \tag{8}$$

for all $x(\cdot) \in W^{1,2}, u(\cdot) \in L^2, x(0) = 0, \dot{x}(t) = A_N(t)x(t) + B_N(t)u(t), u(t) \in U - U$ *a.e. in* $[0, 1]$, *where* $A_N = \nabla_x f(x^N, u^N)$, $B_N = \nabla_u f(x^N, u^N)$, $R_N = \nabla^2_{uu} H(x^N, u^N, \lambda^N)$, $S_N = \nabla^2_{xu} H(x^N, u^N, \lambda^N)$, $Q_N = \nabla^2_{xx} H(x^N, u^N, \lambda^N)$. *Then there exist positive constants* c, δ *and an integer* N^* *such that for every* $N \geq N^*$ *for which*

$$\max_{0 \leq i \leq N-1} |\epsilon_i^N| \leq \delta, \tag{9}$$

there exist an isolated local minimizer $(x^{*N}(\cdot), u^{*N}(\cdot))$ *of the continuous problem* (1) *and a corresponding adjoint variable* $\lambda^{*N}(\cdot)$ *such that*

$$\begin{aligned} &\| x^{*N}(\cdot) - x^N(\cdot) \|_{W^{1,\infty}} + \| \lambda^{*N}(\cdot) - \lambda^N(\cdot) \|_{W^{1,\infty}} \\ &\qquad + \| u^{*N}(\cdot) - u^N(\cdot) \|_{L^\infty} \leq c(h + \max_{0 \leq i \leq N-1} |\epsilon_i^N|). \end{aligned} \tag{10}$$

In the statement of the theorem we use the uniform grid in [0, 1], i.e. with a constant step size h. The same result holds for any regular partition of [0, 1] in which the maximal step size goes to zero. From Theorem 1 we obtain the following corollary.

COROLLARY 1. *Let problem* (1) *have no more than one optimal solution, let*

$$\lim_{N \to \infty} \max_{0 \leq i \leq N-1} |\epsilon_i^N| = 0,$$

and let for every regular partition of $[0, 1]$ *in* N *intervals the sequence* (x^N, u^N, λ^N) *obtained from* (5)–(7) *satisfy coercivity condition* (8). *Then there exists a (unique) optimal control for* (1), *which is Riemann integrable.*

Proof. Theorem 1 implies that problem (1) has a solution (x^*, u^*) that is unique and satisfies (10). Then for every regular partition $\{t_i\}_{i=0}^N$,

$$\max_{0 \leq i \leq N-1} \sup_{t_i \leq t < t_{i+1}} |u^*(t) - u^*(t_i)| \to 0 \quad \text{as } N \to \infty,$$

which implies Riemann integrability of u^*. □

The coercivity condition (8) is a stability-type condition which implies that the inverse of the linearization of (5)–(7) is Lipschitzian around the reference sequence uniformly in N. For instance, it holds when the Hamiltonian is convex in x and strongly convex in u in a neighborhood of a solution and the sequence of successive approximations is in this neighborhood. Practically, during computations one may check whether there exists a constant $\alpha > 0$ such that

$$(x^T, u^T) \begin{pmatrix} Q_N(t) & S_N^T(t) \\ S_N(t) & R_N(t) \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \geq \alpha |u|^2$$

for all N, x, u, and t, which is a computationally available procedure. The consistency condition (9) means that the error of solving the discretized problems is to be sufficiently small when increasing the number N. Note that a finer discretization results in higher dimensions of the discrete problem and, in turn, is more likely to lead to accumulation of various errors accompanying the computations. The practical validity of the error model adopted in Theorem 1 may vary from problem to problem and from computer to computer.

3. Proof of Theorem 1. The proof is based on the following abstract inverse-function-type result for a sequence of set-valued maps.

LEMMA 1. *Let (X, ρ) be a complete metric space, let Y be a linear normed space, and let $B_a(x)$ be the closed ball centered at x with radius a. Let ξ_N be a sequence in X, let Θ_N be a sequence of set-valued maps from X to Y with $0 \in \Theta_N(\xi_N)$ for all N, and let $\phi_N : X \to Y$ be a sequence of functions. Suppose that for each N there exists a function $\Psi_N : Y \to X$ with the following properties.*

(A1) $\xi_N = \Psi_N(0)$ *and there exists a constant $\beta > 0$ such that for every N*

$$\Psi_N(y) \in \Theta_N^{-1}(y) \text{ for every } y \in B_\beta(0).$$

(A2) *There exists a constant $\gamma > 0$ such that Ψ_N is Lipschitz in $B_\beta(0)$ with constant γ for every N.*

(A3) *For every $\epsilon > 0$ there exists $\alpha > 0$ such that for every N*

$$\| \phi_N(u) - \phi_N(v) \| \leq \epsilon \rho(u, v)$$

whenever $u, v \in B_\alpha(\xi_N)$, $N = 1, 2, \ldots$.

Then for every $c > \gamma$ there exists $\Delta > 0$ such that for every N for which

$$\| \phi_N(\xi_N) \| \leq \Delta$$

there exists $\xi_N^* \in (\Theta_N - \phi_N)^{-1}(0)$ satisfying

$$\rho(\xi_N^*, \xi_N) \leq c \| \phi_N(\xi_N) \| . \tag{11}$$

Moreover, if Θ_N^{-1} is single-valued near 0, then there exists exactly one ξ_N^* with the above properties.

Proof of Lemma 1. Let $c > \gamma$ and let $\epsilon > 0$ be so small that

$$0 < \gamma\epsilon < 1 \quad \text{and} \quad \frac{\gamma}{1 - \epsilon\gamma} \leq c. \tag{12}$$

Let $\Delta > 0$ be such that

$$\Delta \leq \min\left\{ \frac{\beta}{c\epsilon + 1}, \frac{\alpha}{c} \right\}, \tag{13}$$

and let

$$r_N = c \| \phi_N(\xi_N) \| . \tag{14}$$

Define the function

$$\Phi_N(x) = \Psi_N(\phi_N(x)).$$

Let N be such that $\| \phi_N(\xi_N) \| \leq \Delta$, and let $x \in B_{r_N}(\xi_N)$. Using (A3), (13), and (14) we have

$$\| \phi_N(x) \| \leq \| \phi_N(x) - \phi_N(\xi_N) \| + \| \phi_N(\xi_N) \| \leq \epsilon r_N + \Delta \leq \Delta(1 + c\epsilon) \leq \beta.$$

Then from (A2), (A3), (12), and (14) we obtain

$$\begin{aligned} \rho(\xi_N, \Phi_N(x)) &= \rho(\Psi_N(0), \Psi_N(\phi_N(x))) \leq \gamma \| \phi_N(x) \| \\ &\leq \gamma \| \phi_N(x) - \phi_N(\xi_N) \| + \gamma \| \phi_N(\xi_N) \| \leq \gamma\epsilon r_N + \gamma r_N / c \leq r_N. \end{aligned}$$

Hence, Φ_N maps $B_{r_N}(\xi_N)$ into itself. Moreover, if $x', x'' \in B_{r_N}(\xi_N)$, then

$$\rho(\Phi_N(x'), \Phi_N(x'')) = \rho(\Psi_N(\phi_N(x')), \Psi_N(\phi_N(x'')))$$
$$\leq \gamma \parallel \phi_N(x') - \phi_N(x'') \parallel \leq \gamma\epsilon\rho(x', x'') < \rho(x', x'').$$

Thus Φ_N is a contraction from $B_{r_N}(\xi_N)$ to $B_{r_N}(\xi_N)$. By the contraction mapping principle for every N there exists $\xi_N^* \in B_{r_N}(\xi_N)$ such that $\xi_N^* = \Phi_N(\xi_N^*)$. The definition of Φ_N yields that $\xi_N^* \in (\Theta_N - \phi_N)^{-1}(0)$. If Θ_N^{-1} is locally single-valued near 0, then ξ_N^* is the unique fixed point of Φ_N in $B_{r_N}(\xi_N)$. This completes the proof of Lemma 1. □

Proof of Theorem 1. We apply Lemma 1 with the specifications

$$X = \{\xi = (x, \lambda, u) \in W^{1,\infty} \times W^{1,\infty} \times L^\infty, x(0) = a\}, \quad Y = L^\infty \times \mathbf{R}^n,$$

$$\phi_N(\xi) = -\begin{pmatrix} \dot{x} - f(x,u) - (\dot{x} - \dot{x}^N) + A_N(x - x^N) + B_N(u - u^N) \\ \dot{\lambda} + \nabla_x H(x,u,\lambda) - (\dot{\lambda} - \dot{\lambda}^N) - A_N^T(\lambda - \lambda^N) - Q_N(x - x^N) - S_N(u - u^N) \\ -\nabla_u H(x,u,\lambda) + \nabla_u H^N + R_N(u - u^N) + S_N^T(x - x^N) + B_N^T(\lambda - \lambda^N) - \kappa^N \\ \sigma^N \end{pmatrix},$$

$$\Theta_N(\xi) = \begin{pmatrix} (\dot{x} - \dot{x}^N) - A_N(x - x^N) - B_N(u - u^N) \\ (\dot{\lambda} - \dot{\lambda}^N) + A_N^T(\lambda - \lambda^N) + Q_N(x - x^N) + S_N(u - u^N) \\ -\nabla_u H^N - R_N(u - u^N) - S_N^T(x - x^N) - B_N^T(\lambda - \lambda^N) + \kappa^N + N(U; u) \\ \lambda(1) - \sigma^N \end{pmatrix},$$

where $\nabla_u H^N(t) = \nabla_u H(x^N(t_i), u^N(t), \lambda^N(t_{i+1}))$ for $t \in [t_i, t_{i+1})$ and $\delta^N, \eta^N, \kappa^N$ are assumed piecewise constant and continuous from the right across the time grid. Clearly, $0 \in \Theta_N(\xi_N)$.

Let $(p, q, r) \in L^\infty, s \in \mathbf{R}^n$, and denote $y = (p, q, r, s)$. Then $\xi \in \Theta_N^{-1}(y)$ if and only if $\xi = (x, \lambda, u)$ satisfies the relations

$$\dot{x} = A_N x + B_N u + p + \alpha_N, \quad x(0) = a; \tag{15}$$

$$\dot{\lambda} = -A_N^T \lambda - Q_N x - S_N u + q + \beta_N, \quad \lambda(1) = s + \sigma^N; \tag{16}$$

$$R_N u + S_N^T x + B_N^T \lambda + r + \gamma_N \in N(U; u) \tag{17}$$

for a.e. $t \in [0, 1]$, where $\alpha_N = \dot{x}^N - A_N x^N - B_N u^N, \beta_N = \dot{\lambda}^N + A_N^T \lambda^N + Q_N x^N + S_N u^N, \gamma_N = \nabla_u H^N - R_N u^N - S_N^T x^N - B_N^T \lambda^N - \kappa^N$. We show first that there exists a function $\Psi_N(y) = (x_N(y), \lambda_N(y), u_N(y)) \in \Theta_N^{-1}(y)$ that satisfies conditions (A1), (A2) of Lemma 1.

Under the coercivity condition (8) the system (15)–(17) is equivalent to the linear-quadratic problem of minimizing

$$-x(1)^T(s+\sigma^N) + 0.5 \int_0^1 [x^T Q_N x + u^T R_N u + 2x^T S_N u - (q+\beta_N)^T x + (r+\gamma_N)^T u] dt \tag{18}$$

subject to

$$\dot{x} = A_N x + B_N u + p + \alpha_N, \quad u(t) \in U \text{ for a.e. } t \in [0, 1], \quad x(0) = a. \tag{19}$$

Let $w_N(p)$ be the solution of

$$\dot{w} = A_N w + p + \alpha_N, \qquad w(0) = a.$$

Changing the state variable $x = z + w_N(p)$ we obtain the following problem equivalent to (18): minimize

$$-z(1)^T(s + \sigma^N) + 0.5 \int_0^1 [z^T Q_N z + u^T R_N u + 2z^T S_N u$$

$$-(q + \beta_N - 2Q_N w_N(p))^T z + (r + \gamma_N + 2S_N^T w_N(p))^T u]dt \tag{20}$$

subject to

$$\dot{z} = A_N z + B_N u, \quad u(t) \in U, \text{ a.e. } t \in [0, 1], \quad z(0) = 0. \tag{21}$$

Let $\mathcal{L}_N : L^2 \to L^2$ be the linear and bounded input–state map for system (21); that is,

$$(\mathcal{L}_N u)(t) = \int_0^t \Phi_N(t, \tau) B_N(\tau) u(\tau) d\tau,$$

where Φ_N is the fundamental matrix solution to $\dot{z} = A_N z$. Note that the conjugate map is

$$(\mathcal{L}_N^T x)(t) = \int_t^1 B_N(t)^T \Phi_N(\tau, t)^T x(\tau) d\tau,$$

and, since the components of A_N and B_N are bounded in L^∞, there exists a constant c_1 independent of N such that the operator norm satisfies

$$\| \mathcal{L}_N \| \leq c_1. \tag{22}$$

Denote

$$c_N[s](t) = -B_N^T(t)\Phi_N(1, t)^T(s + \sigma_N),$$

$$v_N(y) = \mathcal{L}_N^T(-q - \beta_N + 2Q_N w_N(p)) + r + \gamma_N + 2S_N^T w_N(p) + c_N(s),$$

$$\mathcal{R}_N = \mathcal{L}_N^T Q_N \mathcal{L}_N + 2\mathcal{L}_N^T S_N + R_N; \tag{23}$$

and let $\Omega = \{u \in L^2 : u(t) \in U \text{ for a.e. } t \in [0, 1]\}$. Then problem (20) can be rewritten as

$$\text{minimize } 0.5\langle u, \mathcal{R}_N u\rangle + \langle v_N(y), u\rangle \text{ subject to } u \in \Omega, \tag{24}$$

where $\langle \cdot, \cdot \rangle$ is the L^2 scalar product. The set Ω is a closed and convex subset of the space L^2 and the coercivity condition (8) is equivalent to

$$\langle u, \mathcal{R}_N u\rangle \geq \alpha \| u \|_{L^2}^2 \qquad \text{for all } u \in \Omega - \Omega. \tag{25}$$

It is a standard observation that, under (25), for every $y \in L^\infty \times \mathbf{R}^n$ there exists a unique solution $u_N(y)$ of the problem (24). Thus there exist an unique optimal state $x_N(y)$ and an unique optimal adjoint variable $\lambda_N(y)$; that is, the function $\Psi_N(y) = (x_N(y), \lambda_N(y), u_N(y))$ is well defined for all $y \in L^\infty \times \mathbf{R}^n$. Hence assumption (A1) of Lemma 1 holds. To show that

$\Psi_N(y)$ is Lipschitzian we apply a well-known argument; see, e.g., [18, Ch. 2]. For any $y', y'' \in Y$ the corresponding solutions u'_N, u''_N satisfy

$$\langle \mathcal{R}_N u' + v_N(y'), u'' - u' \rangle \geq 0,$$

$$\langle \mathcal{R}_N u'' + v_N(y''), u' - u'' \rangle \geq 0.$$

Combining these two inequalities with (25) we obtain that

$$\| u'_N - u''_N \|_{L^2} \leq \frac{1}{\alpha} \| v_N(y') - v_N(y'') \|_{L^2} . \tag{26}$$

Using (22) and the boundedness in L^∞ of the matrices A_N, B_N, Q_N, S_N we conclude that for every $y \in Y$ there exists a unique optimal control $u_N(y)$ of problem (18) that is Lipschitzian in y from L^∞ to L^2 with a Lipschitz constant independent of N. The Gronwall lemma and (22) applied to the state and adjoint equations yield that the optimal state $x_N(y)$ and the optimal adjoint variable $\lambda_N(y)$ are Lipschitzian with respect to y from $L^\infty \times \mathbf{R}^n$ to $W^{1,2}$ uniformly in N. Furthermore, the coercivity condition (8) implies that for some $\alpha_1 > 0$,

$$u^T R_N(t) u \geq \alpha_1 |u|^2 \quad \text{for a.e. } t \in [0, 1],$$

whenever $u \in U - U$; see [27]. By repeating the argument in deriving (26) for optimality condition (17), we obtain that the optimal control $u_N(y)$ is Lipschitzian in y from L^∞ to L^∞ uniformly in N. The Gronwall lemma applied to the state and adjoint equations yields that $x_N(y)$ and $\lambda_N(y)$ are Lipschitzian in y from $L^\infty \times \mathbf{R}^n$ to $W^{1,\infty}$ uniformly in N. Hence, condition (A2) of Lemma 1 is satisfied.

Let us show that ϕ_N satisfies (A3) at $\xi_N(\cdot) = (x^N(\cdot), \lambda^N(\cdot), u^N(\cdot))$ uniformly in N as a function from $W^{1,\infty} \times W^{1,\infty} \times L^\infty \times \mathbf{R}^n$ to L^∞. Denote $\zeta = (x, u)$ and consider the first component of ϕ_N. From the continuity of the derivative ∇f it follows that for any $\epsilon > 0$ there exists $\alpha > 0$ such that for every $\bar{\zeta}$ with $\| \bar{\zeta} - \zeta_N \|_{L^\infty} \leq \alpha$,

$$\| \nabla f(\bar{\zeta}) - \nabla f(\zeta_N) \|_{L^\infty} \leq \epsilon.$$

If $\zeta_1, \zeta_2 \in B_\alpha(\zeta_N)$, then

$$\| f(\zeta_1) - f(\zeta_2) - \nabla f(\zeta_N)(\zeta_1 - \zeta_2) \|_{L^\infty} \leq \epsilon \| \zeta_1 - \zeta_2 \|_{W^{1,\infty} \times L^\infty} .$$

The proof of (A3) for the remaining components of ϕ_N is completely analogous.

Thus we can apply Lemma 1, obtaining that there exist constants c and Δ such that if $\| \phi_N(\xi_N) \| \leq \Delta$, then there exists $\xi_N^* \in (\Theta_N - \phi_N)^{-1}(0)$ satisfying (11). It remains to estimate $\| \phi_N(\xi_N) \|$. The relations (5)–(7) imply that $\xi_N \in X$ and satisfies

$$\dot{x}(t) = f(x(t_i), u(t)) + \delta^N(t);$$

$$\dot{\lambda}(t) = -\nabla_x H(x(t_i), u(t), \lambda(t_{i+1})) + \eta^N(t), \quad \lambda(1) = \sigma^N;$$

$$\nabla_u H(x(t_i), u(t), \lambda(t_{i+1})) + \kappa^N(t) \in N(U; u(t))$$

for $t \in [t_i, t_{i+1})$, $i = 0, 1, \ldots, N-1$. Applying the Gronwall lemma to the adjoint equation and using the assumption that $x^N(\cdot)$, $u^N(\cdot)$ are bounded in L^∞ and f, g and their derivatives are continuous, we obtain that $\lambda^N(\cdot)$ is bounded in L^∞. Hence, the sequence of the derivatives $(\dot{x}^N(\cdot), \dot{\lambda}^N(\cdot))$ is bounded in L^∞. Then

$$\begin{aligned}\| \phi_N(\xi_N) \|_{L^\infty} &\le \max_{0\le i\le N-1} \sup_{t_i\le t\le t_{i+1}} \{[|f(x^N(t_i), u^N(t)) - f(x^N(t), u^N(t))| \\ &\quad + |\nabla_x H(x^N(t_i), u^N(t), \lambda^N(t_{i+1})) - \nabla_x H(x^N(t), u^N(t), \lambda^N(t))| \\ &\quad + |\nabla_u H(x^N(t_i), u^N(t), \lambda^N(t_{i+1})) - \nabla_u H^N(x(t), u^N(t), \lambda^N(t))|] \\ &\quad + |\kappa_i^N| + |\sigma^N|\} \\ &\le c_2 h + \max_{0\le i\le N-1} |\epsilon_i^N|,\end{aligned}$$

where c_2 is a constant independent of N. Take $\delta > 0$ and N^* such that $c_2/N^* + \delta \le \Delta$. From Lemma 1, if $N \ge N^*$ and $\max_{0\le i\le N-1} |\epsilon_i^N| \le \delta$, then there exist x^{*N}, u^{*N}, and λ^{*N} satisfying the estimate (10) and the maximum principle (2)–(4). The last step of the proof is to show that (x^{*N}, u^{*N}) is an isolated local solution of (1). Clearly, the values of the derivatives of f and H at $(x^{*N}, u^{*N}, \lambda^{*N})$, denoted A_N^*, B_N^*, Q_N^*, S_N^*, R_N^*, are close in L^∞ to the values of the corresponding matrix functions A_N, B_N, Q_N, S_N, R_N defined in the statement of the theorem. Then the operator $\mathcal{R}_N^*$ defined as in (23), where A_N, B_N, Q_N, S_N, R_N are replaced by A_N^*, B_N^*, Q_N^*, S_N^*, R_N^*, satisfies (25). It is known that (8) (or, equivalently, (25)), together with the maximum principle (2)–(4), is a second-order sufficient condition for an isolated local minimum; see, e.g., [21, App. 1]. The proof is complete. □

4. A computational example. If the final state in problem (1) is fixed, e.g. $x(1) = b$, then using Lemma 1 one can easily prove a result completely analogous to that in Theorem 1 on the additional assumption that there exists $c > 0$ such that the ball $B_c(b)$ is in the reachable set of the linearized system (15) for all N (for a related analysis see [22]). As an illustration we consider the problem

$$\text{minimize } 0.5 \int_0^2 u(t)^2 dt$$

subject to

$$\begin{aligned}&\dot{x}_1 = x_2, \\ &\dot{x}_2 = u, \\ &x_1(0) = 0, \qquad x_1(2) = 5/6, \\ &x_2(0) = 0, \qquad x_1(2) = 1/2, \\ &u \ge 0, \quad t \in [0, 2].\end{aligned}$$

The optimal control is a nonsmooth function in time, that is,

$$u^*(t) = \begin{cases} 1 - t & \text{for } t \in [0, 1), \\ 0 & \text{for } t \in [1, 2]; \end{cases}$$

and the first-order optimality conditions result in a boundary-value problem with nonsmooth right-hand side

$$\begin{aligned}&\dot{x}_1 = x_2, \\ &\dot{x}_2 = \max\{0, \lambda_2\}, \\ &\dot{\lambda}_1 = 0,\end{aligned}$$

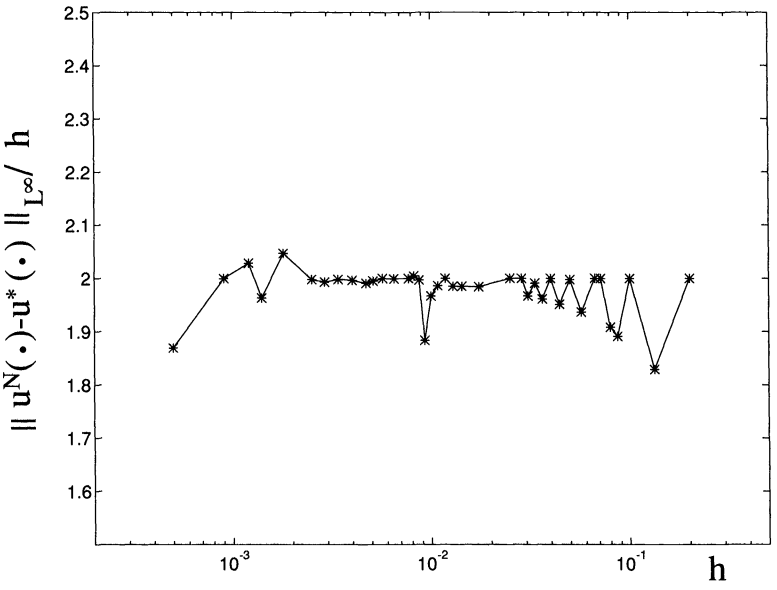


FIG. 1. *The dependence of the relative accuracy on h for* $\epsilon = 10^{-6}$.

$$\dot{\lambda}_2 = -\lambda_1.$$
$$x_1(0) = 0, \qquad x_1(2) = 5/6,$$
$$x_2(0) = 0, \qquad x_1(2) = 1/2, \tag{27}$$

We applied the Euler discretization scheme to the latter problem and solved the resulting finite-dimensional problem by the shooting method with a quadratic penalty. The unconstrained optimization problem is solved with the help of the BFGS code of MATLAB on HP Apollo workstation 715/50.

It turns out that even for the simple linear-quadratic problem considered one observes effects in the line of the theoretical analysis. Figure 1 shows the dependence of the relative accuracy $\| u^*(\cdot)-u^N(\cdot) \|_{L^\infty} / h$ on the mesh spacing h when the tolerance in the stopping test is $\epsilon = 10^{-6}$. We see that this relative accuracy is bounded, hence the convergence rate is $O(h)$ as proved theoretically. Note that the computational time for $h = 5 \times 10^{-4}$ was 1.9542×10^4 sec. In Fig. 2 the same dependence is obtained for the tolerance $\epsilon = 10^{-3}$. In this case, for $h \leq 5 \times 10^{-3}$ the relative accuracy significantly increases; that is, for a tolerance comparable with the mesh spacing, there is no first-order convergence. Figure 3 shows the exact and the approximate optimal controls for $\epsilon = 10^{-6}$ and $h = 0.05$.

An area for future research is error analysis of higher order approximations (e.g., Runge–Kutta schemes) applied to *constrained* optimal control problems. We did some computations for problem (27) with the standard second-order Runge–Kutta scheme. The dependence of the relative accuracy on h is given in Fig. 4; it indicates $O(h^2)$ convergence.

In the author's opinion the abstract result in Lemma 1 can be applied to variational problems with integral state and control constraints, as well as with systems governed by abstract (functional or partial) differential equations. The main steps in such an analysis will be to prove the consistency and stability of the approximation considered. A direct application of the abstract scheme may fail for stiff systems where, in general, the stability condition fails.

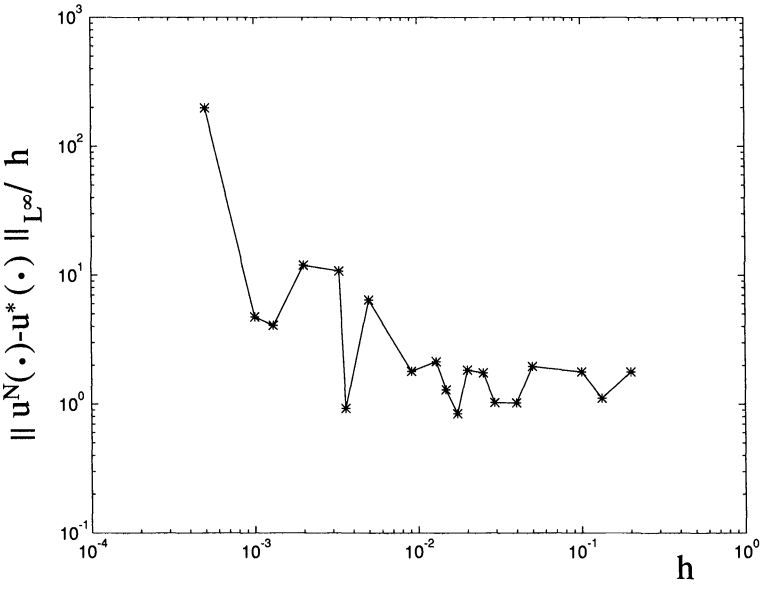


FIG. 2. *The dependence of the relative accuracy on h for* $\epsilon = 10^{-3}$.

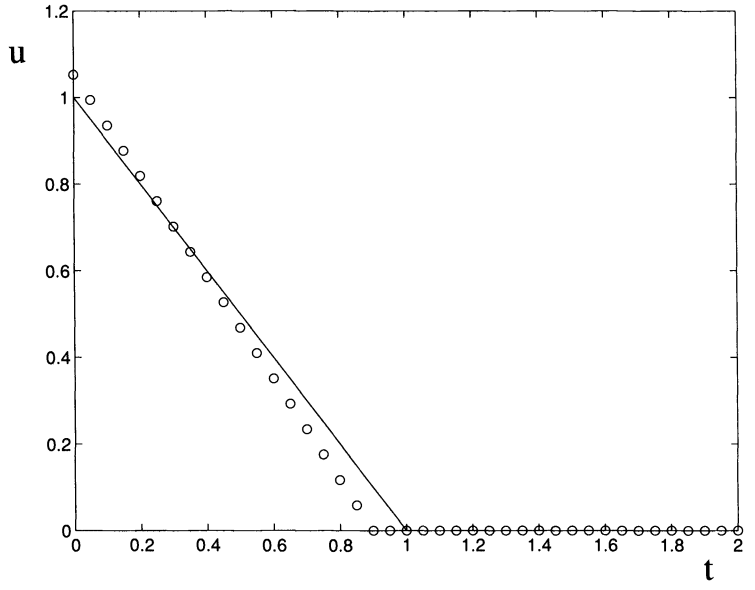


FIG. 3. *The exact and approximate optimal controls for* $h = 0.05$ *and* $\epsilon = 10^{-6}$.

The presence of state constraints considerably complicates the analysis of nonlinear optimal control problems. The main difficulty here is to find a compromise between the requirements for differentiability of the functions involved and the coercivity conditions guaranteeing stability. An $O(h)$ estimate for the L^2 norm of the error for the optimal control was obtained in [17] for a convex problem with linear inequality state and control constraints. We do not know whether such an estimate holds with a stronger norm, for example, in L^∞.

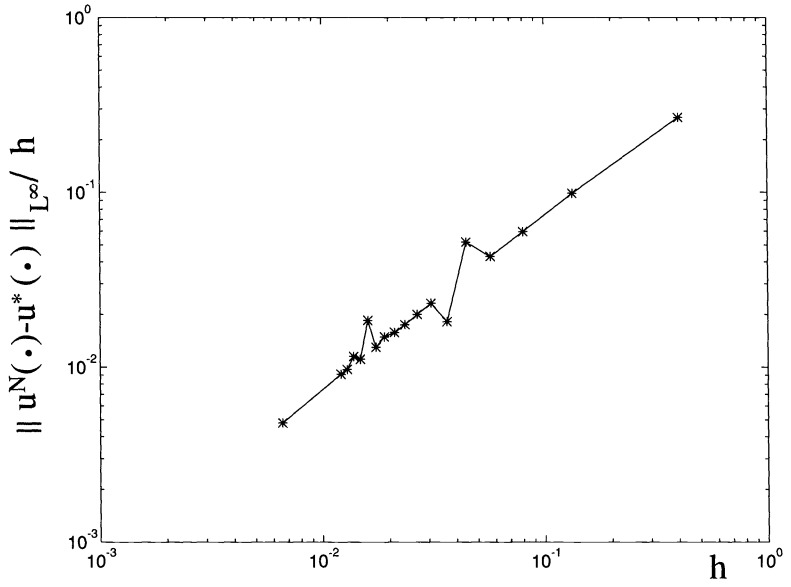


FIG. 4. *The dependence of the relative accuracy on h for the second-order Runge–Kutta method for* $\epsilon = 10^{-6}$.

Acknowledgment. The author thanks Ilya Kolmanovsky for his help in computations.

REFERENCES

[1] W. ALT, *On the approximation of infinite optimization problems with an application to optimal control problems*, Appl. Math. Optim., 12 (1984), pp. 15–27.

[2] E. R. AVAKOV AND F. P. VASIL'EV, *Difference approximation of a maximin problem of optimal control with phase constraints*, Vestnik Moscov. Univ. Ser. XV Vychisl. Mat. Kibernet., 2 (1982), pp. 11–17. (In Russian.)

[3] M. BARDI AND M. FALCONE, *An approximation scheme for the minimum time function*, SIAM J. Control Optim., 28 (1990), pp. 950–965.

[4] J. M. BORWEIN AND A. S. LEWIS, *Duality relationship for entropy-like minimization problems*, SIAM J. Control Optim., 29 (1991), pp. 325–338.

[5] B. M. BUDAK, E. M. BERKOVICH, AND E. N. SOLOV'EVA, *Difference approximations in optimal control problems*, SIAM J. Control, 7 (1969), pp. 18–31. (Originally published in Russian in Vestnik Moskov. Univ. Ser. I Mat. Mekh., 1968, no. 2, pp. 41–55.)

[6] ———, *The convergence of difference approximations in optimal control*, Zh. Vychisl. Mat. i Mat. Fiz., 9 (1969), pp. 533–547. (In Russian.)

[7] B. M. BUDAK AND F. P. VASIL'EV, *Some Computational Aspects of Optimal Control Problems*, Moscow University Press, Moscow, 1975. (In Russian.)

[8] I. CAPUZZO DOLCETTA, *On a discrete approximation of the Hamilton-Jacobi equation of dynamic programming*, Appl. Math. Optim., 10 (1983), pp. 367–377.

[9] I. CAPUZZO DOLCETTA AND M. FALCONE, *Discrete dynamic programming and viscosity solutions of the Bellman equation*, Ann. Inst. H. Poincaré Anal. Non-Lineaire, 6 (1989), suppl., pp. 161–183.

[10] M. G. CRANDALL AND P. L. LIONS, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. of Comp., 43 (1984), pp. 1–19.

[11] J. CULLUM, *Discrete approximations to continuous optimal control problems*, SIAM J. Control, 7 (1969), pp. 32–49.

[12] ———, *An explicit procedure for discretizing continuous, optimal control problems*, J. Optim. Theory Appl., 8 (1971), pp. 15–34.

[13] ———, *Finite-dimensional approximations of state constrained continuous optimal control problems*, SIAM J. Control, 10 (1972), pp. 649–670.

[14] J. W. DANIEL, *On the approximate minimization of functionals*, Math. Comp., 23 (1969), pp. 573–581.

[15] ———, *On the convergence of a numerical method in optimal control*, J. Optim. Theory Appl., 4 (1969), pp. 330–342.

[16] ———, *The Approximate Minimization of Functionals*, Wiley-Interscience, New York 1983.

[17] A. L. DONTCHEV, *Error estimates for a discrete approximation to constrained control problems*, SIAM J. Numer. Anal., 18 (1981), pp. 500–514.
[18] ———, *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, Lecture Notes in Control and Inform. Sci. 52, Springer, New York, 1983.
[19] ———, *Duality methods in constrained best interpolation*, Mathematica Balkanica, 1 (1987), pp. 96–105.
[20] ———, *Best interpolation in a strip*, J. Approx. Theory, 73 (1993), pp. 334–342.
[21] A. L. DONTCHEV AND W. W. HAGER, *Lipschitz stability in nonlinear control and optimization*, SIAM J. Control Optim., 31 (1993), pp. 569–603.
[22] A. L. DONTCHEV, W. W. HAGER, A. B. POORE, AND B. YANG, *Optimality, stability and convergence in nonlinear control*, Appl. Math. Optim., 31 (1995), pp. 297–326.
[23] A. L. DONTCHEV AND W. W. HAGER, *Euler approximation to the feasible set*, Numer. Funct. Anal. Optim., 15 (1994), pp. 245–261.
[24] A. L. DONTCHEV AND F. LEMPIO, *Difference methods for differential inclusions—a survey*, SIAM Rev., 34 (1992), pp. 263–294.
[25] A. L. DONTCHEV AND T. ZOLEZZI, *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer, New York, 1993.
[26] J. C. DUNN, *Diagonally modified conditional gradient method for input constrained optimal control problems*, SIAM J. Control Optim., 24 (1986), pp. 1177–1191.
[27] J. C. DUNN AND T. TIAN, *Variants of the Kuhn-Tucker sufficient conditions in cones of nonnegative functions*, SIAM J. Control Optim., 30 (1992), pp. 1361–1348.
[28] Y. M. ERMOL'EV, V. P. GULENKO, AND T. I. TSARENKO, *Finite element methods in optimal control*, Naukova Dumka, Kiev, 1978. (In Russian.)
[29] M. FALCONE, *A numerical approach to the infinite horizon problem of deterministic control theory*, Appl. Math. Optim., 15 (1987), pp. 1–13. Corrigenda: Appl. Math. Optim., 23 (1991), pp. 213–214.
[30] R. P. FEDORENKO, *Approximate Solution of Optimal Control Problems*, Nauka, Moscow, 1971. (In Russian.)
[31] R. GONZALES AND E. ROFMAN, *On deterministic control problems: An approximation procedure for the optimal cost*, SIAM J. Control Optim., 23 (1985), *Part* I: *The stationary problem*, pp. 242–265, *Part* II: *The nonstationary case*, pp. 267–285.
[32] W. W. HAGER, *The Ritz-Trefftz method for state and control constrained optimal control problems*, SIAM J. Numer. Anal., 12 (1975), pp. 854–867.
[33] ———, *Rate of convergence for discrete approximations to unconstrained control problems*, SIAM J. Numer. Anal., 13 (1976), pp. 449–471.
[34] ———, *Convex control and dual approximations*, Control Cybernet., 8 (1979), part 1: pp. 5–12, part 2: pp. 321–338.
[35] ———, *Multiplier method for nonlinear optimal control*, SIAM J. Numer. Anal., 27 (1990), pp. 1061–1080.
[36] W. W. HAGER AND G. D. IANCULESCU, *Dual approximations in optimal control*, SIAM J. Control Optim., 22 (1984), pp. 423–465.
[37] C. T. KELLEY AND E. W. SACHS, *Mesh independence of the gradient projection method for optimal control problems*, SIAM J. Control Optim., 30 (1992), pp. 477–493.
[38] A. A. LEVIKOV, *Error estimate for solution of the linear optimal control problem*, Avtomat. i Telemekh., 3 (1982), pp. 71–78. (In Russian.)
[39] K. C. P. MACHIELSEN, *Numerical Solution of Optimal Control Problems with State Constraints by Sequential Quadratic Programming in Function Space*, Dissertation, Technische Hogeschool Eindhoven, Eindhoven, 1987.
[40] K. MALANOWSKI, *On convergence of finite-difference approximations to control and state constrained optimal control problems*, Archiwum Aut. i Telem., 24 (1979), pp. 319–337.
[41] ———, *On convergence of finite-difference approximations to optimal control problems for systems with control appearing linearly*, Archiwum Aut. i Telem., 24 (1979), pp. 155–171.
[42] ———, *Convergence of approximations vs. regularity of solutions for convex control-constrained optimal control problems*, Appl. Math. Optim., 8 (1981), pp. 69–95.
[43] B. MORDUKHOVICH, *On difference approximations of optimal control systems*, Appl. Math. Mech., 42 (1978), pp. 452–461.
[44] ———, *Approximation Method in Problems of Optimization and Control*, Nauka, Moscow, 1988. (In Russian.)
[45] ———, *Discrete Approximations and Refined Euler-Lagrange Conditions for Nonconvex Differential Inclusions*, IMA preprint series 1115, March 1993.
[46] H. J. PESCH, *A practical guide to the solution of real-life optimal control problems*, Control Cybernet., 23 (1994), pp. 7–60.
[47] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.
[48] ———, *A historical survey of computations methods in optimal control*, SIAM Rev., 15 (1973), pp. 553–548.
[49] ———, *On the use of consistent approximations in the solution of semi-infinite optimization and optimal control problems*, Math. Programming, 62 (1993), pp. 385–414.
[50] E. POLAK AND LIMIN HE, *Rate-preserving strategies for semi-infinite programming and optimal control*, SIAM J. Control Optim., 30 (1992), pp. 543–572.

[51] E. POLAK, T. H. YANG, AND D. Q. MAYNE, *A method of centers based on barrier functions for solving optimal control problems with continuous state and control constraints*, SIAM J. Control Optim., 31 (1993), pp. 159–179.

[52] G. W. REDDIEN, *Collocation at Gauss points as a discretization in optimal control*, SIAM J. Control Optim., 17 (1979), pp. 298–316.

[53] K. L. TEO, C. J. GOH AND K. H. WANG, *A Unified Computational Approach to Optimal Control Problems*, Longman Sci. Tech., Harlow, 1991.

[54] F. P. VASIL'EV, *Methods for Solving Extremum Problems*, Nauka, Moscow, 1981. (In Russian.)

[55] V. V. VASIN, *Discrete approximation and stability in extremal problems*, Zh. Vychisl. Mat. i Mat. Fiz., 22 (1982), pp. 824–839.

[56] V. M. VELIOV, *Second order discrete approximations to strongly convex differential inclusions*, Systems Control Lett., 13 (1989), pp. 263–269.

[57] ———, *Second order discrete approximations to linear differential inclusions*, SIAM J. Numer. Anal., 29 (1992), pp. 439–451.

[58] ———, *Best approximations of control/uncertain differential systems by means of discrete-time systems*, IIASA WP-91-45, November 1991.

[59] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–145.

[60] P. R. WOLENSKI, *The exponential formula for the reachable set of a Lipschitz differential inclusion*, SIAM J. Control Optim., 28 (1990), pp. 1148–1161.

[61] S. E. WRIGHT, *Consistency of primal-dual approximations for convex optimal control problems*, SIAM J. Control Optim., 33 (1995), pp. 1489–1509.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1329–1341, July 1996

012

A NEVANLINNA–PICK APPROACH TO TIME-DOMAIN CONSTRAINED $\mathcal{H}_\infty$ CONTROL*

HÉCTOR ROTSTEIN†

Abstract. In this paper, generalized Nevanlinna–Pick theory is used to solve a time-domain constrained $\mathcal{H}_\infty$ control problem for linear time-invariant discrete-time systems. First it is shown that if constraints are imposed only over a finite horizon (i.e., only on the first n samples), then the problem reduces to a finite-dimensional convex minimization problem. Subsequently it is shown that if these problems are conveniently modified, then letting the horizon length go to infinity produces a solution to the infinite-horizon problem.

Key words. $\mathcal{H}_\infty$ control, interpolation theory, time-domain constraints

AMS subject classifications. 93B36, 93B28

1. Introduction.

Consider the standard setup for linear time-invariant $\mathcal{H}_\infty$ control illustrated in Fig. 1. The objective is to design a controller that minimizes the worst-case l_2-norm of the controlled variable y, assuming that the disturbance w lies in a (weighted) ball of l_2 signals. Numerous control problems can be reduced to this setup, including problems of disturbance rejection, robust stability, and signal tracking [6, 7]. Suppose, for instance, that one wants to track a given signal r with the output y. In order to formulate this as an $\mathcal{H}_\infty$ problem, one starts by modeling the input by the set

$$\mathcal{W} \doteq \{\hat{r} = Ww \text{ s.t. } \|w\|_2 \leq 1\},$$

where W is a stable weighting function reflecting the frequency content of r, and $r \in \mathcal{W}$. Then, if $T_{e\hat{r}}$ denotes the closed-loop transfer function between $\hat{r}$ and $\hat{r} - y$, the $\mathcal{H}_\infty$ objective becomes the minimization of the $\mathcal{H}_\infty$ norm of $WT_{e\hat{r}}$. (The energy of the control action is usually also penalized to guarantee the properness of the optimal controller.) If the resulting norm is "small," then good tracking can be guaranteed for all signals in $\mathcal{W}$ and hence, in particular, for r. This procedure has two main drawbacks. First, since performance for a *fixed* signal r is replaced by performance for a set of signals $\mathcal{W}$, the result will be generically conservative. This difficulty can be partially alleviated by taking the tracking signal into account when formulating the $\mathcal{H}_\infty$ objective, as proposed in [3]. Second, the $\mathcal{H}_\infty$ norm provides a bound on the l_2-norm of the tracking error, but very little can be inferred about the time response of the signal, in particular over the first few samples.

Specifications for the time responses of the closed-loop system, such as the one considered in the previous paragraph, are not amicable to standard $\mathcal{H}_\infty$ control. Therefore, until recently, one was left with a trial-and-error procedure, by which weighting functions (such as W in the example above) were iteratively adjusted in an attempt to enforce the specifications. In the absence of a formal procedure for constructing weighting functions from time-domain specifications, this trial-and-error procedure may become arduous and time consuming, without guarantees of producing a solution, even if one exists. This problem gave rise to interest in a theory that would include both the $\mathcal{H}_\infty$ objective and time-domain constraints. The first attempts in this direction were to use the Youla lemma [6] to parameterize all feasible (i.e., resulting from stabilizing controllers) closed-loop transfer functions and then compute the free variable in the parameterization through a constrained optimization procedure [2, 13]. Although this constitutes in principle a reasonable approach, it is not completely satisfactory

*Received by the editors June 1, 1993; accepted for publication (in revised form) March 30, 1995.

†Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (hector@ee.technion.ac.il).

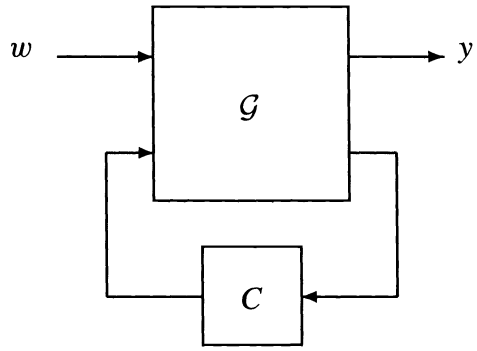


FIG. 1. *Standard $\mathcal{H}_\infty$ setup.*

since it provides little insight into the problem and the computation of a solution becomes very expensive, even for relatively modest problems.

A different approach to the $\mathcal{H}_\infty$ constrained problem was initiated by Helton and Sideris in [8], where Lawson's algorithm for solving the Nehari problem was modified to accommodate time-domain constraints. The resulting program is a combination of Lawson's and quadratic programming steps; it is in the latter that constraints appear. The original idea had several drawbacks, including the lack of a proof of convergence and the fact that it was tailored for a specific $\mathcal{H}_\infty$ problem and cannot be extended directly into the general setup. Also, the approach was based on imposing the constrains only over a finite horizon, i.e., up to a time instant $n-1$. Although this is intuitively plausible, since by stability one expects that the responses will achieve some steady-state value for n sufficiently large, it turned out to be problematic in the subsequent developments. The first two difficulties were largely overcome in [14, 15, 17]. In [17] the problem considered by Helton and Sideris, i.e., a scalar one-block $\mathcal{H}_\infty$ problem with constraints imposed over a finite horizon, was solved by resorting to Nehari's theorem. In addition to providing a better understanding of some of its properties, the approach reduced the problem to a finite-dimensional convex minimization, thus establishing the convergence issue. Also, the special structure of the objective function was exploited to obtain an algorithm for solving the minimization with relatively low computational complexity. Subsequently, in [15], the setup of Fig. 2 was considered. Here the objective is to guarantee an $\mathcal{H}_\infty$ norm bound between w_f and y_f while satisfying some time-domain specifications between w_t and y_t. It was shown that, although technically more involved, the general problem can be reduced to a generalized version of the one considered in [17], and hence it inherits its main properties, including convexity and the bound on the computational complexity. It is interesting to remark that previous approaches to the problem scaled badly to the multivariable case. The original problem has been extended in several different directions; for instance, in [19–21] a mixed $l_\infty/\mathcal{H}_\infty$ problem is solved by using similar techniques.

The aim of the present paper is to solve the time-domain constrained $\mathcal{H}_\infty$ problem by using existing results in interpolation theory. The study was motivated by the facts that a variety of solution techniques have been used to solve the standard $\mathcal{H}_\infty$ problem and that the interplay of the various approaches has provided considerable additional knowledge. In particular, the interested reader may consult [12] for a pointer to relevant literature in interpolation theory and its role in the progress of the field. Since the main motivation of this paper is to obtain further insights into the problem, a simple instance of the setup illustrated in Fig. 2 will be considered, thus sacrificing generality in order to get cleaner results.

The paper is organized as follows. Section 2 is devoted to a review of relevant results in interpolation theory, more specifically a particular generalization of the Nevanlinna–Pick

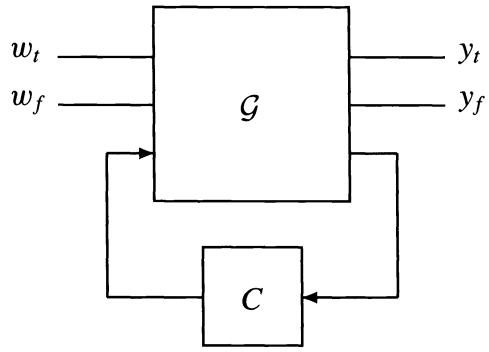


FIG. 2. *Time-domain constrained $\mathcal{H}_\infty$ setup.*

theory borrowed from [1]. Section 3 describes the prototype problem which is studied in this paper. Section 4 contains the solution of the problem when constraints are imposed only over a finite number of sample points, and constitutes the interpolation counterpart of [17]. Section 5 contains the major new contribution of this paper, namely, the construction of a solution when constraints are imposed over an infinite horizon, i.e., for all sampling instants. It is worth stressing that this result does not follow in a straightforward manner from the finite-horizon case considered in [17] and [15]. In fact, it is not clear at this point that just letting the horizon length go to infinity will eventually produce a solution to the original problem, and a special construction is required to guarantee this fact. Section 6 contains the conclusions and a brief discussion of other approaches to the infinite-horizon problem.

2. Preliminaries on interpolation theory. It is well known that interpolation theory can be used to solve numerous linear system problems. For instance, [9, 10] used a result of Nevanlinna–Pick to study the robust stabilization problem, while [11] used the same tool to solve a version of the $\mathcal{H}_\infty$ problem. A common simplifying assumption in these works is that all interpolation points have multiplicity one; since in the present case some multiplicities are assumed to be larger than one (see the discussion below), a result which is more general than the standard is required and reviewed next, adapted from the comprehensive book [1].

Some notation is needed in order to formulate the problem and its solution. Consider the closed unit disk $\mathcal{D} = \{z : |z| \leq 1\}$ and let $\mathcal{H}_\infty$ denote the set of complex functions analytic on the interior of $\mathcal{D}$ and essentially bounded on $\mathcal{D}$. Let $f(z) \in \mathcal{H}_\infty$; then $\|f(z)\|_\infty \doteq \text{ess sup}_\theta |f(e^{j\theta})|$, and the closed unit ball in $\mathcal{H}_\infty$ is denoted by $\overline{\mathcal{BH}}_\infty$. Let $\mathcal{RH}_\infty$ denote the subset of real rational transfer function of $\mathcal{H}_\infty$ and $\mathcal{BRH}_\infty$ ($\overline{\mathcal{BRH}}_\infty$) denote the set of functions in $\mathcal{RH}_\infty$ with norm less than (less than or equal to) 1. Finally, let l_1 denote the set of absolutely summable sequences; i.e., if $f = \{f_1, f_2, \ldots\} \in l_1$, then $\|f\|_1 = \sum_{i=0}^{\infty} |f_i| < \infty$.

In the original Nevanlinna–Pick formulation, the data are given by some points $z_1, z_2, \ldots, z_r$ in the interior of $\mathcal{D}$, $z_i \neq z_j$ for $i \neq j$, and some interpolation values $w_1, w_2, \ldots, w_r$. The aim is then to find conditions under which there exists a function $f \in \mathcal{BRH}_\infty$ (or $\overline{\mathcal{BRH}}_\infty$) such that $f(z_i) = w_i$ and possibly to parameterize all such solutions. A related classical problem is the Carathéodory interpolation problem, in which the data are a polynomial of the form

$$\hat{f}(z) = f_0 + f_1 z + \cdots + f_n z^n$$

and the objective is to find necessary and sufficient conditions for the existence of a function $f \in \mathcal{BRH}_\infty$ (or $\overline{\mathcal{BRH}}_\infty$) such that

$$f(z) = f_0 + f_1 z + \cdots + f_n z^n + z^{n+1} f^n(z)$$

for some $f^n(z) \in \mathcal{RH}_\infty$.

Nevanlinna–Pick and Carathéodory problems may be considered in a unified way by introducing the following formulation.

PROBLEM 1. *Find necessary and sufficient conditions under which there exists an* $f \in \mathcal{BRH}_\infty$ *(or* $\overline{\mathcal{BRH}}_\infty$*) such that*

$$\sum_{z_o \in \mathcal{D}} Res_{z=z_o} f(z) C_-(zI - A)^{-1} = C_+. \tag{1}$$

Problem 1 is a special case of the much broader theory considered in [1] but suffices for the purposes of this paper. The next two examples (see also Examples 18.5.1 and 22.2.1 in [1]) show that the claim that the interpolation condition (1) includes the original Nevanlinna–Pick and Carathéodory problems as special cases is indeed justified. These examples are instrumental in solving the time-domain constrained $\mathcal{H}_\infty$ problem.

Example 2.1. Let $\Gamma = \operatorname{diag}\{z_i\} \in \mathbb{C}^{r\times r}$, and take

$$\begin{aligned} A &= \Gamma, \\ C_- &= \mathbb{1}^T = [1\ 1\ \cdots\ 1] \in \mathbb{R}^r, \\ C_+ &= [w_1\ w_2\ \cdots\ w_r]; \end{aligned}$$

then the interpolation constraint 1 is satisfied if and only if

$$f(z_i) = w_i, \quad i = 1, \ldots, r. \tag{2}$$

Example 2.2. Let $I_{n\times n}$ denote the $n \times n$ identity matrix, and $A_f \in \mathbb{R}^{(n+1)\times(n+1)}$,

$$A_f = \begin{bmatrix} 0 & I_{n\times n} \\ 0 & 0 \end{bmatrix}.$$

Take

$$\begin{aligned} A &= A_f, \\ C_- &= e_1^T = [1\ 0\ \cdots\ 0] \in \mathbb{R}^{n+1}, \\ C_+ &= [f_0\ f_1\ \cdots\ f_n]; \end{aligned}$$

then the interpolation constraint (1) is satisfied if and only if $f(z)$ can be written as

$$f(z) = f_0 + f_1 z + f_2 z^2 + \cdots + f_n z^n + \cdots. \tag{3}$$

Now consider the Stein (or discrete-time Lyapunov) equation

$$M = A^T M A + C_-^T C_- - C_+^T C_+. \tag{4}$$

Under the condition that A has all its eigenvalues in $\mathcal{D}$, a Hermitian solution to (4) exists and is unique. The solution M is called the generalized Pick matrix associated with Problem 1 and plays a key role for finding necessary and sufficient conditions for the solvability of the interpolation problem.

THEOREM 2.3. *Let* M *solve* (4). *Then there exist a transfer function* $f(z) \in \mathcal{BRH}_\infty$ *(*$\overline{\mathcal{BRH}}_\infty$*) such that* (1) *is satisfied if and only if* M *is positive definite (positive semidefinite).*

Proof. See Theorem 18.5.1 of [1] and the note on page 404 therein. □

It is also possible to parameterize all solutions to the interpolation problem cited above, but Theorem 2.3 suffices for the purposes of the present paper. Necessary and sufficient conditions for the existence of a solution to the Nevanlinna–Pick and Carathéodory interpolation problems follow easily from the theorem.

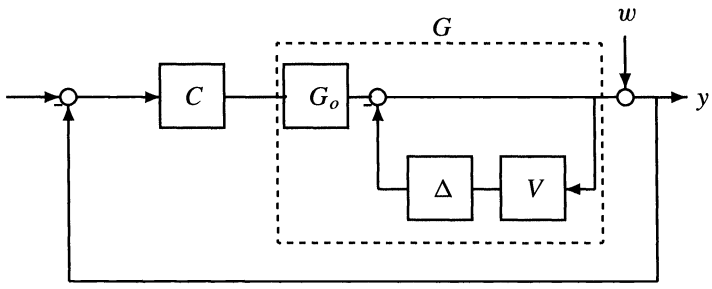


FIG. 3. *The prototype problem.*

COROLLARY 2.4. *Let P be the standard Pick matrix*

$$P = \left[\frac{1 - w_i w_j}{1 - z_i z_j}\right]_{ij}. \tag{5}$$

Then there is a transfer function $f(z) \in \mathcal{BRH}_\infty$ such that (2) is satisfied if and only if P is positive definite.

Proof. Some simple algebra shows that P given by (5) satisfies (4) for A, C_-, and C_+ considered in Example 2.1. □

COROLLARY 2.5. *Let $\mathcal{F}_n$ be the matrix*

$$\mathcal{F}_n = \begin{bmatrix} f_0 & f_1 & \cdots & f_n \\ 0 & f_0 & \cdots & f_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f_0 \end{bmatrix}. \tag{6}$$

Then there is a transfer function $f(z) \in \overline{\mathcal{BRH}}_\infty$ such that (3) is satisfied if and only if $\mathcal{F}_n$ is a contraction (i.e., $\overline{\sigma}(\mathcal{F}_n) \leq 1$).

Proof. Again, some simple algebra shows that $M_c = I - \mathcal{F}_n^T \mathcal{F}_n$ solves (4) for A, C_-, and C_+ considered in Example 2.2. Since this matrix is positive semidefinite if and only if $\overline{\sigma}(\mathcal{F}_n) \leq 1$, the corollary follows from Theorem 2.3. □

3. The control problem and its formulation. The purpose of this section is twofold. First the control problem which will be studied in the rest of the paper is introduced. Then the problem is formulated to resemble an interpolation problem.

3.1. Prototype problem. As anticipated in the introduction, the example to be considered next is not introduced due to its intrinsic significance but rather because it allows a transparent derivation while keeping technical details and notation as simple as possible. More general cases, like problems involving multivariable one-block $\mathcal{H}_\infty$ objectives or time-domain constraints over several transfer functions, can be derived by using a generalized version of the interpolation theory reviewed in the previous section and a more involved notation.

Consider the disturbance rejection problem illustrated in Fig. 3. Here $G = G_o/(1 + \Delta V^{-1})$ is a linear time-invariant discrete-time plant with nominal finite-dimensional transfer function G_o. Δ is assumed to be stable and such that $\|\Delta\| < 1$ but otherwise unknown. V is a weighting function assumed to be finite dimensional and outer (i.e., both V and $V^{-1} \in \mathcal{RH}_\infty$). Let $t(z) = 1/(1 + G_o C)$ denote the *nominal* transfer function from w to y. If C stabilizes the nominal closed loop, then $t(z)$ can be expanded as $t(z) = \sum_{k=0}^{\infty} t_k z^k$.

The prototype control problem considered in this paper is to design a controller $C(z)$ such that

O1) C internally stabilizes the closed loop for all Δ's that satisfy the conditions stated above.

O2) The nominal closed-loop transfer function from w to y is such that, for every k,

$$lb_k \leq t_k \leq ub_k, \tag{7}$$

where the two sequences $lb = \{lb_0, lb_1, \ldots\}$ and $ub = \{ub_0, ub_1, \ldots\}$ are assumed to be in l_1. The control problem is hence a combination of robust stabilization O1 and nominal performance O2. Necessary and sufficient conditions for the existence of a controller satisfying O1 are well documented in the literature.

LEMMA 3.1. *A controller C satisfies* O1 *if and only if*

$$\|V^{-1}/(1 + G_oC)\|_\infty \leq 1. \tag{8}$$

Proof. See, for instance, [5]. □

It follows from the lemma that O1 gives rise to an $\mathcal{H}_\infty$ control problem, while O2 imposes time-domain constraints.

3.2. Problem formulation. Given G_o and V_o as above, the robust stabilization problem is said to be solvable (strictly solvable) if there exists a controller C such that (8) holds (respectively, holds with strict inequality). Necessary and sufficient conditions for solvability using interpolation theory have been known for some time.

LEMMA 3.2. *Let G_o have r_z zeros $z_1, \ldots, z_{r_z}$ and r_p poles $z_{r_z+1}, \ldots, z_{r_z+r_p}$ in the interior of $\mathcal{D}$, $z_i \neq z_j$ for $i \neq j$, and V be as above. Then the robust stabilization problem is solvable (strictly solvable) if and only if the Pick matrix* (5) *is positive semidefinite (positive definite), with $w_i = V^{-1}(z_i)$, $i = 1, \ldots, r_z$, $w_i = 0$, $i = r_z + 1, \ldots, r = r_z + r_p$.*

For further discussion on the relationship between interpolation theory and robust stabilization, see [12].

The next step is to reformulate the control problem into a form that resembles the generalized Nevanlinna–Pick problem reviewed in §2.

THEOREM 3.3. *Let G_o have r_z zeros $z_1, \ldots, z_{r_z}$ and r_p poles $z_{r_z+1}, \ldots, z_{r_z+r_p}$ in the interior of $\mathcal{D}$, $z_i \neq z_j$ for $i \neq j$. Then the control problem considered in the previous section has a solution if and only if there exist $f \in \overline{\mathcal{BRH}}_\infty$ such that*

$$f(z_i) = V^{-1}(z_i) \doteq w_i, \quad i = 1, \ldots, r_z, \tag{9}$$

$$f(z_i) = 0, \quad i = r_z + 1, \ldots, r_z + r_p, \tag{10}$$

$$lb_k \leq t_k \leq ub_k \quad \forall k, \tag{11}$$

and $t(z) = V(z)f(z)$.

Proof. Let $f(z) \doteq V^{-1}(z)\frac{1}{1+G_o(z)C(z)}$ denote the weighted sensitivity transfer function. Then, from Lemma 3.2, $f \in \mathcal{H}_\infty$, $\|f\| \leq 1$, and conditions (9), (10) are equivalent to the robust stability of the closed loop. Finally, condition (11) is a restatement of (7). □

The problem considered in Theorem 3.3 is similar to an interpolation problem, except that the coefficients f_k are not given some specific values (as in a Carathéodory problem) or left unspecified (as in a Nevanlinna–Pick problem) but are constrained by inequalities. For this reason, it is referred to as a *constrained interpolation problem* (CIP). In the next section, it is shown that Theorem 2.3 can be used to obtain necessary and sufficient conditions for the solvability of the CIP if constraints are imposed only over a finite horizon. For future reference let $r = r_z + r_u$.

4. The finite-horizon case. Suppose that one wants to impose the constraint $lb_k \leq t_k \leq ub_k$ only for $k \leq n < \infty$ or, equivalently, that there exists an n in (11) such that $-lb_k$ and ub_k

are arbitrarily large for $k > n$. By the causality of $V(z) = \sum_{i=0}^{\infty} v_i z^i$, (11) may be written as the matrix inequality

$$\begin{bmatrix} lb_0 \\ lb_1 \\ \cdots \\ lb_n \end{bmatrix} \le \begin{bmatrix} v_0 & 0 & \cdots & 0 \\ v_1 & v_0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ v_n & v_{n-1} & \cdots & v_0 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{bmatrix} \le \begin{bmatrix} ub_0 \\ ub_1 \\ \cdots \\ ub_n \end{bmatrix}$$

or, in a compact notation,

$$\mathbf{lb_n} \le \mathbf{V_n f_n} \le \mathbf{ub_n}. \tag{12}$$

The value n is referred to as the horizon length, and the corresponding constrained interpolation problem is called CIP_n. Note that in a CIP_n, constraints are imposed over $n+1$ samples. With a small abuse of notation, the original problem CIP is said to have constraints over an infinite horizon, while the robust stabilization problem is referred to as the unconstrained problem.

From (12), CIP_n has only a finite number of constraints; as shown in the following theorem, this implies that it can be reduced to a finite-dimensional minimization problem.

THEOREM 4.1. *Let G_o, V be as above, and assume that the unconstrained problem is strictly solvable. Then the CIP_n has a solution if and only if there exists a vector $\mathbf{f_n} = [f_0\ f_1 \cdots f_n]^T \in \mathbb{R}^{n+1}$ satisfying (12) such that $\overline{\sigma}(\hat{M}(\mathbf{f_n})) \le 1$, where $\overline{\sigma}(\hat{M}(\mathbf{f_n})$ denotes the largest singular value of the matrix*

$$\hat{M}(\mathbf{f_n}) = \begin{bmatrix} \mathcal{F}_n \\ P^{-1/2}(S^o - S^l \mathcal{F}_n) \end{bmatrix}, \tag{13}$$

with

$$S^o = \begin{bmatrix} 1 & z_1 & z_1^2 & \cdots & z_1^n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & z_{r_z} & z_{r_z}^2 & \cdots & z_{r_z}^n \\ 1 & z_{r_z+1} & z_{r_z+1}^2 & \cdots & z_{r_z+1}^n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & z_r & z_r^2 & \cdots & z_r^n \end{bmatrix},$$

$$S^l = \begin{bmatrix} w_1 & w_1 z_1 & w_1 z_1^2 & \cdots & w_1 z_1^n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ w_{r_z} & w_{r_z} z_{r_z} & w_{r_z} z_{r_z}^2 & \cdots & w_{r_z} z_{r_z}^n \\ & & 0 & & \end{bmatrix},$$

$$\mathcal{F}_n = \begin{bmatrix} f_0 & f_1 & \cdots & f_n \\ 0 & f_0 & \cdots & f_{n-1} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & f_0 \end{bmatrix},$$

and P is the Pick matrix (5).

Proof. Assuming that the f_i's are fixed, CIP_n reduces to the generalized Nevanlinna–Pick interpolation problem of finding $f \in \mathcal{BRH}_\infty$ such that

$$\begin{aligned} f(z_i) &= w_i, \quad i = 1, \ldots, r_z, \\ f(z_i) &= 0, \quad i = r_z + 1, \ldots, r, \\ f(z) &= f_0 + f_1 z + f_2 z^2 + \cdots + f_n z^n + \cdots. \end{aligned}$$

Let $A = \left[\begin{smallmatrix} \Gamma & 0 \\ 0 & A_f \end{smallmatrix}\right]$, where Γ and A_f are as defined in Examples 2.1 and 2.2, respectively. Let

$$\begin{aligned} C_- &= [\mathbb{1}^T \; e_1^T], \\ C_+ &= [w_1 \; w_2 \; \cdots \; w_r \; f_0 \; f_1 \; \cdots \; f_n]. \end{aligned}$$

From Examples 2.1 and 2.2, the interpolation problem has a solution if and only the corresponding generalized Pick matrix $M(\mathbf{f_n})$ in Theorem 2.3 is positive definite. It is claimed that

$$M(\mathbf{f_n}) = \begin{bmatrix} P & S^o - S^l\mathcal{F}_n \\ (S^o - S^l\mathcal{F}_n)^T & I - \mathcal{F}_n^T\mathcal{F}_n \end{bmatrix}. \tag{14}$$

To see this, begin by partitioning M as

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix},$$

compatible with the structure of A. Then if M satisfies the Stein equation (4),

$$M_{11} = \Gamma M_{11}\Gamma + \mathbb{1}\mathbb{1}^T - \mathbf{w}\mathbf{w}^T, \tag{15}$$

$$M_{22} = A_f^T M_{22} A_f + e_1 e_1^T - \mathbf{f_n}\mathbf{f_n}^T, \tag{16}$$

$$M_{12} = \Gamma M_{12} A_f + \mathbb{1} e_1^T - \mathbf{w}\mathbf{f_n}^T, \tag{17}$$

where $\mathbf{w_n} = [w_1 \; \cdots \; w_{r_z} \; 0 \; \cdots 0]^T$. From (15), (16), and Examples 2.1 and 2.2,

$$\begin{aligned} M_{11} &= P, \\ M_{22} &= I - \mathcal{F}_n^T\mathcal{F}_n, \end{aligned}$$

Multiplying the left-hand side of (17) by e_1,

$$M_{12}e_1 = \mathbb{1} - \mathbf{w}f_0,$$

and by e_2,

$$M_{12}e_2 = \Gamma M_{12}e_1 - \mathbf{w}f_1 = \Gamma\mathbb{1} - \Gamma\mathbf{w}f_0 - \mathbf{w}f_1.$$

Continuing this process, one gets

$$M_{12} = S^o - S^l\mathcal{F}_n$$

as required. Note that $M(\cdot)$ depends quadratically on $\mathbf{f_n} = [f_0 \; f_1 \cdots f_n]^T$.

By Lemma 3.2 and the assumption on strict solvability of the unconstrained problem, P is positive definite and hence invertible. M is then positive semidefinite if and only if

$$I - \mathcal{F}_n^T\mathcal{F}_n - (S^o - S^l\mathcal{F}_n)^T P^{-1}(S^o - S^l\mathcal{F}_n) \geq 0,$$

which is equivalent to

$$\rho\left(\mathcal{F}_n^T\mathcal{F}_n - (S^o - S^l\mathcal{F}_n)^T P^{-1}(S^o - S^l\mathcal{F}_n)\right) \geq 1,$$

where ρ denotes the spectral radius, or to

$$\overline{\sigma}\left(\begin{bmatrix} \mathcal{F}_n \\ P^{-1/2}(S^o - S^l\mathcal{F}_n) \end{bmatrix}\right) \leq 1. \tag{18}$$

It follows from these calculations that CIP_n has a solution if and only if there exists a vector $\mathbf{f_n}$ satisfying (12) such that (18) holds. □

The block diagonal terms of M in (14) show that if the CIP is solvable, then the associated Nevanlinna–Pick and Carathéodory problems also have a solution; the condition is *not* sufficient, because of the off-diagonal terms in M.

Theorem 4.1 reduces the solution of an $\mathcal{H}_\infty$ optimization problem with constraints over a finite horizon to a finite-dimensional optimization problem, namely, the minimization of the largest singular value of $N(\mathbf{f_n})$ subject to constraints on $\mathbf{f_n}$. Given the linear dependence of $N(\mathbf{f_n})$ on $\mathbf{f_n}$, and the linearity of the constraints, the optimization problem is convex but in general nondifferentiable. Since, in addition, a large number of variables and constraints are usually involved, solving this problem can be difficult; in particular, general procedures for nondifferentiable programming may not produce a solution at a reasonable cost. An algorithm for solving this type of problems based on the ellipsoid algorithm was outlined in [14, 15, 17].

5. The infinite-horizon case. A first approach to solving the infinite horizon problem would be to take a horizon length n large and compute a vector $\mathbf{f_n}$ that satisfies the constraints (12) and minimizes $\overline{\sigma}(\hat{M}(\mathbf{f_n}))$. Letting $\mu_n \doteq \overline{\sigma}(\hat{M}(\mathbf{f_n}))$, if $\mu_n > 1$, then no solution to CIP exists; but if $\mu_n \leq 1$, then one could construct a solution $f^n(z)$ to CIP_n and then check whether (7) is also satisfied for $k > n$. Note, however, that for $\mu_n < 1$ the solution is highly nonunique and there is no systematic way of selecting $f^n(z)$ so that the violation of the constraints is minimized or driven to zero. To circumvent this difficulty, one could select a particular solution (e.g., the minimum entropy one) to the CIP_n and keep increasing the horizon length in the hope that constraints for all time instants will eventually be satisfied.

Unfortunately, research on the finite-horizon problem using a Nehari approach [15, 17] shows that the infinite-horizon case may not always be addressed directly by taking the limit for the horizon length going to infinity. This is because little can be said about the time-domain behavior after the horizon (i.e., for $k > n$), which turns out to be unacceptable if the optimal solution (in terms of the $\mathcal{H}_\infty$ norm) is pursued. This difficulty was circumvented by solving an approximate problem in which the closed-loop poles are constrained to lie in a disk of radius $\rho < 1$; the fact that all poles have an absolute value strictly less than 1 induces a decay rate on the time responses and hence implies a good (i.e., small) behavior for $k > n$ if the horizon length n is taken to be larger than some precomputable bound. One can then argue that under some suitable continuity condition, letting $\rho \to 1$ would produce in the limit a solution to the infinite-horizon problem. However, since continuity properties are hard to ascertain (note that both the $\mathcal{H}_\infty$ problem *and* the constraints should be continuous in some sense), a more direct approach is pursued in this section.

LEMMA 5.1. *The set of transfer functions $f(z)$ such that $t(z) = V(z)f(z)$ satisfies the time-domain constraints* (7) *is compact as a subset of $\mathcal{H}_\infty$.*

Proof. The proof is based on the fact that both ub and lb are assumed to be in l_1. Let

$$\mathcal{FS} = \left\{ t(z) = \sum_{k=0}^{\infty} t_k z^k \text{ s.t. } lb_k \leq t_k \leq ub_k \ \forall k \right\} \subset \mathcal{H}_\infty.$$

The first step is to show that $\mathcal{FS}$ is compact as a subset of l_1; since the l_1 norm bounds the $\mathcal{H}_\infty$ norm, the lemma will follow.

The space l_1 is complete and $\mathcal{FS}$ is closed, and hence it suffices to show that for every $\epsilon > 0$ it is possible to cover $\mathcal{FS}$ with a finite number of balls of l_1 with radius ϵ [4, p. 59]. Thus, let $\epsilon > 0$ be given. By the assumption $\sum_{k=0}^{\infty} |lb_k| < \infty$, $\sum_{k=0}^{\infty} |ub_k| < \infty$, there exists an N such that

$$\sum_{k=N}^{\infty} |lb_k| < \epsilon/4, \tag{19}$$

$$\sum_{k=N}^{\infty} |ub_k| < \epsilon/4, \tag{20}$$

implying $\sum_{k=N}^{\infty} |t_k| < \epsilon/2 \ \forall\, t \in \mathcal{FS}$. Now consider the set

$$\mathcal{FS}_N \doteq \{[t_0\, t_1\, \cdots\, t_{N-1}] \in \mathbb{R}^N \text{ s.t. } lb_k \le t_k \le ub_k,\ 0 \le k < n\}.$$

$\mathcal{FS}_N$ is a closed and bounded set in $\mathbb{R}^N$, and hence there exist a finite number of balls $\mathcal{B}(\hat{t}^i, \epsilon/2)$, with center at $\hat{t}^i \doteq [t_0^i\, t_1^i\, \cdots\, t_{N-1}^i]$ and radius $\epsilon/2$ that cover $\mathcal{FS}_N$. Define

$$t^i \doteq [t_0^i\, t_1^i\, \cdots\, t_{N-1}^i\, 0\, 0\, \cdots],$$

and consider the balls $\mathcal{B}(t^i, \epsilon) \subset l_1$. Consider now any $t \in \mathcal{FS}$. By construction, there exists an i_o such that $\sum_{k=0}^{N-1} |t_k - t_k^{i_o}| < \epsilon/2$, and then

$$\begin{aligned} \|t - t^{i_o}\|_1 &\le \sum_{k=0}^{N-1} |t_k - t_k^{i_o}| + \sum_{k=N}^{\infty} |t_k| \\ &< \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

Since the finite number of balls $\mathcal{B}(f^i, \epsilon)$ covers $\mathcal{FS}$, the set is compact. Since $V^{-1} \in \mathcal{RH}_\infty$, the proof follows. □

Let $f^{(n)}(z)$ denote the minimum entropy solution to CIP$_n$, and let $\hat{f}^{(n)}(z)$ denote

$$\hat{f}^{(n)} = f_0^n + f_1^n z^1 + f_2^n z^2 + \cdots + f_n^n z^n,$$

i.e., its projection over the space generated by $\{1, z, \ldots, z^n\}$. From Lemma 5.1, an accumulation point for the sequence of transfer functions $\{\hat{f}^{(n)}\}$ constitutes a natural candidate for solving the infinite-horizon problem. Unfortunately, the resulting accumulation point does not necessarily satisfy the interpolation constraints (9), (10), which are satisfied by $f^{(n)}$ and *not* by $\hat{f}^{(n)}$. One could enforce constraints (9), (10) on $\hat{f}^{(n)}$ by modifying the minimization problem to include the constraints explicitly, but it is not clear how this would affect the convergence to the optimal solution. Instead, if the CIP$_n$ are modified so that (9), (10) are only approximately satisfied by $\hat{f}^{(n)}$, it is possible to establish that a solution to the infinite-horizon problem can be computed by letting the horizon length n go to infinity. More explicitly, consider the following modification of the CIP$_n$ problem.

PROBLEM 2 (MCIP$_n$). *Find* $f \in \mathcal{BRH}_\infty$ *such that*

$$\begin{aligned} f(z_i) &= w_i, \qquad i = 1, \ldots, r_z, \\ f(z_i) &= 0, \qquad i = r_z + 1, \ldots, r_z + r_s, \\ lb_k &\le t_k \le ub_k, \quad 0 \le k \le n, \end{aligned}$$

$$\left| \sum_{k=0}^{n} f_k z_i^k - w_i \right| \le \epsilon^n, \qquad i = 1, \ldots, r_z, \tag{21}$$

$$\left| \sum_{k=0}^{n} f_k z_i^k \right| \le \epsilon^n, \qquad i = r_z + 1, \ldots, r, \tag{22}$$

where $\epsilon < 1$.

A careful choice of ϵ now shows that the solvability of the MCIP$_n$ eventually implies the solvability of the infinite-horizon problem.

THEOREM 5.2. *Let* $1 > \epsilon > \max_{i=1,\ldots,r} |z_i|$. *Then the infinite-horizon problem is solvable if and only if there exists an* N *such that the MCIP*$_n$ *is solvable for all* $n > N$.

Proof. Suppose that there exists N such that the MCIP$_n$ is solvable for all $n > N$, and let $f^{(n)}$ denote a solution for the MCIP$_n$. As before, $f^{(n)}$ can be the minimum entropy solution for definiteness. Let $\hat{f}^{(n)}$ be defined as in the proof of Lemma 5.1:

$$\hat{f}_k^{(n)} = \left\{ \begin{array}{ll} f_k^{(n)} & \text{if } k \leq n \\ 0 & \text{if } k > n \end{array} \right\}$$

(i.e., $\hat{f}^{(n)}$ is found by truncating $f^{(n)}$), and consider the sequence of functions $\{\hat{f}^{(n)}\}_n$. It is clear that $\hat{f}^{(n)}$ satisfies the constraint (11) for all n, and hence Lemma 5.1 implies that there exists a convergent subsequence $\{\hat{f}^{n_k}\}$. Let $f \doteq \lim_{k\to\infty} \hat{f}^{n_k}$. Then it is claimed that f solves the infinite-horizon problem. To see this, note that the interpolation constraints (9) and (10), are satisfied in the limit due to the additional constraints (21) and (22) in the MCIP$_n$. The constraint (11) is also satisfied because all $\hat{f}^{(n)}$ do satisfy it, and the conditions on lb and ub imply that $f \in \mathcal{H}_\infty$. Finally $\|f\| \leq 1$ because $\overline{\sigma}(\mathcal{F}_n) \to \|f\|$ for $n \to \infty$.

To prove the converse, assume that f solves the CIP. Then f is a feasible point for the CIP$_n$ for all n, and hence it suffices to show that there exists an N such that f also satisfies (21) and (22) for each $n \geq N$. Since f satisfies (11), $\sum_{j=0}^{\infty} f_j z_i^j = 0$ for $i = r_z + 1, \ldots, r$, and

$$\left| \sum_{j=n+1}^{\infty} f_j z_i^j \right| \leq |z_i|^{n+1} \sum_{j=n}^{\infty} |f_j|, \qquad i = r_z + 1, \ldots, r.$$

Since f satisfies the time-domain constraints, $\sum_{j=0}^{\infty} |f_j| < \infty$, implying that there exists an N such that $\sum_{j=N+1}^{\infty} |f_j| < 1$. Then, if $n > N$,

$$\left| \sum_{n+1}^{\infty} f_k z_i^k \right| \leq |z_i|^n \leq \epsilon^n$$

or

$$\left| \sum_{k=0}^{n} f_k z_i^k \right| \leq \epsilon^n$$

for all $i = r_z, \ldots, r$; hence f satisfies the additional constraint (22). A similar argument proves that f satisfies (21), and hence the MCIP$_n$ is feasible for all $n \geq N$. This concludes the proof. □

The condition in Theorem 5.2 is hard to check, since it involves the solution of MCIP$_n$ for arbitrarily large values of n, and solving the associated optimization problem becomes progressively more expensive as n increases. In practice, this inconvenience can be circumvented as follows. First, solve the optimization problem associated with MCIP$_n$ for n large; the horizon length is related to the location of the singularities z_i, $i = 1, \ldots, r$. Then check if either $\hat{f}^{(n)}$, perturbed to satisfy the interpolation constraints, also satisfies the norm bound or $f^{(n)}$ has a reasonable time behavior after the horizon. The first choice is justified by Theorem 5.2, since it implies that the finite impulse response solution $\hat{f}^{(n)}$ to the corresponding MCIP$_n$ eventually provides an approximate solution to the infinite-horizon problem. On the other hand, numerical experience suggests that picking $f^{(n)}$ often yields a solution to the infinite-horizon problem, specially if μ^n is significantly smaller than one.

6. Conclusions. In this paper the time-domain constrained $\mathcal{H}_\infty$ problem has been studied, using interpolation theory. If constraints are imposed over a finite horizon, then a generalized Nevanlinna–Pick theorem from [1] can be used to reduce the problem into a finite-dimensional, convex minimization. Theorem 4.1 shows that the solution to this problem is less than or equal to 1 if and only if there exists a solution to the original constrained $\mathcal{H}_\infty$ problem. If constraints are imposed over an infinite horizon, then Theorem 5.2 shows that a solution exists if and only if the solution to a sequence of finite-dimensional convex problems is bounded by one. It is worth stressing that the additional insight provided by the interpolation theory approach allowed the solution to the infinite-horizon case, since previous works have concentrated on the finite-horizon problem. After the submission of this manuscript, substantial progress has been achieved in the solution to the infinite-horizon problem, partially fueled by the convergence theorem presented in §5. For instance, in [16] the normal convergence of the finite-horizon solutions is established under no assumptions on the bounds ub and lb and without introducing additional constraints. In [18] the assumptions are stronger, but again convergence is established without resorting to the approximate interpolation constraints.

This work can be extended in several directions. First, although attention has been focused in finding necessary and sufficient conditions for the solvability of the constrained problem, state-space formulas for the computation of an actual solution (i.e., a controller) may be given by using the corresponding results in interpolation theory. Then, the one-block multivariable $\mathcal{H}_\infty$ problem can be solved, and the feasibility of solving the general four-block $\mathcal{H}_\infty$ could be investigated. Also, the problem of imposing time-domain constraints on a transfer function while bounding the $\mathcal{H}_\infty$-norm of another transfer function can be addressed. Although the finite-horizon problem is apparently easy to extend to this case as well, it is not clear at this point if the infinite-horizon result will extend as nicely. This is also true for the multivariable case and is currently under investigation.

Acknowledgments. The author wishes to thank Prof. Athanasios Sideris for his suggestion of using approximate interpolation constraints for the infinite-horizon problem, and the anonymous reviewers for their remarks and suggestions.

REFERENCES

[1] J. BALL, I. GOHBERG, AND L. RODMAN, *Interpolation of Rational Matrix Functions*, Oper. Theory Adv. Appl. 45, Birkhäuser-Verlag, Basel, Boston, Berlin, 1990.

[2] S. BOYD AND C. BARRATT, *Linear Controllers Design—Limits of Performance*, Prentice-Hall, Englewood Cliffs, NJ, 1990.

[3] C. DE SOUZA AND U. SHAKED, *Robust $\mathcal{H}_\infty$ filtering with parametric uncertainty and deterministic input signal*, in Proc. 31st CDC, Tucson, AZ, 1993.

[4] J. DIXMIER, *General Topology*, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1984.

[5] J. DOYLE, B. FRANCIS, AND A. TANNENBAUM, *Feedback Control Theory*, Macmillan, New York, 1992.

[6] B. FRANCIS, *A Course in $\mathcal{H}_\infty$ Control Theory*, Lecture Notes in Control and Inform. Sci., Springer-Verlag, New York, 1987.

[7] B. FRANCIS AND J. DOYLE, *Linear control theory with an $\mathcal{H}_\infty$ optimally criterion*, SIAM J. Control Optim., 25 (1985), pp. 815–844.

[8] J. W. HELTON AND A. SIDERIS, *Frequency response algorithms for $\mathcal{H}_\infty$ optimization with time domain constraints*, IEEE Trans. Automat. Control, 34 (1989), pp. 427–434.

[9] P. KHARGONEKAR AND A. TANNENBAUM, *Non-euclidean metrics and the robust stabilization of systems with parameter uncertainty*, IEEE Trans. Automat. Control, 30 (1985), pp. 1005–1013.

[10] H. KIMURA, *Robust stabilizability for a class of transfer functions*, IEEE Trans. Automat. Control, 29 (1984), pp. 788–793.

[11] ———, *Directional interpolation in the state-space*, Systems Control Lett., 10 (1988), pp. 317–324.

[12] ———, *State space approach to the classical interpolation problem and its application*, in Three Decades of Mathematical System Theory: A Collection of Surveys at the Occasion of the 50th Birthday of Jan C. Willems, Lecture Notes in Control and Information Sciences 135, H. Nijmeijer and J. Schumacher, eds., Springer-Verlag, New York, 1990, pp. 179–218.

[13] E. POLAK AND S. E. SALCUDEAN, *On the design of linear multivariable feedback systems via constrained nondifferentiable optimization in $\mathcal{H}_\infty$ spaces*, IEEE Trans. Automat. Control, 34 (1989), pp. 268–276.

[14] H. ROTSTEIN, *Constrained $\mathcal{H}_\infty$-Optimization for Discrete-Time Control Systems*, Ph.D. thesis, California Institute of Technology, 1992.

[15] H. ROTSTEIN AND A. SIDERIS, *$\mathcal{H}_\infty$ optimization with time domain constraints*, IEEE Trans. Automat. Control, 39 (1994), pp. 762–779.

[16] ———, *$\mathcal{H}_\infty$-control with time domain constraints: The infinite horizon case*, Systems Control Lett., 24 (1995), pp. 251–258.

[17] A. SIDERIS AND H. ROTSTEIN, *Single input-single output $\mathcal{H}_\infty$-control with time domain constraints*, Automatica, 29 (1993), pp. 969–983.

[18] ———, *Constrained $\mathcal{H}_\infty$ optimal control over an infinite horizon*, SIAM J. Control Optim., submitted.

[19] M. SZNAIER, *A mixed $l_\infty/\mathcal{H}_\infty$ approach to robust controller design*, in Proc. 1992 American Control Conference, Chicago, IL, 1992, pp. 727–732.

[20] M. SZNAIER AND F. BLANCHINI, *Mixed $L^\infty/\mathcal{H}^\infty$ suboptimal controllers for siso continuous-time systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 1831–1840.

[21] M. SZNAIER AND H. ROTSTEIN, *Robust controller design for a noncolocated spring-mass system via mixed $l_\infty/\mathcal{H}_\infty$ optimization*, International Journal of Robust and Nonlinear Control, 5 (1994), pp. 53–65.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1342–1364, July 1996

013

PARTIALLY OBSERVED DIFFERENTIAL GAMES, INFINITE-DIMENSIONAL HAMILTON–JACOBI–ISAACS EQUATIONS, AND NONLINEAR H_∞ CONTROL*

M. R. JAMES† AND J. S. BARAS‡

Abstract. This paper presents new results for partially observed nonlinear differential games. Using the concept of *information state*, we solve this problem in terms of an infinite-dimensional partial differential equation, which turns out to be the Hamilton–Jacobi–Isaacs (HJI) equation for partially observed differential games. We give definitions of smooth and viscosity solutions and prove that the value function is a viscosity solution of the HJI equation. We prove a verification theorem, which implies that the optimal controls are separated in that they depend on the observations through the information state. This constitutes a separation principle for partially observed differential games. We also present some new results concerning the certainty equivalence principle under certain standard assumptions. Our results are applied to a nonlinear output feedback H_∞ robust control problem.

Key words. partially observed differential games, infinite-dimensional partial differential equations, viscosity solutions, nonlinear H_∞ robust control

AMS subject classifications. 90D25, 93B36, 93C10, 93C41, 49L25, 35R15

1. Introduction.

The nonlinear H_∞ robust control problem has generated considerable activity in recent years, and important contributions have been made by a number of authors; see [1]–[3], [5], [8], [9], [12], [16], [20]–[24], [26], [29]–[35]. The state feedback problem is reasonably well understood, although the issue of controller synthesis for continuous-time systems remains outstanding. This is because the value functions solving the various partial differential equations (PDEs) that have been proposed need not be smooth—a standard difficulty even for simple deterministic optimal control problems. The output feedback problem is much more difficult, and various approaches have been suggested in the literature. Perhaps the most general of these approaches was initiated in [24], [25], where the concept of *information state* was used to solve a partially observed dynamic game, and applied in [23] to solve the output feedback H_∞ problem (see also the discussion in [6]). The results in [24], [23] are presented in the discrete-time context for technical simplicity, although the system-theoretic ideas are valid in continuous-time also; ideed, the key equations were presented in [25], [32] and later in [6]. The purpose of this paper is to commence the task of developing a mathematical theory for continuous-time partially observed differential games and output feedback H_∞ robust control.

The information state $p_t = p_t(x)$ is the solution of a first-order PDE and takes values in a suitable infinite-dimensional Banach space $p_t \in \mathcal{X}$ (here, $x \in \mathbf{R}^n$ is the state of the system being controlled, so $\mathcal{X}$ is a space of real-valued functions of x). The partially observed differential game that we consider can be transformed into an equivalent game with full state information, and this leads via dynamic programming to a value or optimal cost function $W(p, t)$ that "solves" a PDE on $\mathcal{X} \times [0, T]$. This PDE is a nonlinear first-order equation and is the correct Hamilton–Jacobi–Isaacs (HJI) equation for partially observed differential games. This HJI equation appears to be new, and we are not aware of any results in the literature concerning this type of infinite-dimensional PDE. It is not clear what if anything the results in [7] have to say about this HJI equation. In the case of partially observed stochastic control,

*Received by the editors August 24, 1994; accepted for publication (in revised form) April 1, 1995. This research was supported by the funding of the activities of the Cooperative Research Centre for Robust and Adaptive Systems by the Australian Commonwealth Government under the Cooperative Research Centers Program.

†Department of Engineering, Faculty of Engineering and Information Technology, Australian National University, Canberra, ACT 0200, Australia.

‡Martin Marietta Chair in Systems Engineering, Department of Electrical Engineering and Institute for Systems Research, University of Maryland, College Park, MD 20742. The research of this author was supported in part by NSF grant NSFD CDR 8803012 through the Engineering Research Centers Program.

the idea of information state is familiar, a theory has been developed [14], [17], [27], and the dynamic programming equation is an infinite-dimensional nonlinear second-order PDE.

The particular class of problems that we consider in this paper is presented in §2. This class should be regarded as a prototype class, and the ideas and principles we develop are expected to apply in much more general contexts. The relevant information state is defined in §3, and some if its properties are analyzed for use in later sections. In particular, the key representation theorem is given. In §4, the value function and HJI equation are defined and studied. Definitions of smooth and viscosity solutions are given. We prove that the value function is a viscosity solution of the HJI equation. We do not know a proof of a uniqueness or comparison theorem for equations of this type, and consequently our definition of viscosity solution should be regarded as a provisional one. While in general it is not expected that smooth solutions will exist, a verification theorem is proven in §5 assuming a smooth solution exists, yielding that the optimal control is a separated control in the sense that it depends on the observations via the information state. The certainty equivalence principle proposed in [5] and [9] is considered in §6. We explain how this principle fits into the general information state framework and show that, under a generalization of the standard assumptions, the certainty equivalence controller can be optimal at certain values of the information state. The standard assumptions are very stringent and are unlikely to hold in general, and we explain what can happen in such an event. In §7, we apply our results to a relatively simple nonlinear H_∞ control problem, viz., finite horizon disturbance attenuation. The solution is expressed in terms of two PDEs, a finite-dimensional one for the information state and an infinite-dimensional equation for the value function. Infinite horizon H_∞ problems are closely related to the theory of dissipative systems [18], [36], and we present the relevant partial differential inequality (PDI) for the output feedback problem. Finally, we make some comments concerning more general cases.

2. Problem formulation. We consider the class of nonlinear partially observed deterministic systems described by the state space equations

$$\begin{cases} \dot{x}(t) = f(x(t), u(t)) + g(x(t), w(t)), \\ y(t) = h(x(t)) + w(t). \end{cases} \tag{2.1}$$

Here, $x(t) \in \mathbf{R}^n$ denotes the state of the system and is not directly measurable; instead, an output quantity $y(t) \in \mathbf{R}^p$ is observed. The control input is $u(t) \in U \subset \mathbf{R}^m$, and $w(t) \in \mathbf{R}^p$ is regarded as an opposing disturbance input. The functions $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$, $g : \mathbf{R}^n \times \mathbf{R}^p \to \mathbf{R}^{n\times m}$, and $h : \mathbf{R}^n \to \mathbf{R}^n$ are assumed bounded and smooth with bounded derivatives of orders up to three, say. The set U is compact.

Most of this paper is concerned with a differential game problem on a finite-time horizon $[0, T]$, and we use the following type of admissible strategies. The admissible disturbances are the square integrable functions

$$w \in \mathcal{W}(t) = L_2([t, T], \mathbf{R}^p),$$

while the admissible controls are the nonanticipating (causal) maps

$$\mathbf{u} : \mathcal{Y}(t) \to \mathcal{U}(t),$$

where

$$\mathcal{U}(t) = L_2([t, T], U), \quad \mathcal{Y}(t) = L_2([t, T], \mathbf{R}^p).$$

The nonanticipating property means that if $y_1, y_2 \in \mathcal{Y}(t)$ and $y_1(r) = y_2(r)$ a.e. $r \in [t, s]$, then $\mathbf{u}[y_1](r) = \mathbf{u}[y_2](r)$ a.e. $r \in [t, s]$ (cf. the Elliott–Kalton notion of strategy [10], [11]).

We will denote by $\mathbf{U}(t)$ the class of such nonanticipating strategies for which (2.1) and (3.2) (for $u = \mathbf{u}[y]$) have unique solutions.

We next introduce several function spaces that will be used in the sequel. The Banach space of continuous functions with at most linear growth is denoted

$$\mathcal{X} = \{p \in C(\mathbf{R}^n) \ : \| \ p \ \| < \infty\},$$

where the norm is defined by

$$\| \ p \ \| = \sup_{x \in \mathbf{R}^n} \frac{|p(x)|}{1 + |x|}.$$

Denote by

$$\mathcal{X}^1 = \{p \in C^1(\mathbf{R}^n) \ : \| \ p \ \|_1 < \infty\}$$

the Banach space of continuously differentiable functions with bounded derivatives equipped with norm

$$\| \ p \ \|_1 = \sup_{x \in \mathbf{R}^n} \frac{|p(x)|}{1 + |x|} + \sup_{x \in \mathbf{R}^n} |\nabla_x p(x)|,$$

where $\nabla_x p$ is the gradient of p. Also, we need to define the function space

$$\mathcal{D} = \{p \in C(\mathbf{R}^n) \ : \ p(x) \leq -c_1|x| + c_2 \ \forall x \in \mathbf{R}^n, \text{ for some } c_1 > 0, c_2 \in \mathbf{R}\}.$$

Note that the subsets $\mathcal{D} \cap \mathcal{X} \subset \mathcal{X}$ and $\mathcal{D} \cap \mathcal{X}^1 \subset \mathcal{X}^1$ are open in their respective topologies. As sets, $\mathcal{X}^1 \subset \mathcal{X}$, but $\mathcal{X}^1$ is not a subspace of $\mathcal{X}$ as Banach spaces.

The minimax differential game is defined as follows. The payoff is

$$J(\mathbf{u}, w, x_0) = \alpha(x_0) + \int_0^T [L(x(t), \mathbf{u}[y](t)) - \gamma^2 \ell(w(t))] \, dt + \Phi(x(T)),$$

where the initial state $x(0) = x_0$ is in general unknown. The functions $L : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$, $\ell : \mathbf{R}^p \to \mathbf{R}$, and $\Phi : \mathbf{R}^n \to \mathbf{R}$ are assumed bounded and smooth with bounded derivatives of orders up to three, and $\alpha \in \mathcal{D} \cap \mathcal{X}$. The assumptions imply that J is well defined and bounded uniformly in $w \in \mathcal{W}(0)$ and $\mathbf{u} \in \mathbf{U}(0)$. We will also assume that $L \geq 0$, $\ell \geq 0$, $\Phi \geq 0$. The controller's objective is to minimize J, while the disturbances attempt to maximize J. For $\mathbf{u} \in \mathbf{U}(0)$ define the functional

$$J(\mathbf{u}) = \sup_{w \in \mathcal{W}(0), \ x_0 \in \mathbf{R}^n} J(\mathbf{u}, w, x_0).$$

The problem addressed in this paper is that of minimizing $J(\mathbf{u})$ over $\mathbf{u} \in \mathbf{U}(0)$. This is a partially observed minimax differential game. Note that x_0 is regarded as an unknown opponent also.

3. Information state. The key to solving the partially observed game is to replace it by an equivalent one with full state information. The difficulty is that the new state is infinite dimensional in general.

To this end, for fixed output path $y \in \mathcal{Y}(0)$ and control $u \in \mathcal{U}(0)$, define the *information state* by

$$p_t(x) = \alpha(x_0) + \int_0^t [L(x(s), u(s)) - \gamma^2 \ell \, (y(s) - h(x(s)))] \, ds, \tag{3.1}$$

where $x(\cdot)$ is the solution of

$$\dot{x}(s) = f(x(s), u(s)) + g(x(s), y(s) - h(x(s))), \qquad 0 < s < t, \tag{3.2}$$

with terminal condition $x(t) = x$. This quantity describes the worst-case performance up to time t using the control u, which is consistent with the observed output and the constraint $x(t) = x$. It summarizes the observed information in a way that is suitable for fulfilling the control objective. The information state evolves according to the dynamics

$$\begin{cases} \dot{p}_t = F(p_t, u(t), y(t)), \\ p_0 = \alpha, \end{cases} \tag{3.3}$$

where F is the differential operator

$$F(p, u, y) = -\nabla_x p \cdot (f(\cdot, u) + g(\cdot, y - h)) + L(\cdot, u) - \gamma^2 \ell(y - h), \tag{3.4}$$

defined on a domain in $\mathcal{X} \times \mathbf{R}^m \times \mathbf{R}^p$ mapping into $\mathcal{X}$. (Note that F is not continuous on $\mathcal{X} \times \mathbf{R}^m \times \mathbf{R}^p$ but is continuous from $\mathcal{X}^1 \times \mathbf{R}^m \times \mathbf{R}^p$ to $\mathcal{X}$.)

The smoothness of $p_t(x)$ and consequently the sense in which (3.3) is to be understood depends on the smoothness of the initial data α (the other data are assumed smooth) and on the regularity of $u(\cdot)$ and $y(\cdot)$.

DEFINITION 3.1. *We say that a function $p_t(x)$ is a* smooth solution *of the dynamics* (3.3) *if $\alpha \in \mathcal{D} \cap \mathcal{X}^1$; and when $u \in \mathcal{U}(0)$ and $y \in \mathcal{Y}(0)$ are continuous,*

(i) *$p_t(x)$ is of class $C^1(\mathbf{R}^n \times [0, T])$ and*

(ii) *$p_t(x)$ satisfies* (3.3) *in $\mathbf{R}^n \times (0, T)$ in the usual sense.*

LEMMA 3.2. *If $\alpha \in \mathcal{D} \cap \mathcal{X}^1$, then $p_t(x)$ is the unique smooth solution of* (3.3), *and for any $u \in \mathcal{U}(0)$ and $y \in \mathcal{Y}(0)$, the information state p_t evolves in $\mathcal{D} \cap \mathcal{X}^1$ as*

$$p_t \in \mathcal{D} \cap \mathcal{X}^1, \qquad t \in [0, T], \tag{3.5}$$

whenever $\alpha \in \mathcal{D} \cap \mathcal{X}^1$. Moreover, the map $t \mapsto p_t$ from $[0, T]$ into $\mathcal{D} \cap \mathcal{X}^1$ is continuous, with modulus of continuity independent of $u \in \mathcal{U}(0)$, $y \in \mathcal{Y}(0)$.

Proof. 1. Let $\alpha \in \mathcal{D} \cap \mathcal{X}^1$, $u \in \mathcal{U}(0) \cap C([0, T], \mathbf{R}^m)$, and $y \in \mathcal{Y}(0) \cap C([0, T], \mathbf{R}^p)$. Then by the method of characteristics (see, e.g., [13]), we see that (3.3) has a unique solution given by (3.1) that is of class C^1, and moreover, the gradient has the representation

$$\begin{aligned} \nabla_x p_t(x) &= \nabla_x \alpha(x(0)) \\ &\quad + \int_0^t [\nabla_x L(x(s), u(s)) + \gamma^2 \nabla_w \ell\,(y(s) - h(x(s)))\, \nabla_x h(x(s))] \Xi(s)\, ds, \end{aligned} \tag{3.6}$$

where $x(\cdot)$ is the solution of (3.2) and

$$\begin{aligned} \dot{\Xi}(s) = [&\nabla_x f(x(s), u(s)) + \nabla_x g(x(s), y(s) - h(x(s))) \\ &- \nabla_w g(x(s), y(s) - h(x(s)) \nabla_x h(x(s))] \Xi(s), \end{aligned} \tag{3.7}$$

$0 \le s \le t$, where $\Xi(t) = I$.

2. These formulas are also valid for $u \in \mathcal{U}(0)$, $y \in \mathcal{Y}(0)$, by an approximation argument using continuous functions, as follows. Let $u^i \to u$ in $\mathcal{U}(0)$ and $y^i \to y$ in $\mathcal{Y}(0)$ as $i \to \infty$, where each u^i and y^i is continuous. We claim that

$$\lim_{i \to \infty} \sup_{0 \le t \le T} \| p_t^i - p_t \|_1 = 0, \tag{3.8}$$

where $p_\cdot^i$ and $p_\cdot$ denote the corresponding solutions of (3.3) with initial data α.

To prove this, let $x^i(\cdot)$ and $x(\cdot)$ denote the corresponding solutions of (3.2) with terminal data $x^i(t) = x, x(t) = x$. Then a standard estimate using the Gronwall and Holder inequalities gives

$$|x^i(s) - x(s)| \leq K(\| u^i - u \|_{L_2} + \| u^i - u \|_{L_2}),$$

for $s \in [0, t]$, where $K > 0$ is independent of $t \in [0, T]$, $x \in \mathbf{R}^n$. This implies

$$\begin{aligned}
|p_t^i(x) - p_t(x)| &\leq |\alpha(x^i(0)) - \alpha(x(0))| \\
&+ \int_0^t [|L(x^i(s), u^i(s)) - L(x(s), u(s))| + |\ell(y^i(s) - h(x^i(s))) - \ell(y(s) - h(x(s)))|]\, ds \\
&\leq \rho_\alpha(|x^i(0) - x(0)|) + K \int_0^t [|x^i(s) - x(s)| + |u^i(s) - u(s)| + |y^i(s) - y(s)|]\, ds \\
&\leq K(\| u^i - u \|_{L_2} + \| y^i - y \|_{L_2}),
\end{aligned}$$

uniformly, where ρ_α is a modulus of continuity function for α. A similar estimate for $|\nabla_x p_t^i(x) - \nabla_x p_t(x)|$ can be obtained using (3.6). This proves (3.8).

3. Note that by assumption and (3.6), $\nabla_x p_t(x)$ is bounded uniformly in $(x, t) \in \mathbf{R}^n \times [0, T]$. This implies $p_t \in \mathcal{X}^1$. The fact that $p_t \in \mathcal{D} \cap \mathcal{X}^1$ follows from the estimate (3.13).

4. Finally, we claim that there exists a modulus function ρ depending on α but independent of u and y such that

$$\| p_{t^2} - p_{t^1} \|_1 \leq \rho(|t^2 - t^1|). \tag{3.9}$$

To prove (3.9), assume $t^1 \leq t^2$. Let $x^i(\cdot)$ $(i = 1, 2)$ denote the solution of (3.2) with terminal data $x^i(t^i) = x$. Now

$$|x^2(t^1) - x| \leq K|t^2 - t^1|,$$

and for $0 \leq s \leq t^1$, using Gronwall's inequality,

$$|x^1(s) - x^2(s)| \leq K|x^1(t^1) - x^2(t^1)| \leq K|t^2 - t^1|.$$

Therefore

$$\begin{aligned}
|p_{t^2}(x) - p_{t^1}(x)| &\leq |\alpha(x^2(0)) - \alpha(x^1(0))| + \int_0^{t^1} [|L(x^2(s), u(s)) - L(x^1(s), u(s))| \\
&\quad + K|\ell(y(s) - h(x^2(s))) - \ell(y(s) - h(x^1(s)))|]\, ds \\
&\quad + \int_{t^1}^{t^2} [|L(x^2(s), u(s))| + K|g(x^2(s), y(s) - h(x^2(s)))|]\, ds \\
&\leq \rho_\alpha(|x^2(0) - x^1(0)|) + K \int_0^{t^1} |x^2(s) - x^1(s)|\, ds + K|t^2 - t^1| \\
&\leq \rho_\alpha(K|t^2 - t^1|) + K|t^2 - t^1|,
\end{aligned}$$

uniformly in x, u, y. A similar estimate for $|\nabla_x p_{t^2}(x) - \nabla_x p_{t^1}(x)|$ can be obtained using (3.6). This completes the proof. □

If α is not differentiable, then (3.3) can be interpreted in the viscosity sense [13], [28].

DEFINITION 3.3. *We say that a function $p_t(x) \in C(\mathbf{R}^n \times [0, T])$ is a* viscosity subsolution *of the dynamics* (3.3) *if $\alpha \in \mathcal{D} \cap \mathcal{X}$, and if for all $\phi \in C^\infty(\mathbf{R}^n)$, $\psi \in L^1[0, T]$, whenever there*

exists $(x', t') \in \mathbf{R}^n \times (0, T)$ *with* $p_{t'}(x') + \int_0^{t^1} \psi(s)\,ds - \phi(x') = \max_{(x,t)\in\mathbf{R}^n\times[0,T]}(p_t(x) + \int_0^t \psi(s)\,ds - \phi(x))$, *then*

$$\lim_{\delta\to 0} \inf_{|t-t'|<\delta} \operatorname{ess\,inf}\{\psi(t) - \lambda\cdot(f(x,u(t)) + g(x,y(t)-h(x))) + L(x,u(t)) - \gamma^2\ell(y(t)-h(x)) \;:\; |x-x'|<\delta,\ |\lambda - \nabla_x\phi(x')|\le\delta\} \ge 0; \tag{3.10}$$

or a viscosity supersolution *of the dynamics* (3.3) *if* $\alpha \in \mathcal{D}\cap\mathcal{X}$, *and if for all* $\phi \in C^\infty(\mathbf{R}^n)$, $\psi \in L^1[0,T]$, *whenever there exists* $(x', t') \in \mathbf{R}^n\times(0,T)$ *with* $p_{t'}(x') + \int_0^{t^1}\psi(s)\,ds - \phi(x') = \min_{(x,t)\in\mathbf{R}^n\times[0,T]}(p_t(x) + \int_0^t \psi(s)\,ds - \phi(x))$, *then*

$$\lim_{\delta\to 0} \sup_{|t-t'|<\delta} \operatorname{ess\,sup}\{\psi(t) - \lambda\cdot(f(x,u(t)) + g(x,y(t)-h(x))) + L(x,u(t)) - \gamma^2\ell(y(t)-h(x)) \;:\; |x-x'|<\delta,\ |\lambda - \nabla_x\phi(x')|\le\delta\} \le 0; \tag{3.11}$$

or a viscosity solution *if it is both a subsolution and a supersolution.*

LEMMA 3.4. *If* $\alpha \in \mathcal{D}\cap\mathcal{X}$, *then* $p_t(x)$ *is the unique viscosity solution of* (3.3). *Moreover, for any* $u \in \mathcal{U}(0)$ *and* $y \in \mathcal{Y}(0)$, *the information state* p_t *evolves in* $\mathcal{D}\cap\mathcal{X}$ *as*

$$p_t \in \mathcal{D}\cap\mathcal{X}, \qquad t \in [0,T], \tag{3.12}$$

whenever $\alpha \in \mathcal{D}\cap\mathcal{X}$.

Proof. Let $\alpha \in \mathcal{D}\cap\mathcal{X}$, $u \in \mathcal{U}(0)$, and $y \in \mathcal{Y}(0)$. From the formula (3.1) and from well-known continuity properties of ODEs, it is not hard to show that $p_t(x) \in C(\mathbf{R}^n\times[0,T])$, and we omit the details. The fact that $p_t(x)$ is the unique viscosity solution of (3.3) follows from the results in [28].

To show that $p_t \in \mathcal{D}\cap\mathcal{X}$, we must prove the estimate

$$-\underline{c}_1|x| - \underline{c}_2 \le p_t(x) \le -\bar{c}_1|x| + \bar{c}_2 \quad \text{for all } x \in \mathbf{R}^n,\ 0 \le t \le T, \tag{3.13}$$

where the constants are independent of $u \in \mathcal{U}(0)$, $y \in \mathcal{Y}(0)$ but may depend on α. To this end, let $x(\cdot)$ be the solution of (3.2). Then for $0 \le r \le s \le t \le T$, we have

$$|x(s)| \le |x(r)| + K|s-r| \tag{3.14}$$

for some constant $K > 0$. A similar inequality holds if $s \le r$. Therefore,

$$\begin{aligned} p_t(x) &\le \alpha(x(0)) + KT \\ &\le -c_1|x(0)| + c_2 + KT \\ &\le -c_1|x| + c_1KT + c_2 + KT. \end{aligned}$$

This proves the upper bound in (3.13). The lower bound is proven similarly. □

LEMMA 3.5. *The map* $(p,t) \mapsto p_T$ *from* $\mathcal{D}\cap\mathcal{X}\times[0,T]$ *into* $\mathcal{D}\cap\mathcal{X}$ *is continuous, where* p_T *denotes the solution of* (3.3) *at time* T *with initial data* $p_t = p$. *In addition, if* $u \in \mathcal{U}(0)$ *and* $y \in \mathcal{Y}(0)$ *are continuous, then the map* $t \mapsto p_t$ *from* $[0,T]$ *into* $\mathcal{X}$ *is* $\mathcal{X}$*-Frechet differentiable with continuous derivative* $t \mapsto F(p_t, u(t), y(t))$.

Proof. 1. Proof of first assertion. Fix $p^1 \in \mathcal{D}\cap\mathcal{X}$, and consider $p^2 \in \mathcal{D}\cap\mathcal{X}$ and $0 \le t^1 \le t^2 \le T$. If $u \in \mathcal{U}(t^1)$, $y \in \mathcal{Y}(t^1)$, then the natural truncations of u and y belong to $\mathcal{U}(t^2)$ and $\mathcal{Y}(t^2)$, respectively. Similarly, if $u \in \mathcal{U}(t^2)$, $y \in \mathcal{Y}(t^2)$ are given, then one can extend them to elements of $\mathcal{U}(t^1)$ and $\mathcal{Y}(t^1)$ by setting them equal to arbitrary but fixed elements of U and $\mathbf{R}^p$, respectively, on $[t^1, t^2]$. We use this convention to avoid any ambiguity in the sequel. We claim that there exist a constant $K > 0$ and modulus function ρ that may depend on p^1 such that

$$\| p_T^1 - p_T^2 \| \le K \| p^1 - p^2 \| + \rho(|t^2 - t^1|), \tag{3.15}$$

where p_T^i $(i = 1, 2)$ denotes the solution of (3.3) at time T with initial data $p_{t^i}^i = p^i$, and inputs $u(\cdot)$, $y(\cdot)$, with interpretations as explained above. This inequality implies that $(p, t) \mapsto p_T$ is continuous at $(p^1, t^1) \in \mathcal{D} \cap \mathcal{X} \times [0, T]$ (since (3.15) holds also if $t^2 \leq t^1$).

Fix x and let $x^i(\cdot)$ denote the corresponding solutions of (3.2) with terminal data $x^i(T) = x$. Then $x^1(s) = x^2(s)$ for $t^2 \leq s \leq T$ (in particular $x^1(t^2) = x^2(t^2)$),

$$|x^1(t^1) - x^1(t^2)| \leq K|t^2 - t^1|$$

and

$$|x^2(t^2)| \leq |x| + K.$$

Using these estimates,

$$\begin{aligned}
|p_T^1(x) - p_T^2(x)| &\leq |p^1(x^1(t^1)) - p^2(x^2(t^2))| \\
&\quad + \int_{t^1}^{t^2} |L(x^1(s), u(s)) - \gamma^2 \ell(y(s) - h(x^1(s)))|\, ds \\
&\leq |p_1(x^1(t^1)) - p_1(x^2(t^2))| + |p_1(x^2(t^2)) - p^2(x^2(t^2))| + K|t^2 - t^1| \\
&\leq \rho_{p^1}(|x^1(t^1) - x^2(t^2))|) + \| p^1 - p^2 \| (1 + |x^2(t^2)) + K|t^2 - t^1| \\
&\leq \rho_{p^1}(K|t^2 - t^1|) + K \| p^1 - p^2 \| (1 + |x|) + K|t^2 - t^1|.
\end{aligned}$$

This estimate implies (3.15) with $\rho(s) = \rho_{p^1}(K|s|) + K|s|$.

2. Proof of second assertion. Let $\alpha \in \mathcal{D} \cap \mathcal{X}^1$, and assume that u and y are continuous. We must prove that

$$\lim_{\delta \to 0} \frac{\| p_{t+\delta} - p_t - F(p_t, u(t), y(t))\delta \|}{\delta} = 0. \tag{3.16}$$

Since $p_t(x)$ is of class C^1, we have

$$\begin{aligned}
&\frac{|p_{t+\delta}(x) - p_t(x) - F(p_t, u(t), y(t))(x)\delta|}{\delta} \\
&\quad = \left| \frac{1}{\delta} \int_t^{t+\delta} [F(p_s, u(s), y(s))(x) - F(p_t, u(t), y(t))(x)]\, ds \right| \\
&\quad \leq \frac{1}{\delta} \int_t^{t+\delta} [|\nabla_x p_s(x) - \nabla_x p_t(x)||f(x, u(s)) + g(x, y(s) - h(x))| \\
&\qquad + |\nabla_x p_t(x)||f(x, u(s)) - f(x, u(t)) + g(x)(y(s) - y(t))| \\
&\qquad + |L(x, u(s)) - L(x, u(t))| + \gamma^2 K|y(s) - y(t)|]\, ds \\
&\quad \leq \frac{1}{\delta} \int_t^{t+\delta} [K|\nabla_x p_s(x) - \nabla_x p_t(x)| \\
&\qquad + K|f(x, u(s)) - f(x, u(t))| + K(1 + \gamma^2)|y(s) - y(t)| \\
&\qquad + |L(x, u(s)) - L(x, u(t))|]\, ds \\
&\quad \leq \frac{1}{\delta} \int_t^{t+\delta} C\rho(|s - t|)\, ds \ \to 0,
\end{aligned}$$

as $\delta \to 0$ uniformly, where ρ denotes a suitable modulus of continuity function. This follows because of our assumptions on the data and using (3.6), and is enough to prove (3.16). □

Using the definition (3.1), we have the following key representation theorem.

THEOREM 3.6. *For any* $\mathbf{u} \in \mathbf{U}(0)$ *we have*

$$J(\mathbf{u}) = \sup_{y \in \mathcal{Y}(0)} \{(p_T, \Phi) \,:\, p_0 = \alpha\}, \tag{3.17}$$

where $(p, \Phi) = \sup_{x \in \mathbf{R}^n}(p(x) + \Phi(x))$ *is the "sup-pairing"* [24].

Proof. For any $w \in \mathcal{W}(0)$, an output $y \in \mathcal{Y}(0)$ can be defined by solving the ODE (2.1) with $u(t) = \mathbf{u}[y](t)$, $0 \le t \le T$. Conversely, given any $y \in \mathcal{Y}(0)$, a disturbance $w \in \mathcal{W}(0)$ is defined by solving the ODE (3.2) with $u(t) = \mathbf{u}[y](t)$, $0 \le t \le T$, and setting $w(t) \triangleq -h(x(t)) + y(t)$, $0 \le t \le T$. Therefore there is a natural bijection between $\mathcal{W}(0)$ and $\mathcal{Y}(0)$ (for each $\mathbf{u}$). Consequently,

$$\begin{aligned}
J(\mathbf{u}) &= \sup_{w \in \mathcal{W}(0),\ x_0 \in \mathbf{R}^n} \left\{ \alpha(x_0) + \int_0^T [L(x(t), \mathbf{u}[y](t)) - \gamma^2 \ell(w(t))]\, dt + \Phi(x(T)) \right\} \\
&= \sup_{y \in \mathcal{Y}(0),\ x_0 \in \mathbf{R}^n} \left\{ \alpha(x_0) + \int_0^T [L(x(t), \mathbf{u}[y](t)) - \gamma^2 \ell(y(t) - h(x(t)))]\, dt + \Phi(x(T)) \right\} \\
&= \sup_{y \in \mathcal{Y}(0),\ x \in \mathbf{R}^n} \left\{ \alpha(x_0) + \int_0^T [L(x(t), \mathbf{u}[y](t)) - \gamma^2 \ell(y(t) - h(x(t)))]\, dt + \Phi(x(T)) \right. \\
&\qquad\qquad \left. :\, x(T) = x \right\} \\
&= \sup_{y \in \mathcal{Y}(0),\ x \in \mathbf{R}^n} \{p_T(x) + \Phi(x)\} \\
&= \sup_{y \in \mathcal{Y}(0)} \{(p_T, \Phi) \,:\, p_0 = \alpha\}. \qquad \square
\end{aligned}$$

The equivalent differential game with full state information is to minimize the right-hand side of (3.17) over $\mathbf{u} \in \mathbf{U}(0)$ subject to the infinite-dimensional dynamics (3.3).

We conclude this section with a brief discussion of an "adjoint" information state q_t, which runs backward in time and has the interesting property that the sup-pairing (p_t, q_t) is constant [24]. The *adjoint information state* is defined for fixed $u \in \mathcal{U}(t)$ and $y \in \mathcal{Y}(t)$ by

$$q_t(x) = \int_t^T [L(x(s), u(s)) - \gamma^2 \ell\,(y(s) - h(x(s)))]\, ds + \Phi(x(T)), \tag{3.18}$$

where $x(\cdot)$ is the solution of (3.2) on $[t, T]$ with initial data $x(t) = x$. The dynamics for the adjoint information state are

$$\begin{cases} \dot{q}_s &= -F(-q_s, u(s), y(s)), \qquad s \in [t, T], \\ q_T &= \Phi. \end{cases} \tag{3.19}$$

THEOREM 3.7. *The sup-pairing of the information state and the adjoint information state is constant and expressed as*

$$(p_t, q_t) \text{ is independent of } t \in [0, T]. \tag{3.20}$$

Proof. The assertion can be verified easily by combining the definitions (3.1) and (3.18). Alternatively, suppose p_t and q_t are smooth solutions of (3.3) and (3.19), respectively. Define $v(t) = (p_t, q_t) = p_t(\bar{x}(t)) + q(\bar{x}(t))$. Then $\nabla_x p_t(\bar{x}(t)) = -\nabla_x q_t(\bar{x}(t))$ and

$$\begin{aligned}\dot{v}(t) &= \frac{\partial p_t}{\partial t}(\bar{x}(t)) + \frac{\partial q_t}{\partial t}(\bar{x}(t)) \\ &= -\nabla_x p_t(\bar{x}(t)) \cdot (f(\bar{x}(t), u(t)) + g(\bar{x}(t), y(t) - h(\bar{x}(t)))) \\ &\quad + L(\bar{x}(t), u(t)) - \gamma^2 \ell(y - h(\bar{x}(t))) \\ &\quad - \nabla_x q_t(\bar{x}(t)) \cdot (f(\bar{x}(t), u(t)) + g(\bar{x}(t), y(t) - h(\bar{x}(t)))) \\ &\quad - L(\bar{x}(t), u(t)) + \gamma^2 \ell(y - h(\bar{x}(t))) = 0.\end{aligned}$$

This shows that $v(t) = (\alpha, q_0) = (p_T, \Phi)$ is constant, as required. □

4. Value function and the HJI equation. Given Theorem 3.6, one can now apply dynamic programming methods to solve the equivalent problem and, hence, the original partially observed problem. The value function is defined for $(p, t) \in \mathcal{D} \cap \mathcal{X} \times [0, T]$ by

$$W(p, t) = \inf_{\mathbf{u} \in \mathbf{U}(t)} \sup_{y \in \mathcal{Y}(t)} \{(p_T, \Phi) : p_t = p\}. \tag{4.1}$$

This function is finite, as the following lemma shows.

LEMMA 4.1. *For all* $(p, t) \in \mathcal{D} \cap \mathcal{X} \times [0, T]$ *we have*

$$(p, 0) - K \leq W(p, t) \leq K + (p, 0) \tag{4.2}$$

for some constant $K > 0$.

Proof. For any $\mathbf{u} \in \mathbf{U}(t)$ and $y \in \mathcal{Y}(t)$ we have

$$\begin{aligned}p_T(x(T)) + \Phi(x(T)) &= p(x(t)) + \int_t^T [L(x(s), \mathbf{u}[y](s)) - \gamma^2 \ell(y(s) - h(x(s)))]\, ds + \Phi(x(T)) \\ &\leq p(x(t)) + K,\end{aligned}$$

where $K > 0$ does not depend on $\mathbf{u}$, y, and hence

$$(p_T, \Phi) = \sup_{x(T)} \{p(x(T)) + \Phi(x(T))\} \leq \sup_{x(t)} \{p(x(t)) + K\} = (p, 0) + K.$$

This proves the upper bound in (4.2).

To obtain the lower estimate in (4.2), select $x \in \operatorname{argmax} p$. Then

$$(p_T, \Phi) \geq p(x) + \int_t^T [L(x(s), \mathbf{u}[y](s)) - \gamma^2 \ell(y(s) - h(x(s)))]\, ds + \Phi(x(T)) \geq (p, 0) - K,$$

where $x(\cdot)$ is the solution of (2.1) with initial data $x(t) = x$. □

In the next lemma, $B(0, R)$ denotes the ball of radius R centered at 0 in $\mathbf{R}^n$.

LEMMA 4.2. *Fix* $p^1 \in \mathcal{D} \cap \mathcal{X}$. *Then there exist* $\delta^1 > 0$ *and* $R^1 > 0$ *such that* $\| p^2 - p^1 \| < \delta^1$ *implies that* $\operatorname{argmax}_{x \in \mathbf{R}^n} (p^2(x) + \Phi(x)) \subset B(0, R^1)$.

Proof. Since $p^1 \in \mathcal{D}$ and Φ is bounded, $p^1(x) + \Phi(x) \leq -c_1|x| + c_2 + K$. Then for $p^2 \in \mathcal{X}$,

$$p^2(x) + \Phi(x) = p^1(x) + \Phi(x) + (p^2(x) - p^1(x)) \leq -c_1|x| + c_2 + K + \| p^2 - p^1 \| (1 + |x|).$$

Set $\delta^1 = c_1/2$. Then for $\| p^2 - p^1 \| < \delta^1$,

$$p^2(x) + \Phi(x) \leq -c_1'|x| + c_2', \tag{4.3}$$

where $c_1' = c_1/2$, $c_2' = c_2 + K + \delta^1$.

Next, select a sequence x_i such that $\lim_{i\to\infty} p^2(x_i) = \sup_{x\in\mathbf{R}^n} p^2(x) = (p^2, 0) < +\infty$. Fix $\varepsilon > 0$. Then for all large i,

$$(p^2, \Phi) - \varepsilon \le p^2(x_i) \le -c_1'|x_i| + c_2',$$

and hence

$$|x_i| \le R^1 \tag{4.4}$$

for some constant $R^1 > 0$ depending on p^1 and δ^1. Thus the sequence x_i is bounded, and any limit point x^2 satisfies $|x^2| \le R^1$. Hence $\operatorname{argmax}_{x\in\mathbf{R}^n}(p^2(x) + \Phi(x)) \subset B(0, R^1)$. □

THEOREM 4.3. *The value function $W(p, t)$ defined by* (4.1) *is continuous, denoted*

$$W \in C(\mathcal{D} \cap \mathcal{X} \times [0, T]).$$

Proof. Fix $(p^1, t^1) \in \mathcal{D} \cap \mathcal{X} \times [0, T]$. Given $\varepsilon > 0$, we will show that there exists $\delta > 0$ (depending on p^1) such that $\| p^2 - p^1 \| < \delta$ and $|t^2 - t^1| < \delta$ imply

$$|W(p^1, t^1) - W(p^2, t^2)| \le \varepsilon. \tag{4.5}$$

The proof of this assertion is based on the proof of [11, Thm. 3.2].

Assume that $0 \le t^1 \le t^2 \le T$ and $0 < \delta < \delta^1$, where δ^1, R^1 are as in Lemma 4.2, and that $\| p^2 - p^1 \| < \delta$, $|t^2 - t^1| < \delta$.

Choose $\mathbf{u} \in \mathbf{U}(t^1)$ such that

$$W(p^1, t^1) \ge \sup_{y\in\mathcal{Y}(t^1)} \{(p_T^1, \Phi)\} - \varepsilon/3,$$

where p_s^1 is the solution of (3.3) with initial data $p_{t^1}^1 = p^1$ and using this $\mathbf{u}$ and any y. For any $y \in \mathcal{Y}(t^2)$ define $\tilde{y} \in \mathcal{Y}(t^1)$ by

$$\tilde{y}(s) = \begin{cases} 0, & t^1 \le s < t^2, \\ y(s), & t^2 \le s \le T. \end{cases}$$

Define $\tilde{\mathbf{u}} \in \mathbf{U}(t^2)$ by

$$\tilde{\mathbf{u}}[y] = \mathbf{u}[\tilde{y}] \quad \text{for all } y \in \mathcal{Y}(t^2).$$

Select $y \in \mathcal{Y}(t^2)$ such that

$$W(p^2, t^2) \le (p_T^2, \Phi) + \varepsilon/3,$$

where p_s^2 is the solution of (3.3) with initial data $p_{t^2}^2 = p^2$ and using $\tilde{\mathbf{u}}$ and y. Then

$$W(p^2, t^2) - W(p^1, t^1) \le (p_T^2, \Phi) - (p_T^1, \Phi) + 2\varepsilon/3 = p_T^2(x^2) - p_T^1(x^2) + 2\varepsilon/3, \tag{4.6}$$

where $x^2 \in \operatorname{argmax}\{p^2 + \Phi\}$. By Lemma 4.2, since Φ is bounded, $|x^2| \le R^1$. Then using $u(s) = \mathbf{u}[\tilde{y}](s)$ and $\tilde{y}(s)$, $s \in [t^1, T]$, and $s \in [t^2, T]$, inequality (3.15) of Lemma 3.5 implies that

$$p_T^2(x^2) - p_T^1(x^2) \le (K\delta + \rho(|t^2 - t^1|))(1 + |x^2|) \le (K\delta + \rho(|t^2 - t^1|))(1 + R^1).$$

Therefore there exists $\delta^2 < \delta^1$ such that $\delta < \delta^2$ implies, using (4.6),

$$W(p^2, t^2) - W(p^1, t^1) \le \varepsilon. \tag{4.7}$$

The proof of the opposite inequality is similar. Choose $\mathbf{u} \in \mathbf{U}(t^2)$ such that

$$W(p^2, t^2) \geq \sup_{y \in \mathcal{Y}(t^2)} \{(p_T^2, \Phi)\} - \varepsilon/3,$$

where p_s^2 is the solution of (3.3) with initial data $p_{t^2}^2 = p^2$ and using $\mathbf{u}$ and any y. For all $y \in \mathcal{Y}(t^1)$, define $\tilde{y} \in \mathcal{Y}(t^2)$ by $\tilde{y} = y$ on $[t^2, T]$. Fix $u_0 \in U$. Define $\tilde{\mathbf{u}} \in \mathbf{U}(t^1)$ by

$$\tilde{\mathbf{u}}[y] = \begin{cases} u_0, & t^1 \leq s \leq t^2, \\ \mathbf{u}[\tilde{y}](s), & t^2 \leq s \leq T. \end{cases}$$

Now choose $y \in \mathcal{Y}(t^1)$ such that

$$W(p^1, t^1) \leq (p_T^1, \Phi) + \varepsilon/3,$$

where p_s^1 is the solution of (3.3) with initial data $p_{t^1}^1 = p^1$ and using $\tilde{\mathbf{u}}$ and y. Therefore,

$$W(p^1, t^1) - W(p^2, t^2) \leq (p_T^1, \Phi) - (p_T^2, \Phi) + 2\varepsilon/3,$$

and proceeding as above there exists $\delta^3 \leq \delta^2$ such that $\delta < \delta^3$ implies

$$W(p^1, t^1) - W(p^2, t^2) \leq \varepsilon. \tag{4.8}$$

Inequalities (4.7) and (4.8) are both valid for $\delta < \delta^3$, hence (4.5). □

The principle of optimality (dynamic programming principle) for this problem is as follows.

THEOREM 4.4. *For any* $0 \leq t \leq r \leq T$ *we have*

$$W(p, t) = \inf_{\mathbf{u} \in \mathbf{U}(t)} \sup_{y \in \mathcal{Y}(t)} \{W(p_r, r) \,:\, p_t = p\}. \tag{4.9}$$

Proof. The proof uses the same methods as in [10], [11].

Indeed, let $R(p, t)$ denote the right-hand side of (4.9), and fix $\varepsilon > 0$. Choose $\mathbf{u}^1 \in \mathbf{U}(p, t)$ such that

$$R(p, t) \geq \sup_{y \in \mathcal{Y}(t)} \{W(p_r, r)\} - \varepsilon.$$

For any $q \in \mathcal{D} \cap \mathcal{X}$ there exists $\mathbf{u}^2 \in \mathbf{U}(q, r)$ such that

$$W(q, r) \geq \sup_{y \in \mathcal{Y}(t)} \{(p_T, \Phi)\} - \varepsilon,$$

where $p_r = q$. Define $\mathbf{u}^3 \in \mathbf{U}(p, t)$ by

$$\mathbf{u}^3(y)(s) = \begin{cases} \mathbf{u}^1(p, y)(s), & t \leq s \leq r, \\ \mathbf{u}^2(p_r, y)(s), & r \leq s \leq T. \end{cases}$$

Then for any $y \in \mathcal{Y}(t)$ we have, using the control $\mathbf{u}^3$,

$$\begin{aligned} R(p, t) &\geq W(p_r, r) - \varepsilon \\ &\geq (p_T, \Phi) - 2\varepsilon; \end{aligned}$$

hence

$$(p_T, \Phi) \leq R(p, t) + 2\varepsilon \quad \text{for all } y \in \mathcal{Y}(t).$$

Therefore

$$\sup_{y\in\mathcal{Y}(t)}\{(p_T,\Phi)\}\le R(p,t)+2\varepsilon \quad (\text{using } \mathbf{u}^3).$$

This implies

$$W(p,t)\le R(p,t)+2\varepsilon. \tag{4.10}$$

To prove the opposite inequality, choose $\mathbf{u}\in\mathbf{U}(p,t)$ such that

$$W(p,t)\ge \sup_{y\in\mathcal{Y}(t)}\{(p_T,\Phi)\}-\varepsilon. \tag{4.11}$$

Then

$$R(p,t)\le \sup_{y\in\mathcal{Y}(t)}\{W(p_r,r)\},$$

and there exists $y^1\in\mathcal{Y}(t)$ such that

$$R(p,t)\le W(p_r,r)+\varepsilon.$$

For each $y\in\mathcal{Y}(r)$, define $\tilde{y}\in\mathcal{Y}(t)$ by

$$\tilde{y}(s)=\begin{cases} y^1(s), & t\le s\le r,\\ y(s), & r\le s\le T.\end{cases}$$

Then define $\tilde{\mathbf{u}}\in\mathbf{U}(q,r)$ ($q=p_r$ results from $\mathbf{u}$, y^1, and $p_t=p$) by $\tilde{\mathbf{u}}(y)(s)=\mathbf{u}(\tilde{y})(s)$, $r\le s\le T$. Then

$$W(p_r,r)\le \sup_{y\in\mathcal{Y}(r)}\{(p_T,\Phi)\ :\ p_r=q\},$$

and there exists $y^2\in\mathcal{Y}(r)$ such that

$$W(p_r,r)\le (p_T,\Phi)+\varepsilon.$$

Define $y^3\in\mathcal{Y}(t)$ by

$$y^3(s)=\begin{cases} y^1(s), & t\le s\le r,\\ y^2(s), & r\le s\le T.\end{cases}$$

Therefore we have

$$R(p,t)\le W(p_r,r)+\varepsilon\le (p_T,\Phi)+2\varepsilon,$$

which implies, by (4.11),

$$R(p,t)\le \sup_{y\in\mathcal{Y}(t)}\{(p_T,\Phi)\}+2\varepsilon\le W(p,t)+3\varepsilon. \tag{4.12}$$

Since $\varepsilon>0$ was arbitrary, inequalities (4.10) and (4.12) imply (4.9). □

Equation (4.9) leads to the dynamic programming equation (DPE)

$$\begin{cases} \dfrac{\partial W}{\partial t}+\inf\limits_{u\in U}\sup\limits_{y\in\mathbf{R}^p}\langle\nabla_p W(p,t),F(p,u,y)\rangle = 0 \quad \text{in } \mathcal{D}\cap\mathcal{X}^1\times(0,T),\\ W(p,T) = (p,\Phi) \quad \text{in } \mathcal{D}\cap\mathcal{X}.\end{cases} \tag{4.13}$$

In (4.13), $\nabla_p W(p,t)$ denotes the gradient of W with respect to p and, if it exists, belongs to the dual space $\mathcal{X}^*$ and $\langle\lambda,p\rangle$ denotes the value of $\lambda\in\mathcal{X}^*$ at $p\in\mathcal{X}$. In view of the structure of F (see (3.4)), the order of inf and sup in (4.13) is immaterial; i.e., the Isaacs condition holds.

The DPE (4.13) is the appropriate Hamilton–Jacobi–Isaacs (HJI) equation for the partially observed differential game.

We will make use of two classes $\mathcal{C}^1 \subset \mathcal{C} \subset C(\mathcal{D} \cap \mathcal{X} \times [0, T])$ of test functions. We take $\phi \in \mathcal{C}$ to mean that

(i) ϕ is $\mathcal{X}$-Frechet differentiable, with derivative denoted $(\nabla_p \phi, \frac{\partial \phi}{\partial t})$;

(ii) the Frechet derivative $(\nabla_p \phi, \frac{\partial \phi}{\partial t})$ is continuous on $\mathcal{D} \cap \mathcal{X} \times [0, T]$; and $\phi \in \mathcal{C}^1$ means that in addition

(iii) the Frechet derivative $(\nabla_p \phi, \frac{\partial \phi}{\partial t})$ is continuous on $\mathcal{D} \cap \mathcal{X}^1 \times [0, T]$.

These classes of functions will be used to define smooth and viscosity solutions of (4.13).

LEMMA 4.5. *Let $p \in \mathcal{D} \cap \mathcal{X}^1$ so that $p_r(x)$ is a smooth solution of* (3.3) *on $[t, T]$ with initial data $p_t = p$, and let $\phi \in \mathcal{C}$. Then we have the following version of the fundamental theorem of calculus:*

$$\phi(p_r, r) = \phi(p_t, t) + \int_t^r \left[\frac{\partial \phi}{\partial t}(p_s, s) + \langle \nabla_p \phi(p_s, s), F(p_s, u(s), y(s)) \rangle \right] ds. \tag{4.14}$$

Proof. Let u and y be continuous. Then the function $r \mapsto \phi(p_r, r)$ is continuously differentiable, and so by the usual fundamental theorem of calculus, (4.14) holds.

By an approximation argument, (4.14) holds for all $u \in \mathcal{U}(t)$, $y \in \mathcal{Y}(t)$ as follows. Let $u^i \to u$ in $\mathcal{U}(0)$ and $y^i \to y$ in $\mathcal{Y}(0)$ as $i \to \infty$, where each u^i and y^i is continuous. Then

$$\phi(p_r^i, r) = \phi(p_t^i, t) + \int_t^r \left[\frac{\partial \phi}{\partial t}(p_s^i, s) + \langle \nabla_p \phi(p_s^i, s), F(p_s^i, u^i(s), y^i(s)) \rangle \right] ds. \tag{4.15}$$

We claim that

$$\lim_{i \to \infty} \phi(p_r^i, r) = \phi(p_r, r), \qquad \lim_{i \to \infty} \phi(p_t^i, t) = \phi(p_t, t), \tag{4.16}$$

and

$$\begin{aligned} &\lim_{i \to \infty} \int_t^r \left[\frac{\partial \phi}{\partial t}(p_s^i, s) + \langle \nabla_p \phi(p_s^i, s), F(p_s^i, u^i(s), y^i(s)) \rangle \right] ds \\ &\quad = \int_t^r \left[\frac{\partial \phi}{\partial t}(p_s, s) + \langle \nabla_p \phi(p_s, s), F(p_s, u(s), y(s)) \rangle \right] ds, \end{aligned} \tag{4.17}$$

where $p_\cdot^i$ and $p_\cdot$ denote the corresponding solutions of (3.3) with initial data $p \in \mathcal{D} \cap \mathcal{X}^1$ at time t. By (3.8) in Lemma 3.2, (4.16) follows directly by continuity. Thus it remains to prove (4.17). This can be done by showing that

$$\begin{aligned} &\left| \int_t^r \left[\frac{\partial \phi}{\partial t}(p_s^i, s) + \langle \nabla_p \phi(p_s^i, s), F(p_s^i, u^i(s), y^i(s)) \rangle \right. \right. \\ &\qquad \left. \left. - \frac{\partial \phi}{\partial t}(p_s, s) - \langle \nabla_p \phi(p_s, s), F(p_s, u(s), y(s)) \rangle \right] ds \right| \\ &\quad \le \int_t^r \left| \frac{\partial \phi}{\partial t}(p_s^i, s) - \frac{\partial \phi}{\partial t}(p_s, s) \right| + |\langle \nabla_p \phi(p_s^i, s) - \nabla_p \phi(p_s, s), F(p_s, u(s), y(s)) \rangle| \\ &\qquad + |\langle \nabla_p \phi(p_s^i, s), F(p_s^i, u^i(s), y^i(s)) - F(p_s, u(s), y(s)) \rangle|] \, ds \\ &\quad \le \int_t^r \left| \frac{\partial \phi}{\partial t}(p_s^i, s) - \frac{\partial \phi}{\partial t}(p_s, s) \right| + \| \nabla_p \phi(p_s^i, s) - \nabla_p \phi(p_s, s) \|_* \| F(p_s, u(s), y(s)) \| \\ &\qquad + \| \nabla_p \phi(p_s^i, s) \|_* \| F(p_s^i, u^i(s), y^i(s)) - F(p_s, u(s), y(s)) \|] \, ds \to 0 \end{aligned} \tag{4.18}$$

as $i \to \infty$. Here, $\| \cdot \|$ denotes the norm on $\mathcal{X}$ as in §2 and $\| \cdot \|_*$ indicates the norm on the dual space $\mathcal{X}^*$. We treat each term as follows.

Given $\varepsilon > 0$, the compactness of $[t, r]$, the assumed continuity of the partial derivatives, and the uniform convergence (3.8) imply that for i sufficiently large,

$$\begin{aligned} &\sup_{t \le s \le r} \| \nabla_p \phi(p_s^i, s) - \nabla_p \phi(p_s, s) \|_* \le \varepsilon, \\ &\sup_{t \le s \le r} \left| \frac{\partial \phi}{\partial t}(p_s^i, s) - \frac{\partial \phi}{\partial t}(p_s, s) \right| \le \varepsilon. \end{aligned} \tag{4.19}$$

This also implies

$$\sup_{t \le s \le r} \| \nabla_p \phi(p_s^i, s) \|_* \le K. \tag{4.20}$$

Next, because of (3.8) and the assumed bounds on the problem data, we have

$$\sup_{t \le s \le r} \| F(p_s, u(s), y(s)) \| \le K. \tag{4.21}$$

Finally, it is not hard to verify the estimate

$$\begin{aligned} &\int_r^t \| F(p_s^i, u^i(s), y^i(s)) - F(p_s, u(s), y(s)) \| \, ds \\ &\le K \sup_{t \le s \le r} \| p_s^i - p_s \|_1 + K \left(1 + \sup_{t \le s \le r} \| p_s^i \|_1 \right) \\ &\quad \cdot \int_t^r [|u^i(s) - u(s)| + |y^i(s) - y(s)|] \, ds \end{aligned} \tag{4.22}$$

Therefore using (4.19), (4.20), (4.21), (4.22) in (4.18), we have

$$(4.18) \le T\varepsilon + TK\varepsilon + K \int_t^r [|u^i(s) - u(s)| + |y^i(s) - y(s)|] \, ds \le K\varepsilon$$

for all i sufficiently large, for a suitable constant $K > 0$ not depending on i. This completes the proof. □

DEFINITION 4.6. *A function* $W : \mathcal{D} \cap \mathcal{X} \times [0, T] \to \mathbf{R}$ *is called a* smooth solution *of the DPE* (4.13) *if*

(i) $W \in C^1$;

(ii) W *satisfies* (4.13) *in* $\mathcal{D} \cap \mathcal{X}^1 \times (0, T)$ *in the usual sense.*

In general, it is too much to expect that the DPE will have smooth solutions, and so one must appeal to a weaker notion of solution. To this end, we will show that the value function W is a viscosity solution of (4.13) in a suitable sense. The definition we provide below is consistent with our definition of smooth solution and is a generalization of it. We do not know a proof of uniqueness, and it may be the case that the definition has to be modified to achieve this. Consequently, Definition 4.7 is a provisional one. An abstract formulation of the viscosity solution definition is given in [13]. It is not clear at present how our definition relates to those appearing in [7], [17], [27]; indeed, the PDE (4.13) does not appear to be covered by existing theory.

DEFINITION 4.7 (provisional viscosity solution). *We say that a function* $W \in C(\mathcal{D} \cap \mathcal{X} \times [0, T])$ *is a* viscosity subsolution *of the DPE* (4.13) *if for all* $\phi \in C$, *whenever there exists* $(p', t') \in \mathcal{D} \cap \mathcal{X}^1 \times (0, T)$ *with* $W(p', t') - \phi(p', t') = \max_{(p,t) \in \mathcal{D} \cap \mathcal{X} \times [0,T]} (W(p, t) - \phi(p, t)) = 0$, *then*

$$\frac{\partial \phi}{\partial t}(p', t') + \inf_{u \in U} \sup_{y \in \mathbf{R}^p} \langle \nabla_p \phi(p', t'), F(p', u, y) \rangle \ge 0; \tag{4.23}$$

a viscosity supersolution *of the DPE* (4.13) *if for all* $\phi \in \mathcal{C}$, *whenever there exists* $(p', t') \in \mathcal{D} \cap \mathcal{X}^1 \times (0, T)$ *with* $W(p', t') - \phi(p', t') = \min_{(p,t)\in\mathcal{D}\cap\mathcal{X}\times[0,T]}(W(p,t) - \phi(p,t)) = 0$, *then*

$$\frac{\partial \phi}{\partial t}(p', t') + \inf_{u\in U} \sup_{y\in\mathbf{R}^p} \langle \nabla_p \phi(p', t'), F(p', u, y)\rangle \leq 0; \tag{4.24}$$

and a viscosity solution *if it is both a subsolution and a supersolution.*

The proof that W is a viscosity solution of (4.13) requires the following result (cf. [11, Lem. 4.3]).

LEMMA 4.8. *Let* $\phi \in \mathcal{C}$. *Assume that* ϕ *satisfies*

$$\frac{\partial \phi}{\partial t}(p', t') + \inf_{u\in U} \sup_{y\in\mathbf{R}^p} \langle \nabla_p \phi(p', t'), F(p', u, y)\rangle \leq -\theta, \tag{4.25}$$

where $\theta > 0$, $p' \in \mathcal{D} \cap \mathcal{X}^1$, $t' \in [0, T]$. *Then there exists* $\delta_0 > 0$, $\mathbf{u} \in \mathbf{U}(t')$ *such that for all* $\delta < \delta_0$, $y \in \mathcal{Y}(t')$

$$\int_{t'}^{t'+\delta} \left[\frac{\partial \phi}{\partial t}(p_s, s) + \langle \nabla_p \phi(p_s, s), F(p_s, \mathbf{u}[y](s), y(s))\rangle\right] ds \leq -\delta\theta/2. \tag{4.26}$$

Similarly, if ϕ *satisfies*

$$\frac{\partial \phi}{\partial t}(p', t') + \inf_{u\in U} \sup_{y\in\mathbf{R}^p} \langle \nabla_p \phi(p', t'), F(p', u, y)\rangle \geq \theta, \tag{4.27}$$

where $\theta > 0$, $p' \in \mathcal{D} \cap \mathcal{X}^1$, $t' \in [0, T]$, *then there exists* $\delta_1 > 0$, $y \in \mathcal{Y}(t')$ *such that for all* $\delta < \delta_1$, $\mathbf{u} \in \mathbf{U}(t')$

$$\int_{t'}^{t'+\delta} \left[\frac{\partial \phi}{\partial t}(p_s, s) + \langle \nabla_p \phi(p_s, s), F(p_s, \mathbf{u}[y](s), y(s))\rangle\right] ds \geq \delta\theta/2. \tag{4.28}$$

Proof. Write

$$\Lambda(p, t, u, y) = \frac{\partial \phi}{\partial t}(p, t) + \langle \nabla_p \phi(p, t), F(p, u, y)\rangle.$$

Since $\phi \in \mathcal{C}$, $\Lambda : \mathcal{D} \cap \mathcal{X}^1 \times [0, T] \times U \times \mathbf{R}^p \to \mathbf{R}$ is continuous. In fact, for $\| p - p' \|_1 < \nu$, $|t - t'| < \nu$ ($\nu > 0$ small), we have the estimate

$$\begin{aligned} |\Lambda(p, t, u, y) - \Lambda(p', t', u', y')| \leq{}& \left|\frac{\partial \phi}{\partial t}(p, t) - \frac{\partial \phi}{\partial t}(p', t')\right| \\ &+ K \| \nabla_p \phi(p, t) - \nabla_p \phi(p', t') \|_* + K(\| p - p' \|_1 + |u - u'| + |y - y'|), \end{aligned} \tag{4.29}$$

where $K > 0$ depends on p', t', and ν.

By (4.25), $\inf_u \sup_y \Lambda(p', t', u, y) \leq -\theta$. Select $u_0 \in \operatorname{argmin}_u \sup_y \Lambda(p', t', u, y)$, which does not depend on y because the Isaacs condition holds. Therefore

$$\Lambda(p', t', u_0, y) \leq -\theta \quad \text{for all } y \in \mathbf{R}^p.$$

Define $\mathbf{u} \in \mathbf{U}(t')$ by $\mathbf{u}[y] \equiv u_0$. Let $y \in \mathcal{Y}(t')$. As in (3.9), Lemma 3.2, the map $s \mapsto p_s$ from $[t', T]$ into $\mathcal{D} \cap \mathcal{X}^1$ is continuous, with modulus of continuity independent of u, y. Thus there exists $\delta_0 > 0$ such that if $\delta < \delta_0$ and $t' \leq s \leq t' + \delta$, then

$$\Lambda(p_s, s, \mathbf{u}[y](s), y(s)) \leq -\theta/2.$$

Integrating from t' to $t' + \delta$ gives (4.26).

To prove (4.28), we note that (4.27) implies the existence of $y_0 \in \mathbf{R}^p$ (independent of u) such that

$$\Lambda(p', t', u, y_0) \geq \theta \quad \text{for all } u \in U.$$

Define $y \in \mathcal{Y}(t')$ by $y \equiv y_0$, and let $\mathbf{u} \in \mathbf{U}(t')$. Then by continuity, with $\delta < \delta_1$ (some $\delta_1 > 0$), $t' \leq s \leq t' + \delta$ implies

$$\Lambda(p_s, s, \mathbf{u}[y](s), y(s)) \geq \theta/2.$$

Integrating from t' to $t' + \delta$ gives (4.28). □

THEOREM 4.9. *The value function $W(p, t)$ defined by* (4.1) *is a continuous viscosity solution of the DPE* (4.13).

Proof. To show that $W(p, t)$ is a viscosity subsolution, assume there exist $\phi \in \mathcal{C}$, $(p', t') \in \mathcal{D} \cap \mathcal{X}^1 \times (0, T)$ with $W(p', t') - \phi(p', t') = \max_{(p,t)\in\mathcal{D}\cap\mathcal{X}\times[0,T]}(W(p, t) - \phi(p, t)) = 0$. We must show that ϕ satisfies (4.23). If not, then there exists $\theta > 0$ such that (4.25) holds. By Lemma 4.8, (4.26) holds, which implies

$$(4.30)\quad \inf_{\mathbf{u}\in\mathbf{U}(t')} \sup_{y\in\mathcal{Y}(t')} \left\{ \int_{t'}^{t'+\delta} \left[\frac{\partial \phi}{\partial t}(p_s, s) + \langle \nabla_p \phi(p_s, s), F(p_s, \mathbf{u}[y](s), y(s)) \rangle \right] ds \right\} \leq -\delta\theta/2.$$

Now

$$W(p', t') = \phi(p', t') \quad \text{and} \quad W(p, t) \leq \phi(p, t);$$

hence the dynamic programming principle (4.9) with $t = t'$, $r = t' + \delta$ implies

$$0 \leq \inf_{\mathbf{u}\in\mathbf{U}(t')} \sup_{y\in\mathcal{Y}(t')} \{\phi(p_{t'+\delta}, t' + \delta) - \phi(p', t')\}.$$

Since $\phi \in \mathcal{C}$, Lemma 4.5 and this inequality imply

$$(4.31)\quad 0 \leq \inf_{\mathbf{u}\in\mathbf{U}(t')} \sup_{y\in\mathcal{Y}(t')} \left\{ \int_{t'}^{t'+\delta} \left[\frac{\partial \phi}{\partial t}(p_s, s) + \langle \nabla_p \phi(p_s, s), F(p_s, \mathbf{u}[y](s), y(s)) \rangle \right] ds \right\}.$$

But (4.31) contradicts (4.30), hence (4.23) is valid. Therefore $W(p, t)$ is a viscosity subsolution of (4.13).

Now suppose there exists $\phi \in \mathcal{C}$, $(p', t') \in \mathcal{D} \cap \mathcal{X}^1 \times (0, T)$ with

$$W(p', t') - \phi(p', t') = \min_{(p,t)\in\mathcal{D}\cap\mathcal{X}\times[0,T]} (W(p, t) - \phi(p, t)) = 0.$$

If (4.24) does not hold, then there exists $\theta > 0$ such that (4.27) holds. Then by Lemma 4.8, (4.28) holds, implying

$$(4.32)\quad \inf_{\mathbf{u}\in\mathbf{U}(t')} \sup_{y\in\mathcal{Y}(t')} \left\{ \int_{t'}^{t'+\delta} \left[\frac{\partial \phi}{\partial t}(p_s, s) + \langle \nabla_p \phi(p_s, s), F(p_s, \mathbf{u}[y](s), y(s)) \rangle \right] ds \right\} \geq \delta\theta/2.$$

Now

$$W(p', t') = \phi(p', t') \quad \text{and} \quad W(p, t) \geq \phi(p, t),$$

and by dynamic programming,

$$0 \geq \inf_{\mathbf{u}\in\mathbf{U}(t')} \sup_{y\in\mathcal{Y}(t')} \{\phi(p_{t'+\delta}, t' + \delta) - \phi(p', t')\}.$$

This implies

$$(4.33)\quad 0 \geq \inf_{\mathbf{u}\in\mathbf{U}(t')} \sup_{y\in\mathcal{Y}(t')} \left\{ \int_{t'}^{t'+\delta} \left[\frac{\partial \phi}{\partial t}(p_s, s) + \langle \nabla_p \phi(p_s, s), F(p_s, \mathbf{u}[y](s), y(s)) \rangle \right] ds \right\},$$

contradicting (4.32). Therefore $W(p, t)$ is a viscosity supersolution of (4.13). □

5. Verification theorem. The main reason for defining value functions and using dynamic programming is to determine optimal controls. Typically, some type of smoothness is required. The following theorem says essentially that if both (3.3) and (4.13) have smooth solutions, then the optimal control is obtained by finding the control value $\mathbf{u}^*(p,t)$ that attains the minimum in (4.13) as

$$\mathbf{u}^*[y](t) = \mathbf{u}^*(p[y]_t, t). \tag{5.1}$$

This control is an *information state feedback* controller and depends on the output y via the information state. This is a type of *separation principle* for this partially observed differential game.

THEOREM 5.1. *Assume that there exists a smooth solution $\tilde{W}$ of the DPE* (4.13). *If there exist $\mathbf{u}^* \in \mathbf{U}(t)$, $y^* \in \mathcal{Y}(t)$ such that*

$$\mathbf{u}^*(s) \in \operatorname*{argmin}_{u\in U}\{\langle \nabla_p W(p_s, s), -\nabla_x p_s \cdot f(\cdot, u) + L(\cdot, u)\rangle\}, \tag{5.2}$$

$$y^*(s) \in \operatorname*{argmax}_{y\in \mathbf{R}^p}\{\langle \nabla_p W(p_s, s), -\nabla_x p_s \cdot g(\cdot, y-h) - \gamma^2 \ell(y-h)\rangle\} \tag{5.3}$$

for a.e. $s \in [t,T]$, then $\mathbf{u}^*$ is optimal for the initial data $(p,t) \in \mathcal{D}\cap\mathcal{X}^1 \times [0,T]$ and $\tilde{W}(p,t) = W(p,t)$, where $(p,t) \in \mathcal{D}\cap\mathcal{X}^1$. In particular, for $(p,t) = (\alpha, 0) \in \mathcal{D}\cap\mathcal{X}^1 \times [0,T]$ the control $\mathbf{u}^*$ solves the partially observed minimax differential game.

Proof. Since $p_t(x)$ and $\tilde{W}(p,t)$ are smooth solutions ($t \le r \le T$, $p_t = p \in \mathcal{D}\cap\mathcal{X}^1$), Lemma 4.5 implies

$$\tilde{W}(p_r, r) = \tilde{W}(p_t, t) + \int_t^r \left[\frac{\partial \tilde{W}}{\partial t}(p_s, s) + \langle \nabla_p \tilde{W}(p_s, s), F(p_s, u(s), y(s))\rangle\right] ds. \tag{5.4}$$

Fix $y^* \in \mathcal{Y}(t)$ as in (5.3). Then for any $\mathbf{u} \in \mathbf{U}(t)$, we have, using the DPE (4.13) and (5.4) with $r = T$,

$$\begin{aligned}\tilde{W}(p,t) &= -\int_t^T \left[\frac{\partial \tilde{W}}{\partial t}(p_s, s) + \langle \nabla_p \tilde{W}(p_s, s), F(p_s, \mathbf{u}[y^*](s), y^*(s))\rangle\right] ds + (p_T, \Phi) \\ &\le (p_T, \Phi),\end{aligned}$$

with equality for $\mathbf{u} = \mathbf{u}^*$ as in (5.2). Therefore $\tilde{W}(p,t) \le W(p,t)$.

Conversely, let $\mathbf{u} = \mathbf{u}^*$ and $\varepsilon > 0$. Then there exists $y \in \mathcal{Y}(t)$ such that

$$W(p,t) \le (p_T, \Phi) + \varepsilon.$$

Using (5.4) and $r = T$, this gives

$$\begin{aligned}W(p,t) &\le \tilde{W}(p,t) + \int_t^T \left[\frac{\partial \tilde{W}}{\partial t}(p_s, s) + \langle \nabla_p \tilde{W}(p_s, s), F(p_s, u^*(s), y(s))\rangle\right] ds + \varepsilon \\ &\le \tilde{W}(p,t) + \int_t^T \left[\frac{\partial \tilde{W}}{\partial t}(p_s, s) + \langle \nabla_p \tilde{W}(p_s, s), F(p_s, u^*(s), y(s))\rangle\right] ds + \varepsilon \\ &\le \tilde{W}(p,t) + \varepsilon.\end{aligned}$$

Hence $W(p,t) \le \tilde{W}(p,t)$.

We conclude that $W(p,t) = \tilde{W}(p,t)$, and in fact

$$\tilde{W}(\alpha, 0) = J(\mathbf{u}^*) = \inf_{\mathbf{u}\in\mathbf{U}(0)} J(\mathbf{u}),$$

proving the optimality of $\mathbf{u}^*$. □

6. Relation to certainty equivalence. In this section we explain how the certainty equivalence principle [5], [9], [6] fits into the general framework developed in this paper. This issue was treated in discrete-time in [24], [21] and in the case of continuous-time bilinear systems in [32], [33]; see also [6]. The certainty equivalence principle is as follows.

Consider a state feedback differential game with value function $V(x,t)$ satisfying the DPE

$$\begin{cases} \dfrac{\partial V}{\partial t} + \inf\limits_{u\in U}\sup\limits_{w\in\mathbf{R}^p}\{\nabla_x V\cdot(f(x,u)+g(x,w)) - \gamma^2\ell(w) + L(x,u)\} = 0 \text{ in } \mathbf{R}^n\times(0,T), \\ V(x,T) = \Phi(x) \quad \text{in } \mathbf{R}^n. \end{cases} \tag{6.1}$$

Equation (6.1) is a nonlinear first-order PDE and need not possess smooth solutions; thus (6.1) must also be interpreted in the viscosity sense in general. The value function $V(x,t)$ is the unique viscosity solution of (6.1) and is bounded Lipschitz continuous. Let $u^s(x,t)$ denote the optimal state feedback control (if it exists), i.e., the control value attaining the minimum in (6.1). The *minimum stress estimate* is defined by

$$\bar{x}(t) \in \operatorname*{argmax}_{x\in\mathbf{R}^n}(p_t(x) + V(x,t)). \tag{6.2}$$

In [9] it is proven (for a closely related problem) that the *certainty equivalence* controller

$$u^{ce}(t) = u^s(\bar{x}(t),t) \tag{6.3}$$

is optimal provided (i) $p_t(x)$ is a smooth solution of (3.3), (ii) $V(x,t)$ is a smooth solution of (6.1), and, most significantly, (iii) $\bar{x}(t)$ is *unique*.

The following theorem provides a new interpretation of the certainty equivalence controller (see also [6], [21], [24], [32], [33]).

THEOREM 6.1. *Fix a point* $(p^1,t^1)\in\mathcal{D}\cap\mathcal{X}^1\times(0,T)$, *and assume that*

(i) $V(x,t) = V_t(x)$ *is a smooth solution of* (6.1);

(ii) *the quantity*

$$\bar{x}_{t^1}(p^1) = \operatorname*{argmax}_{x\in\mathbf{R}^n}(p^1(x) + V_{t^1}(x)) \tag{6.4}$$

is unique (i.e., the maximum is attained at a unique point).
Then the function $\tilde{W}\in C(\mathcal{D}\cap\mathcal{X}\times[0,T])$ *defined by*

$$\tilde{W}(p,t) = (p, V_t) \tag{6.5}$$

(sup-pairing) is $\mathcal{X}$*-Gateaux differentiable at* (p^1,t^1) *and satisfies the DPE* (4.13) *at* (p^1,t^1). *Further, the optimal control at the point* (p^1,t^1) *is given by*

$$\mathbf{u}^*_{t^1}(p^1) = u^s(\bar{x}_{t^1}(p^1),t^1) = u^{ce}(t^1). \tag{6.6}$$

Proof. 1. We claim first that the Gateaux derivative $\partial_p\tilde{W}(p^1,t^1)$ is given by

$$\partial_p\tilde{W}(p^1,t^1) = E_{\bar{x}_{t^1}(p^1)}, \tag{6.7}$$

where $E_x\in\mathcal{X}^*$ is the evaluation map

$$\langle E_x, q\rangle = q(x), \quad (q\in\mathcal{X}). \tag{6.8}$$

For brevity, write $\bar{x} = \bar{x}_{t^1}(p^1)$, $\phi(p) = (p,V)$, $V(x) = V_{t^1}(x)$, and let $q\in\mathcal{X}$, $\varepsilon>0$. Since $p^1\in\mathcal{D}\cap\mathcal{X}$, $\phi(p^1+\varepsilon q)$ is finite for all ε sufficiently small, and $\phi(p^1+\varepsilon q)\to f(p^1)$,

$\bar{x}^\varepsilon \to \bar{x}$ as $\varepsilon \to 0$ by hypothesis (iii), where $\bar{x}^\varepsilon \in \operatorname{argmax}_{x\in\mathbf{R}^n}(p^1(x) + \varepsilon q(x) + V_{t^1}(x))$. We must show that

$$\lim_{\varepsilon\to 0} \frac{\phi(p^1 + \varepsilon q) - \phi(p^1)}{\varepsilon} = q(\bar{x}). \tag{6.9}$$

Now

$$\phi(p^1 + \varepsilon q) - \phi(p^1) \geq p^1(\bar{x}) + \varepsilon q(\bar{x}) + V(\bar{x}) - (p^1(\bar{x}) + V(\bar{x})) = \varepsilon q(\bar{x}),$$

and hence

$$\liminf_{\varepsilon\to 0} \frac{\phi(p^1 + \varepsilon q) - \phi(p^1)}{\varepsilon} \geq q(\bar{x}).$$

Similarly,

$$\phi(p^1 + \varepsilon q) - \phi(p^1) \leq p^1(\bar{x}^\varepsilon) + \varepsilon q(\bar{x}^\varepsilon) + V(\bar{x}^\varepsilon) - (p^1(\bar{x}^\varepsilon) + V(\bar{x}^\varepsilon)) = \varepsilon q(\bar{x}^\varepsilon),$$

and hence

$$\limsup_{\varepsilon\to 0} \frac{\phi(p^1 + \varepsilon q) - \phi(p^1)}{\varepsilon} \leq q(\bar{x}).$$

These two inequalities prove (6.9), establishing the claim.

2. It follows from Danskin's theorem (see [5, App.]) that

$$\begin{aligned} \frac{\partial \tilde{W}}{\partial t}(p^1, t^1) &= \frac{\partial V}{\partial t}(\bar{x}, t^1), \\ \nabla_x V(\bar{x}, t^1) &= -\nabla_x p^1(\bar{x}). \end{aligned} \tag{6.10}$$

The gradient $\nabla_x p^1$ is well defined since $p^1 \in \mathcal{X}^1$.

3. Next substitute the derivatives calculated above into the left-hand side of DPE (4.13) to yield

$$\begin{aligned} &\frac{\partial \tilde{W}}{\partial t}(p^1, t^1) + \inf_{u\in U} \sup_{y\in\mathbf{R}^p} \langle \partial_p \tilde{W}(p^1, t^1), F(p^1, u, y)\rangle \\ &= \frac{\partial V}{\partial t}(\bar{x}, t^1) + \inf_{u\in U} \sup_{y\in\mathbf{R}^p} \langle E_{\bar{x}}, -\nabla_x p^1 \cdot (f(\cdot, u) + g(\cdot, y-h)) + L(\cdot, u) - \gamma^2 \ell(h-y)\rangle \\ &= -\inf_{u\in U} \sup_{w\in\mathbf{R}^p} \left[\nabla_x V(\bar{x}, t^1) \cdot (f(\bar{x}, u) + g(\bar{x}, w)) - \gamma^2 \ell(w) + L(\bar{x}, u)\right] \\ &\quad + \inf_{u\in U} \sup_{y\in\mathbf{R}^p} \left[\nabla_x V(\bar{x}, t^1) \cdot (f(\bar{x}, u) + g(\bar{x}, y-h)) - \gamma^2 \ell(y-h) + L(\bar{x}, u)\right] \\ &= 0. \end{aligned} \tag{6.11}$$

This proves that $\tilde{W}$ satisfies the DPE at (p^1, t^1).

4. Finally, the above calculation yields explicitly the formula (6.6). □

Remark 6.2. If the minimum stress estimate (6.4) is not unique, i.e. contains more than one point, then in general the function $\tilde{W}(p, t)$ defined by (6.5) is not a solution of (4.13). To see this, we know from Lemma 4.2 that $\bar{x} = \bar{x}_{t^1}(p^1)$ is compact. The proof of Theorem 6.1 shows that

$$\partial_p \tilde{W}(p^1, t^1) \geq \max_{x\in\bar{x}_{t^1}(p^1)} E_x. \tag{6.12}$$

Hence in place of (6.11), Theorem 6.1, we have (for $x \in \bar{x}$)

$$\begin{aligned}
&\frac{\partial \tilde{W}}{\partial t}(p^1, t^1) + \inf_{u\in U} \sup_{y\in \mathbf{R}^p} \langle \partial_p \tilde{W}(p^1, t^1), F(p^1, u, y)\rangle \\
&\geq \frac{\partial V}{\partial t}(x, t^1) + \inf_{u\in U} \sup_{y\in \mathbf{R}^p} \sup_{x\in \bar{x}} \langle E_x, -\nabla_x p^1 \cdot (f(\cdot, u) + g(\cdot, y-h)) + L(\cdot, u) - \gamma^2 \ell(h-y)\rangle \\
&= -\inf_{u\in U} \sup_{w\in \mathbf{R}^p} \left[\nabla_x V(x, t^1)\cdot (f(x,u) + g(x,w)) - \gamma^2 \ell(w) + L(x,u)\right] \\
&\quad + \inf_{u\in U} \sup_{y\in \mathbf{R}^p} \sup_{x\in \bar{x}} \left[\nabla_x V(x, t^1)\cdot (f(x,u) + g(x,y-h)) - \gamma^2 \ell(y-h) + L(x,u)\right] \\
&\geq -\inf_{u\in U} \sup_{w\in \mathbf{R}^p} \left[\nabla_x V(x, t^1)\cdot (f(x,u) + g(x,w)) - \gamma^2 \ell(w) + L(x,u)\right] \\
&\quad + \sup_{x\in \bar{x}} \inf_{u\in U} \sup_{y\in \mathbf{R}^p} \left[\nabla_x V(x, t^1)\cdot (f(x,u) + g(x,y-h)) - \gamma^2 \ell(y-h) + L(x,u)\right] \\
&\geq 0.
\end{aligned} \tag{6.13}$$

This inequality can be strict in general. This calculation suggests that $\tilde{W}(p, t)$ is a *subsolution* of (4.13), but not in general a solution, and consequently

$$W(p, t) \geq \tilde{W}(p, t). \tag{6.14}$$

7. H_∞ control. As an application of the above results, we consider a relatively simple nonlinear H_∞ control problem, viz. finite-horizon disturbance attenuation. Some comments on the infinite-horizon problem will be made in §8.2. We follow closely the approach initiated in [4], [23]. We emphasize that we provide both necessary and sufficient conditions in terms of two PDEs—one defined on a finite-dimensional space $\mathbf{R}^n$, the other defined on an infinite-dimensional space $\mathcal{D} \cap \mathcal{X}$.

Associated with the system (2.1) is the performance output z (not measured) given by

$$z(t) = l(x(t), u(t)). \tag{7.1}$$

To maintain consistency with earlier notation, we set $L(x, u) = \frac{1}{2}|l(x, u)|^2$ and $\Phi = 0$. We assume that zero is an equilibrium; that is, $f(0, 0) = 0$, $g(0, 0) = 0$, $h(0) = 0$, $l(0, 0) = 0$.

Given $\gamma > 0$ and a fixed time interval $[0, T]$, the disturbance attenuation H_∞ problem is to find an output feedback control $\mathbf{u} \in \mathbf{U}(0)$ such that the resulting closed loop system $\Sigma^{\mathbf{u}}$ is *finite gain* $[0, T]$, i.e.,

$$\left\{\frac{1}{2}\int_0^T |z(t)|^2\, dt \leq \frac{\gamma^2}{2}\int_0^T \ell(w(t))\, dt + \beta(x_0) \quad \text{for all } w \in \mathcal{W}(0),\right. \tag{7.2}$$

for some function $\beta(x) \geq 0$, $\beta(0) = 0$, $-\beta \in \mathcal{D} \cap \mathcal{X}$.

Clearly, $\Sigma^{\mathbf{u}}$ is finite gain on $[0, T]$ if and only if

$$J(\mathbf{u}) \leq 0 \text{ for } p_0 = \alpha = -\beta. \tag{7.3}$$

THEOREM 7.1. *If there exists a solution of the finite-time H_∞ problem, then there exist solutions of the PDEs* (3.3) *and* (4.13) *such that $p_0 = -\beta$ and $W(-\beta, 0) = 0$ for some $\beta(x) \geq 0$ with $\beta(0) = 0$. Conversely, if there exist smooth solutions of the PDEs* (3.3) *and* (4.13) *such that $p_0 = -\beta$ and $W(-\beta, 0) = 0$, for some $\beta(x) \geq 0$ with $\beta(0) = 0$, then the controller $\mathbf{u}^*$ defined by* (5.1), (5.2) *solves the finite-time H_∞ problem.*

Proof. 1. Assume that a control $\mathbf{u}^o \in \mathbf{U}(0)$ solves the finite-time H_∞ problem, and set $p_0 = -\beta^{\mathbf{u}^o}$. Then

$$J(\mathbf{u}^o) \leq 0,$$

and in fact

$$0 = (-\beta^{\mathbf{u}^o}, 0) \leq W(-\beta^{\mathbf{u}^o}, 0) \leq 0,$$

as in Lemma 4.1. Thus $W(-\beta^{\mathbf{u}^o}, 0) = 0$. Solutions to the PDEs (3.3) and (4.13) exist according to the results in earlier sections in the viscosity sense.

2. Conversely, if (3.3) and (4.13) have smooth solutions, then the verification Theorem 5.1 implies that the control $\mathbf{u}^*$ given by (5.1), (5.2) is optimal. Therefore, with $p_0 = -\beta$,

$$J(\mathbf{u}^*) = W(-\beta, 0) = 0,$$

which implies that $\Sigma^{\mathbf{u}^*}$ is finite gain on $[0, T]$. □

8. Remarks.

8.1. General partially observed differential games. We expect that the results developed in this paper will extend to much more general situations. However, additional technical difficulties arise. For instance, suppose (2.1) is replaced by

$$\begin{cases} \dot{x}(t) = f(x(t), u(t)) + g(x(t), w(t)), \\ y(t) = h(x(t)) + v(t), \end{cases} \tag{8.1}$$

where $v(\cdot)$ is a second independent and unknown disturbance input. In this case, the function $F(p, u, y)$ governing the information state dynamics is nonlinear:

$$F(p, u, y) = \sup_w \{-\nabla_x p \cdot (f(\cdot, u) + g(\cdot, w)) + L(\cdot, u) - \gamma^2 \ell(w, y - h)\}. \tag{8.2}$$

A consequence of this is that (3.3) does not in general have smooth solutions (even if α is smooth). This complicates substantially the proof that the value function $W(p, t)$ is a viscosity solution of the corresponding HJI equation (4.13).

8.2. Infinite-horizon H_∞ control. The theory of dissipative systems [18], [36] provides a powerful framework for treating infinite-horizon H_∞ problems, and many of the articles listed in the reference section make use of this theory. In the state feedback case, one is led to a partial differential inequality (PDI); see, e.g., [2], [18], [20], [34], [36]. In particular, it is shown in [2] and [20] that the PDI can be interpreted in the viscosity sense.

In the discrete-time case, the infinite-horizon output feedback H_∞ problem was discussed in [23], and an infinite-dimensional dissipation inequality was used. The continuous-time analogue of this inequality is an infinite-dimensional PDI, closely related to the steady-state version of (4.13). The PDI is

$$\begin{cases} \inf_{u \in U} \sup_{y \in \mathbf{R}^p} \langle \nabla_p W, F(p, u, y) \rangle \leq 0 \text{ in } \mathcal{D} \cap \mathcal{X}^1, \\ W(p) \geq (p, 0) \text{ in } \mathcal{D} \cap \mathcal{X}, \\ W(-\beta) = 0. \end{cases} \tag{8.3}$$

A theory of infinite-horizon H_∞ control can be developed using this type of equation; see [16]. It is possible to prove the existence of a function $W(p)$ satisfying (8.3) in the viscosity sense (i.e., as a viscosity supersolution), using a stationary version of the definition given in §4. An explicit storage function for the closed-loop system is defined in [15].

Acknowledgment. We wish to thank Professor W. H. Fleming for some interesting and useful discussions.

REFERENCES

[1] J. A. BALL AND J. W. HELTON, *Nonlinear H_∞ control theory for stable plants*, Math. Control Signals Systems, 5 (1992), pp. 233–261.

[2] ———, *Viscosity Solutions of Hamilton-Jacobi Equations Arising in Nonlinear H_∞ Control*, Math. Control Signals Systems, to appear.

[3] J. A. BALL, J. W. HELTON, AND M. L. WALKER, *H_∞ control for nonlinear systems with output feedback*, IEEE Trans. Automat. Control, AC–38 (4) (1993), pp. 546–559.

[4] J. S. BARAS AND M. R. JAMES, *Nonlinear H_∞ control*, 33rd IEEE CDC, Orlando, December 1994, pp. 1435–1438.

[5] T. BASAR AND P. BERNHARD, *H^∞–Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, Birkhäuser, Boston, 1991.

[6] P. BERNHARD, *Discrete and continuous time partial information minimax control*, preprint, July 1994.

[7] M. G. CRANDALL AND P. L. LIONS, *Hamilton-Jacobi equations in infinite dimensions, Part* I, *Uniqueness of solutions*, J. Funct. Anal., 62 (1985), pp. 339–396; *Part* II, *Existence of solutions*, 65 (1986), pp. 368–405; *Part* III, 68 (1986), pp. 214–247; *Part* IV, *Unbounded linear terms*, 90 (1990), pp. 137–283; *Part* V, *B-Continuous solutions*, 97 (1991), pp. 417–465.

[8] J. C. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State–space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (8) (1989), pp. 831–847.

[9] G. DIDINSKY, T. BASAR, AND P. BERNHARD, *Structural properties of minimax policies for a class of differential games arising in nonlinear H_∞ control and filtering*, 32nd IEEE CDC, San Antonio, December 1993.

[10] R. J. ELLIOTT AND N. J. KALTON, *The Existence of Value in Differential Games*, Mem. Amer. Math. Soc., 126 American Mathematical Society, Providence, RI, 1972.

[11] L. C. EVANS AND P. E. SOUGANIDIS, *Differential games and representation formulas for solutions of Hamilton-Jacobi-Isaacs equations*, Indiana Univ. Math. J., 33 (5) (1984), pp. 773–797.

[12] W. H. FLEMING AND W. M. MCENEANEY, *Risk sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.

[13] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[14] W. H. FLEMING AND E. PARDOUX, *Existence of optimal controls for partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 261–283.

[15] J. W. HELTON AND M. R. JAMES, *An information state approach to nonlinear J-inner/outer factorization*, 33rd IEEE CDC, Orlando, December 1994, pp. 2565–2571.

[16] ———, *Nonlinear H_∞ control, J-inner/outer factorization, and information states*, in preparation.

[17] O. HIJAB, *Partially observed control of Markov processes, Part* I, Stochastics, 28 (1989), pp. 123–144; *Part* II, Stochastics, (1989), pp. 247–262; *Part* III, Ann. Probab., 18 (1990), pp. 1099–1125.

[18] D. J. HILL AND P. J. MOYLAN, *The stability of nonlinear dissipative systems*, IEEE Trans. Automat. Control, 21 (1976), pp. 708–711.

[19] A. ISIDORI AND A. ASTOLFI, *Disturbance attenuation and H_∞ control via measurment feedback in nonlinear systems*, IEEE Trans. Automat. Control, 37 (9) (1992), pp. 1283–1293.

[20] M. R. JAMES, *A partial differential inequality for dissipative nonlinear systems*, Systems Control Lett., 21 (1993), pp. 315–320.

[21] ———, *On the certainty equivalence principle and the optimal control of partially observed dynamic games*, IEEE Trans. Automat. Control, 39 (11) (1994), pp. 2321–2324.

[22] ———, *Recent developments in nonlinear H_∞ control*, IFAC Conf. Nonlinear Control and Applications, Lake Tahoe, June 1995, pp. 578–589.

[23] M. R. JAMES AND J. S. BARAS, *Robust H_∞ output feedback control for nonlinear systems*, IEEE Trans. Automat. Control, 40 (6) (1995), pp. 1007–1017.

[24] M. R. JAMES, J. S. BARAS, AND R. J. ELLIOTT, *Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems*, IEEE Trans. Automat. Control, AC-39-4 (1994), pp. 780–792.

[25] ———, *Output feedback risk-sensitive control and differential games for continuous-time nonlinear systems*, 32nd IEEE CDC, San Antonio, December 1993, pp. 3357–3360.

[26] A. J. KRENER, *Necessary and sufficient conditions for nonlinear worst case (H-infinity) control and estimation*, preprint, 1994.

[27] P. L. LIONS, *Viscosity solutions and optimal stochastic control in infinite dimensions, Part* I, Acta. Math., 161 (1988), pp. 243–278. *Part* II, Lecture Notes in Mathematics, 1390 (1988), pp. 147–170; *Part* III, J. Funct. Anal., 86 (1989), pp. 1–18.

[28] P. L. LIONS AND B. PERTHAME, *Remarks on Hamilton-Jacobi equations with measurable time-dependent Hamiltonians*, Nonlinear Anal., 11 (5) (1987), pp. 613–621.

[29] W. M. MCENEANEY, *Uniqueness for viscosity solutions of nonstationary HJB equations under some a priori conditions (with applications)*, SIAM J. Control Optim., 33 (1995), pp. 1560–1576.

[30] ———, *Robust control and differential games on a finite time horizon*, Math. Control Signals Systems, to appear.

[31] P. SORAVIA, H_∞ *control of nonlinear systems: Differential games and viscosity solutions*, SIAM J. Control Optim., 34 (1996), pp. 1071–1097.

[32] C. A. TEOLIS, *Robust* H_∞ *Output Feedback Control for Nonlinear Systems*, Ph.D. Dissertation, University of Maryland, 1994.

[33] C. A. TEOLIS, S. YULIAR, M. R. JAMES, AND J. S. BARAS, *Robust* H_∞ *output feedback control of bilinear systems*, 33rd IEEE CDC, Orlando, December 1994, pp. 1421–1426.

[34] A. J. VAN DER SCHAFT, L_2 *gain analysis of nonlinear systems and nonlinear state feedback* H_∞ *control*, IEEE Trans. Automat. Control, AC-37 (6) (1992), pp. 770–784.

[35] ———, *Nonlinear state space* H_∞ *control theory*, in Essays on Control, H. L. Trentelman and J. C. Willems, eds., Progr. Systems Control Theory, Birkhäuser, Boston, 1993.

[36] J. C. WILLEMS, *Dissipative dynamical systems, Part* I: *general theory*, Arch. Rational Mech. Anal., 45 (1972), pp. 321–351.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1365–1391, July 1996

014

INVERSE OPTIMALITY IN ROBUST STABILIZATION*

R. A. FREEMAN† AND P. V. KOKOTOVIC†

Abstract. The concept of a robust control Lyapunov function (rclf) is introduced, and it is shown that the existence of an rclf for a control-affine system is equivalent to robust stabilizability via continuous state feedback. This extends Artstein's theorem on nonlinear stabilizability to systems with disturbances. It is then shown that every rclf satisfies the steady-state Hamilton–Jacobi–Isaacs (HJI) equation associated with a meaningful game and that every member of a class of pointwise min-norm control laws is optimal for such a game. These control laws have desirable properties of optimality and can be computed directly from the rclf without solving the HJI equation for the upper value function.

Key words. nonlinear systems, robust stabilization, control Lyapunov functions, input-to-state stability, differential games

AMS subject classifications. 93D09, 93D22, 49N50

1. Introduction.

The relationship between stability and optimality has been a central issue in the optimal stabilization problem ever since the advent of the steady-state Hamilton–Jacobi–Bellman (HJB) equation. Optimal feedback systems enjoy many desirable properties beyond stability, provided the optimality is meaningful, that is, provided the associated cost functional places suitable penalty on the state and control. For example, linear-quadratic optimal control systems have favorable gain and phase margins and reduced sensitivity [1]. Similar robustness properties have been shown to hold also for nonlinear control systems that are optimal with respect to meaningful cost functionals [14]. Another consequence of optimality is that control effort is not wasted to counteract beneficial nonlinearities. Optimality is thus a discriminating measure by which to select from among the entire set of stabilizing control laws those with desirable properties. Unfortunately, its usefulness as a synthesis tool for nonlinear systems is hampered by the computational burden associated with the HJB equation. In this paper we explore the links between stability and optimality with the aim of developing a synthesis strategy that achieves the desirable properties of optimality but avoids HJB computations.

An important link between stability and optimality is well known; the value function for a meaningful optimal stabilization problem is also a Lyapunov function for the closed-loop system. In short,

- *Every meaningful value function is a Lyapunov function.*

One purpose of this paper is to establish the converse link. We show that every Lyapunov function for every stable closed-loop system is also a value function for a meaningful optimal stabilization problem. In short,

- *Every Lyapunov function is a meaningful value function.*

While the first link has implications for the *analysis* of optimal feedback control systems, this converse link will have implications for their *synthesis*.

For systems with control inputs, the property of interest is stabilizability rather than stability. We therefore base our theory on the *control Lyapunov function* (clf) [2, 39, 34] rather than the Lyapunov function. Although quite old, the clf concept was not formalized until Artstein proved that the existence of a clf for a control-affine system is equivalent to

*Received by the editors November 18, 1993; accepted for publication (in revised form) April 17, 1995. This research was supported by Department of Energy grant DE-FG-02-88-ER-13939, Air Force Office of Scientific Research grant F49620-92-J-0495, and a National Science Foundation Graduate Research Fellowship.

†Center for Control Engineering and Computation, Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106.

stabilizability via continuous (ordinary) state feedback [2]. The above converse link between stability and optimality can be restated as

- *Every* clf *is a meaningful value function.*

This will be a corollary of a more general result that we prove for the *robust* stabilizability of systems with both control and disturbance inputs. We first extend Artstein's theorem to such systems by introducing the *robust control Lyapunov function* (rclf) and showing that its existence is equivalent to robust stabilizability via continuous state feedback. We then show that every rclf is an upper value function for a meaningful two-person zero-sum differential game, the opposing players being the control and the disturbance. In short,

- *Every* rclf *is a meaningful upper value function.*

To prove this, we show that every rclf solves the steady-state Hamilton–Jacobi–Isaacs (HJI) equation associated with a meaningful game. As a consequence of this result, if an rclf is known, we can construct a feedback law that is optimal with respect to a meaningful cost functional. Moreover, we can accomplish this *without* solving the HJI equation for the upper value function. In fact, we do not even need to construct the cost functional because we can calculate the optimal feedback directly from the rclf without recourse to the HJI equation. Indeed, we provide a formula that generates a class of such optimal control laws and that involves only the rclf, the system equations, and design parameters. The control laws given by our formula are called *pointwise min-norm* control laws, and each one inherits the desirable properties of optimality because

- *Every pointwise min-norm control law is optimal for a meaningful game.*

Let us now summarize the synthesis strategy suggested by our results. The first step is to find an rclf for our system. Fortunately, techniques for the systematic construction of rclfs for many classes of nonlinear systems are beginning to appear [28, 25, 32, 7, 29, 12, 26]. Even feedback linearization can be viewed as such a technique for systems with no disturbances [13]. Once an rclf is known, we choose design parameters and use our formula to generate a pointwise min-norm control law. This control law will have the desirable properties of optimality even though its construction is independent of any cost functional or HJI equation.

Our results represent a missing ingredient in Lyapunov design. We provide a systematic, optimality-based method for choosing a control law once an rclf is known. To fully appreciate the importance of this contribution, one should recall that other methods for choosing the control law, based on the cancellation or domination of nonlinear terms, do not necessarily possess the desirable properties of optimality and may lead to poor robustness and wasted control effort. Let us illustrate this point on an elementary example. Suppose we wish to robustly stabilize the first-order system

$$\dot{x} = -x^3 + u + wx, \tag{1}$$

where u is the control input and w is the disturbance input satisfying $|w(t)| \leqslant 1$. A control law suggested by feedback linearization would cancel the nonlinearity $-x^3$ as

$$u = x^3 - 2x. \tag{2}$$

This control law obviously (globally) asymptotically stabilizes the equilibrium at $x = 0$ for any admissible disturbance $w(t)$. However, it is an absurd choice because the term x^3 in (2) represents control effort wasted to cancel a beneficial nonlinearity. Moreover, this term is actually *positive* feedback that increases the risk of instability. It is easy to find a better control law for this simple system, but what we desire is a *systematic* method for choosing the control law that prevents wasteful mistakes like (2). The method we propose in this paper is based on the pointwise min-norm control law. For this example, the simplest pointwise min-norm

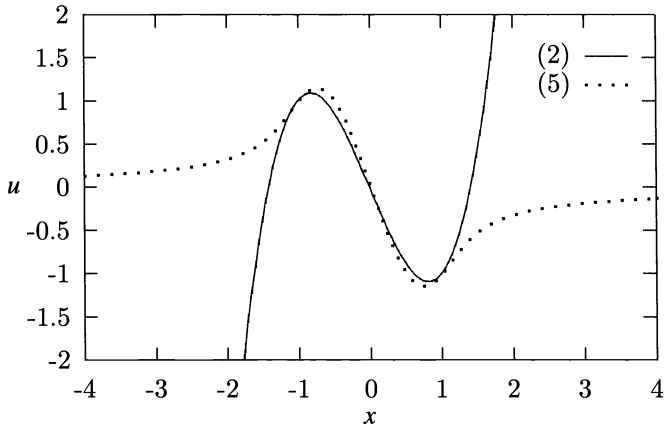


FIG. 1. *A comparison between the control laws* (2) *and* (5).

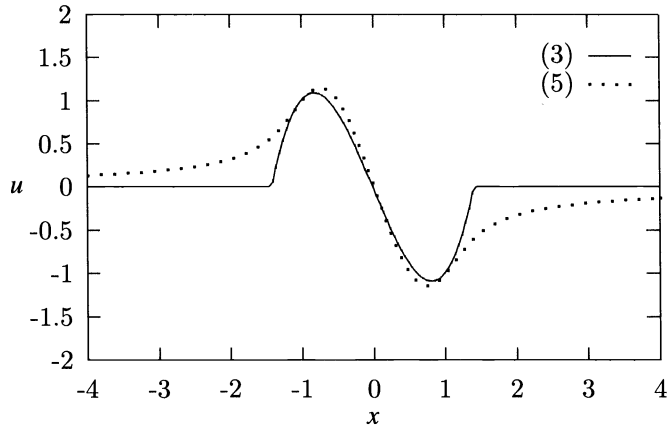


FIG. 2. *A comparison between the control laws* (3) *and* (5).

control law is

$$u = \begin{cases} x^3 - 2x & \text{when } x^2 < 2, \\ 0 & \text{when } x^2 \geqslant 2. \end{cases} \tag{3}$$

For the sake of comparison, we also solved the HJI equation associated with the system (1) and the cost functional

$$J = \int_0^\infty [x^2 + u^2]dt. \tag{4}$$

The resulting optimal feedback law is

$$u = x^3 - x - x\sqrt{x^4 - 2x^2 + 2}. \tag{5}$$

The three control laws (2), (3), and (5) are plotted in Figs. 1 and 2. We see that the control laws (3) and (5), both of which are optimal with respect to a meaningful cost functional, are very similar. They both recognize the benefit of the nonlinearity $-x^3$ in (1) and accordingly expend little control effort for large signals; moreover, these control laws are never positive

feedback. The main difference between them lies in their synthesis. The pointwise min-norm control law (3) came from a simple formula, while the control law (5) required the solution of an HJI equation. In general, the pointwise min-norm calculation is feasible, but the HJI calculation is not.

We can interpret our results in this paper as a solution to an *inverse optimal stabilization problem* in a differential game setting. The first inverse problem to be formulated and solved was for linear time-invariant systems [17, 1], where the authors identified those stabilizing gain matrices that were also optimal with respect to some quadratic cost. Inverse problems for nonlinear systems have since been considered but with more limited success. As described in the survey paper [14], the task is to choose a candidate value function and then to construct a meaningful cost functional so that the corresponding HJB equation holds. For open-loop stable nonlinear systems, one can obtain a solution by choosing the candidate value function to be a Lyapunov function for the open-loop system [16, 14]. In this paper, we extend this result to open-loop *unstable* systems by choosing the candidate value function to be a clf for the system. We provide a further generalization by solving an inverse optimal *robust* stabilization problem for systems with disturbances; we show that every rclf is an upper value function for a meaningful game.

This paper is organized as follows. We briefly list some notation and terminology in §2. This includes a review of continuity concepts for set-valued maps and statement of Michael's selection theorem. We use set-valued maps throughout this paper for three reasons. First, they arise naturally in optimization problems as the subdifferentials of convex functions; second, they allow us to weaken some assumptions; and third, they generate more concise notation.

In §3 we formulate the robust stabilization problem of achieving the *guaranteed global stability* of the closed-loop system in the presence of any admissible disturbance. In this formulation, we assume knowledge of a *bounded* set in which the disturbance is allowed to take its values. The guaranteed stability formulation has a long history (see for example [20, 15, 11, 6, 8] and the references therein), and it includes as a special case the *quadratic stability* of linear systems (see the recent survey [10] for references). In §4 we introduce the rclf and prove that its existence is sufficient for a system to be robustly stabilizable via continuous state feedback. In §5 we show that the existence of an rclf is also necessary for robust stabilizability. This necessity result relies on a converse Lyapunov theorem recently proved in [24]. Many Lyapunov-based controller designs for guaranteed stability can be cast in our rclf framework [20, 15, 11, 6–8, 28, 25, 32, 29, 12]; in particular, the control Lyapunov function defined in [31] for quadratic stability is a special kind of rclf.

In §6 we do *not* assume knowledge of a bounded set in which the disturbance is allowed to take its values. We instead consider the design objective of achieving the *input-to-state stability* of the closed-loop system (with the disturbance regarded as the input). This type of stability was introduced in [33] and is becoming a popular tool in the Lyapunov design and analysis of nonlinear systems. In §6 we show that input-to-state stabilizability is equivalent to the existence of an rclf for an auxiliary system.

In §7 we define the class of pointwise min-norm control laws described above. Our formula for this class of control laws generates those values of the control that minimize the instantaneous control effort while maintaining some desired negativity of the worst case Lyapunov derivative. In §§8–10 we show that each pointwise min-norm control law is optimal for a meaningful game and that every rclf solves the steady-state HJI equation associated with such a game. In §§8 and 9 we consider cost functionals defined over the infinite horizon, in which case we require the disturbances to vanish at the equilibrium. In §10 we allow persistent disturbances and consider cost functionals defined over a finite horizon determined by the time required to reach a given target set.

2. Notation and terminology. In what follows, Z and Y are metric spaces with both metrics denoted by $d(\cdot, \cdot)$ and X is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ (or $\cdot^T \cdot$) and norm $\| \cdot \|$.

For a subset $C \subset Z$ and a point $z \in Z$, we let either $d(z, C)$ or $|z|_C$ denote the number $\inf\{d(z, \xi) : \xi \in C\}$. We let $\overline{C}$ denote the closure of C, we let $\operatorname{int} C$ denote the interior of C, and we let $\partial C = \overline{C} \setminus \operatorname{int} C$ denote the boundary of C. For a sequence $\{C_i\}$ of subsets of Z, we let $\operatorname{Limsup} C_i$ denote the set $\{z \in Z : \liminf d(z, C_i) = 0\}$ of cluster points of sequences $z_i \in C_i$.

By $A : Z \rightsquigarrow Y$ we mean a set-valued map A mapping points in Z to subsets of Y. For such maps we let $\operatorname{Graph}(A)$ denote the set $\{(z, y) \in Z \times Y : y \in A(z)\}$. We say A is upper semicontinuous (usc) at a point $z \in Z$ when for every open set $N \subset Y$ satisfying $A(z) \subset N$ there exists an open neighborhood M of z such that $A(\xi) \subset N$ for all $\xi \in M$. We say A is lower semicontinuous (lsc) at a point $z \in Z$ when for every open set $N \subset Y$ satisfying $A(z) \cap N \neq \emptyset$ there exists an open neighborhood M of z such that $A(\xi) \cap N \neq \emptyset$ for all $\xi \in M$ (where $\emptyset$ denotes the empty set). An equivalent definition of lower semicontinuity is as follows ([5, Def. 1.4.2; 19, Thm. II.2.9]). A is lsc at a point $z \in Z$ when for every sequence $\{z_i\} \in Z$ converging to z and every $y \in A(z)$ there exists a sequence $\{y_i\} \in Y$ converging to y and $N \geqslant 1$ such that $y_i \in A(z_i)$ for all $i \geqslant N$. We say A is continuous at a point $z \in Z$ when it is both usc and lsc at z. We say A is locally Lipschitz when for each $z \in Z$ there exists an open neighborhood M of z and a constant $L \geqslant 0$ such that for all $\xi_1, \xi_2 \in M$ we have $y \in A(\xi_1)$ implies $d(y, A(\xi_2)) \leqslant L \cdot d(\xi_1, \xi_2)$. Michael's selection theorem [19, Thm. II.4.1] states that if Y is a Banach space and A is lsc with nonempty closed convex values, then A admits a continuous selection, that is, there exists a continuous (single-valued) function $a : Z \to Y$ such that $a(z) \in A(z)$ for all $z \in Z$.

We let B_X denote the closed unit ball of X, abbreviated B when no confusion is likely to arise. For a subset $D \subset X$ we let $\|D\|$ denote the number $\sup\{\|x\| : x \in D\}$. For $\lambda \in R$ we let λD denote the set $\{\lambda x : x \in D\}$. We let $\operatorname{co}(D)$ denote the convex hull of D, and we let $\overline{\operatorname{co}}(D)$ denote the closed convex hull of D. We let $D^\perp$ denote the set $\{x \in X : \langle x, D \rangle = \{0\}\}$. If D is convex and $x \in D$, we let $N_D(x) \subset X$ denote the normal cone to D at x. For $x \in X \setminus \{0\}$ we let $\operatorname{sgn}(x)$ denote the vector $x/\|x\|$. If W is another normed space, we let $\mathcal{B}(X, W)$ denote the normed space of bounded linear operators from X to W. For a convex function $h : X \to R$ we let $\partial h : X \rightsquigarrow X$ denote the subdifferential of h. We let R_+ denote the closed interval $[0, \infty)$. We say a function $\chi : R_+ \to R_+$ is of class $\mathcal{K}$ when χ is continuous, strictly increasing, and $\chi(0) = 0$. We say χ is of class $\mathcal{K}_\infty$ when χ is of class $\mathcal{K}$ and satisfies $\chi(r) \to \infty$ as $r \to \infty$. We say a function $\beta : R_+ \times R_+ \to R_+$ is of class $\mathcal{KL}$ when $\beta(\cdot, t)$ is of class $\mathcal{K}$ for each fixed $t \in R_+$ and $\beta(r, t)$ decreases to zero as $t \to \infty$ for each fixed $r \in R_+$. By C^1 we mean having a continuous (first) derivative, and by smooth we mean having continuous derivatives of any order.

3. Robust stabilizability. Let us consider three finite-dimensional Euclidean spaces: the *state space* $\mathcal{X}$, the *control space* $\mathcal{U}$, and the *disturbance space* $\mathcal{W}$. Given a continuous function $f : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to \mathcal{X}$ we define the differential equation

$$\dot{x} = f(x, u, w), \tag{6}$$

where $x(t)$ is the state trajectory taking values in $\mathcal{X}$ and satisfying an initial condition $x(0) = x_0$, $u(t)$ is the control input taking values in $\mathcal{U}$, and $w(t)$ is the disturbance input taking values in $\mathcal{W}$. Associated with the control and disturbance are constraints given by set-valued maps $U : \mathcal{X} \rightsquigarrow \mathcal{U}$ and $W : \mathcal{X} \rightsquigarrow \mathcal{W}$, respectively. The *system* Σ is (6) together with these constraints, $\Sigma := (f, U, W)$. Our standing assumptions on Σ are as follows.

A1. For each $(x, w) \in \mathcal{X} \times \mathcal{W}$, the mapping $u \mapsto f(x, u, w)$ is affine.

A2. The control constraint U is lsc with nonempty closed convex values.

A3. The disturbance constraint W is usc with nonempty compact values.

Our most restrictive assumption is A1, which requires that the system be affine in the control variable. It may be possible to remove this assumption by considering *relaxed controls* as in [2].

A *control law* for Σ is a continuous function $k : \mathcal{X} \to \mathcal{U}$ that satisfies the control constraint U; that is, for each $x \in \mathcal{X}$ we have $k(x) \in U(x)$. Assumption A2 together with Michael's theorem guarantee the existence of a control law for Σ. Each control law k for Σ defines a closed-loop differential equation

$$\dot{x} = f(x, k(x), w). \tag{7}$$

We let $\mathcal{C}(\Sigma)$ denote the set of all functions $u_g : \mathcal{X} \times R_+ \to \mathcal{U}$ such that $u_g(x, t)$ is continuous in x for each fixed $t \in R_+$ and locally bounded and (Lebesgue) measurable in t for each fixed $x \in \mathcal{X}$. These functions u_g are called *generic controls*, and they need not satisfy the control constraint U. We give the obvious meaning to the abuse of notation $k(x) \in \mathcal{C}(\Sigma)$ for a control law k defined on $\mathcal{X}$ so that the set of control laws for Σ can be regarded as a subset of the set of generic controls $\mathcal{C}(\Sigma)$. Similarly, we let $\mathcal{D}(\Sigma)$ denote the set of all functions $w_a : \mathcal{X} \times \mathcal{U} \times R_+ \to \mathcal{W}$ such that $w_a(x, u, t)$ is continuous in (x, u) for each fixed $t \in R_+$ and measurable in t for each fixed $(x, u) \in \mathcal{X} \times \mathcal{U}$, and furthermore $w_a(x, u, t) \in W(x)$ for all $(x, u, t) \in \mathcal{X} \times \mathcal{U} \times R_+$. These functions w_a are called *admissible disturbances*, and they are required to satisfy the disturbance constraint W. Again, we allow the abuse of notation $w_a(t) \in \mathcal{D}(\Sigma)$ and $w_a(x, u) \in \mathcal{D}(\Sigma)$ for functions defined on R_+ and $\mathcal{X} \times \mathcal{U}$, respectively. Implicit throughout this paper is the assumption that $\mathcal{D}(\Sigma)$ is nonempty.

It follows from assumption A3 [19, Prop. II.2.3] and standard existence theorems that for every generic control $u_g \in \mathcal{C}(\Sigma)$ and every admissible disturbance $w_a \in \mathcal{D}(\Sigma)$, solutions to the differential equation

$$\dot{x} = f\big(x,\ u_g(x, t),\ w_a(x, u_g(x, t), t)\big) \tag{8}$$

exist locally from every initial condition and bounded solutions can be extended for all $t \geqslant 0$. Note that we are *not* requiring the uniqueness of solutions. To summarize, we have defined *generic controls* that need not satisfy the constraint U, and we have defined *control laws* and *admissible disturbances* that are required to satisfy the constraints U and W, respectively.

Our formulation allows the constraint sets U and W to vary with the state x as in [3]. Sometimes we can parameterize these set-valued maps to obtain an equivalent formulation with *constant* constraints. For example, suppose $\mathcal{X} = \mathcal{W} = R$ and $W(x) = [-x^2, x^2]$ for all $x \in R$. We can parameterize W by setting $w = \widehat{w}\, x^2$ with $\widehat{w} \in [-1, 1]$; we then obtain an equivalent formulation with a constant disturbance constraint $\widehat{W} := [-1, 1]$. Continuous parameterizations always exist for continuous set-valued maps with nonempty compact convex values [4, Thm. 1.7.2]. However, even if such a parameterization for U exists, it may lead to an equivalent formulation that violates assumption A1. Furthermore, by allowing state-varying control constraints, we can sometimes satisfy assumption A1 by redefining the control *without* introducing less practical relaxed controls. For example, suppose we have $\mathcal{X} = \mathcal{U} = R$ and $\dot{x} = f_0(x, w) + (x + u)^3$ where $U = [-1, 1]$ is a constant control constraint. This system is not affine in the control u. However, if we redefine the control by setting $\widehat{u} := (x + u)^3$, we obtain a system that is now affine in the control $\widehat{u}$ but that has a state-varying control constraint $\widehat{U}(x) := \big[(x - 1)^3,\ (x + 1)^3\big]$.

Given a control law k for Σ, we will consider two types of robust stability for the closed-loop system (7). If $f(0, k(0), w) = 0$ for every $w \in W(0)$, then the system has a robust equilibrium solution $x(t) \equiv 0$ and we can possibly achieve the global uniform asymptotic

stability of this equilibrium solution. Otherwise we can only hope to achieve convergence to some compact *residual set* Ω containing the point $x = 0$.

DEFINITION 3.1. *Let Ω be a compact subset of $\mathcal{X}$ such that $0 \in \Omega$. The solutions of* (7) *are* robustly globally uniformly asymptotically stable with respect to Ω (RGUAS-Ω) *when there exists a class $\mathcal{KL}$ function β such that for any initial condition $x_0 \in \mathcal{X}$ and any admissible disturbance $w = w_a \in \mathcal{D}(\Sigma)$, all solutions $x(t)$ starting from x_0 exist for all $t \geqslant 0$ and satisfy $|x(t)|_\Omega \leqslant \beta(|x_0|_\Omega, t)$ for all $t \geqslant 0$. The solutions of* (7) *are* RGUAS *when they are* RGUAS-$\{0\}$.

This type of robust stability, defined for example in [24], is uniform with respect to admissible disturbances. Note that RGUAS-Ω implies that the residual set Ω is (robustly) positively invariant. We next define three types of stabilizability for the system Σ in the order of decreasing restrictiveness.

DEFINITION 3.2. *The system Σ is* robustly asymptotically stabilizable (RAS) *when there exists a control law k such that the solutions of* (7) *are* RGUAS. *The system Σ is* robustly practically stabilizable (RPS) *when for every $\varepsilon > 0$ there exists a control law k and a compact set $\Omega \subset \mathcal{X}$ satisfying $0 \in \Omega \subset \varepsilon B$ such that the solutions of* (7) *are* RGUAS-Ω. *The system Σ is* robustly stabilizable (RS) *when there exists a control law k and a compact set $\Omega \subset \mathcal{X}$ satisfying $0 \in \Omega$ such that the solutions of* (7) *are* RGUAS-Ω.

Clearly RAS $\Rightarrow$ RPS $\Rightarrow$ RS. The difference between RPS and RS is that if the system is RPS, then the residual set Ω can be made arbitrarily small by choice of the control law, whereas if the system is only RS, then there is a lower limit on the size of Ω. However, even if the system is only RS, we can often make Ω arbitrarily small in some directions by allowing it to grow in others.

Although we have described our system in finite-dimensional spaces $\mathcal{X}, \mathcal{U}$, and $\mathcal{W}$, many of the results in this paper can easily be modified to be valid also in infinite dimensions. For example, the results in §4 can be extended to the case where $\mathcal{X}$ and $\mathcal{U}$ are Banach spaces and $\mathcal{W}$ is a metric space, provided one assumes or can guarantee the existence of solutions to (8). Likewise, the results in §§7–10 can be extended to the case where $\mathcal{X}$ is a Banach space and $\mathcal{W}$ is a metric space.

4. Robust control Lyapunov functions. Our main tool is the *robust control Lyapunov function* (rclf) defined for a system Σ. The rclf represents an extension of the control Lyapunov function (clf) [2, 39, 34] to systems with disturbances. In this section we show that the existence of an rclf implies robust stabilizability, and in §5 we show that its existence is necessary for robust stabilizability. These results extend Artstein's theorem [2, Thm. 5.1] to systems with disturbances.

Let $\mathcal{A}(\mathcal{X})$ denote the class of continuous positive definite functions on $\mathcal{X}$, that is, continuous functions $\alpha : \mathcal{X} \to R_+$ such that $\alpha(0) = 0$ and $\alpha(x) > 0$ for $x \in \mathcal{X} \setminus \{0\}$. Let $\mathcal{A}_\kappa(\mathcal{X})$, respectively $\mathcal{A}_\infty(\mathcal{X})$, denote the set of those functions $\alpha \in \mathcal{A}(\mathcal{X})$ for which there exists a class $\mathcal{K}$, respectively class $\mathcal{K}_\infty$, function χ such that $\alpha(x) \geqslant \chi(\|x\|)$ for all $x \in \mathcal{X}$. To each C^1 function $V \in \mathcal{A}(\mathcal{X})$ we associate the *Lyapunov derivative* $L_f V : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to R$ defined by $L_f V(x, u, w) := \nabla V(x) \cdot f(x, u, w)$. Because V has a minimum at $x = 0$, we have $L_f V(0, u, w) = 0$ for all $(u, w) \in \mathcal{U} \times \mathcal{W}$. Given $c \in R_+$, we let $\Omega_c(V) := \{x \in \mathcal{X} : V(x) \leqslant c\}$ denote the c-sublevel set of V. Note that if $V \in \mathcal{A}_\infty(\mathcal{X})$, then $\Omega_c(V)$ is compact for all $c \in R_+$.

DEFINITION 4.1. *A C^1 function $V \in \mathcal{A}_\infty(\mathcal{X})$ is an* rclf *for the system Σ when there exists $c \in R_+$ such that*

$$\inf_{u \in U(x)} \sup_{w \in W(x)} L_f V(x, u, w) < 0 \tag{9}$$

for all $x \in \mathcal{X} \setminus \Omega_c(V)$. We let $c_V \in R_+$ denote the smallest value of c for which (9) *is satisfied for all $x \in \mathcal{X} \setminus \Omega_c(V)$.*

Definition 4.1 reduces to the usual clf definition when $c = 0$, $U(x)$ is constant, and $W(x) \equiv \{0\}$ (no disturbances). The main result of this section is the following.

THEOREM 4.2. *Let* Σ *satisfy assumptions* A1–A3. *If there exists an* rclf V *for* Σ, *then* Σ *is* RS. *If furthermore* $c_V = 0$, *then* Σ *is* RPS.

We will prove this theorem through a series of simple propositions. Let V be an rclf for Σ; then $L_f V$ is continuous, which together with A3 implies that we can define the *worst case Lyapunov derivative* $D_f V : \mathcal{X} \times \mathcal{U} \to R$ by the maximum

$$D_f V(x,u) := \max_{w \in W(x)} L_f V(x,u,w). \tag{10}$$

Note that $D_f V(0,u) = 0$ for all $u \in \mathcal{U}$.

PROPOSITION 4.3. $D_f V$ *is usc and for each* $x \in \mathcal{X}$, *the mapping* $u \mapsto D_f V(x,u)$ *is convex. Furthermore, there exists a continuous function* $\alpha : \mathcal{X} \to R_+$ *such that* $\alpha(0) = 0$ *and for all* $x \in \mathcal{X} \setminus \Omega_{c_V}(V)$ *we have*

$$\inf_{u \in U(x)} D_f V(x,u) < -\alpha(x) < 0. \tag{11}$$

Proof. The upper semicontinuity of $D_f V$ follows from A3 and [5, Thm. 1.4.16]. Now for each fixed $x \in \mathcal{X}$, the function $D_f V(x,u)$ is the pointwise maximum of the family of affine functions $\{L_f V(x,u,w) : w \in W(x)\}$ (assumption A1), and it follows that $D_f V(x,u)$ is convex in u. Next, it follows from [5, Thm. 1.4.16] and (9) that the left-hand side of (11) is usc and strictly negative on $\mathcal{X} \setminus \Omega_{c_V}(V)$. The existence of a function α satisfying (11) can then be deduced from [18, Prob. 5X]. □

At each point $x \in \mathcal{X} \setminus \Omega_{c_V}(V)$, the negative number $-\alpha(x)$ in Proposition 4.3 represents a certain level of negativity of the worst case Lyapunov derivative that can be guaranteed by values of the control in $U(x)$. This *negativity margin* α is not unique and will be regarded as a design parameter in later sections. Note that $\alpha \in \mathcal{A}(\mathcal{X})$ when $c_V = 0$. Given an rclf V for Σ and a negativity margin α as in Proposition 4.3, we define set-valued maps $L_V, K_V : \mathcal{X} \rightsquigarrow \mathcal{U}$ as

$$L_V(x) := \{u \in \mathcal{U} : D_f V(x,u) \leqslant -\alpha(x)\}, \tag{12}$$

$$K_V(x) := U(x) \cap L_V(x). \tag{13}$$

At each $x \in \mathcal{X}$, the set $K_V(x) \subset U(x) \subset \mathcal{U}$ is the set of all possible values of the control that satisfy the control constraint *and* make the worst-case Lyapunov derivative at least as negative as $-\alpha(x)$. Therefore, if we can find a control law k for Σ such that $k(x) \in K_V(x)$ for all $x \in \mathcal{X} \setminus \Omega_c(V)$, then k will render the solutions of (7) RGUAS-$\Omega_c(V)$. Our goal is to show that such a k exists.

PROPOSITION 4.4. K_V *is lsc with nonempty closed convex values on* $\mathcal{X} \setminus \Omega_{c_V}(V)$.

Proof. We define the set-valued map $K_V^\circ : \mathcal{X} \rightsquigarrow \mathcal{U}$ by

$$K_V^\circ(x) := U(x) \cap \{u \in \mathcal{U} : D_f V(x,u) < -\alpha(x)\}. \tag{14}$$

From Proposition 4.3 and [4, Prop. 1.10.4] we see that K_V° is lsc with nonempty convex values on $\mathcal{X} \setminus \Omega_{c_V}(V)$. Fix $x_0 \in \mathcal{X} \setminus \Omega_{c_V}(V)$. Let $\mathcal{I}(u)$ denote the convex indicator function of the set $U(x_0)$; then $\mathcal{I}$ is closed and proper (in the sense of convex analysis) because $U(x_0)$ is closed and nonempty. It follows from [30, Thms. 9.3 and 9.4] that the mapping $u \mapsto D_f V(x_0,u) + \mathcal{I}(u)$ is closed and proper. It then follows from [30, Thm. 7.6] that $\overline{K_V^\circ(x_0)} = \overline{\{u \in \mathcal{U} : D_f V(x_0,u) + \mathcal{I}(u) < -\alpha(x_0)\}} = \{u \in \mathcal{U} : D_f V(x_0,u) + \mathcal{I}(u) \leqslant -\alpha(x_0)\} =$

$K_V(x_0)$. Now any set-valued map whose values are the closures of the values of a lsc map is itself lsc, and thus K_V is lsc with nonempty closed convex values on $\mathcal{X} \setminus \Omega_{c_V}(V)$. □

PROPOSITION 4.5. *For every $c > c_V$ there exists a control law k for Σ such that $k(x) \in K_V(x)$ for all $x \in \mathcal{X} \setminus \Omega_c(V)$.*

Proof. It follows from Proposition 4.4 and Michael's theorem that there exists a continuous function $\phi : \mathcal{X} \setminus \Omega_{c_V}(V) \to \mathcal{U}$ such that $\phi(x) \in K_V(x)$ for every $x \in \mathcal{X} \setminus \Omega_{c_V}(V)$. Fix $c > c_V$ and consider the set-valued map $K_c : \mathcal{X} \rightsquigarrow \mathcal{U}$ defined by

$$K_c(x) := \begin{cases} U(x) & \text{when } x \in \operatorname{int} \Omega_c(V), \\ \{\phi(x)\} & \text{when } x \in \mathcal{X} \setminus \operatorname{int} \Omega_c(V). \end{cases} \tag{15}$$

It follows from A2 that K_c is lsc with nonempty closed convex values and thus admits a continuous selection k (Michael's theorem). Now $k(x) \in U(x)$ for all $x \in \mathcal{X}$, which means k is a control law for Σ; moreover, $k(x) = \phi(x) \in K_V(x)$ for $x \in \mathcal{X} \setminus \Omega_c(V)$. □

Proof of Theorem 4.2. Fix $c > c_V$ and let k be as in Proposition 4.5. Then for all $x \in \mathcal{X} \setminus \Omega_c(V)$ we have $D_f V(x, k(x)) \leqslant -\alpha(x)$. It follows that $\dot{V}(t) \leqslant -\alpha(x(t))$ (for almost all t) along solutions $x(t)$ of (7) that lie outside $\Omega_c(V)$, and standard Lyapunov arguments show that the solutions of (7) are RGUAS-$\Omega_c(V)$. We thus conclude that Σ is RS. If $c_V = 0$, then for any $\varepsilon > 0$ we can choose $c > 0$ above such that $\Omega_c(V) \subset \varepsilon B$, and we then conclude that Σ is RPS. □

When $c_V = 0$, we can prove that Σ is RAS rather than RPS provided V has the following additional property [2, 39, 34].

DEFINITION 4.6. *An* rclf V *for* Σ *satisfies the* continuous control property (ccp) *when* $c_V = 0$ *and there exists* $u^\circ \in U(0)$ *such that for every* $\varepsilon > 0$ *there exists* $\delta > 0$ *such that* $0 < \|x\| < \delta$ *implies*

$$\inf \left\{ D_f V(x, u) \,:\, u \in U(x),\ \|u - u^\circ\| < \varepsilon \right\} < 0. \tag{16}$$

V *satisfies the* small control property (scp) *when it satisfies the* ccp *with* $u^\circ = 0$.

COROLLARY 4.7. *If there is an* rclf *for* Σ *that satisfies the* ccp, *then* Σ *is* RAS.

Proof. Let ρ be a class $\mathcal{K}$ function such that $\rho(\varepsilon) < \delta$ for each (ε, δ)-pair described in Definition 4.6. We define a new control constraint $\widehat{U} : \mathcal{X} \rightsquigarrow \mathcal{U}$ as

$$\widehat{U}(x) := \begin{cases} U(x) & \text{when } \|x\| > \rho(1), \\ U(x) \cap \left[u^\circ + \rho^{-1}(\|x\|)B\right] & \text{when } \|x\| \leqslant \rho(1). \end{cases} \tag{17}$$

It follows from Definition 4.6 that $\widehat{U}$ is lsc with nonempty closed convex values (see [19, Prop. II.2.4]) and that V is an rclf for the new system $\widehat{\Sigma} := (f, \widehat{U}, W)$ with $c_V = 0$. The corresponding set-valued map $\widehat{K}_V$ is then lsc with nonempty closed convex values on all of $\mathcal{X}$, and it follows from Michael's theorem that there exists a control law k for $\widehat{\Sigma}$ such that $k(x) \in \widehat{K}_V(x)$ for all $x \in \mathcal{X}$. Now k is also a control law for the original system Σ, and standard Lyapunov arguments show that the solutions of the resulting closed-loop system are RGUAS. □

Note that the control law k given in the above proof satisfies $k(0) = u^\circ$. If V satisfies the scp, then by taking $u^\circ = 0$ we obtain a robustly asymptotically stabilizing control law k with $k(0) = 0$.

We conclude this section by mentioning that if the control constraint U has nonempty interior for all $x \in \mathcal{X}$ (in addition to satisfying assumption A2), then the control law k in Theorem 4.2 and Corollary 4.7 can be chosen to be *smooth* except possibly at the point $0 \in \mathcal{X}$. The proof of this fact is similar to Artstein's original proof in [2].[1] However, such smoothness cannot be achieved in general when $\mathcal{X}$ is an infinite-dimensional Banach space.

[1] We thank the anonymous reviewer for showing us this smooth version of our results.

5. Necessary conditions. Theorem 4.2 and Corollary 4.7 provide sufficient conditions for robust stabilizability in terms of the existence of an rclf. In this section we explore the necessity of these conditions using the recent converse Lyapunov theorem of [24], which was proved under the assumption that the closed-loop system is locally Lipschitz. We make this assumption in the following theorem, but we believe that it can be removed by modifying the proof in [24].

THEOREM 5.1. *Suppose the disturbance constraint W is locally Lipschitz with nonempty compact convex values, and suppose Σ is* RS *via a control law k that renders locally Lipschitz the mapping $(x, w) \to f(x, k(x), w)$. Then there is a smooth* rclf V *for Σ. If in addition Σ is* RAS *via such a control law k, then there is a smooth* rclf V *for Σ that satisfies the* ccp.

Proof. It follows from [5, Thm. 9.6.2] that we may assume $W(x) \equiv B$ without loss of generality. Let Ω denote the compact residual set in the definition of RS. It follows from [24, Thm. 2] that there exists a smooth function $V_0 : \mathcal{X} \to R_+$ and class $\mathcal{K}_\infty$ functions χ_1, χ_2, and χ_3 such that $\chi_1(|x|_\Omega) \leqslant V_0(x) \leqslant \chi_2(|x|_\Omega)$ for all $x \in \mathcal{X}$ and furthermore $D_f V_0(x, k(x)) \leqslant -\chi_3(|x|_\Omega)$ for all $x \in \mathcal{X} \setminus \Omega$. Because Ω is compact, there exists $c > 0$ and a smooth function $V \in \mathcal{A}_\infty(\mathcal{X})$ such that $V(x) = V_0(x)$ for all $x \in \mathcal{X} \setminus \Omega_c(V)$. It follows that V is an rclf for Σ. If Σ is RAS via such a control law k, then we take $\Omega = \{0\}$ and $V \equiv V_0$ to obtain a smooth rclf that satisfies the ccp. □

Note that in this theorem, assumptions A1–A3 have been replaced by an assumption on the disturbance constraint W alone. This theorem establishes the necessity of the existence of an rclf for both robust stabilizability and robust asymptotic stabilizability. The key part of our formulation that leads to this necessity is that there are no restrictions (other than measurability) on the time-variation of the admissible disturbances. In contrast, adaptive control formulations require the disturbances to be constant or slowly varying. If we were to so restrict the time-variation of the admissible disturbances, then the existence of an rclf would become a sufficient, but not necessary, condition for robust stabilizability.

6. Input-to-state stabilizability. In this section we describe an alternative problem formulation using the definition of *input-to-state stability* in [33]. Rather than consider disturbances that take values in a known compact set, we instead assume no knowledge of such a set and look for a control law that renders the closed-loop system input-to-state stable (ISS) with respect to the disturbance input. The ISS property is stronger than bounded-input/bounded-state stability together with the global asymptotic stability of the nominal (zero-input) system [36]. It is closely related to Lyapunov stability, and many authors have used the ISS property in conjunction with the Lyapunov-based design and analysis of nonlinear systems [35, 24, 27, 38, 40, 37]. In this section we show that ISS-stabilizability is equivalent to the existence of an rclf for an auxiliary system.

In Definition 6.1 and Corollaries 6.2 and 6.3, we consider disturbances $w(t)$ in the set $\mathcal{L}_\infty(\mathcal{W})$ of all measurable essentially bounded functions from R_+ to $\mathcal{W}$, with the essential supremum denoted by $\|\cdot\|_\infty$. Given a continuous function f describing the system dynamics and a control constraint U, we define the ISS-stabilizability of the pair (f, U) as follows.

DEFINITION 6.1. *The closed-loop system* (7) *is* globally input-to-state stable (ISS) *when there exists a class $\mathcal{KL}$ function β and a class $\mathcal{K}$ function χ such that for any $x_0 \in \mathcal{X}$ and any disturbance $w \in \mathcal{L}_\infty(\mathcal{W})$, all solutions $x(t)$ starting from x_0 exist for all $t \geqslant 0$ and satisfy $\|x(t)\| \leqslant \beta(\|x_0\|, t) + \chi(\|w\|_\infty)$ for all $t \geqslant 0$. The pair (f, U) is* ISS-stabilizable *when there exists a control law k such that the closed-loop system* (7) *is* ISS.

Given a class $\mathcal{K}_\infty$ function ρ, we let $W_\rho : \mathcal{X} \rightsquigarrow \mathcal{W}$ denote the set-valued map defined by $W_\rho(x) := \rho(\|x\|)B$; note that W_ρ satisfies A3. We associate with the pair (f, U) the *auxiliary system* $\Sigma_\rho := (f, U, W_\rho)$. Our first result is a consequence of Corollary 4.7 and states that the existence of an rclf for some such auxiliary system Σ_ρ implies the ISS-stabilizability of (f, U).

COROLLARY 6.2. *Let the pair* (f, U) *satisfy assumptions* A1 *and* A2. *Suppose there exists a class* $\mathcal{K}_\infty$ *function* ρ *such that the system* $\Sigma_\rho = (f, U, W_\rho)$ *has an* rclf V *that satisfies the* ccp. *Then the pair* (f, U) *is* ISS-*stabilizable.*

Proof. It follows from the proof of Corollary 4.7 that there exists a control law k for Σ_ρ and a negativity margin $\alpha \in \mathcal{A}(\mathcal{X})$ such that $D_f V(x, k(x)) \leqslant -\alpha(x)$ for all $x \in \mathcal{X}$. In other words, $w \in W_\rho(x)$ implies $L_f V(x, k(x), w) \leqslant -\alpha(x)$. From the definition of W_ρ, we know that $w \in W_\rho(x)$ if and only if $\|w\| \leqslant \rho(\|x\|)$. Thus for all $(x, w) \in \mathcal{X} \times \mathcal{W}$, $\|x\| \geqslant \rho^{-1}(\|w\|)$ implies $L_f V(x, k(x), w) \leqslant -\alpha(x)$. In the terminology of [24, 21], the function V is an ISS-Lyapunov function for (7), and it follows from [21, Prop. 3.1.7] that the closed-loop system (7) is ISS. □

We next use [24, Thm. 3] to prove a converse of the previous result.

COROLLARY 6.3. *Suppose the pair* (f, U) *is* ISS-*stabilizable via a control law* k *that renders locally Lipschitz the mapping* $(x, w) \to f(x, k(x), w)$. *Then there exists a class* $\mathcal{K}_\infty$ *function* ρ *such that the system* $\Sigma_\rho := (f, U, W_\rho)$ *has a smooth* rclf V *that satisfies the* ccp.

Proof. It follows from the converse ISS-Lyapunov theorem [24, Thm. 3] that there exists an ISS-Lyapunov function for (7), namely, that there exists a smooth function $V \in \mathcal{A}_\infty(\mathcal{X})$ and class $\mathcal{K}_\infty$ functions χ_3 and ζ such that for all $(x, w) \in \mathcal{X} \times \mathcal{W}$, $\|x\| \geqslant \zeta(\|w\|)$ implies $L_f V(x, k(x), w) \leqslant -\chi_3(\|x\|)$. If we define the class $\mathcal{K}_\infty$ function $\rho := \zeta^{-1}$, then $w \in W_\rho(x)$ implies $L_f V(x, k(x), w) \leqslant -\chi_3(\|x\|)$. It follows that $D_f V(x, k(x)) \leqslant -\chi_3(\|x\|)$ for all $x \in \mathcal{X}$. We conclude that V is an rclf for Σ_ρ that satisfies the ccp. □

7. Pointwise min-norm control laws. We have shown that the existence of an rclf implies robust stabilizability, but how do we use our knowledge of an rclf to construct a robustly stabilizing control law k? For systems without disturbances, constructive proofs of Artstein's theorem with explicit formulas for k are given in [39, 34, 23, 22, 21]. In these papers, different formulas are given for different *constant* control constraints (a formula for the unconstrained control case $U(x) \equiv \mathcal{U}$ is given in [39, 34], a formula for the control constraint $U(x) \equiv B$ is given in [23], and a formula for the constraint of positive controls is given in [22]). In this section our goal is to provide a formula for k that results in a *robustly* stabilizing control law for systems with disturbances, works for a general nonconstant control constraint $U(x)$, and naturally incorporates the negativity margin α as a design parameter. Moreover, we will show in following sections that the control laws generated by our formula, called pointwise min-norm control laws, are in fact optimal for meaningful games. We introduce the following additional assumptions on the system Σ:

A4. The control constraint U is such that Graph(U) is closed.

A5. The disturbance constraint W is lsc.

Let V be an rclf for Σ, and let L_V and K_V be the set-valued maps defined in (12) and (13). It follows from Proposition 4.4 that we can define $m_V : \mathcal{X} \to \mathcal{U}$ by

$$m_V(x) := \begin{cases} \arg\min \{\|u\| : u \in K_V(x)\} & \text{when } x \in \mathcal{X} \setminus \Omega_{c_V}(V), \\ 0 & \text{when } x \in \Omega_{c_V}(V). \end{cases} \tag{18}$$

This is called a *minimal selection for* V. It is not unique because the definition of K_V depends on the choice of the negativity margin α in Proposition 4.3.

PROPOSITION 7.1. *Let* Σ *satisfy assumptions* A1–A5, *and let* V *be an* rclf *for* Σ. *Then every minimal selection* m_V *for* V *is continuous on* $\mathcal{X} \setminus \Omega_{c_V}(V)$. *If furthermore* V *satisfies the* scp, *then there is a minimal selection* m_V *for* V *that is continuous on* $\mathcal{X}$.

Proof. It follows from A5, Proposition 4.3, and [5, Thm. 1.4.16] that $D_f V$ is continuous. Thus from A4 we see that Graph(K_V) = Graph$(U) \cap$ Graph(L_V) is closed. Now from Proposition 4.4 we know that K_V is lsc on $\mathcal{X} \setminus \Omega_{c_V}(V)$, and it follows from [5, Prop. 9.3.2] that Graph(m_V) is closed relative to $\mathcal{X} \setminus \Omega_{c_V}(V) \times \mathcal{U}$. Also, it follows from [5, Lem. 9.3.1] that

the mapping $x \mapsto \|m_V(x)\|$ is usc on $\mathcal{X} \setminus \Omega_{c_V}(V)$, which implies that m_V is locally bounded on $\mathcal{X} \setminus \Omega_{c_V}(V)$. Because the dimension of $\mathcal{U}$ is finite, the closedness of Graph(m_V) and the local boundedness of m_V on $\mathcal{X} \setminus \Omega_{c_V}(V)$ imply the continuity of m_V on $\mathcal{X} \setminus \Omega_{c_V}(V)$. Next suppose V satisfies the scp; then from the the proof of Corollary 4.7 there exists a choice for α in Proposition 4.3 such that K_V admits a continuous selection k with $k(0) = 0$. From (18) we have $0 \leqslant \|m_V(x)\| \leqslant \|k(x)\|$ for all $x \in \mathcal{X}$, and therefore m_V is continuous at $x = 0$. □

As a result of Proposition 7.1, minimal selections m_V for V can be used to generate robustly stabilizing control laws for Σ.

DEFINITION 7.2. *Let V be an* rclf *for Σ. If V satisfies the* scp, *then every control law k for Σ that is also a minimal selection for V is called* pointwise min-norm for V. *If V does not satisfy the* scp, *then every control law k for Σ that satisfies $k(x) = m_V(x)$ for all $x \in \mathcal{X} \setminus \Omega_c(V)$ for some $c > c_V$ and some minimal selection m_V for V is called pointwise min-norm for V.*

Pointwise min-norm control laws are so named because at each point x (except possibly inside some sublevel set of V), their value is the unique element of $\mathcal{U}$ of minimum norm that satisfies the control constraint $U(x)$ and makes the worst case Lyapunov derivative at least as negative as $-\alpha(x)$. These control laws naturally incorporate the negativity margin α as a design parameter. Note that if V satisfies the scp, then only those choices for the negativity margin α that lead to *continuous* minimal selections m_V for V will generate pointwise min-norm control laws for V.

We can compute the value of a pointwise min-norm control law at any point x by solving the static minimization problem (18). This is a convex programming problem on the control space $\mathcal{U}$ and is completely determined by the data Σ, V, and α. One of the constraints in this problem depends on $D_f V$, and the calculation of $D_f V(x, u)$ in (10) for any fixed $(x, u) \in \mathcal{X} \times \mathcal{U}$ is itself a static nonlinear programming problem on the disturbance space $\mathcal{W}$. We will show in the next sections that every pointwise min-norm control law is optimal for a meaningful differential game, and therefore our formula (18) allows us to compute such a control law by solving a static rather than dynamic programming problem. Furthermore, this static programming problem has a simple explicit solution in the following special cases.

Jointly affine systems. Suppose the function f is jointly affine in the control and disturbance; that is, suppose the system (6) can be written as

$$\dot{x} = f_0(x) + f_1(x)\,u + f_2(x)\,w \tag{19}$$

for continuous functions $f_0 : \mathcal{X} \to \mathcal{X}$, $f_1 : \mathcal{X} \to \mathcal{B}(\mathcal{U}, \mathcal{X})$, and $f_2 : \mathcal{X} \to \mathcal{B}(\mathcal{W}, \mathcal{X})$. Suppose also that the control and disturbance constraints are given by $U(x) \equiv \mathcal{U}$ and $W(x) \equiv B$, respectively. Let V be an rclf for this system; then from (10) we have

$$D_f V(x, u) = \nabla V(x) \cdot f_0(x) + \nabla V(x) \cdot f_1(x)\,u + \left\|\nabla V(x) \cdot f_2(x)\right\|. \tag{20}$$

We choose a negativity margin α according to (11) and use (20) to write

$$D_f V(x, u) + \alpha(x) = \psi_0(x) + \psi_1^T(x)\,u \tag{21}$$

with $\psi_0(x) := \nabla V(x) \cdot f_0(x) + \left\|\nabla V(x) \cdot f_2(x)\right\| + \alpha(x)$ and $\psi_1(x) := \left[\nabla V(x) \cdot f_1(x)\right]^T$. Then from (13) we have

$$K_V(x) = \left\{u \in \mathcal{U} : \psi_0(x) + \psi_1^T(x)\,u \leqslant 0\right\}. \tag{22}$$

It now follows from (18) and the projection theorem that

$$m_V(x) = \begin{cases} -\dfrac{\psi_0(x)\,\psi_1(x)}{\psi_1^T(x)\,\psi_1(x)} & \text{when } x \in \mathcal{X} \setminus \Omega_{c_V}(V) \text{ and } \psi_0(x) > 0, \\ 0 & \text{when } x \in \Omega_{c_V}(V) \text{ or } \psi_0(x) \leqslant 0. \end{cases} \tag{23}$$

This explicit formula for m_V depends on the negativity margin α through the function ψ_0 and has the continuity properties described in Proposition 7.1. Note that there is never division by zero in this expression because the set $K_V(x)$ is nonempty for all $x \in \mathcal{X} \setminus \Omega_{c_V}(V)$ (Proposition 4.4). Because of the symmetry of the unit ball, this expression is also valid under control constraints of the form $U(x) = p(x)B$ for a continuous function $p : \mathcal{X} \to R_+$.

This formula (23) generates the control law (3) for the system (1) in §1 as follows. For this system we have $f_0(x) = -x^3$, $f_1(x) = 1$, and $f_2(x) = x$. We choose $V(x) = \frac{1}{2}x^2$ so that $\nabla V(x) = x$, and we choose $\alpha(x) = x^2$. Then we obtain $\psi_0(x) = -x^4 + 2x^2$ and $\psi_1(x) = x$, and the formula (23) yields the control law (3).

Feedback linearizable systems. We now apply the formula (23) to the class of (globally) feedback linearizable systems with no disturbances. Suppose there are coordinates in which our system is

$$\dot{x} = Fx + G\big[\ell_0(x) + \ell_1(x)\,u\big], \tag{24}$$

where the matrix pair (F, G) is stabilizable and the continuous functions $\ell_0 : \mathcal{X} \to \mathcal{U}$ and $\ell_1 : \mathcal{X} \to \mathcal{B}(\mathcal{U}, \mathcal{U})$ are such that $\ell_0(0) = 0$ and $\ell_1(x)$ is nonsingular for all $x \in \mathcal{X}$. Suppose also that the control constraint is $U(x) \equiv \mathcal{U}$. The control law suggested by feedback linearization would cancel nonlinearities and apply linear feedback as

$$k(x) = \big[\ell_1(x)\big]^{-1}\big[-\ell_0(x) + Kx\big], \tag{25}$$

where K is a stabilizing gain matrix for the pair (F, G). Unfortunately, this control law (25) may not have good robustness properties and might waste control effort to counteract beneficial nonlinearities. We instead construct a pointwise min-norm control law by using the formula (23) as follows. We first choose symmetric positive definite matrices P and Q such that the Lyapunov matrix equation

$$P\big(F + GK\big) + \big(F + GK\big)^T P = -Q \tag{26}$$

is satisfied. Then $V(x) := x^T Px$ is a clf that satisfies the scp, and an appropriate choice for the negativity margin is $\alpha(x) := \varepsilon x^T Qx$ for some $\varepsilon \in (0, 1)$. In this case, the function ψ_0 above is

$$\psi_0(x) = 2x^T P\big[Fx + G\,\ell_0(x)\big] + \varepsilon x^T Qx \tag{27}$$

and the formula (23) becomes

$$m_V(x) = \begin{cases} -\dfrac{\psi_0(x)\,\ell_1^T(x)G^T Px}{2x^T PG\,\ell_1(x)\,\ell_1^T(x)G^T Px} & \text{when } \psi_0(x) > 0, \\ 0 & \text{when } \psi_0(x) \leqslant 0. \end{cases} \tag{28}$$

The design parameters in this expression are ε, P, and Q; and their choices are constrained by $\varepsilon \in (0, 1)$ and the Lyapunov matrix equation (26). This minimal selection m_V is continuous everywhere (Proposition 7.1) and, therefore, defines a pointwise min-norm control law $k(x) := m_V(x)$. Although both control laws (28) and (25) globally asymptotically stabilize the system (24), the control law (28) is optimal with respect to a meaningful cost functional (as shown in the next sections) whereas the control law (25) is not (in general). The potential advantage of (28) over (25) was illustrated in the comparison of the control laws (3) and (2) in §1.

We end this section with a discussion of the role of the negativity margin α. This function represents the desired negativity of the Lyapunov derivative and can be adjusted to achieve a trade-off between the control effort and the rate of convergence of the state to zero. For example, if α is a negativity margin, then so is $\varepsilon\alpha$ for every $\varepsilon \in (0, 1)$. For each such ε we then obtain a different pointwise min-norm control law k for Σ. In general, smaller values of ε will lead to smaller control magnitudes and slower convergence. Moreover, by adjusting the shape of α we can place more cost on some states and less cost on others. Thus the function α should be regarded as a design parameter to be adjusted according to design specifications.

8. Inverse optimal robust stabilization. Our goal in the next three sections is to show that every pointwise min-norm control law is optimal for a meaningful game. We accomplish this by showing that every rclf solves the steady-state HJI equation associated with such a game. These results represent a solution to an inverse optimal robust stabilization problem for nonlinear systems with disturbances. As a consequence of these results, we can use formulas (18) and (23) to compute optimal robustly stabilizing control laws *without* solving the HJI equation for the upper value function, provided an rclf is known.

In this section and the next, we assume that the scp is satisfied, which in particular implies that the disturbances vanish at the equilibrium point. In this case our system is RAS and we can therefore consider cost functionals defined over the infinite horizon. In §10 we remove the scp assumption and allow persistent disturbances, in which case we must consider cost functionals defined over a finite horizon determined by the time required to reach a given target set.

We assume that the system Σ satisfies A1–A5 plus the following assumption on the control constraint.

A6. There exists a continuous function $\pi : \mathcal{X} \to R_+$ such that $\pi(x) > 0$ and $\pi(x)B \subset U(x)$ for all $x \in \mathcal{X}$.

One can show that, under assumption A2, assumption A6 is equivalent to the assumption that $0 \in \operatorname{int} U(x)$ for all $x \in \mathcal{X}$. Our cost functionals will be characterized by functions $q : \mathcal{X} \to R_+$ and $r : \mathcal{X} \times \mathcal{U} \to R_+$, which satisfy the following specifications.

S1. q satisfies $q \in \mathcal{A}_\kappa(\mathcal{X})$.

S2. r is continuous, and for each fixed $x \in \mathcal{X}$ we have $r(x, u) = \gamma_x(\|u\|_x)$ for some convex class $\mathcal{K}$ function γ_x and some norm $\|\cdot\|_x$

Using such a pair (q, r), we form a two-person zero-sum differential game $\mathcal{G}(q, r)$ by considering a cost functional J parameterized by the initial condition $x_0 \in \mathcal{X}$. In this game, the control tries to minimize J and the disturbance tries to maximize J. Given $x_0 \in \mathcal{X}$, $u_g \in \mathcal{C}(\Sigma)$, and $w_a \in \mathcal{D}(\Sigma)$ we define the cost

$$J(u_g, w_a, x; x_0) := \int_0^\infty [q(x) + r(x, u_g)]dt, \tag{29}$$

where the integration is taken along the solution $x(t)$ of the differential equation (8) starting from the initial condition x_0. Because such solutions are not necessarily unique, we included in our notation $J(u_g, w_a, x; x_0)$ the dependence on the particular state trajectory x along which we integrate. If the solution $x(t)$ cannot be extended for all $t \geqslant 0$, then we set $J(u_g, w_a, x; x_0) := \infty$. Also, because q is bounded from below by a class $\mathcal{K}$ function of the norm, $J < \infty$ implies $x(t) \to 0$ as $t \to \infty$. We define the *upper value function* $\bar{J} : \mathcal{X} \to R_+ \cup \{\infty\}$ of the game by the equation

$$\bar{J}(x_0) := \inf_{u_g} \sup_{w_a} \sup_{x} J(u_g, w_a, x; x_0). \tag{30}$$

The first supremum is taken over all solutions $x(t)$ of (8) starting from x_0 (this supremum is superfluous if solutions are unique), the second supremum is taken over all admissible

disturbances $w_a \in \mathcal{D}(\Sigma)$, and the infimum is taken over all generic controls $u_g \in \mathcal{C}(\Sigma)$. When this infimum is achieved by some $u_g^* \in \mathcal{C}(\Sigma)$ for some $x_0 \in \mathcal{X}$ such that $J(x_0) < \infty$, we say that u_g^* is *optimal from* x_0. We say a control law k^* for Σ is *optimal for* $\mathcal{G}(q, r)$ when k^* is optimal from every $x_0 \in \mathcal{X}$. Such an optimal control law minimizes the worst case cost for every initial condition. Also, because $J < \infty$ only if $x(t) \to 0$ as $t \to \infty$, every optimal control law drives the state (robustly) to zero from any initial condition.

Before we proceed, we introduce the notion of a strong rclf. If V is an rclf for Σ, then Proposition 4.3 guarantees the existence of a negativity margin α that is nonzero outside some compact set. We say that V is a *strong* rclf when α can be chosen to be *bounded away from zero* outside some compact set. If $c_V = 0$, this means that we can choose $\alpha \in \mathcal{A}_\kappa(\mathcal{X})$ instead of merely $\alpha \in \mathcal{A}(\mathcal{X})$. We will see below that this stronger property will lead to a function q in (29) that belongs to $\mathcal{A}_\kappa(\mathcal{X})$ rather than just to $\mathcal{A}(\mathcal{X})$, thus guaranteeing that $J < \infty$ only if $x(t) \to 0$ as $t \to \infty$. The restriction to strong rclf's is not important in practice; it follows from results in [33] that if an rclf is known, then the construction of a strong rclf is straightforward.

Let V be a strong rclf for Σ that satisfies the scp, and let k^* be a pointwise min-norm control law for V associated with a negativity margin $\alpha \in \mathcal{A}_\kappa(\mathcal{X})$. Theorem 8.1 states that there exists a pair (q, r) satisfying S1 and S2 such that k^* is optimal for $\mathcal{G}(q, r)$ with V being the corresponding upper value function. In simple terms, every pointwise min-norm control law is optimal and every (strong) rclf is an upper value function. This theorem is only of interest when such a game $\mathcal{G}(q, r)$ is meaningful, and we claim that this is indeed the case. First of all, it follows from S1 and S2 that the integrand in (29) is bounded below by a class $\mathcal{K}$ function of $\|x\|$; thus $J < \infty$ only if the objective of driving the state to zero is achieved. Furthermore, for each fixed $x \in \mathcal{X}$ the integrand is a convex function of u with a global minimum at the point $u = 0$; thus there is always a higher penalty for values of u further away from zero and there are no local minima other than $u = 0$. In fact, integrands satisfying S1 and S2 are a generalization of the familiar quadratic integrand $x^T Q x + u^T R u$ for symmetric positive definite matrices Q and R. Next, the disturbance w is given two advantages in the game consistent with the goal of robust stabilization. First, there is no direct cost on w in (29); and second, w is allowed to base its strategy on knowledge of the strategy of the control u (we consider the *upper* value of the game). Finally, the control law k^* is optimal with respect to *all* generic controls $u_g \in \mathcal{C}(\Sigma)$, not just those that satisfy the control constraint; in other words, we do not allow the possibility of reducing the guaranteed cost by relaxing the control constraint.

To prove optimality rather than suboptimality, we need to assume that the effect of a certain *worst case disturbance* (one that maximizes the Lyapunov derivative) can be approximated arbitrarily closely by admissible disturbances. A worst case disturbance is a function $w^* : \mathcal{X} \times \mathcal{U} \to \mathcal{W}$ such that $L_f V(x, u, w^*(x, u)) = D_f V(x, u)$ and $w^*(x, u) \in W(x)$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$. Such a function w^* may be discontinuous and thus not belong to the set $\mathcal{D}(\Sigma)$ of admissible disturbances. Our approximation condition is as follows.

DC. For every $x_0 \in \mathcal{X}$, every $u_g \in \mathcal{C}(\Sigma)$, and every $\Delta > 0$ there exists $w_\Delta \in \mathcal{D}(\Sigma)$ and a solution $x_\Delta(t)$ of (8) starting from x_0 (with $w_a = w_\Delta$) such that either $J(u_g, w_\Delta, x_\Delta; x_0) = \infty$ or for every $T \geqslant 0$ we have

$$\int_0^T L_f V(x_\Delta, u_g, w_\Delta)\, dt \geqslant \int_0^T D_f V(x_\Delta, u_g)\, dt - \Delta. \tag{31}$$

If this condition is not true for our system, then we can only prove suboptimality. We now state one of the main results of this paper.

THEOREM 8.1. *Let* Σ *satisfy assumptions* A1–A6, *let* V *be a strong* rclf *for* Σ *which satisfies the* scp, *and let* k^* *be a pointwise min-norm control law for* V *associated with a*

negativity margin $\alpha \in \mathcal{A}_\kappa(\mathcal{X})$. *Then there exists a pair* (q, r) *satisfying* S1 *and* S2 *such that* $J(k^*, w_a, x; x_0) \leqslant V(x_0)$ *for every* $x_0 \in \mathcal{X}$, *every* $w_a \in \mathcal{D}(\Sigma)$, *and every solution* $x(t)$ *of the closed-loop system*

$$\dot{x} = f(x, k^*(x), w_a) \tag{32}$$

starting from x_0. *If, furthermore,* DC *is true, then* $\bar{J}(x_0) = V(x_0)$ *for all* $x_0 \in \mathcal{X}$, *which means* k^* *is optimal for* $\mathcal{G}(q, r)$.

The proof of this theorem involves several steps and will be presented in the next section. The main idea is to construct the functions q and r in such a way that V satisfies the steady-state HJI equation

$$0 = \min_{u \in \mathcal{U}} \max_{w \in W(x)} [q(x) + r(x, u) + L_f V(x, u, w)] \tag{33}$$

for all $x \in \mathcal{X}$. We accomplish this by constructing a continuous function r that satisfies S2 and that has two additional properties. First, r must be such that the function q defined by

$$0 = q(x) + r(x, k^*(x)) + D_f V(x, k^*(x)) \tag{34}$$

satisfies S1; and second, r must be such that the equality

$$\min_{u \in \mathcal{U}} [r(x, u) + D_f V(x, u)] = r(x, k^*(x)) + D_f V(x, k^*(x)) \tag{35}$$

holds for all $x \in \mathcal{X}$. Once we find such a function r, it will follow from (34) and (35) that the HJI equation (33) is true.

9. Proof of Theorem 8.1. We begin with a careful construction of the function r in (29). We first define a set-valued map $C : \mathcal{X} \rightsquigarrow \mathcal{U}$ as

$$C(x) = \operatorname{co}\big([\pi(x)B] \cup \{k^*(x), -k^*(x)\} \big). \tag{36}$$

The values of C are compact and convex, and for each $x \in \mathcal{X}$ we have $0 \in \operatorname{int} C(x)$ and $C(x) = -C(x)$. Furthermore, it follows from [19, Thms. II.2.7 and II.2.10] that C is continuous on $\mathcal{X}$. Associated with C is the Minkowski distance functional $\sigma : \mathcal{X} \times \mathcal{U} \to R_+$ defined as

$$\sigma(x, u) := \inf \{\lambda \geqslant 0 : u \in \lambda C(x)\}. \tag{37}$$

PROPOSITION 9.1. *For each* $x \in \mathcal{X}$, $\sigma(x, \cdot)$ *is a norm on* $\mathcal{U}$. *Also,* σ *is continuous on* $\mathcal{X} \times \mathcal{U}$.

Proof. The first statement follows from [30, Thm. 15.2], and we have left to prove that σ is continuous. Fix $(x_0, u_0) \in \mathcal{X} \times \mathcal{U}$ and let $\{(x_i, u_i)\} \in \mathcal{X} \times \mathcal{U}$ be a sequence converging to (x_0, u_0). Define $\sigma_i := \sigma(x_i, u_i)$ and $\sigma_0 := \sigma(x_0, u_0)$; we need to show that $\sigma_i \to \sigma_0$ or, equivalently, that every subsequence of $\{\sigma_i\}$ has in turn a subsequence that converges to σ_0. Let $\{\sigma_{i_j}\}$ be a subsequence of $\{\sigma_i\}$, and define $C_j := C(x_{i_j})$ and $C_0 := C(x_0)$. It follows from (37) that there exists a sequence $\{w_j\} \in \mathcal{U}$ such that $w_j \in \partial C_j$ and $u_{i_j} = \sigma_{i_j} w_j$ for all $j \geqslant 1$. It then follows from [19, Thm. II.2.2] that the sequence $\{w_j\}$ has a subsequence $\{w_{j_k}\}$ that converges to some $w_0 \in C_0$. We claim that $w_0 \in \partial C_0$. Indeed, because $w_{j_k} \in \partial C_{j_k}$, we know there exists a sequence of unit vectors $\{v_k\} \in \mathcal{U}$ such that $v_k \in N_{C_{j_k}}(w_{j_k})$ for all $k \geqslant 1$. The unit sphere in $\mathcal{U}$ is compact, and thus Limsup $N_{C_{j_k}}(w_{j_k})$ contains a unit vector. It then follows from [5, Cor. 7.6.5] that $N_{C_0}(w_0) \neq \{0\}$, which means $w_0 \in \partial C_0$. We can now show that $\sigma_{i_{j_k}} \to \sigma_0$. First suppose $u_0 = 0$; then $\sigma_0 = 0$ and $u_{i_{j_k}} = \sigma_{i_{j_k}} w_{j_k} \to 0$, and because $w_{j_k} \to w_0 \neq 0$, it follows that $\sigma_{i_{j_k}} \to 0 = \sigma_0$. Next suppose $u_0 \neq 0$. By the continuity of the inner product and the

continuity of $\operatorname{sgn}(\cdot)$ away from zero, we have $\langle \operatorname{sgn}(w_0), \operatorname{sgn}(u_{i_{j_k}})\rangle = \langle \operatorname{sgn}(w_0), \operatorname{sgn}(w_{j_k})\rangle \to \langle \operatorname{sgn}(w_0), \operatorname{sgn}(w_0)\rangle = 1$ and $\langle \operatorname{sgn}(w_0), \operatorname{sgn}(u_{i_{j_k}})\rangle \to \langle \operatorname{sgn}(w_0), \operatorname{sgn}(u_0)\rangle$, and it follows that $u_0 = \lambda w_0$ for some $\lambda > 0$. Because $w_0 \in \partial C_0$, we have $\lambda = \sigma_0$, and thus $u_{i_{j_k}} \to u_0$ implies $\sigma_{i_{j_k}} w_{j_k} \to \sigma_0 w_0$. Now because $w_{j_k} \to w_0 \neq 0$, we have $\sigma_{i_{j_k}} \to \sigma_0$ as desired. □

We next define a set-valued map $D : \mathcal{X} \rightsquigarrow \mathcal{U}$ as

$$D(x) := \partial_u D_f V(x, k^*(x))$$

$$= \{w \in \mathcal{U} : D_f V(x,u) \geqslant D_f V(x, k^*(x)) + \langle w, u - k^*(x)\rangle \quad \forall u \in \mathcal{U}\}. \tag{38}$$

Thus $D(x)$ is the partial (convex) subdifferential with respect to u of the worst case Lyapunov derivative $D_f V$, evaluated at $(x, k^*(x))$. It follows from Lemma 12.1 and [19, Cor. II.2.1] that D is usc on $\mathcal{X}$ and has nonempty, convex, compact values. The next two propositions follow from the pointwise min-norm property of k^*.

PROPOSITION 9.2. *If $x \in \mathcal{X}$ is such that $k^*(x) \neq 0$, then $D_f V(x, k^*(x)) = -\alpha(x)$ and $0 \notin D(x)$.*

Proof. Recall that k^* is the minimal selection for V associated with α. Fix $x \in \mathcal{X}$ and suppose $k^*(x) \neq 0$. It follows from A6, (13), and (18) that $D_f V(x, 0) > -\alpha(x)$. Because the mapping $u \mapsto D_f V(x,u)$ is continuous and $U(x)$ is convex, the intermediate value theorem gives $D_f V(x, k^*(x)) = -\alpha(x)$. Suppose $0 \in D(x)$; then from (38) we have $-\alpha(x) = D_f V(x, k^*(x)) \leqslant D_f V(x,u)$ for all $u \in \mathcal{U}$, but this contradicts (11) because $k^*(x) \neq 0$ implies $x \in \mathcal{X} \setminus \Omega_{c_V}(V)$. □

PROPOSITION 9.3. *There exist continuous functions $\mu, \nu : \mathcal{X} \to R_+$ such that $\mu(x) \leqslant d(0, D(x))$ and $\|D(x)\| < \nu(x)$ for all $x \in \mathcal{X}$, with μ having the additional property that $\mu(x) > 0$ if and only if $k^*(x) \neq 0$.*

Proof. Because D is usc with nonempty bounded values, it follows from [5, Lem. 9.3.1] that the map $x \mapsto \|D(x)\|$ is usc on $\mathcal{X}$ and the map $x \mapsto d(0, D(x))$ is lsc on $\mathcal{X}$. The existence of ν then follows from [18, Prob. 5X]. Let $G \subset \mathcal{X}$ denote the open set $G = \{x \in \mathcal{X} : k^*(x) \neq 0\}$; it then follows from Proposition 9.2 and [18, Prob. 5X] that there exists a continuous function $\mu_0 : G \to R_+$ such that $0 < \mu_0(x) \leqslant d(0, D(x))$ for all $x \in G$. We define μ by setting $\mu(x) := 0$ when $k^*(x) = 0$ and otherwise

$$\mu(x) := \frac{\|k^*(x)\|}{1 + \|k^*(x)\| + \mu_0(x)}\, \mu_0(x). \tag{39}$$

This function μ has the desired properties. □

These functions μ and ν should be regarded as continuous "lower" and "upper" bounds (respectively) on the set-valued map D. We will now use them to construct the function r in the cost functional (29). Let $\alpha_0 \in \mathcal{A}(\mathcal{X})$ be such that $(\alpha - \alpha_0) \in \mathcal{A}_\kappa(\mathcal{X})$; for example, take $\alpha_0 := \varepsilon\alpha$ for some $\varepsilon \in (0,1)$. We next construct continuous functions $a, b : \mathcal{X} \to R_+$ as

$$a(x) := \begin{cases} \min\left\{\mu(x)\pi(x),\ \dfrac{\alpha_0(x)}{\sigma(x, k^*(x))}\right\} & \text{when } k^*(x) \neq 0, \\ 0 & \text{when } k^*(x) = 0; \end{cases} \tag{40}$$

$$b(x) := \nu(x) \max\left\{\|k^*(x)\|,\ \pi(x)\right\}. \tag{41}$$

It follows from Propositions 9.1 and 9.3 that a is continuous with $a(x) > 0$ whenever $k^*(x) \neq 0$. Also, we see from Proposition 9.3 that b is continuous with $b(x) > 0$ for all $x \in \mathcal{X}$. Furthermore, for all $x \in \mathcal{X}$ we have $0 \leqslant a(x) \leqslant \mu(x)\pi(x) < \nu(x)\pi(x) \leqslant b(x)$. We use a and b to define a function $\gamma : \mathcal{X} \times R_+ \to R_+$ as

$$\gamma(x,s) := \begin{cases} a(x)s & \text{when } 0 \leqslant s \leqslant \sigma(x, k^*(x)), \\ b(x)s + \sigma(x, k^*(x))\big[a(x) - b(x)\big] & \text{when } \sigma(x, k^*(x)) < s. \end{cases} \tag{42}$$

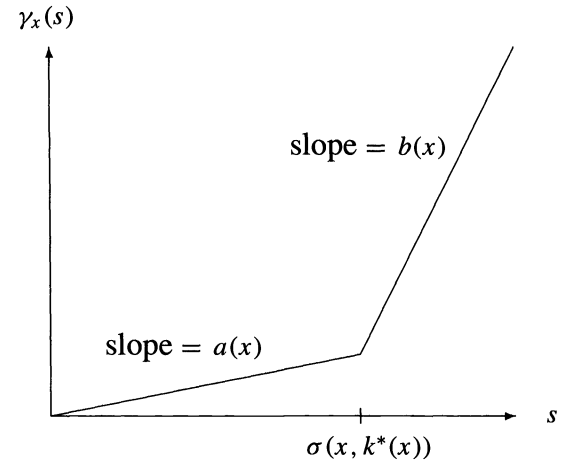


FIG. 3. *The function $\gamma_x(s)$ when $k^*(x) \neq 0$.*

It follows from the continuity of a, b, σ, and k^* that γ is continuous. Now fix $x \in \mathcal{X}$ and consider the function $\gamma_x : R_+ \to R_+$ given by $\gamma_x(s) = \gamma(x, s)$. Suppose $k^*(x) = 0$; then for all $s \geqslant 0$ we have $\gamma_x(s) = b(x)s$ with $b(x) > 0$, and thus γ_x is a (linear) convex class $\mathcal{K}$ function. Now suppose $k^*(x) \neq 0$; then $0 < a(x) < b(x)$ and we see from (42) that γ_x is a (piecewise-linear) convex class $\mathcal{K}$ function (see Fig. 3). Our choice for γ is not unique; in fact, redefining γ for $s \geqslant \sigma(x, k^*(x)) + 1$ to be anything preserving convexity and continuity (for example, replacing the linear growth with quadratic growth) will not alter our results.

We now choose the functions q and r in (29) as

$$r(x, u) := \gamma(x, \sigma(x, u)), \tag{43}$$

$$q(x) := -r(x, k^*(x)) - D_f V(x, k^*(x)) \tag{44}$$

for all $(x, u) \in \mathcal{X} \times \mathcal{U}$. Clearly q and r are continuous, and it follows from Proposition 9.1 and the properties of γ discussed above that r satisfies S2. We next show that q satisfies S1. We see from (42) and (40) that $r(x, k^*(x)) = a(x)\sigma(x, k^*(x)) \leqslant \alpha_0(x)$ for all $x \in \mathcal{X}$. Now $D_f V(x, k^*(x)) \leqslant -\alpha(x)$ for all $x \in \mathcal{X}$, and it follows from (44) that $q(x) \geqslant \alpha(x) - \alpha_0(x)$ for all $x \in \mathcal{X}$. By inspection we have $q(0) = 0$, and because $(\alpha - \alpha_0) \in \mathcal{A}_\kappa(\mathcal{X})$, it follows that $q \in \mathcal{A}_\kappa(\mathcal{X})$. We now show that the function r satisfies the key equation (35).

PROPOSITION 9.4. *It is true for all $x \in \mathcal{X}$ that*

$$\min_{u \in \mathcal{U}} [r(x, u) + D_f V(x, u)] = r(x, k^*(x)) + D_f V(x, k^*(x)). \tag{45}$$

Proof. We fix $x \in \mathcal{X}$. Because $r + D_f V$ is convex in u, (45) is true if and only if $0 \in \partial_u(r + D_f V)(x, k^*(x))$, where ∂_u denotes the partial subdifferential with respect to u. From [30, Thm. 23.8] and (38) we see that this condition is equivalent to the condition $0 \in \partial_u r(x, k^*(x)) + D(x)$.

First suppose $k^*(x) = 0$. Then from (36) we have $C(x) = \pi(x)B$, and it follows from (37) that $\sigma(x, u) = \|u\|/\pi(x)$ for all $u \in \mathcal{U}$. Thus $\partial_u \sigma(x, 0) = (1/\pi(x))B$. Now we have $\sigma(x, k^*(x)) = \sigma(x, 0) = 0$, which means $\gamma(x, s) = b(x)s$ for all $s \geqslant 0$; and it follows from (43) and (41) that $\partial_u r(x, k^*(x)) = b(x)\,\partial_u \sigma(x, k^*(x)) = (b(x)/\pi(x))B = \nu(x)B$. Now $\|D(x)\| < \nu(x)$ and so $0 \in [\nu(x)B] + D(x) = \partial_u r(x, k^*(x)) + D(x)$ as desired.

Next suppose $k^*(x) \neq 0$. We need to compute $\partial_u r(x, k^*(x))$. Now $\sigma(x, k^*(x)) > 0$, and it follows immediately from (42) (see Fig. 3) that

$$\partial_s \gamma\big(x, \sigma(x, k^*(x))\big) = \big[a(x),\ b(x)\big], \tag{46}$$

where ∂_s denotes the partial subdifferential with respect to s. We next compute $\partial_u \sigma(x, k^*(x))$. Let $E := \big\{u \in \mathcal{U} : \sigma(x, u) \leqslant \sigma(x, k^*(x))\big\}$ denote the sublevel set of σ at $k^*(x)$, and let N_E denote the normal cone to E at $k^*(x)$. From (36) and (37) we have $E \subset \|k^*(x)\| B$, and it follows that $k^*(x) \in N_E$. It then follows from [30, Cor. 23.7.1] that there exists $\lambda > 0$ such that $\lambda \operatorname{sgn}(k^*(x)) \in \partial_u \sigma(x, k^*(x))$. We now calculate the value of λ. From the definition of the subdifferential, it follows that $\sigma(x, \xi) \geqslant \sigma(x, k^*(x)) + \big\langle \lambda \operatorname{sgn}(k^*(x)),\ \xi - k^*(x)\big\rangle$ for all $\xi \in \mathcal{U}$. Setting $\xi = 0$ we obtain $0 \geqslant \sigma(x, k^*(x)) + \big\langle \lambda \operatorname{sgn}(k^*(x)),\ -k^*(x)\big\rangle$, which implies $\sigma(x, k^*(x)) \leqslant \lambda \|k^*(x)\|$. Setting $\xi = 2k^*(x)$ we obtain $2\sigma(x, k^*(x)) \geqslant \sigma(x, k^*(x)) + \big\langle \lambda \operatorname{sgn}(k^*(x)),\ k^*(x)\big\rangle$, which implies $\sigma(x, k^*(x)) \geqslant \lambda \|k^*(x)\|$. We have thus shown that $\lambda = \sigma(x, k^*(x))/\|k^*(x)\|$. Now if $\|k^*(x)\| \leqslant \pi(x)$ we have $\sigma(x, k^*(x)) = \|k^*(x)\|/\pi(x)$ and thus $\lambda = 1/\pi(x)$. Otherwise $\|k^*(x)\| > \pi(x)$, which means $\sigma(x, k^*(x)) = 1$ and we have $\lambda = 1/\|k^*(x)\|$. It follows from (41) that $\lambda = \nu(x)/b(x)$, and thus from Lemma 12.2 we obtain

$$\partial_u \sigma(x, k^*(x)) = N_E \cap \left[\frac{\nu(x)}{b(x)} \operatorname{sgn}(k^*(x)) + \big\{k^*(x)\big\}^{\perp}\right]. \tag{47}$$

It then follows from the projection theorem that $d(0, \partial_u \sigma(x, k^*(x))) = \nu(x)/b(x)$. We next show that $\|\partial_u \sigma(x, k^*(x))\| \leqslant 1/\pi(x)$. If $\|k^*(x)\| \leqslant \pi(x)$, then $\partial_u \sigma(x, k^*(x))$ is a singleton, which means $\|\partial_u \sigma(x, k^*(x))\| = d(0, \partial_u \sigma(x, k^*(x))) = \nu(x)/b(x) = 1/\pi(x)$. Next suppose $\|k^*(x)\| > \pi(x)$ and let $w \in \partial_u \sigma(x, k^*(x))$; then from (47) we have $w \in N_E \backslash \{0\}$, which means $H := \big[k^*(x) + \{w\}^{\perp}\big]$ is a supporting hyperplane of E. Let $v = d(0, H) \operatorname{sgn}(w)$; then from the projection theorem we have $v \in H$, which means $H = \big[v + \{w\}^{\perp}\big]$. Because $\pi(x) B \subset E$ we have $\|v\| = d(0, H) \geqslant \pi(x) > 0$, which means $H = \big[v + \{v\}^{\perp}\big]$. Now $\nu(x)/b(x) = 1/\|k^*(x)\|$ for $\|k^*(x)\| > \pi(x)$, and it follows from (47) that $w = \operatorname{sgn}(k^*(x))/\|k^*(x)\| + w_1$ for some $w_1 \in \big\{k^*(x)\big\}^{\perp}$. Thus $\big\langle w,\ k^*(x)\big\rangle = \big\langle \operatorname{sgn}(k^*(x))/\|k^*(x)\|,\ k^*(x)\big\rangle = 1$. Also, because $k^*(x) \in H$, we have $k^*(x) = v + v_1$ for some $v_1 \in \{v\}^{\perp}$ and so $\big\langle v,\ k^*(x)\big\rangle = \langle v, v\rangle = \|v\|^2$. Now $\operatorname{sgn}(w) = \operatorname{sgn}(v)$, which means $\big\langle w,\ k^*(x)\big\rangle / \|w\| = \big\langle v,\ k^*(x)\big\rangle / \|v\|$, and substituting from above we obtain $1/\|w\| = \|v\|$. Because $\|v\| \geqslant \pi(x)$, we have $\|w\| \leqslant 1/\pi(x)$; and because w was arbitrary we conclude that $\|\partial_u \sigma(x, k^*(x))\| \leqslant 1/\pi(x)$. We summarize these results as

$$\frac{\nu(x)}{b(x)} = d(0, \partial_u \sigma(x, k^*(x))) \leqslant \|\partial_u \sigma(x, k^*(x))\| \leqslant \frac{1}{\pi(x)}. \tag{48}$$

It follows from (43), (46), and the chain rule [9, Thm. 2.3.9] that

$$\begin{aligned} \partial_u r(x, k^*(x)) &= \overline{\operatorname{co}}\{\eta\zeta : \eta \in \partial_s \gamma(x, \sigma(x, k^*(x))),\ \zeta \in \partial_u \sigma(x, k^*(x))\} \\ &= \overline{\operatorname{co}}\{\eta\zeta : \eta \in \big[a(x),\ b(x)\big],\ \zeta \in \partial_u \sigma(x, k^*(x))\}. \end{aligned} \tag{49}$$

We now use (48) and (49) to show that

$$N_E \cap \{u \in \mathcal{U} : \mu(x) \leqslant \|u\| \leqslant \nu(x)\} \subset \partial_u r(x, k^*(x)). \tag{50}$$

Let $w \in N_E$ be such that $\mu(x) \leqslant \|w\| \leqslant \nu(x)$. Because $k^*(x) \neq 0$, we have $\mu(x) > 0$, which means $w \neq 0$. It then follows from [30, Cor. 23.7.1] that there exists $\delta > 0$ such that

$\delta w \in \partial_u \sigma(x, k^*(x))$. From (48) we have $\delta\mu(x) \leqslant \delta\|w\| \leqslant 1/\pi(x)$, and it follows from (40) that $(1/\delta) \geqslant \mu(x)\pi(x) \geqslant a(x)$. Also from (48) we have $\nu(x)/b(x) \leqslant \delta\|w\| \leqslant \delta\nu(x)$, which means $(1/\delta) \leqslant b(x)$. Thus $(1/\delta) \in [a(x), b(x)]$, and it follows from (49) that $w = (1/\delta)\delta w \in \partial_u r(x, k^*(x))$. The choice for w was arbitrary, and thus (50) is true.

Now $k^*(x) \neq 0$, and it follows from Proposition 9.2 that $D_f V(x, k^*(x)) = -\alpha(x)$. It thus follows from (12) that $L_V(x) = \{u \in \mathcal{U} : D_f V(x, u) \leqslant D_f V(x, k^*(x))\}$ is the sublevel set of $D_f V$ at $k^*(x)$. Let N_L denote the normal cone to $L_V(x)$ at $k^*(x)$, let N_U denote the normal cone to $U(x)$ at $k^*(x)$, and let N_K denote the normal cone to $K_V(x)$ at $k^*(x)$. Now it follows from (11) that $0 \in \operatorname{int}[L_V(x) - U(x)]$, and so from (13) and [5, Table 4.3] we have $N_K = N_U + N_L$. Recall that $k^*(x)$ is the element of $K_V(x)$ of minimum norm; this together with the projection theorem implies $-k^*(x) \in N_K$. Also, it follows from A6 and (36) that $E \subset C(x) \subset U(x)$, and therefore $N_U \subset N_E$. It then follows that $-k^*(x) \in N_E + N_L$. Thus there exist $v_E \in N_E$ and $v_L \in N_L$ such that $-k^*(x) = v_E + v_L$. Our goal is to show that $-v_L \in N_E \setminus \{0\}$. Now $-k^*(x) \notin N_E$, which means $v_L \neq 0$. Suppose $v_E = 0$; then $-v_L = k^*(x) \in N_E$. Otherwise we use the linearity of the inner product to obtain $\langle v_E, \operatorname{sgn}(k^*(x))\rangle + \langle v_L, \operatorname{sgn}(k^*(x))\rangle = \langle -k^*(x), \operatorname{sgn}(k^*(x))\rangle = -\|k^*(x)\|$, and thus

$$\begin{aligned}\langle \operatorname{sgn}(-v_L), \operatorname{sgn}(k^*(x))\rangle &= \frac{\|v_E\|\langle \operatorname{sgn}(v_E), \operatorname{sgn}(k^*(x))\rangle + \|k^*(x)\|}{\|v_L\|} \\ &\geqslant \frac{\|v_E\|\langle \operatorname{sgn}(v_E), \operatorname{sgn}(k^*(x))\rangle + \|k^*(x)\|}{\|v_E\| + \|k^*(x)\|} \\ &\geqslant \langle \operatorname{sgn}(v_E), \operatorname{sgn}(k^*(x))\rangle.\end{aligned}$$

Thus the angle between $-v_L$ and $k^*(x)$ is smaller than the angle between v_E and $k^*(x)$. Now $v_E \in N_E$, and it follows from the convexity of N_E and the symmetry of N_E around $k^*(x)$ that $-v_L \in N_E$. We have thus shown that $-v_L \in N_E \setminus \{0\}$.

Now $v_L \in N_L \setminus \{0\}$, and it follows from [30, Cor. 23.7.1] that there exists $\delta > 0$ such that $\delta v_L \in D(x)$. From Proposition 9.3 we have $\mu(x) \leqslant \|\delta v_L\| \leqslant \nu(x)$; and because $-\delta v_L \in N_E$, we have from (50) that $-\delta v_L \in \partial_u r(x, k^*(x))$. Therefore $0 = -\delta v_L + \delta v_L \in \partial_u r(x, k^*(x)) + D(x)$, and the proof is complete. □

An immediate consequence of (44) and Proposition 9.4 are the equalities

$$0 = q(x) + r(x, k^*(x)) + D_f V(x, k^*(x)), \tag{51}$$

$$= q(x) + \min_{u \in \mathcal{U}} [r(x, u) + D_f V(x, u)], \tag{52}$$

$$= \min_{u \in \mathcal{U}} \max_{w \in W(x)} [q(x) + r(x, u) + L_f V(x, u, w)], \tag{53}$$

which hold for all $x \in \mathcal{X}$. Therefore, the rclf V satisfies the steady-state HJI equation associated with the cost functional J. We are now ready to prove Theorem 8.1.

Let $J_T(u_g, w_a, x; x_0)$ denote the cost J in (29) truncated at time T, again setting $J_T := \infty$ when the solution does not exist over $[0, T]$:

$$J_T(u_g, w_a, x; x_0) := \int_0^T [q(x) + r(x, u_g)]dt. \tag{54}$$

Fix $x_0 \in \mathcal{X}$ and $w_a \in \mathcal{D}(\Sigma)$, and let $x(t)$ be a solution of the closed-loop system (32) starting from x_0. Because k^* renders the solutions of (32) RGUAS, the solution $x(t)$ exists for all

$t \geqslant 0$ and furthermore $x(t) \to 0$ as $t \to \infty$. Thus we can integrate $L_f V$ along $x(t)$ and use (10) and (51) to obtain for all $T \geqslant 0$

$$\begin{aligned} V(x_0) &= V(x(T)) - \int_0^T L_f V(x, k^*(x), w_a)\, dt \\ &\geqslant V(x(T)) - \int_0^T D_f V(x, k^*(x))\, dt \\ &\geqslant V(x(T)) + \int_0^T [q(x) + r(x, k^*(x))] dt. \end{aligned} \tag{55}$$

Thus $V(x(T)) + J_T(k^*, w_a, x; x_0) \leqslant V(x_0)$ for all $T \geqslant 0$; and because $V(x(T)) \to 0$ as $T \to \infty$, we can take the limit to obtain $J(k^*, w_a, x; x_0) \leqslant V(x_0)$.

Next fix $x_0 \in \mathcal{X}$, $u_g \in \mathcal{C}(\Sigma)$, and $\Delta > 0$, and suppose condition DC is true. Then there exists $w_\Delta \in \mathcal{D}(\Sigma)$ and a solution $x_\Delta(t)$ of (8) starting from x_0 (with $w_a = w_\Delta$) such that either $J(u_g, w_\Delta, x_\Delta; x_0) = \infty$ or for every $T \geqslant 0$ we have

$$\int_0^T L_f V(x_\Delta, u_g, w_\Delta)\, dt \geqslant \int_0^T D_f V(x_\Delta, u_g)\, dt - \Delta. \tag{56}$$

If $J(u_g, w_\Delta, x_\Delta; x_0) = \infty$, then trivially we have $J(u_g, w_\Delta, x_\Delta; x_0) \geqslant V(x_0) - \Delta$. Otherwise $x_\Delta(t) \to 0$ as $t \to \infty$ and we can integrate $L_f V$ along the solution x_Δ as above and use (56) and (52) to obtain for all $T \geqslant 0$

$$\begin{aligned} V(x_0) &= V(x_\Delta(T)) - \int_0^T L_f V(x_\Delta, u_g, w_\Delta)\, dt \\ &\leqslant V(x_\Delta(T)) + \Delta - \int_0^T D_f V(x_\Delta, u_g)\, dt \\ &\leqslant V(x_\Delta(T)) + \Delta + \int_0^T [q(x_\Delta) + r(x_\Delta, u_g)] dt. \end{aligned} \tag{57}$$

Thus $V(x_\Delta(T)) + J_T(u_g, w_\Delta, x_\Delta; x_0) \geqslant V(x_0) - \Delta$ for all $T \geqslant 0$; and because $V(x_\Delta(T)) \to 0$ as $T \to \infty$, we obtain $J(u_g, w_\Delta, x_\Delta; x_0) \geqslant V(x_0) - \Delta$ in the limit. Because Δ was arbitrary, it follows from (30) that $\bar{J}(x_0) \geqslant V(x_0)$. Recall from above that k^* guarantees $J \leqslant V(x_0)$; it follows that $\bar{J}(x_0) = V(x_0)$ and that k^* is optimal from x_0. The initial condition x_0 was arbitrary, and the proof is complete.

10. Inverse optimal robust stabilization: finite horizon. We next extend the results of the previous two sections to the case where the scp is *not* satisfied. In this case, persistent disturbances may make convergence to the origin impossible, and so to obtain finite costs we introduce a *terminal time* for our game. Given a nonempty bounded *target set* $\Lambda \subset \mathcal{X}$ and a solution $x(t)$ of the differential equation (8), we define the terminal time T_Λ as

$$T_\Lambda := \inf\{T \geqslant 0 : x(t) \in \Lambda \text{ for all } t \geqslant T\}. \tag{58}$$

Thus T_Λ represents the first time the solution enters the target set Λ *without ever again leaving*. If the solution $x(t)$ does not exist for all $t \geqslant 0$ or is not eventually contained in Λ, then we set $T_\Lambda := \infty$. In addition to a terminal time for our game, we introduce a terminal cost $q_f : \mathcal{X} \to R_+$ according to the following specification.

S3. q_f satisfies $q_f \in \mathcal{A}_\kappa(\mathcal{X})$.

Using a triple (q, r, q_f) satisfying S1–S3, we will define a game $\mathcal{G}_\Lambda(q, r, q_f)$ for each nonempty bounded set $\Lambda \subset \mathcal{X}$ by considering a cost functional J_Λ parameterized by the initial condition $x_0 \in \mathcal{X}$. Given $x_0 \in \mathcal{X}$, $u_g \in \mathcal{C}(\Sigma)$, and $w_a \in \mathcal{D}(\Sigma)$ we define

$$J_\Lambda(u_g, w_a, x; x_0) := q_f(x(T_\Lambda)) + \int_0^{T_\Lambda} [q(x) + r(x, u_g)]dt, \tag{59}$$

where the integration is taken along the (possibly nonunique) solution $x(t)$ of the differential equation (8) starting from the initial condition x_0. If $T_\Lambda = \infty$, then we set $J_\Lambda(u_g, w_a, x; x_0) := \infty$. We define the upper value function $\bar{J}_\Lambda : \mathcal{X} \to R_+ \cup \{\infty\}$ of the game by the equation

$$\bar{J}_\Lambda(x_0) := \inf_{u_g} \sup_{w_a} \sup_{x} J_\Lambda(u_g, w_a, x; x_0). \tag{60}$$

The first supremum is taken over all solutions $x(t)$ of (8) starting from x_0 (this supremum is superfluous if solutions are unique), the second supremum is taken over all admissible disturbances $w_a \in \mathcal{D}(\Sigma)$, and the infimum is taken over all generic controls $u_g \in \mathcal{C}(\Sigma)$. When the infimum in (60) is achieved by some $u_g^* \in \mathcal{C}(\Sigma)$ for some $x_0 \in \mathcal{X}$ such that $\bar{J}_\Lambda(x_0) < \infty$, we say that u_g^* is *optimal from* x_0. We say a control law k^* for Σ is *optimal for* $\mathcal{G}_\Lambda(q, r, q_f)$ when k^* is optimal from every $x_0 \in \mathcal{X}$. Note that every optimal control law is robustly stabilizing in the sense that every closed-loop trajectory is eventually contained in Λ.

Let V be a strong rclf for Σ. We say a nonempty bounded set $\Lambda \subset \mathcal{X}$ is an *admissible target set for* V when there exists $c > c_V$ such that $\Omega_c(V) \subset \Lambda$. If $c_V = 0$, then any bounded set Λ containing a neighborhood of the origin is an admissible target set for V; this is consistent with practical stabilizability. For each admissible target set Λ^* for V, we define a nonempty set $K(V, \Lambda^*)$ of pointwise min-norm control laws as follows. We say a control law k for Σ belongs to $K(V, \Lambda^*)$ when it is pointwise min-norm for V with the associated constant $c > c_V$ satisfying $\Omega_c(V) \subset \Lambda^*$ and the associated negativity margin α being bounded away from zero outside some compact set. Theorem 10.1 states that for every control law $k^* \in K(V, \Lambda^*)$ there exists a triple (q, r, q_f) satisfying S1–S3 such that k^* is optimal for $\mathcal{G}_\Lambda(q, r, q_f)$ for every $\Lambda \supset \Lambda^*$, with V being the corresponding upper value function. Each game $\mathcal{G}_\Lambda(q, r, q_f)$ is meaningful for the same reasons the game of §8 was meaningful; the difference here is the terminal time T_Λ and the terminal cost q_f. Our choice for the terminal time in (58) is consistent with the stabilization objective: the game ends when we enter the target set Λ permanently. If we enter and then leave again, the game continues. We must allow for the possibility of the trajectory leaving Λ because there does not necessarily exist a control that renders Λ positively invariant. Our technical condition on the admissible disturbances now depends on Λ as follows:

DC$_\Lambda$. For every $x_0 \in \mathcal{X}$, every $u_g \in \mathcal{C}(\Sigma)$, and every $\Delta > 0$ there exists $w_\Delta \in \mathcal{D}(\Sigma)$ and a solution $x_\Delta(t)$ of (8) starting from x_0 (with $w_a = w_\Delta$) such that either $J_\Lambda(u_g, w_\Delta, x_\Delta; x_0) = \infty$ or

$$\int_0^{T_\Lambda} L_f V(x_\Delta, u_g, w_\Delta)\, dt \geqslant \int_0^{T_\Lambda} D_f V(x_\Delta, u_g)\, dt - \Delta. \tag{61}$$

As in the previous section, if DC$_\Lambda$ is not true, then we may achieve suboptimality rather than optimality. We now state the main result of this section.

THEOREM 10.1. *Let* Σ *satisfy assumptions* A1–A6, *let* V *be a strong* rclf *for* Σ, *let* Λ^* *be an admissible target set for* V, *and let* $k^* \in K(V, \Lambda^*)$. *Then there exists a triple* (q, r, q_f) *satisfying* S1–S3 *such that* $J_\Lambda(k^*, w_a, x; x_0) \leqslant V(x_0)$ *for every admissible target set* $\Lambda \supset \Lambda^*$,

every $x_0 \in \mathcal{X}$, *every* $w_a \in \mathcal{D}(\Sigma)$, *and every solution* $x(t)$ *of* (32) *starting from* x_0. *If furthermore* DC_Λ *is true, then* $\bar{J}_\Lambda(x_0) = V(x_0)$ *for all* $x_0 \in \mathcal{X}$, *which means* k^* *is optimal for* $\mathcal{G}_\Lambda(q, r, q_f)$.

The proof of this theorem is similar to the proof of Theorem 8.1. From Definition 7.2 and the definition of $K(V, \Lambda^*)$ above, there exist $c > c_V$ and a minimal selection m_V of V such that $\Omega_c(V) \subset \Lambda^*$ and $k^*(x) = m_V(x)$ for all $x \in \mathcal{X} \setminus \Omega_c(V)$, with the associated negativity margin α being bounded away from zero outside some compact set. We choose $q_f(x) := V(x)$ for all $x \in \mathcal{X}$; this choice satisfies S3. We define $C(x)$, $\sigma(x, u)$, and $D(x)$ as in (36)–(38). Proposition 9.1 remains true, but Propositions 9.2 and 9.3 become weaker as follows (we omit their proofs).

PROPOSITION 10.2. *If* $x \in \mathcal{X} \setminus \Omega_c(V)$ *and* $k^*(x) \neq 0$, *then* $D_f V(x, k^*(x)) = -\alpha(x)$ *and* $0 \notin D(x)$.

PROPOSITION 10.3. *There exist continuous functions* $\mu, \nu : \mathcal{X} \to R_+$ *such that* $\mu(x) \leqslant d(0, D(x))$ *for all* $x \in \mathcal{X} \setminus \Omega_c(V)$ *and* $\|D(x)\| < \nu(x)$ *for all* $x \in \mathcal{X}$, *with* μ *having the additional property that* $\mu(x) > 0$ *if and only if* $k^*(x) \neq 0$.

Because α is bounded away from zero outside a compact set, it follows from (11) that there exist $\alpha_0 \in \mathcal{A}(\mathcal{X})$ and $\widehat{\alpha} > 0$ such that $\alpha(x) - \alpha_0(x) \geqslant \widehat{\alpha}$ for all $x \in \mathcal{X} \setminus \Omega_c(V)$. We then define $a(x)$, $b(x)$, $\gamma(x, s)$, and $r(x, u)$ as in (40)–(43). It follows that r satisfies S2. Our next task is to construct the function q. We first show that there exist continuous functions $d_1, d_2 : \mathcal{X} \to R$ such that $d_1(x) > 0$ for all $x \in \mathcal{X}$ and

$$r(x, u) + D_f V(x, u) \geqslant d_1(x)\,\|u\| + d_2(x) \tag{62}$$

for all $(x, u) \in \mathcal{X} \times \mathcal{U}$. First, it follows from (37) and (41) that $b(x)\,\sigma(x, u) \geqslant \nu(x)\,\|u\|$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$. It then follows from (42) and (43) that

$$\begin{aligned} r(x, u) &\geqslant b(x)\,\sigma(x, u) + \sigma(x, k^*(x))\big[a(x) - b(x)\big] \\ &\geqslant \nu(x)\,\|u\| + \sigma(x, k^*(x))\big[a(x) - b(x)\big] \end{aligned} \tag{63}$$

for all $(x, u) \in \mathcal{X} \times \mathcal{U}$. For each $x \in \mathcal{X}$ choose $w_x \in D(x)$; then it follows from (38) and Proposition 10.3 that

$$\begin{aligned} D_f V(x, u) &\geqslant D_f V(x, k^*(x)) + \langle w_x, u - k^*(x)\rangle \\ &\geqslant D_f V(x, k^*(x)) - \|w_x\| \cdot \|u\| - \|w_x\| \cdot \|k^*(x)\| \\ &\geqslant D_f V(x, k^*(x)) - \|D(x)\| \cdot \|u\| - \nu(x)\,\|k^*(x)\| \end{aligned} \tag{64}$$

for all $(x, u) \in \mathcal{X} \times \mathcal{U}$. We add the inequalities (63) and (64) to obtain

$$\begin{aligned} r(x, u) + D_f V(x, u) \geqslant\ & [\nu(x) - \|D(x)\|]\|u\| + \sigma(x, k^*(x))\big[a(x) - b(x)\big] \\ & + D_f V(x, k^*(x)) - \nu(x)\,\|k^*(x)\| \end{aligned} \tag{65}$$

for all $(x, u) \in \mathcal{X} \times \mathcal{U}$. It follows from Proposition 10.3, [18, Prob. 5X], and the lower semicontinuity of the mapping $x \mapsto \nu(x) - \|D(x)\|$ that there exists a continuous function $d_1 : \mathcal{X} \to R$ such that $0 < d_1(x) \leqslant \nu(x) - \|D(x)\|$ for all $x \in \mathcal{X}$. Thus if we define $d_2(x) := \sigma(x, k^*(x))\big[a(x) - b(x)\big] + D_f V(x, k^*(x)) - \nu(x)\,\|k^*(x)\|$ for all $x \in \mathcal{X}$, then d_2 is continuous and (65) implies (62).

We next define a function $\omega : \mathcal{X} \to R_+$ for all $x \in \mathcal{X}$ as

$$\omega(x) := \frac{r(x, k^*(x)) + D_f V(x, k^*(x)) - d_2(x)}{d_1(x)}. \tag{66}$$

Thus ω is continuous and it follows from (62) that $\|k^*(x)\| \leqslant \omega(x)$ for all $x \in \mathcal{X}$. Also, from (62) we see that for all $(x, u) \in \mathcal{X} \times \mathcal{U}$ such that $\|u\| > \omega(x)$ we have $r(x, u) + D_f V(x, u) \geqslant d_1(x)\|u\| + d_2(x) > r(x, k^*(x)) + D_f V(x, k^*(x))$. Therefore

$$\begin{aligned} \inf_{u \in \mathcal{U}} [r(x, u) + D_f V(x, u)] &= \inf_{u \in \omega(x)B} [r(x, u) + D_f V(x, u)] \\ &= \min_{u \in \omega(x)B} [r(x, u) + D_f V(x, u)] \end{aligned} \tag{67}$$

for all $x \in \mathcal{X}$, where the second line follows from the continuity of $r + D_f V$ and the compactness of $\omega(x)B$. Now [5, Thm. 1.4.16] implies that the right-hand side of (67) is continuous on $\mathcal{X}$, and so the mapping

$$x \mapsto \min_{u \in \mathcal{U}} [r(x, u) + D_f V(x, u)] \tag{68}$$

is well defined and continuous on $\mathcal{X}$. Let $\alpha_1 \in \mathcal{A}_\kappa(\mathcal{X})$ be such that $\alpha_1(x) \leqslant \widehat{\alpha}$ for all $x \in \mathcal{X}$. We define q for all $x \in \mathcal{X}$ as

$$q(x) := \max \left\{ \alpha_1(x), \ -\min_{u \in \mathcal{U}} [r(x, u) + D_f V(x, u)] \right\}. \tag{69}$$

Thus q is continuous with $q(0) = 0$ and $q \geqslant \alpha_1$, and it follows that $q \in \mathcal{A}_\kappa(\mathcal{X})$ as required in S1. With this choice for q, it is true for all $x \in \mathcal{X}$ that

$$q(x) + \min_{u \in \mathcal{U}} [r(x, u) + D_f V(x, u)] \geqslant 0. \tag{70}$$

One can use Propositions 10.2 and 10.3 in the proof of Proposition 9.4 to show that (45) is true for $x \in \mathcal{X} \setminus \Omega_c(V)$, and so we have

$$q(x) = \max\{\alpha_1(x), \ -r(x, k^*(x)) - D_f V(x, k^*(x))\} \tag{71}$$

for all such x. Now $-r(x, k^*(x)) - D_f V(x, k^*(x)) \geqslant \alpha(x) - \alpha_0(x) \geqslant \widehat{\alpha} \geqslant \alpha_1(x)$ for all such x, which means

$$q(x) = -r(x, k^*(x)) - D_f V(x, k^*(x)) \tag{72}$$

for all such x. Therefore, it is true for all $x \in \mathcal{X} \setminus \Omega_c(V)$ that

$$0 = q(x) + r(x, k^*(x)) + D_f V(x, k^*(x)) \tag{73}$$

$$= q(x) + \min_{u \in \mathcal{U}} [r(x, u) + D_f V(x, u)]. \tag{74}$$

Thus the HJI *inequality* (70) is true for all $x \in \mathcal{X}$, whereas the HJI *equality* (74) is true whenever $x \in \mathcal{X} \setminus \Omega_c(V)$. We are now ready to prove Theorem 10.1.

Proof of Theorem 10.1. Let $\Lambda \supset \Lambda^*$ be an admissible target set for V, let $x_0 \in \mathcal{X}$, let $w_a \in \mathcal{D}(\Sigma)$, and let $x(t)$ be a solution of the closed-loop system (32) starting from x_0. Because k^* renders the solutions of (32) RGUAS-$\Omega_c(V)$ and $\Omega_c(V) \subset \Lambda$, it follows from (58) that $T_\Lambda < \infty$. Thus we can integrate $L_f V$ along $x(t)$ to obtain

$$0 = V(x_0) - V(x(T_\Lambda)) + \int_0^{T_\Lambda} L_f V(x, k^*(x), w_a)\, dt. \tag{75}$$

Because $q_f = V$, we can add this zero quantity (75) to (59) and use (10) to obtain

$$J_\Lambda(k^*, w_a, x; x_0) = V(x_0) + \int_0^{T_\Lambda} [q(x) + r(x, k^*(x)) + L_f V(x, k^*(x), w_a)]dt$$

$$\leqslant V(x_0) + \int_0^{T_\Lambda} [q(x) + r(x, k^*(x)) + D_f V(x, k^*(x))]dt. \tag{76}$$

Now k^* renders the set $\Omega_c(V)$ robustly positively invariant, and so from (58) we have $x(t) \in \mathcal{X} \setminus \Omega_c(V)$ for all $t \in [0, T_\Lambda)$. It then follows from (73) that the integrand in (76) is zero for all $t \in [0, T_\Lambda)$, and thus we have $J_\Lambda(k^*, w_a, x; x_0) \leqslant V(x_0)$.

Next fix $x_0 \in \mathcal{X}$, $u_g \in \mathcal{C}(\Sigma)$, and $\Delta > 0$, and suppose condition DC_Λ is true. Then there exists $w_\Delta \in \mathcal{D}(\Sigma)$ and a solution $x_\Delta(t)$ of (8) starting from x_0 (with $w_a = w_\Delta$) such that either $J_\Lambda(u_g, w_\Delta, x_\Delta; x_0) = \infty$ or

$$\int_0^{T_\Lambda} L_f V(x_\Delta, u_g, w_\Delta)\, dt \geqslant \int_0^{T_\Lambda} D_f V(x_\Delta, u_g)\, dt - \Delta. \tag{77}$$

If $J_\Lambda(u_g, w_\Delta, x_\Delta; x_0) = \infty$, then trivially we have $J_\Lambda(u_g, w_\Delta, x_\Delta; x_0) \geqslant V(x_0) - \Delta$. Otherwise $T_\Lambda < \infty$ and we can integrate $L_f V$ along the solution x_Δ as above and use (77) and (70) to obtain

$$J_\Lambda(u_g, w_\Delta, x_\Delta; x_0) = V(x_0) + \int_0^{T_\Lambda} [q(x_\Delta) + r(x_\Delta, u_g) + L_f V(x_\Delta, u_g, w_\Delta)]dt$$

$$\geqslant V(x_0) - \Delta + \int_0^{T_\Lambda} [q(x_\Delta) + r(x_\Delta, u_g) + D_f V(x_\Delta, u_g)]dt$$

$$\geqslant V(x_0) - \Delta. \tag{78}$$

Because Δ was arbitrary, it follows from (60) that $\bar{J}_\Lambda(x_0) \geqslant V(x_0)$. Recall from above that k^* guarantees $J_\Lambda \leqslant V(x_0)$; it follows that $\bar{J}_\Lambda(x_0) = V(x_0)$ and that k^* is optimal from x_0. The initial condition x_0 was arbitrary, and the proof is complete. □

11. Conclusion. We have introduced the robust control Lyapunov function and shown that its existence is equivalent to robust stabilizability. Also, we have solved an inverse optimal robust stabilization problem by showing that every rclf is an upper value function for a meaningful game and that every pointwise min-norm control law is optimal for such a game. Our formulas (18) and (23) can be used to generate control laws that have the desirable properties of optimality but do not require the solution of an HJI equation.

Our results motivate further research in the development of methods for the construction rclf 's for uncertain nonlinear systems. Recent breakthroughs in this area include recursive backstepping techniques [25, 32, 29]. In this paper we provided a method for choosing a reasonable control law given an rclf, but the overall system performance will ultimately be determined by the choice of the rclf itself. This point is illustrated in [12], where it is shown that an improper choice of the rclf can lead to undesirable behavior no matter how the associated control law is chosen. Methods for improving the choices of rclf 's, when combined with the results in this paper on choosing control laws, will constitute a promising strategy for the design of controllers for uncertain nonlinear systems.

12. Appendix. We include here two simple lemmas whose proofs we could not find elsewhere.

LEMMA 12.1. *Let Z be a metric space, let Y be a finite-dimensional Hilbert space, and let $h : Z \times Y \to R$ be continuous. If the mapping $y \mapsto h(z, y)$ is convex for every $z \in Z$,*

then the partial subdifferential $\partial_y h : Z \times Y \rightsquigarrow Y$ is usc on $Z \times Y$ and has nonempty, convex, compact values.

Proof. It follows from [30, Thm. 23.4] that $\partial_y h$ has nonempty, convex, compact values. To prove upper semicontinuity, fix $(z_0, y_0) \in Z \times Y$ and let $\{(z_i, y_i)\} \in Z \times Y$ converge to (z_0, y_0). Let $\{w_i\} \in Y$ be any sequence such that $w_i \in \partial_y h(z_i, y_i)$ for all $i \geqslant 1$. It follows from [19, Thm. II.2.2] that we need only show that $\{w_i\}$ has a subsequence converging to some $w_0 \in \partial_y h(z_0, y_0)$. First we show that $\{w_i\}$ is bounded. From the definition of $\partial_y h$ we have $\langle w_i, v - y_i \rangle \leqslant h(z_i, v) - h(z_i, y_i)$ for all $v \in Y$ and all $i \geqslant 1$. It follows from the continuity of h and the compactness of the unit sphere in Y that there exists $M \in R$ such that $\langle w_i, e \rangle \leqslant M$ for all unit vectors $e \in Y$ and all $i \geqslant 1$, and we conclude that $\{w_i\}$ is bounded. Let $\{w_{i_j}\}$ be a convergent subsequence of $\{w_i\}$ with limit $w_0 \in Y$. Fix $v \in Y$; then for all $j \geqslant 1$ we have $\langle w_{i_j}, v - y_{i_j} \rangle \leqslant h(z_{i_j}, v) - h(z_{i_j}, y_{i_j})$. It follows from the continuity of the inner product and h that $\langle w_0, v - y_0 \rangle \leqslant h(z_0, v) - h(z_0, y_0)$. This holds for all $v \in Y$, and so $w_0 \in \partial_y h(z_0, y_0)$ as desired. □

LEMMA 12.2. *Let Y be a finite-dimensional Hilbert space, and let $h : Y \to R$ be a sublinear functional. For each $y \in Y$, let $E_y = \{\xi \in Y : h(\xi) \leqslant h(y)\}$ denote the sublevel set of h at y and $N_E(y)$ denote the normal cone to E_y at y. Then $\partial h(y) \neq \emptyset$ for all $y \in Y$, and for every $y \in Y$ such that $h(y) \neq 0$ we have*

$$\partial h(y) = N_E(y) \cap \left[p + \{y\}^{\perp} \right], \tag{79}$$

where p is any member of $\partial h(y)$.

Proof. It follows from [30, Thm. 23.4] that $\partial h(y)$ is nonempty for all $y \in Y$. Let $y \in Y$ be such that $h(y) \neq 0$, and let $p \in \partial h(y)$. We first show that $\langle w, y \rangle = h(y)$ for all $w \in \partial h(y)$. Indeed, $w \in \partial h(y)$ implies $h(\xi) \geqslant h(y) + \langle w, \xi - y \rangle$ for all $\xi \in Y$. Taking $\xi = 0$ we obtain $0 \geqslant h(y) + \langle w, -y \rangle$, which means $\langle w, y \rangle \geqslant h(y)$, and taking $\xi = 2y$ we obtain $h(2y) = 2h(y) \geqslant h(y) + \langle w, y \rangle$, which means $\langle w, y \rangle \leqslant h(y)$. It then follows that for any $w \in \partial h(y)$ we have $\langle w - p, y \rangle = h(y) - h(y) = 0$, which means $w \in [p + \{y\}^{\perp}]$. Now because h is sublinear and $h(y) \neq 0$, we know that y does not minimize h, and it follows from [30, Cor. 23.7.1] that $\partial h(y) \subset N_E(y)$. We have thus shown that $\partial h(y) \subset N_E(y) \cap [p + \{y\}^{\perp}]$.

Next suppose $w \in N_E(y) \cap [p + \{y\}^{\perp}]$. Because $\langle p, y \rangle = h(y) \neq 0$, we have $0 \notin [p + \{y\}^{\perp}]$, which means $w \neq 0$. It then follows from [30, Cor. 23.7.1] that $\lambda w \in \partial h(y)$ for some $\lambda > 0$. From above we have $\lambda w \in [p + \{y\}^{\perp}]$, and it follows that $(\lambda w - w) \in \{y\}^{\perp}$ and so $(\lambda - 1)\langle w, y \rangle = 0$. Now $\langle w, y \rangle = \langle \lambda w, y \rangle / \lambda = h(y)/\lambda \neq 0$, and it follows that $\lambda = 1$. Therefore $w = \lambda w \in \partial h(y)$, and we have shown that $N_E(y) \cap [p + \{y\}^{\perp}] \subset \partial h(y)$. □

Acknowledgments. This work was initiated in part through discussions with Tamer Başar, and we thank him for his useful comments. Also, we thank Eduardo Sontag for providing us with an early copy of the paper [24], the results of which enabled us to prove the necessary conditions in §§5 and 6.

REFERENCES

[1] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
[2] Z. ARTSTEIN, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.
[3] J.-P. AUBIN, *Contingent Isaacs equations of a differential game*, in Differential Games and Applications, T. Başar and P. Bernhard, eds., Springer-Verlag, Berlin, 1989, pp. 51–61.
[4] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.
[5] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.
[6] B. R. BARMISH, M. J. CORLESS, AND G. LEITMANN, *A new class of stabilizing controllers for uncertain dynamical systems*, SIAM J. Control Optim., 21 (1983), pp. 246–255.
[7] Y. H. CHEN, *A new matching condition for robust control design*, in Proc. 1993 American Control Conference, San Francisco, CA, June 1993, pp. 122–126.
[8] Y. H. CHEN AND G. LEITMANN, *Robustness of uncertain systems in the absence of matching assumptions*, Int. J. Control, 45 (1987), pp. 1527–1542.

[9] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Society for Industrial and Applied Mathematics, Philadelphia, 1990.
[10] M. J. CORLESS, *Robust stability analysis and controller design with quadratic Lyapunov functions*, in Variable Structure and Lyapunov Control, A. Zinober, ed., Springer-Verlag, Berlin, 1993.
[11] M. J. CORLESS AND G. LEITMANN, *Continuous state feedback guaranteeing uniform ultimate boundedness for uncertain dynamic systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 1139–1144.
[12] R. A. FREEMAN AND P. V. KOKOTOVIĆ, *Design of 'softer' robust nonlinear control laws*, Automatica, 29 (1993), pp. 1425–1437.
[13] ———, *Optimal nonlinear controllers for feedback linearizable systems*, in Proc. Workshop on Robust Control via Variable Structure & Lyapunov Techniques, Benevento, Italy, September 1994, pp. 286–293.
[14] S. T. GLAD, *Robustness of nonlinear state feedback—a survey*, Automatica, 23 (1987), pp. 425–435.
[15] S. GUTMAN, *Uncertain dynamical systems—Lyapunov min-max approach*, IEEE Trans. Automat. Control, 24 (1979), pp. 437–443.
[16] D. H. JACOBSON, *Extensions of Linear-Quadratic Contol, Optimization and Matrix Theory*, Academic Press, London, 1977.
[17] R. E. KALMAN, *When is a linear control system optimal?*, Trans. ASME Ser. D J. Basic Engrg., 86 (1964), pp. 1–10.
[18] J. L. KELLEY, *General Topology*, D. Van Nostrand Company, Princeton, NJ, 1955.
[19] M. KISIELEWICZ, *Differential Inclusions and Optimal Control*, PWN—Polish Scientific Publishers, Warszawa, Poland, 1991.
[20] G. LEITMANN, *Guaranteed ultimate boundedness for a class of uncertain linear dynamical systems*, IEEE Trans. Automat. Control, 23 (1978), pp. 1109–1110.
[21] Y. LIN, *Lyapunov Function Techniques for Stabilization*, Ph.D. thesis, Rutgers, The State University of New Jersey, 1992.
[22] Y. LIN AND E. D. SONTAG, *Further universal formulas for Lyapunov approaches to nonlinear stabilization*, in Proc. Conference on Information Science and Systems, Johns Hopkins University, March 1991, pp. 541–546.
[23] ———, *A universal formula for stabilization with bounded controls*, Systems Control Lett., 16 (1991), pp. 393–397.
[24] Y. LIN, E. D. SONTAG, AND Y. WANG, *Recent results on Lyapunov-theoretic techniques for nonlinear stability*, in Proc. 1994 American Control Conference, Baltimore, MD, June 1994, pp. 1771–1775.
[25] R. MARINO AND P. TOMEI, *Robust stabilization of feedback linearizable time-varying uncertain nonlinear systems*, Automatica, 29 (1993), pp. 181–189.
[26] F. MAZENC AND L. PRALY, *Adding an integration and global asymptotic stabilization of feedforward systems*, IEEE Trans. Automat. Control, to appear.
[27] L. PRALY AND Z. P. JIANG, *Stabilization by output feedback for systems with ISS inverse dynamics*, Systems Control Lett., 21 (1993), pp. 19–33.
[28] Z. QU, *Global stabilization of nonlinear systems with a class of unmatched uncertainties*, Syst. Control Lett., 18 (1992), pp. 301–307.
[29] ———, *Robust control of nonlinear uncertain systems under generalized matching conditions*, Automatica, 29 (1993), pp. 985–998.
[30] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[31] M. A. ROTEA AND P. P. KHARGONEKAR, *Stabilization of uncertain systems with norm bounded uncertainty: A control Lyapunov function approach*, SIAM J. Control Optim., 27 (1989), pp. 1462–1476.
[32] J. J. E. SLOTINE AND K. HEDRICK, *Robust input-output feedback linearization*, Int. J. Control, 57 (1993), pp. 1133–1139.
[33] E. D. SONTAG, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.
[34] ———, *A 'universal' construction of Artstein's theorem on nonlinear stabilization*, Systems Control Lett., 13 (1989), pp. 117–123.
[35] ———, *Feedback stabilization of nonlinear systems*, in Robust Control of Linear Systems and Nonlinear Control, M. A. Kaashoek, J. H. van Schuppen, and A. C. M. Ran, eds., Birkhäuser, Boston, 1990, pp. 61–81.
[36] ———, *Further facts about input to state stabilization*, IEEE Trans. Automat. Control, 35 (1990), pp. 473–476.
[37] A. R. TEEL, *A nonlinear small gain theorem for the analysis of control systems with saturation*, IEEE Trans. Automat. Control, to appear.
[38] A. R. TEEL AND L. PRALY, *Tools for semiglobal stabilization by partial state and output feedback*, SIAM J. Control Optim., 33 (1995), pp. 1443–1488.
[39] J. TSINIAS, *Sufficient Lyapunov-like conditions for stabilization*, Math. Control Signals Systems, 2 (1989), pp. 343–357.
[40] ———, *Sontag's 'input-to-state stability condition' and global stabilization using state detection*, Systems Control Lett., 20 (1993), pp. 219–226.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1392–1404, July 1996

015

THE STRUCTURED SINGULAR VALUE FOR LINEAR INPUT/OUTPUT OPERATORS*

HARI BERCOVICI†, CIPRIAN FOIAS†, AND ALLEN TANNENBAUM‡

Abstract. In this paper, we employ our lifting method to study the structured singular value applied to input/output operators of control systems. We moreover give a new criterion which guarantees that the structured singular value equals its upper bound defined by D-scalings.

Key words. structured singular value, time-varying perturbations, lifting, robust control

AMS subject classifications. 93B35, 93C05

1. Introduction.

Let A be a linear operator on a Hilbert space $\mathcal{E}$, and let Δ be an algebra of operators on $\mathcal{E}$. The structured singular value of A (relative to Δ) is the number

$$\mu_\Delta(A) = 1/\inf\{\|X\| : X \in \Delta,\ -1 \in \sigma(AX)\}\,.$$

This quantity was introduced by Doyle and Safonov [6, 12] under a more restrictive context, and it has proved to be a powerful tool in robust system analysis and design. In system analysis, the structured singular value gives a measure of robust stability with respect to certain perturbation measures. Unfortunately, $\mu_\Delta(A)$ is very difficult to calculate, and in practice an upper bound for it is used. This upper bound is defined by

$$\widehat{\mu}_\Delta(A) := \inf\{\|XAX^{-1}\| : X \in \Delta',\ X \text{ invertible}\},$$

where Δ' is the commutant of the algebra Δ.

In [1, 5], we formulated a lifting technique for the study of the structured singular value. The basic idea is that $\widehat{\mu}_\Delta(A)$ can be shown to be equal to the structured singular value of an operator on a bigger Hilbert space. (In [1] this was done for finite-dimensional Hilbert spaces, and then in [5] this was extended to the infinite-dimensional case.) The problem with these results is that the size of the ampliation necessary to get $\widehat{\mu}_\Delta(A)$ equal to a structured singular value was equal to the dimension of the underlying Hilbert space. Hence in the infinite-dimensional case we needed an infinite ampliation. In this work, we will show that in fact one can always get by with a finite lifting. (Note that in this paper we will be using the terms "ampliation" and "lifting" interchangeably.) For the block diagonal algebras of interest in robust control, the ampliation only depends on the number of blocks of the given perturbation structure. (See Theorem 4.1 below.) We moreover give a new result when $\widehat{\mu}_\Delta(A) = \mu_\Delta(A)$, that is, when no lifting is necessary and so $\widehat{\mu}_\Delta(A)$ gives a nonconservative measure of robustness. (See Theorem 5.3.) This is then used to derive an elegant result of Shamma [13, 14] on Toeplitz operators. See also [7, 9, 10] for related work in this area.

We now briefly sketch the contents of this paper. In §2, we give some background results which will be needed in the proof of Theorem 4.1. In §3, we derive a number of useful facts about the relative numerical range. Then in §4, we state and prove our new version of the lifting theorem relating the structured singular value and its upper bound. In §5, we give new conditions when $\mu = \hat{\mu}$. These are applied in §6 to give a new proof of the aforementioned result of Shamma. Finally, in §7, we give a system-theoretic interpretation of our lifting methodology.

*Received by the editors June 1, 1994; accepted for publication (in revised form) April 24, 1995. This research was supported in part by grants from the Research Fund of Indiana University, National Science Foundation grants DMS-8811084 and ECS-9122106, Air Force Office of Scientific Research grant AFOSR AF/F49620-94-1-00S8DEF, and Army Research Office grants DAAH04-94-G-0054 and DAAH04-93-G-0332.

†Department of Mathematics, Indiana University, Bloomington, IN 47405.

‡Department of Electrical Engineering, University of Minnesota, Minneapolis, MN 55455.

2. Preliminary results. Denote by $\mathcal{L}(\mathcal{E})$ the algebra of all bounded linear operators on the (complex, separable) Hilbert space $\mathcal{E}$. Fix an operator $A \in \mathcal{L}(\mathcal{E})$ and a subalgebra $\Delta \subset \mathcal{L}(\mathcal{E})$. The numbers $\mu_\Delta(A)$ and $\widehat{\mu}_\Delta(A)$ have already been defined in the Introduction. Observe that $\Delta \subset \Delta''$ and $\Delta''' = (\Delta'')' = \Delta'$ so that we have the inequalities

$$\mu_\Delta(A) \le \mu_{\Delta''}(A), \quad \widehat{\mu}_\Delta(A) = \widehat{\mu}_{\Delta''}(A) .$$

Observe that the algebras Δ considered in [6] consisted of block diagonal matrices, so our approach is more general in this respect. In the following proposition we summarize some of the elementary properties of μ_Δ; see Doyle [6] or [1] for proofs. We will denote by $\|T\|_{\mathrm{sp}}$ the spectral radius of the operator T.

LEMMA 2.1. (i) $\mu_\Delta(A) = \sup\{\|AX\|_{\mathrm{sp}} : X \in \Delta, \ \|X\| \le 1\}$;

(ii) μ_Δ *is upper semicontinuous*;

(iii) *if* $\mathcal{E}$ *is finite dimensional, then* μ_Δ *is continuous*;

(iv) $\mu_\Delta(A) \le \widehat{\mu}_\Delta(A)$.

In our study we will need further singular values which we now define. For $n \in \{1, 2, \ldots, \infty\}$ we denote by $\mathcal{E}^{(n)}$ the orthogonal sum of n copies of $\mathcal{E}$ and by $T^{(n)}$ the orthogonal of n copies of $T \in \mathcal{L}(\mathcal{E})$. Operators on $\mathcal{E}^{(n)}$ can be represented as $n \times n$ matrices of operators in $\mathcal{L}(\mathcal{E})$, and $T^{(n)}$ is represented by a diagonal matrix, with diagonal entries equal to T.

Denote by Δ_n the algebra of all operators on $\mathcal{E}^{(n)}$ whose matrix entries belong to Δ, and observe that $(\Delta_n)'' = (\Delta'')_n$ and $(\Delta_n)' = (\Delta')^{(n)} = \{T^{(n)} : T \in \Delta'\}$. Therefore we will denote these algebras by Δ_n'' and Δ_n', respectively.

LEMMA 2.2. *For every finite number n we have*

$$\mu_\Delta(A) \le \mu_{\Delta_n}(A^{(n)}) \le \mu_{\Delta_{n+1}}(A^{(n+1)}) \le \mu_{\Delta_\infty}(A^{(\infty)}) \le \widehat{\mu}_\Delta(A)$$

and

$$\mu_{\Delta''}(A) \le \mu_{\Delta_n''}(A^{(n)}) \le \mu_{\Delta_{n+1}''}(A^{(n+1)}) \le \mu_{\Delta_\infty''}(A^{(\infty)}) \le \widehat{\mu}_\Delta(A) .$$

Proof. It is clearly sufficient to prove the first sequence of inequalities. Observe that for every $X \in \Delta_n$ and for $m > n$ we can define an operator $Y \in \Delta_m$ by $Y = X \oplus 0$. Clearly $\sigma(A^{(n)}X) = \sigma(A^{(m)}Y) \cup \{0\}$ and hence $-1 \in \sigma(A^{(n)}X)$ implies $-1 \in \sigma(A^{(m)}Y)$. Since $\mu_\Delta(A) = \mu_{\Delta_1}(A^{(1)})$, this proves the first three inequalities. The last one follows because $\mu_{\Delta_\infty}(A^{(\infty)}) \le \widehat{\mu}_{\Delta_\infty}(A^{(\infty)}) = \widehat{\mu}_\Delta(A)$. □

We will now state (without proof) several results from [1–5] which we will need.

LEMMA 2.3 (see [5]). *Let A be a finite-dimensional* $\mathbf{C}^*$*-algebra. Then A has only finitely many equivalence classes of cyclic representations.*

LEMMA 2.4 (see [5]). *Let the sequence* Y_j *of operators on* $\mathcal{H}$ *and the sequence* $h_j \in \mathcal{H}$ *satisfy*

(i) $\sup_j \operatorname{rank} Y_j < \infty$, $\sup_j \|Y_j\| < \infty$;

(ii) $\lim_{j\to\infty} \|(Y_j - I)h_j\| = 0$;

(iii) $\lim_{j\to\infty} \|h_j\| = 1$.

Then $\liminf_{n\to\infty} \|Y_j\|_{\mathrm{sp}} \ge 1$.

LEMMA 2.5 (see [4]). *Let* $\mathcal{H}$ *be a Hilbert space,* $T \in \mathcal{L}(\mathcal{H})$*, and* $D_j \in \mathcal{L}(\mathcal{H})$ *be invertible so that*

$$T_0 = \lim_{j\to\infty} D_j T D_j^{-1}.$$

If the set $\{D_j, D_j^{-1} : j = 1, 2, \ldots\}$ *is contained in a finite-dimensional subspace, then* $\|T_0\|_{\mathrm{sp}} = \|T\|_{\mathrm{sp}}$.

3. Relative numerical range. We will also need some results in what follows about the relative numerical range. Let $\mathcal{H}$ be a complex separable Hilbert space, and let $\mathcal{L}(\mathcal{H})$ denote the set of bounded linear operators on $\mathcal{H}$. Let $T_1, \dots, T_m, Q \in \mathcal{L}(\mathcal{H})$. Then we define the following *relative numerical ranges*:

$$W_Q(T_1, \dots, T_m) := \Big\{\lambda \in \mathbf{C}^n,\ \lambda = \lim_{n\to\infty} (\langle T_j h_n, h_n\rangle)_{j=1}^m :$$
$$h_n \in \mathcal{H}, \|h_n\| = 1, \lim_{n\to\infty} \|Qh_n\| = 0\Big\}$$

and

$$W_Q^0(T_1, \dots, T_m) := \Big\{\lambda \in \mathbf{C}^n,\ \lambda = \lim_{n\to\infty} (\langle T_j h_n, h_n\rangle)_{j=1}^m :$$
$$h_n \in \mathcal{H}, \|h_n\| = 1, \lim_{n\to\infty} \|Qh_n\| = 0, h_n \to 0 \text{ weakly}\Big\}.$$

LEMMA 3.1. *$W_Q^0(T_1, \dots, T_m)$ is a compact convex subset of $\mathbf{C}^m$.*

Proof. The compactness is immediate since $W_Q^0(T_1, \dots, T_m)$ is a closed bounded subset of $\mathbf{C}^m$. As for the convexity, let $\lambda = (\lambda_1, \dots, \lambda_m)$, $\mu = (\mu_1, \dots, \mu_m) \in W_Q^0(T_1, \dots, T_m)$, and let the sequences of unit vectors

$$\{h_n\}_{n=1}^\infty,\ \{k_n\}_{n=1}^\infty \subset \mathcal{H}$$

satisfy

$$\lambda_j = \lim_{n\to\infty} \langle T_j h_n, h_n\rangle,$$
$$\mu_j = \lim_{n\to\infty} \langle T_j k_n, k_n\rangle, \qquad j = 1, \dots, m,$$
$$\lim_{n\to\infty} \|Qh_n\| = 0 = \lim_{n\to\infty} \|Qk_n\|,\ h_n \to 0,\ k_n \to 0 \text{ weakly}.$$

Next for n fixed choose $N_n \geq n$ such that

$$|\langle h_n, k_{N_n}\rangle| \leq \frac{1}{n},\ |\langle T_j h_n, k_{N_n}\rangle| + |\langle T_j^* h_n, k_{N_n}\rangle| \leq \frac{1}{n}, \qquad j = 1, 2, \dots, m.$$

Then for any $\theta \in [0, 1]$,

$$g_n := \sqrt{\theta} h_n + \sqrt{1-\theta} k_{N_n}, \qquad n \geq 2,$$

satisfies the following conditions:

$$\|g_n\|^2 = 1 + 2\sqrt{\theta(1-\theta)}\Re\langle h_n, k_{N_n}\rangle \to 1, \qquad \|g_n\|^2 \geq \frac{1}{2},$$
$$g_n \to 0 \text{ weakly}, \qquad \|Qg_n\| \to 0,$$
$$|\langle T_j g_n, g_n\rangle - \theta\lambda_j - (1-\theta)\mu_j| \leq \theta|\langle T_j h_n, h_n\rangle - \lambda_j| + (1-\theta)|\langle T_j k_{N_n}, k_{N_n}\rangle - \mu_j|$$
$$+\sqrt{\theta(1-\theta)}\frac{1}{n} \to 0 \text{ as } n \to \infty.$$

Thus replacing the g_n by $g_n/\|g_n\|$ we immediately conclude that

$$\theta\lambda + (1-\theta)\mu \in W_Q^0(T_1, \dots, T_m),$$

as required. □

LEMMA 3.2. *$W_Q(T_1, \ldots, T_m)$ is the union of all segments*

$$\{\theta\lambda + (1-\theta)\mu : 0 \le \theta \le 1\},$$

where $\lambda \in W_Q^0(T_1, \ldots, T_m)$ and $\mu = (\langle T_j h, h\rangle)_{j=1}^m$ for some $h \in \ker Q$, $\|h\| = 1$.

Proof. Let $\theta, \lambda = (\lambda_j)_{j=1}^m, \mu = (\mu_j)_{j=1}^m$ be as above and let the sequence $\{h_n\}_{n=1}^\infty \subset \mathcal{H}$ satisfy $\|h_n\| = 1$, $h_n \to 0$ weakly, $Qh_n \to 0$ strongly, and $\langle T_j h_n, h_n\rangle \to \lambda_j$ for $j = 1, \ldots, m$. Then as in the proof of Lemma 3.1, we obtain that

$$g_n = \sqrt{\theta} h_n + \sqrt{1-\theta} h, \qquad n = 1, 2, \ldots$$

satisfies the conditions

$$\|g_n\| \to 1, \qquad \|Qg_n\| \to 0,$$

and

$$\langle T_j g_n, g_n\rangle \to \theta\lambda_j + (1-\theta)\mu_j, \qquad j = 1, \ldots, m.$$

Therefore

$$\theta\lambda + (1-\theta)\mu \in W_Q(T_1, \ldots, T_m).$$

Conversely, if $\psi = (\psi_j)_{j=1}^m \in W_Q(T_1, \ldots, T_m)$, then

$$\psi_j = \lim_{n\to\infty} \langle T_j g_n, g_n\rangle, \qquad j = 1, \ldots, m,$$

for some sequence $\{g_n\}_{n=1}^\infty \subset \mathcal{H}$ such that $\|g_n\| = 1$, $\|Qg_n\| \to 0$. Without loss of generality we can assume that g_n converges weakly to some $h' \in \ker Q$.

If $\|h'\| = 1$, then g_n converges strongly to h' and $\psi_j = \langle T_j h', h'\rangle$, $j = 1, 2, \ldots, m$. Clearly then

$$\psi = \theta\lambda + (1-\theta)(\langle T_j h, h\rangle)_{j=1}^m$$

with $h = h', \theta = 0$, and $\lambda \in W_Q^0(T_1, \ldots, T_m)$. If $\|h'\| = 0$, then ψ belongs to $W_Q^0(T_1, \ldots, T_m)$ and hence $\psi = \theta\psi + (1-\theta)\mu$ with $\theta = 1$ and μ arbitrary. Finally we consider the case when $h' \neq 0$ and $\|h'\| \neq 1$. The vectors $h_n = (g_n - h')/\|g_n - h'\|$ converge weakly to zero, $\|Qh_n\| \to 0$, and $h = h'/\|h'\|$ is a unit vector in $\ker Q$. Clearly then

$$\psi_j = (1 - \|h\|^2) \lim_{n\to\infty} \langle T_j h_n, h_n\rangle + \|h\|^2 \langle T_j h, h\rangle, \qquad j = 1, \ldots, m,$$

and therefore

$$\psi \in \{\theta\lambda + (1-\theta)(\langle T_j h, h\rangle)_{j=1}^m : 0 \le \theta \le 1\},$$

where

$$\lambda := \lim_{n\to\infty} (\langle T_j h_n, h_n\rangle)_{j=1}^m \in W_Q^0(T_1, \ldots, T_m).$$

This concludes the proof. □

COROLLARY 3.3. *For all $T, Q \in \mathcal{L}(\mathcal{H})$, the set*

$$W_Q(T) = \left\{\lambda = \lim_{n\to\infty} \langle Th_n, h_n\rangle : h_n \in H\ ,\ \|h_n\| = 1\ ,\ \lim_{n\to\infty} \|Qh_n\| = 0\right\}$$

is a compact convex set.

Proof. First notice that by an application of the classical Toeplitz–Hausdorff theorem to $T_Q := PT|\ker Q$, where P denotes the orthogonal projection of $\mathcal{H}$ onto $\ker Q$, we see that the set

$$W(T_Q) = \{\langle Th, h\rangle : \|h\| = 1,\ h \in \ker Q\}$$

is compact and convex. Therefore the convex hull of

$$W_Q^0(T) \cup W(T_Q)$$

is the union of all segments

$$\{\theta\lambda + (1-\theta)\mu : 0 \leq \theta \leq 1\},$$

where λ and μ run over $W_Q^0(T)$ and $W(T_Q)$, respectively. But according to Lemma 3.2, this union is precisely $W_Q(T)$. □

Remark. Corollary 3.3 was proven in [3] using a completely different argument that was based on an approximation lemma, which is of independent interest.

Finally, for the proof of our lifting theorem (to be given in §4), we will need the following elementary fact.

LEMMA 3.4. *Let $\mathcal{Z}$ denote a finite-dimensional normed space and S be a set of linear functionals on $\mathcal{Z}$. Suppose that for every $z \in \mathcal{Z}$ there exists a sequence $\ell_n \in S$ such that $\lim_{n\to\infty} \ell_n(z) = 0$. Then there exists a sequence ℓ_n in the convex hull of S such that $\lim_{n\to\infty} \|\ell_n\| = 0$.*

Proof. Since $\mathcal{Z}$ is finite dimensional, S is contained in the dual $\mathcal{Z}'$ of $\mathcal{Z}$. We may also assume that S is a convex set. To prove the lemma we must show that the closure of S contains zero. If it did not then the Hahn–Banach theorem would imply the existence of a vector $z \in \mathcal{Z}$ and of a number $\varepsilon > 0$ such that $\Re\ell(z) > \varepsilon$ for all $\ell \in S$. This is contrary to the assumption of the lemma. □

4. Ampliations of perturbations. In this section, we will formulate and prove a new lifting result relating $\mu_\Delta(A)$ and $\widehat{\mu}_\Delta(A)$. For finite-dimensional $\mathcal{E}$, a lifting result of this type was first proven in [1]. The result was then generalized to the infinite-dimensional case in [5]. (For another proof of this type of lifting result in finite dimensions, see [7].) In these theorems, the lifting or ampliation of the operator A and perturbation structure Δ depends on the dimension of $\mathcal{E}$. Thus if $\mathcal{E}$ is infinite dimensional, we get an infinite lifting. In the new result proven below, we only have to lift up to the dimension of Δ', which in the cases of interest in the control applications of this theory only depends on the number of blocks of the given perturbation structure.

The notation will be that used in §2.

THEOREM 4.1. *Assume that Δ' is a $*$-algebra of finite dimension n. Then*

$$\widehat{\mu}_\Delta(A) = \mu_{\Delta_n''}(A^{(n)}),$$

for every $A \in \mathcal{L}(\mathcal{E})$.

Proof. The argument starts as in the proof of Theorem 3 in [1] and of Theorem 1 of [5]. Without loss of generality, we may assume that $\widehat{\mu}_\Delta(A) = 1$. We must show that $\mu_{\Delta_n''}(A^{(n)}) \geq 1$. Choose a sequence of invertible operators $X_j \in \Delta'$ such that $\|X_j A X_j^{-1}\| \to \widehat{\mu}_\Delta(A)$. Since $X_j A X_j^{-1}$ belongs to the finite-dimensional space generated by $\Delta' A \Delta'$, we may assume that the sequence $X_j A X_j^{-1}$ converges to some operator A_0 such that $\|A_0\| = 1$. Obviously $\|X A_0 X^{-1}\| \geq \|A_0\|$ for every invertible operator $X \in \Delta'$. In particular we have

$$\|(I - X)A_0(I + X + X^2 + \cdots)\| \geq 1,$$

for $X \in \Delta'$ with $\|X\| < 1$. Fix an operator $X \in \Delta'$ and a sequence $\varepsilon_j > 0$ converging to zero. There exist vectors $h_j \in \mathcal{E}$ with $\|h_j\| = 1$ such that

$$\|(I - \varepsilon_j X)A_0(I + \varepsilon_j X + \varepsilon_j^2 X^2 + \cdots)h_j\|^2 \geq 1 - \varepsilon_j^2 .$$

This can be rewritten as

$$\langle A_0^* A_0 h_j, h_j\rangle + 2\varepsilon_j \Re\langle A_0^*(A_0 X - X A_0)h_j, h_j\rangle + O(\varepsilon_j^2) \geq 1 - \varepsilon_j^2$$

or, equivalently,

$$2\varepsilon_j \Re\langle A_0^*(A_0 X - X A_0)h_j, h_j\rangle + O(\varepsilon_j^2) \geq \langle (I - A_0^* A_0)h_j, h_j\rangle - \varepsilon_j^2 \geq -\varepsilon_j^2.$$

Dividing by ε_j and letting $\varepsilon_j \to 0$ as $j \to \infty$, we see from the last equation that

$$\langle (I - A_0^* A_0)h_j, h_j\rangle \to 0, \tag{1}$$

$$\liminf_{j\to\infty} \Re\langle A_0^*(A_0 X - X A_0)h_j, h_j\rangle \geq 0. \tag{2}$$

We easily conclude that

$$\liminf_{j\to\infty} \Re\langle (X - A_0^* X A_0)h_j, h_j\rangle \geq 0. \tag{3}$$

Set

$$Q = I - A_0^* A_0, \quad T = X - A_0^* X A_0.$$

Then from (1), (3), we see that

$$Qh_j \to 0, \quad \liminf_{j\to\infty} \Re\langle T h_j, h_j\rangle \geq 0. \tag{4}$$

Applying the above argument to ζX for any $\zeta \in \partial\mathbf{D}$ (the unit circle), we see that there exists a sequence $h_j^{(\zeta)}$, $\|h_j^{(\zeta)}\| = 1$, such that

$$Qh_j^{(\zeta)} \to 0, \quad \liminf_{j\to\infty} \Re\zeta\langle T h_j, h_j\rangle \geq 0. \tag{5}$$

We claim that $0 \in W_{Q,0}(T)$. Indeed, if this were not the case, Corollary 3.3 would imply the existence of $\zeta \in \partial\mathbf{D}$ such that

$$\liminf_{j\to\infty} \Re\zeta\langle T h_j, h_j\rangle < 0,$$

for all sequences of unit vectors h_j such that $Qh_j \to 0$, which would contradict (5).

Thus, we have shown that for each $X \in \Delta'$, there exists a sequence of unit vectors $h_j \in \mathcal{E}$ such that

$$(I - A_0^* A_0)h_j \to 0 \text{ and } \langle (X - A_0^* X A_0)h_j, h_j\rangle \to 0. \tag{6}$$

Let

$$\Delta'_{\text{sa}} := \{X = X^* : X \in \Delta'\}.$$

Consider now a subspace $D \subset \Delta'_{\text{sa}}$ of real dimension $n - 1$ such that $\Delta'_{\text{sa}} = D + \mathbf{R}I$. Set $\mathcal{Z} = \{X - A_0^* X A_0 : X \in D\}$, and for every unit vector $h \in \mathcal{E}$ define a linear functional

$\ell(h)$ on $\mathcal{Z}$ by $\ell(h)(T) = \langle Th, h\rangle$, $T \in \mathcal{Z}$. Then Lemma 3.4 applied to the set $S_k = \{\ell(h) : \|(I - A_0^*A_0)h\| \leq 1/k\}$ implies the existence of linear functionals ℓ_k in the convex hull of S_k such that $\|\ell_k\| \leq 1/k$. Observe furthermore that the real dimension of $\mathcal{Z}$ is at most $n - 1$. Then from a standard result (see, e.g., [11, p. 73], each ℓ_k is a convex combination of at most n functionals $\ell(h)$, say $\ell_k = \sum_{j=1}^n \alpha_j^{(k)} \ell(h_j^{(k)})$, where $\alpha_j^{(k)} \geq 0$, $\sum_{j=1}^n \alpha_j^{(k)} = 1$, and the $h_j^{(k)}$ are unit vectors in $\mathcal{E}$, such that

$$\|(I - A_0^*A_0)h_j^{(k)}\| \leq 1/k.$$

Let us define unit vectors $u_k \in \mathcal{E}^{(n)}$ by

$$u_k = \bigoplus_{j=1}^n (\alpha_j^{(k)})^{1/2} h_j^{(k)} \tag{7}$$

and observe that $\lim_{k\to\infty}\langle (X^{(n)} - A_0^{*(n)}X^{(n)}A_0^{(n)})u_k, u_k\rangle = 0$, for every $X \in \Delta'$. Taking $X = Y^*Y$ we obtain

$$\lim_{k\to\infty} (\|Y^{(n)}A_0^{(n)}u_k\| - \|Y^{(n)}u_k\|) = 0, \tag{8}$$

for every $Y \in \Delta'$.

Consider now the spaces $\mathcal{H}_k = \Delta_n' A_0^{(n)} u_k$ and $\mathcal{K}_k = \Delta_n' u_k$. Lemma 2.3 implies that, by passing to appropriate subsequences, we may assume that all the representations $X \to X^{(n)}|\mathcal{H}_k$ (respectively $X \to X^{(n)}|\mathcal{K}_k$) are unitarily equivalent. It follows that we can find partial isometries U_k, V_k in Δ_n'' such that $U_k\mathcal{H}_k = \mathcal{H}_1$ and $V_k\mathcal{K}_k = \mathcal{K}_1$. Dropping again to appropriate subsequences, we may assume that the limits $u = \lim_{k\to\infty} U_k A_0^{(n)} u_k$ and $v = \lim_{k\to\infty} V_k u_k$ exist. Then (8) implies that

$$\|Y^{(n)}u\| = \|Y^{(n)}v\|,$$

for every $Y \in \Delta'$. Therefore there exists a partial isometry $W \in \Delta_n''$ such that

$$WY^{(n)}u = Y^{(n)}v,$$

for every $Y \in \Delta'$. Of course, W can be chosen to be equal to zero on the orthogonal complement of $\Delta_n' u$ and thus to have finite rank at most n. The partial isometries $R_k := V_k^* W U_k$ are in Δ_n'', they have uniformly bounded rank, and

$$\lim_{k\to\infty} (R_k A_0^{(n)} - I)u_k = \lim_{k\to\infty} V_k^*(WU_k A_0^{(n)} u_k - V_k u_k) = 0.$$

Therefore Lemma 2.4 implies that

$$\liminf_{k\to\infty} \|R_k A_0^{(n)}\|_{\mathrm{sp}} \geq 1.$$

Finally, since R_k commutes with $X^{(n)}$, $X \in \Delta'$, and we have

$$X_j^{(n)} R_k A^{(n)} X_j^{(n)-1} \to R_k A_0^{(n)}$$

in norm as $j \to \infty$. Lemma 2.5 shows that

$$\|R_k A_0^{(n)}\|_{\mathrm{sp}} = \|R_k A^{(n)}\|_{\mathrm{sp}}.$$

Consequently, we have

$$\liminf_{k\to\infty} \|R_k A^{(n)}\|_{\mathrm{sp}} = \liminf_{k\to\infty} \|R_k A_0^{(n)}\| \geq 1.$$

Thus,

$$\mu_{\Delta_n''}(A^{(n)}) \geq \liminf_{k\to\infty} \|A^{(n)} X_k\|_{\mathrm{sp}} \geq 1 = \widehat{\mu}_{\Delta}(A),$$

which completes the proof of the theorem. □

Remark. In the cases of interest in control,

$$\Delta'' = \Delta,$$

and so one has from Theorem 4.1 that

$$\mu_{\Delta_n}(A) = \hat{\mu}_{\Delta}(A).$$

5. Conditions for $\mu = \hat{\mu}$. In this section, we will discuss some new conditions when $\mu = \hat{\mu}$ without any need for lifting or ampliation. In the finite-dimensional case, there have been some results of this kind, the most famous of which is that of Doyle [6], who showed that no lifting is necessary for perturbation structures with three or fewer blocks.

We begin by noting that in the proof of Theorem 4.1, we established a useful property of the critical operators A_0 in the closed Δ' similarity orbit

$$\overline{\mathcal{O}_{\Delta'}(A)} = \overline{\{XAX^{-1} : X \in \Delta'\}}$$

of A. Namely, if we call *critical* any $A_0 \in \overline{\mathcal{O}_{\Delta'}(A)}$ satisfying

$$\limsup_{\epsilon\downarrow 0} \|(I - \epsilon X)A_0(I - \epsilon X)^{-1}\| \geq \|A_0\|, \ \forall X \in \Delta',$$

then the first part of the proof of Theorem 4.1 establishes the following.

LEMMA 5.1. *If A_0 is a critical operator in $\overline{\mathcal{O}_{\Delta'}(A)}$, then it enjoys the following property $(\mathcal{O})$:*

$$0 \in W_Q(\|A_0\|^2 X - A_0^* X A_0), \qquad X \in \Delta',$$

where $Q = \|A_0\|^2 I - A_0^* A_0$.

Indeed, property $(\mathcal{O})$ is a reformulation of equation (6) in the case in which the norm of A_0 may be different from 1.

The next lemma is the key step in adapting the proof of Theorem 4.1 to show that

$$\mu_{\Delta}(A) = \widehat{\mu}_{\Delta}(A)$$

in several interesting cases.

LEMMA 5.2. *Let A_0 be an operator on $\mathcal{E}$ which satisfies the essential version of property $(\mathcal{O})$, property $(\mathcal{O}^0)$, namely,*

$$0 \in W_Q^0(\|A_0\|^2 X - A_0^* X A_0), \qquad X \in \Delta',$$

where $Q = \|A_0\|^2 I - A_0^* A_0$. Then there exists a sequence $\{h_k\}_{k=1}^{\infty} \subset \mathcal{E}$, $\|h_k\| = 1$, $k = 1, 2, \ldots$, such that

$$Qh_k \to 0 \text{ strongly and } \langle(\|A_0\|^2 X - A_0^* X A_0)h_k, h_k\rangle \to 0,$$

for all $X \in \Delta'$.

Proof. Without loss of generality we can assume that $\|A_0\| = 1$. Let $X_1, \dots, X_n$ be an algebraic basis of Δ'. (Note that Δ' is finite dimensional.) Set $T_j := X_j - A_0^* X_j A_0$, $j = 1, \dots, n$. Then by virtue of Lemma 3.1, $W_Q^0(T_1, \dots, T_n)$ is convex and compact. If $0 \notin W_Q^0(T_1, \dots, T_n)$, there exists $\psi = (\psi_1, \dots, \psi_n) \in \mathbf{C}^n$ and $\epsilon > 0$ such that

$$\Re \sum_{j=1}^{n} \psi_j \lambda_j \geq \epsilon, \quad \forall \lambda = (\lambda_1, \dots, \lambda_n) \in W_Q^0(T_1, \dots, T_n).$$

Set

$$T = \sum_{j=1}^{n} \psi_j T_j.$$

Property $(\mathcal{O}^0)$ implies that there exists a sequence $\{g_k\}_{k=1}^{\infty} \subset \mathcal{E}$, $\|g_k\| = 1$, $g_k \to 0$ weakly such that $\langle T g_k, g_k \rangle \to 0$. Without loss of generality (by passing to a subsequence if necessary), we can assume that

$$\langle T_j g_k, g_k \rangle \to \lambda_j, \quad j = 1, \dots, n,$$

for $k \to \infty$. Thus

$$\lambda = (\lambda_1, \dots, \lambda_n) \in W_Q^0(T_1, \dots, T_n).$$

Hence

$$0 \leftarrow \Re \langle T g_k, g_k \rangle = \Re \sum_{j=1}^{n} \psi_j \langle T_j g_k, g_k \rangle \to \Re \sum_{j=1}^{n} \psi_j \lambda_j \geq \epsilon,$$

which is a contradiction. We therefore conclude that $0 \in W_Q^0(T_1, \dots, T_n)$, i.e., there exists a sequence $\{h_k\}_{k=1}^{\infty} \subset \mathcal{E}$ satisfying the properties $\|h_k\| = 1$, $k = 1, 2, \dots$, $\|Q h_k\| \to 0$, $h_k \to 0$ weakly, and

$$\langle (X_j - A_0^* X_j A_0) h_k, h_k \rangle = \langle T_j h_k, h_k \rangle \to 0,$$

for all $j = 1, 2, \dots, n$. This implies that

$$\langle (X - A_0^* X A_0) h_k, h_k \rangle \to 0,$$

for all $X \in \Delta'$. □

We can now state the second main result of this paper.

THEOREM 5.3. *If there exists a critical operator A_0 satisfying property $\mathcal{O}^0$ in the closed Δ'-orbit of A, then*

$$\mu_{\Delta''}(A) = \hat{\mu}_{\Delta}(A).$$

Proof. We only have to note that because of Lemma 5.2, we need not take direct sums in the proof of Theorem 4.1. More precisely, referring to equation (7) in the proof of Theorem 4.1, we can take $u_k = h_k$, where

$$\{h_k\}_{k=1}^{\infty}$$

is the sequence provided by Lemma 5.2. The proof then proceeds exactly as in Theorem 4.1 with A_0 replacing $A_0^{(n)}$, X replacing $X^{(n)}$, and Y replacing $Y^{(n)}$. □

Remark. Under the hypotheses of Theorem 5.3, when $\Delta'' = \Delta$ (which happens in all cases of interest in control), we have that

$$\mu_\Delta(A) = \hat{\mu}_\Delta(A).$$

Let $L(\Delta' A \Delta')$ denote the linear space generated by

$$\Delta' A \Delta' = \{XAY : X, Y \in \Delta'\}.$$

Obviously $L(\Delta' A \Delta')$ is finite dimensional and therefore closed. Hence $\overline{\mathcal{O}_{\Delta'}(A)} \subset L(\Delta' A \Delta')$.

COROLLARY 5.4. *If for every $B \in L(\Delta' A \Delta')$, $B \neq 0$, the norm of B is not attained (that is, there is no $h \in \mathcal{H}$ such that $\|Bh\| = \|B\|\|h\| \neq 0$), then*

$$\mu_{\Delta''}(A) = \hat{\mu}_\Delta(A).$$

Proof. The critical operator A_0 constructed in the first part of the proof of Theorem 4.1 belongs to $L(\Delta' A \Delta')$, and therefore its norm is not attained. However in equation (6), we can assume that the sequence $\{h_j\}_{j=1}^\infty$ is weakly convergent, say $h_j \to h$ weakly. Without loss of generality, we may assume that $\|A_0\| = 1$. Then (6) shows that

$$(I - A_0^* A_0)h = 0.$$

Therefore if $h \neq 0$, we would have

$$\|A_0 h\|^2 = \|h\|^2 = \|A_0\|^2 \|h\|^2 \neq 0,$$

and so the norm of A_0 would be attained. We conclude that $h_j \to 0$ weakly, and so A_0 satisfies property $(\mathcal{O}^0)$. The required result now follows by Theorem 5.3. □

Remark. Note that Corollary 5.4 applies only to infinite-dimensional Hilbert spaces $\mathcal{E}$.

Example. We would like to give an explicit example to which Corollary 5.4 applies. Let A_j be an operator on a Hilbert space $\mathcal{E}_j$ $(j = 1, \ldots, n)$ for which the norm is not attained. (For example, take $\mathcal{E}_j = L^2((0, 1))$ and let A_j be the muliplication operator $f(x) \mapsto xf(x)$ for $x \in (0, 1)$ and $f \in L^2((0, 1))$.) Set

$$\mathcal{E} := \mathcal{E}_1 \oplus \cdots \oplus \mathcal{E}_n,$$

and let Δ be the algebra of operators on $\mathcal{E}$ of the form

$$\begin{bmatrix} X_1 & 0 & \ldots & 0 \\ 0 & X_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & X_n \end{bmatrix},$$

with $X_j \in \mathcal{L}(\mathcal{E}_j)$, $j = 1, 2, \ldots, n$. Then Δ' is formed by the diagonal operators

$$\begin{bmatrix} \lambda_1 I_{\mathcal{E}_1} & 0 & \ldots & 0 \\ 0 & \lambda_2 I_{\mathcal{E}_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n I_{\mathcal{E}_n} \end{bmatrix},$$

for $\lambda_j \in \mathbf{C}$, $j = 1, 2, \ldots, n$, and $I_{\mathcal{E}_j}$ denotes the identity operator on $\mathcal{E}_j$, $j = 1, 2, \ldots, n$. Let A be any operator on $\mathcal{E}$, the $n \times n$ block matrix representation of which has entries in the set $\{0, A_1, \ldots, A_n\}$ with only one nonzero entry in each row and column. Then it is easy to check that $L(\Delta' A \Delta')$ has the property required in Corollary 5.4, and therefore

$$\mu_{\Delta''}(A) = \hat{\mu}_\Delta(A).$$

6. Toeplitz operators. In this section, we want to use our lifting methodology to derive a beautiful result of Shamma [13, 14] on the structured singular value of a Toeplitz operator, i.e., a linear time-invariant system.

Accordingly, set $\mathcal{E} = H^2(\mathbf{C}^n)$ and let A denote the multiplication (analytic Toeplitz) operator on $\mathcal{E}$ defined by

$$(Ah)(z) = A(z)h(z), \qquad |z| < 1, \ h \in \mathcal{E},$$

where

$$A(z) = [a_{jk}]_{j,k=1}^{n}, \qquad |z| < 1,$$

has H^∞ entries. Let Δ' be any $*$-subalgebra of $\mathcal{L}(\mathbf{C}^n)$, the elements of which are regarded as multiplication operators on $\mathcal{E}$. Note that in this case, $\Delta'' = \Delta$ is the algebra generated by operators of the form

$$(Bh)(z) = B(z)h(z), \qquad |z| < 1, \ h \in \mathcal{E},$$

with $B(z)X = XB(z)$, $|z| < 1$, $X \in \Delta'$, as well as of the form

$$B \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix} = \begin{bmatrix} Yh_1 \\ Yh_2 \\ \vdots \\ Yh_n \end{bmatrix},$$

with $Y \in \mathcal{L}(H^2(\mathbf{C}))$ arbitrary. We can now state the following lemma.

LEMMA 6.1. *Let A_0 be an analytic Toeplitz operator. Then if A_0 has property $(\mathcal{O})$, it also has property $(\mathcal{O}^0)$.*

Proof. Without loss of generality we may assume $\|A_0\| = 1$. Let $X \in \Delta'$ and let h_j, $j = 1, 2, \ldots$ be a sequence of unit vectors satisfying

$$\|(I - A_0^* A_0)h_j\|^2 \to 0, \quad \langle (X - A_0^* X A_0)h_j, h_j \rangle \to 0. \tag{9}$$

Note that since $I - A_0^* A_0 \geq 0$, the first condition in (9) is equivalent to

$$\langle (I - A_0^* A_0)h_j, h_j \rangle \to 0.$$

Let U denote the canonical unilateral shift on $\mathcal{E} = H^2(\mathbf{C}^n)$, that is,

$$(Uh)(z) := zh(z), \qquad |z| < 1, \ h \in \mathcal{E}.$$

As is well known, we can view $H^2(\mathbf{C}^n)$ as a subspace of $L^2(\mathbf{C}^n)$. In particular, in this representation the relations (9) are equivalent to

$$\int_0^{2\pi} (\|h_j(e^{it})\|^2 - \|A_0(e^{it})h_j(e^{it})\|^2)dt \to 0,$$

$$\int_0^{2\pi} [\langle Xh_j(e^{it}), h_j(e^{it}) \rangle - \langle XA_0(e^{it})h_j(e^{it}), h_j(e^{it}) \rangle]dt \to 0. \tag{10}$$

Note that X is an $n \times n$ matrix with constant coefficients. Therefore in (10), h_j can be replaced by $U^k h_j$ for any $k \geq 0$ without changing the values of the integrals. We infer that

$$\|(I - A_0^* A_0)U^{k_j} h_j\|^2 \to 0, \qquad \langle (X - A_0^* X A_0)U^{k_j} h_j, U^{k_j} h_j \rangle \to 0,$$

for any sequence $\{k_j\}_{j=1}^{\infty}$ of natural numbers. Since for any $g, h \in \mathcal{E}$,

$$|\langle U^k g, h\rangle| = |\langle g, U^{*k} h\rangle| \leq \|g\|\|U^{*k} h\| \to 0, \quad k \to \infty,$$

we can choose k_j sufficiently large to guarantee that

$$|\langle U^{k_j} h_j, h\rangle| \leq \frac{1}{2^j},$$

for any h of the form

$$h = (z^m \delta_{pk})_{k=1}^n, \qquad 0 \leq m \leq j,\ 1 \leq p \leq n, \tag{11}$$

where δ_{pk} is the Kronecker delta. Thus

$$\langle U^{k_j} h_j, h\rangle \to 0 \text{ as } j \to \infty,$$

for all vectors of the form (11). Since these vectors form an orthonormal basis of $\mathcal{E}$, we see that $U^{k_j} h_j \to 0$ weakly, which concludes the proof of the lemma. □

COROLLARY 6.2 (see [13, 14]). *For A and Δ' as above, we have that*

$$\mu_{\Delta}(A) = \hat{\mu}_{\Delta}(A).$$

Proof. First, note that any operator B in $L(\Delta' A \Delta')$ is also an analytic Toeplitz operator. In particular, the critical operator A_0 obtained in the proof of Theorem 4.1 is a multiplication operator given by

$$A_0(z) = [a_{jk}^0(z)]_{j,k=1}^n, \qquad |z| < 1.$$

By Lemma 5.1, the operator A_0 has property $(\mathcal{O})$ and thus also property $(\mathcal{O}^0)$, by virtue of Lemma 6.1. The conclusion now follows from Theorem 5.3. □

7. Structured singular value of input/output operators. In this section, we will put some of the above results into a system-theoretic framework. Accordingly, let ℓ_+^2 be the space of square summable one-sided sequences in $\mathbf{C}$, let $\mathcal{C}$ denote the set of all bounded linear operators on ℓ_+^2. Further, let $A : \ell_+^2(\mathbf{C}^n) \to \ell_+^2(\mathbf{C}^n)$ be an arbitrary bounded linear operator. Thus A defines a (possibly) time-varying system. (Here $\ell_+^2(\mathbf{C}^n)$ denotes the space of square summable sequences in $\mathbf{C}^n$, i.e., the space of finite energy vector-valued signals with n components.) Then we want to interpret $\hat{\mu}_{\Delta}(A)$ as a structured singular value on an extended space with an enhanced perturbation structure. Note $\mathcal{E}$ in this case is the Hilbert space $\ell_+^2(\mathbf{C}^n)$.

Define the algebra of perturbations

$$\Delta := \left\{ \begin{bmatrix} \delta_1 & 0 & \dots & 0 \\ 0 & \delta_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \delta_n \end{bmatrix} : \delta_i \in \mathcal{C}, i = 1, \dots, n \right\}.$$

Then the commutant of Δ is the finite-dimensional C^*-algebra

$$\Delta' := \left\{ \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_n \end{bmatrix} : d_i \in \mathbf{C}, i = 1, \dots, n \right\}.$$

Note that a constant $d \in \mathbf{C}$ defines an operator on ℓ_+^2 via multiplication.

From Theorem 4.1, it follows that the μ upper bound given by the infimum of $\|XAX^{-1}\|$ over all constant X-scales equals $\bar{\mu}_\Delta(A)$. We now have the following interpretation of $\widehat{\mu}_\Delta(A)$. We lift A to $A^{(n)} : \mathcal{E}^{(n)} \to \mathcal{E}^{(n)}$. Then

$$(\Delta_n)'' \cong \left\{ \begin{bmatrix} \tilde{\Delta}_1 & 0 & 0 & 0 \\ 0 & \tilde{\Delta}_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \tilde{\Delta}_n \end{bmatrix} : \tilde{\Delta}_j \in \Delta \right\}.$$

$(\Delta_n)''$ is a space of time-varying perturbations and we have from Theorem 4.1 that

$$\widehat{\mu}_\Delta(A) = \mu_{\Delta_n''}(A^{(n)}).$$

This is true for *arbitrary time-varying systems* A. When A is Toeplitz, i.e., the system is time invariant, then as we have seen (Corollary 6.2, [13, 14]),

$$\widehat{\mu}_\Delta(A) = \mu_\Delta(A).$$

REFERENCES

[1] H. BERCOVICI, C. FOIAS, AND A. TANNENBAUM, *Structured interpolation theory*, Oper. Theory Adv. Appl., 47 (1990), pp. 195–220.

[2] ———, *A spectral commutant lifting theorem*, Trans. Amer. Math. Soc., 325 (1991), pp. 741–763.

[3] ———, *A relative Toeplitz-Hausdorff theorem*, Oper. Theory Adv. Appl., 71 (1994), pp. 29–32.

[4] ———, *Continuity of the spectrum on closed similarity orbits*, Integral Equations Oper. Theory, 18 (1994), pp. 242–246.

[5] H. BERCOVICI, C. FOIAS, P. KHARGONEKAR, AND A. TANNENBAUM, *On a lifting theorem for the structured singular value*, J. Math. Anal. Appl., 187 (1994), pp. 617–627.

[6] J. C. DOYLE, *Analysis of feedback systems with structured uncertainties*, IEEE Proc., 129 (1982), pp. 242–250.

[7] M. FAN, *A Lifting Result on Structured Singular Values*, Technical report, Georgia Institute of Technology, Atlanta, GA, 1992.

[8] P. HALMOS, *A Hilbert Space Problem Book*, Springer-Verlag, New York, 1982.

[9] M. KHAMMASH AND J. B. PEARSON, *Performance robustness of discrete-time systems with structured uncertainty*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 398–412.

[10] A. MAGRETSKI, *Power Distribution Approach in Robust Control*, Technical report, Royal Institute of Technology, Stockholm, Sweden, 1992.

[11] W. RUDIN, *Functional Analysis*, 2nd ed., McGraw-Hill, New York, 1991.

[12] M. G. SAFONOV, *Stability Robustness of Multivariable Feedback Systems*, MIT Press, Cambridge, MA, 1980.

[13] J. SHAMMA, *Robust stability with time-varying structured uncertainty*, IEEE Trans. Automat. Control, 39 (1994), pp. 714–724.

[14] ———, *Robust stability for time-varying systems*, in 31st Proc. IEEE Conference on Decision and Control, Tucson, AZ, 1992, pp. 3163–3168.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1405–1419, July 1996

016

A TURNPIKE THEORY FOR INFINITE-HORIZON OPEN-LOOP COMPETITIVE PROCESSES*

D. CARLSON† AND A. HAURIE‡

Abstract. This paper deals with a class of dynamic discrete time open-loop games played over an infinite time horizon. The equilibrium concept is defined in the sense of overtaking optimal responses by the players to the program choices of the opponents. We extend to this dynamic game framework the results obtained by Rosen for concave static games. We prove existence, uniqueness, and asymptotic stability (also called the turnpike property) of overtaking equilibrium programs for a class of games satisfying a strong concavity assumption (strict diagonal concavity).

Key words. dynamic games, overtaking equilibrium, infinite-horizon dynamic games, open-loop equilibrium

AMS subject classifications. 49J99, 49N40, 90D06, 90D10, 90D50

1. Introduction.

This paper deals with a class of dynamic competitive process models defined over an infinite time horizon. We use overtaking optimality to deal with unbounded payoffs. This optimality concept has been used to study optimal economic growth models in [28], [21], [3], [7] and in dynamic games [4], [18], [25], [20]. We prove existence and uniqueness of open-loop equilibria and asymptotic stability of the equilibrium programs under a *strict diagonal concavity–convexity* assumption for the Hamiltonians associated with the extremal dynamic programs of each player. This assumption is very close to the one introduced by Rosen [26] in his study of static concave games.

The theory of asymptotic optimality of dynamical systems has been mainly motivated by the modeling of optimal economic growth processes. Von Weizäcker [28] was the first to propose the use of the overtaking optimality concept to compare accumulation programs over infinite time horizons. These economic growth models developed under two formulations: either in continuous time (see, e.g., [9], [10], [12], [5]) or in discrete time (see, e.g., [3], [21]). A major result in these theories is the *turnpike property* which links existence as well as necessary and sufficiency conditions with the asymptotic stability of the solutions of an associated Hamiltonian system. This property is easily derived from the convexity–concavity of the Hamiltonian [23], when the stream of payoffs is undiscounted. For the discounted case, Cass and Shell [12], Brock and Scheinkman [6], and Rockafellar [24] have obtained the same property under a stricter *curvature assumption*. The whole theory extends to optimal control, both in the deterministic and stochastic frameworks, as exposed in [7], [17].

The class of competitive processes considered encompasses the oligopoly paradigm, representing the competition among a few firms on a single market, which is a cornerstone of microeconomic theory (see, e.g., [15]). The Cournot equilibrium [13], extended by Nash [22] to a general game format, has been used in many economic models. Rosen [26], in a seminal paper gave a condition, called *strict diagonal concavity*, for existence, uniqueness, and stability of a gradient-path following method for Nash–Cournot equilibria in static m-player games. Linking game theory and asymptotic economic growth models has been attempted in several papers. Brock [4] proposed the first study, in continuous time, of differential games with infinite time horizon. In a more general setting, the asymptotic stability of the solutions

*Received by the editors March 24, 1994; accepted for publication (in revised form) April 25, 1995. This research was supported by FNRS-Switzerland, FCAR-Québec, NSERC-Canada, a Summer Research Fellowship from the University of Toledo, and a travel grant from the Department of Management Studies of the University of Geneva.

†Department of Mathematics, University of Toledo, Toledo, OH 46306-3390.

‡Department of Management Studies, University of Geneva, 102 Carl-Vogt, CH-1211, Geneva, Switzerland, and Groupe d'Étude et de Recherche en Analyse des Décision–École des Hautes Études Commerciales, Montréal, PQ H3T 1V6, Canada.

of the *pseudo Hamiltonian* systems associated with open-loop differential games was studied in [18]. More recently, in [19], a continuous time piecewise deterministic differential game model of oligopoly has been studied and conditions for asymptotic stability have been stated.

In the present paper we consider a dynamic competitive process modeled in discrete time and we provide a complete theory of existence, uniqueness, and asymptotic stability. We obtain these results under conditions very similar to Rosen's strict diagonal concavity. In a sense, the present paper can be seen as an extension of Rosen's work to a dynamic setting.

We would like now to address those "mathematical economists" who don't read further as soon as they have detected that a paper deals with an open-loop equilibrium concept on an infinite horizon. The recent polarization of interest on *subgame perfectness* in dynamic games (see [27] for a definition of this concept) and the famous *Folk theorem* for infinite-horizon games (see [16] for a discussion of modern game theory) have indeed diminished the interest of economists for the open-loop equilibrium solution concept. The first concept is close to, in spirit, but more general than the *feedback-Nash* equilibrium solution (see [1] for a complete discussion of this concept); it is of limited applicability in a truly dynamical setting (i.e., not a repeated game setting) since a complete and general theory of feedback-Nash equilibria is still to be developed. The Folk theorem (see [20] for a statement in the realm of multistage games) establishes that, on an infinite time horizon, any solution dominating the feedback-Nash equilibrium solution can be transformed into a subgame perfect equilibrium; this result is based on an information structure allowing the players to remember a preplay agreement and to retaliate if the agreed upon trajectory is not implemented. Both concepts present attractive features. We claim, however, that the exploration of the properties of open-loop solutions deserves continuing interest.

To give full mathematical respectability to open-loop equilibria we only have to assume an information structure whereby each player knows only the initial state of the process and has to choose a control for the entire infinite time horizon. Then the open-loop equilibrium solution is similar to a static game concept, except that the players' actions are now described as functions of time.

In [19] it is shown how this paradigm can be used in a more complex information structure, called *piecewise deterministic games*, where the information on the state occurs at "random stopping time" as the outcome of a random jump process. The game studied in [19] is akin to a *sequential game* with continuous state and action spaces. It has been shown in [19] that the dynamic programming solution of this type of game is closely linked with the turnpike property of an associated class of infinite-horizon open-loop games.

This gives already two good reasons for studying this interesting class of equilibria. A third one is more operational. Most of the large-scale *computable* market models which have been developed in operations research studies have been based on the open-loop equilibrium concept. This has been the case in the modeling of energy markets and this is the current trend when dealing with the modeling of environmental management. Of course, the feedback-Nash concept usually does not lend itself to a numerical computation with the exception of two extreme cases: the *linear quadratic* game structure (with some difficulties for the infinite-horizon case, see, e.g., [8]) and the *affine dynamics* one (see [2] for a discussion of the difficulties in solving dynamic programming equilibrium equations). Therefore the operational representation of competition through capital accumulation or capacity expansion for large-scale models remains, for the moment, in the realm of open-loop games.

As in the economic growth problem, the justification for considering an infinite time horizon stems from the necessity to give, at the end of any finite horizon, a bequest value to the accumulated stocks (capital goods). These values should express the utility of these stocks in further economic activity, hence providing a justification for a never-ending process. From

an operational viewpoint, the theory developed in the present paper would give a theoretical justification for adopting as approximate terminal conditions for a long but finite-horizon model the steady-state equilibrium, which should be the attractor if the horizon is infinite.

The paper is organized as follows. In §2 we explore the turnpike property for a general class of competitive process models, permitting, as in [21] for the single-player case, a nonautonomous dynamic. In §3 we show that the conditions used to insure asymptotic stability also imply existence of overtaking equilibria. In §4 we show that, in addition we get uniqueness. Finally in §5 we present a class of examples to which the results given below are applicable.

2. Turnpikes for discrete time overtaking equilibria. In this section we define the concept of an *overtaking equilibrium* for a class of games which represent infinite-horizon dynamic competition among m firms. We give conditions under which all the overtaking equilibrium programs, emanating from different initial states, bunch together at infinity.

2.1. Competitive programs. We consider a competitive process defined by the following data:

(i) There is an infinite sequence of stages or time periods $k = 0, 1, 2, \ldots$.

(ii) A set $M \doteq \{1, \ldots, m\}$ of m players (or firms) is represented at period k by a state $x_j^k \in \mathbb{R}^{n_j}$, where n_j is a given positive integer (we denote $n \doteq n_1 + \cdots + n_m$). This state is, e.g., the production capacity of firm j.

(iii) For each $j \in M$, a program for player j is defined as an infinite sequence $\mathbf{x}_j = (x_j^k \in \mathbb{R}^{n_j} : k = 0, 1, \ldots)$. An M-program is defined as $\mathbf{x} = (x^k : k = 0, 1, \ldots) \doteq ((x_j{}^k)_{j\in M} : k = 0, 1, \ldots)$.

(iv) Along an M-program, the reward accumulation process for player j is defined as

$$\phi_j^K(\mathbf{x}) = \sum_{k=0}^{K-1} (\beta_j)^k L_j^k(x^k, \Delta x_j^k), \quad K = 1, 2, \ldots, \tag{1}$$

where $\beta_j \in (0, 1]$ is the discount factor for player j ($(\beta_j)^k$ denotes the kth power of β_j), $L_j^k : \mathbb{R}^n \times \mathbb{R}^{n_j} \mapsto \mathbb{R} \cup \{-\infty\}$, $j \in M$, are given functions and $\Delta x_j^k = x_j^{k+1} - x_j^k$. The expression $L_j^k(x^k, \Delta x_j^k)$ represents, e.g., the net income to firm j when the market price is a function of total supply $\sum_{j\in M} x_j^k$ minus the cost for the capacity adjustment Δx_j^k.

2.2. Optimality. Given an M-program $\mathbf{x}^*$ we denote by $[\mathbf{x}^{*(j)}; \mathbf{x}_j]$ the M-program obtained when player j changes unilaterally his program to $\mathbf{x}_j$.

DEFINITION 2.1. *An M-program $\mathbf{x}^*$ is an* equilibrium *at x^o if*

1. $x^{0^*} = x^o$,
2. $\lim_{K\to\infty} \phi_j^K(\mathbf{x}^*) < \infty$ *for all* $j \in M$,
3. $\liminf_{K\to\infty}(\phi_j^K(\mathbf{x}^*) - \phi_j^K([\mathbf{x}^{*(j)}; \mathbf{x}_j])) \geq 0$ *for all programs* $\mathbf{x}_j$ *such that* $x_j^0 = x_j^o$ *for all* $j \in M$.

If only the first and third conditions hold, the M-program $\mathbf{x}^*$ is called an overtaking equilibrium *at x^o.*

2.3. Optimality conditions. Let us introduce for $j \in M$ and $p_j \in \mathbb{R}^{n_j}$ the *Hamiltonians* $H_j^k : \mathbb{R}^n \times \mathbb{R}^{n_j} \mapsto \mathbb{R} \cup \{\infty\}$, defined as

$$H_j^k(x, p_j) = \sup_{z_j}\{L_j^k(x, z_j) + p_j' z_j\}. \tag{2}$$

Here p_j is called a *j-supporting price* vector. A sequence

$$\mathbf{p} \doteq (p^k : k = 0, 1, \ldots) \doteq ((p_j{}^k)_{j\in M} : k = 0, 1, \ldots)$$

will be called an *M-price schedule.*

Assumption 2.2. We assume that the Hamiltonians H_j^k are concave in x_j, convex in p_j.

For an (overtaking) equilibrium the following necessary conditions hold.

THEOREM 2.3. *Under Assumption* 2.2, *if* $\mathbf{x}^*$ *is an* overtaking equilibrium *at initial state* x^o, *then there exists an* M*-price schedule* $\mathbf{p}^*$ *such that*

$$\Delta x_j^{k^*} \in \partial_{p_j} H_j^k(x^{k^*}, p_j^{k+1^*}), \tag{3}$$

$$\Delta p_j^{k^*} + \alpha_j\, p_j^{k+1^*} \in -\partial_{x_j} H_j^k(x^{k^*}, p_j^{k+1^*}), \tag{4}$$

for all $j \in M$, *where we have denoted* $\alpha_j = (1 - \frac{1}{\beta_j})$.

Proof. See [3], [21] for a proof of Theorem 2.3. □

Equations (3), (4) were called pseudo-Hamiltonian systems in [18]. These conditions are incomplete since only initial conditions are specified for the M-programs and not their associated M-price schedules. In the single-player case, this system is made complete by invoking the turnpike property, which provides an asymptotic transversality condition. Due to the coupling among the players, the system (3), (4) considered here does not fully enjoy the rich geometric structure found in the classical optimization setting (for instance the saddle point behavior of Hamiltonian systems in the autonomous case). In the next two sections we provide conditions under which the turnpike property holds for these pseudo-Hamiltonian systems.

2.4. A Turnpike result for the undiscounted case. Let us first consider the case where $\beta_j \equiv 1$ for all $j \in M$. Then, clearly, we have to deal with overtaking equilibria since the sequence of accumulated rewards can be unbounded.

Assumption 2.4 (strict diagonal concavity–convexity assumption (SDCCA)). We assume that the combined Hamiltonian $\sum_{j\in M} H_j^k(x, p_j)$ is strictly diagonally concave in x, convex in p. That is,

$$\sum_{j\in M}\Big[(\hat{p}_j - \tilde{p}_j)'(\hat{\pi}_j - \tilde{\pi}_j) + (\hat{x}_j - \tilde{x}_j)'(\hat{\xi}_j - \tilde{\xi}_j)\Big] > 0, \tag{5}$$

for all $(\hat{x}_j, \tilde{x}_j, \hat{p}_j, \tilde{p}_j)$ and $(\hat{\pi}_j, \tilde{\pi}_j, \hat{\xi}_j, \tilde{\xi}_j)$, such that

$$\hat{\pi}_j \in \partial_{p_j} H_j^k(\hat{x}, \hat{p}_j), \quad \tilde{\pi}_j \in \partial_{p_j} H_j^k(\tilde{x}, \tilde{p}_j), \tag{6}$$

$$\hat{\xi}_j \in -\partial_{x_j} H_j^k(\hat{x}, \hat{p}_j), \quad \tilde{\xi}_j \in -\partial_{x_j} H_j^k(\tilde{x}, \tilde{p}_j). \tag{7}$$

LEMMA 2.5. *Assume* $\beta_j \equiv 1$ *for all* $j \in M$. *Let* $\hat{\mathbf{x}}$ *and* $\tilde{\mathbf{x}}$ *be two (overtaking) equilibria at* $\hat{x}^o$ *and* $\tilde{x}^o$, *respectively, with their respective associated* M*-price schedules* $\hat{\mathbf{p}}$ *and* $\tilde{\mathbf{p}}$. *Then under Assumptions* 2.2 *and* 2.4 *the following inequality holds:*

$$\sum_{j\in M}\Big[(\Delta\hat{p}_j^k - \Delta\tilde{p}_j^k)'(\hat{x}_j^k - \tilde{x}_j^k) + (\hat{p}_j^{k+1} - \tilde{p}_j^{k+1})'(\Delta\hat{x}_j^k - \Delta\tilde{x}_j^k)\Big] > 0. \tag{8}$$

Proof. According to Theorem 2.3 we have

$$\Delta\hat{x}_j^k \in \partial_{p_j} H_j^k(\hat{x}^k, \hat{p}_j^{k+1}), \tag{9}$$

$$\Delta\hat{p}_j^k \in -\partial_{x_j} H_j^k(\hat{x}^k, \hat{p}_j^{k+1}), \tag{10}$$

$$\Delta\tilde{x}_j^k \in \partial_{p_j} H_j^k(\tilde{x}^k, \tilde{p}_j^{k+1}), \tag{11}$$

$$\Delta\tilde{p}_j^k \in -\partial_{x_j} H_j^k(\tilde{x}^k, \tilde{p}_j^{k+1}). \tag{12}$$

Then (8) follows directly from Assumption 2.4. □

We now prove the turnpike theorem, under Assumption 2.4, for the class of *strongly diagonally supported* (overtaking) equilibria defined below.

DEFINITION 2.6. *We say that the (overtaking) equilibrium M-program* $\hat{\mathbf{x}}$ *is strongly diagonally supported by the M-price schedule* $\hat{\mathbf{p}}$ *if, for every* $\varepsilon > 0$, *there exists a* $\delta > 0$ *such that for all* $k = 0, 1, \ldots, \|x - \hat{x}^k\| + \|p - \hat{p}^k\| > \varepsilon$ *implies*

$$\sum_{j \in M} \left[(\hat{p}_j^{k+1} - p_j)'(\Delta \hat{x}_j^k - \pi_j) + (\hat{x}_j^k - x_j)'(\Delta \hat{p}_j^k - \xi_j) \right] > \delta, \tag{13}$$

for all (x_j, p_j) *and* (π_j, ξ_j), *such that*

$$\pi_j \in \partial_{p_j} H_j^k(x, p_j), \tag{14}$$

$$\xi_j \in -\partial_{x_j} H_j^k(x, p_j). \tag{15}$$

Remark 2.7. In the autonomous case, treated in more detail in §2.7, the stricter inequality (13) is obtained as a consequence of Assumption 2.4 or inequality (5) when the state variable x remains in a compact set (this is known as the Atsumi lemma in the single-player case). In the general nonautonomous case, the condition (13) is certainly more restrictive and not always easy to verify. The case of discounted payoffs is a special case of the general nonautonomous case and it exhibits a special structure that permits these strong support properties to be modified in such as way as to be useful. This is particularly true if this nonautonomy arises only as a result of the discounting. We pursue these developments in §§2.6 and 2.7 below.

THEOREM 2.8. *Suppose Assumptions* 2.2 *and* 2.4 *hold. Assume* $\beta_j \equiv 1$ *for all* $j \in M$. *Let* $\hat{\mathbf{x}}$, *with its associated M-price schedule* $\hat{\mathbf{p}}$, *be a strongly diagonally supported (overtaking) equilibrium at* $\hat{x}^o$ *such that*

$$\limsup_{k \to \infty} \|(\hat{x}^k, \hat{p}^k)\| < \infty.$$

Let $\tilde{\mathbf{x}}$ *be another (overtaking) equilibrium at* $\tilde{x}^o$ *with* $\tilde{\mathbf{p}}$ *its associated M-price schedule such that*

$$\limsup_{k \to \infty} \|(\tilde{x}^k, \tilde{p}^k)\| < \infty.$$

Then

$$\lim_{k \to \infty} \|(\hat{x}^k - \tilde{x}^k, \hat{p}^k - \tilde{p}^k)\| = 0. \tag{16}$$

Proof. Assume (16) does not hold. Then, according to equation (13) of Definition 2.6 we have

$$\lim_{K \to \infty} \sum_{k=0}^{K} \sum_{j \in M} \left[(\Delta \hat{p}_j^k - \Delta \tilde{p}_j^k)'(\hat{x}_j^k - \tilde{x}_j^k) + (\hat{p}_j^{k+1} - \tilde{p}_j^{k+1})'(\Delta \hat{x}_j^k - \Delta \tilde{x}_j^k) \right] = \infty. \tag{17}$$

However the left-hand side of the above expression is also equal to $\lim_{K \to \infty} V(K)$, where

$$V(K) \doteq \sum_{j \in M} \left[(\hat{p}_j^{K+1} - \tilde{p}_j^{K+1})'(\hat{x}_j^{K+1} - \tilde{x}_j^{K+1}) - (\hat{p}_j^0 - \tilde{p}_j^0)'(\hat{x}_j^0 - \tilde{x}_j^0) \right]. \tag{18}$$

If the M-programs and their associated M-price schedules are bounded, then so should be $V(K)$ for all K. This contradicts (17). □

Remark 2.9. This turnpike result is very much in the spirit of McKenzie [21] since it is established for nonautonomous systems as well as a nonconstant turnpike. The special case of autonomous systems will be considered in more detail in §2.7.

2.5. Conditions for SDCCA.

LEMMA 2.10. *Assume $L_j^k(x, z_j)$ is concave in (x_j, z_j) and assume that the total reward function $\sum_{j\in M} L_j^k(x, z_j)$ is diagonally strictly concave in (x, z), i.e., it verifies*

$$\sum_{j\in M}[(z_j^1 - z_j^0)'(\zeta_j^1 - \zeta_j^0) + (x_j^1 - x_j^0)(\eta_j^1 - \eta_j^0)] < 0, \tag{19}$$

for all

$$\eta_j^1 \in \partial_{x_j} L_j^k(x^1, z_j^1), \quad \zeta_j^1 \in \partial_{z_j} L_j^k(x^1, z_j^1),$$
$$\eta_j^0 \in \partial_{x_j} L_j^k(x^0, z_j^0), \quad \zeta_j^0 \in \partial_{z_j} L_j^k(x^0, z_j^0).$$

Then Assumption 2.4 *holds true.*

Proof. The concavity of $L_j^k(x, z_j)$ in (x_j, z_j) implies that each Hamiltonian

$$H_j^k(x, p_j) = \sup_{z_j}\{L_j^k(x, z_j) + p_j' z_j\}$$

defined in (2) is convex in p_j, for $j \in M$. This property with (19) imply that (5) holds true for all $(\hat{x}_j, \tilde{x}_j, \hat{p}_j, \tilde{p}_j)$ and $(\hat{\pi}_j, \tilde{\pi}_j, \hat{\xi}_j, \tilde{\xi}_j)$, such that (6) and (7) are satisfied. □

Remark 2.11. In the special case when the functions $L_j^k(\cdot, \cdot)$ are twice continuously differentiable the strict diagonal concavity in (x, z) of the total reward function $\sum_{j\in M} L_j^k(x, z)$ may be verified by applying the conditions given in Rosen [26, Thm. 6, p. 528].

2.6. The turnpike property with discounting. We have seen that concavity–convexity of the Hamiltonians and SDCCA, when combined with the strong support property, imply the turnpike property for overtaking equilibria when the infinite-horizon payoffs are not discounted. Discounting the payoffs could be considered as a form of a nonautonomous system. However, in that case, the condition of strong diagonal support will not be easily satisfied. To give a workable turnpike theorem for the discounted case, i.e., when $\beta_j < 1$ for at least one j, we must strengthen the support property of Definition 2.6 by a curvature condition.

THEOREM 2.12. *Suppose Assumptions* 2.2 *and* 2.4 *hold. Let $\hat{\mathbf{x}}$ and $\tilde{\mathbf{x}}$ be two overtaking equilibria at $\hat{x}^o$ and $\tilde{x}^o$, respectively, with associated M-price schedules $\hat{\mathbf{p}}$ and $\tilde{\mathbf{p}}$ such that*

$$\limsup_{k\to\infty} \| (\hat{x}^k, \hat{p}^k) \| < \infty \text{ and } \limsup_{k\to\infty} \| (\tilde{x}^k, \tilde{p}^k) \| < \infty$$

and additionally satisfy the following property.

For each $\epsilon > 0$ there exists $\delta > 0$ so that whenever

$$\| (\hat{x}^k - \tilde{x}^k, \hat{p}^k - \tilde{p}^k) \| > \epsilon,$$

one has

$$\sum_{j\in M}\left[\left(\hat{p}_j^{k+1} - \tilde{p}_j^{k+1}\right)'\left(\Delta\hat{x}_j^k - \Delta\tilde{x}_j^k\right) + \left(\hat{x}_j^k - \tilde{x}_j^k\right)'\left(\hat{\xi}_j^k - \tilde{\xi}_j^k\right)\right]$$
$$> \delta + \sum_{j\in M}\alpha_j\left(\hat{x}_j^k - \tilde{x}_j^k\right)'\left(\hat{p}_j^k - \tilde{p}_j^k\right) \tag{20}$$

for all

$$\hat{\xi}_j^k \in -\partial_{x_j} H_j^k(\hat{x}^k, \hat{p}_j^{k+1}), \tag{21}$$

$$\tilde{\xi}_j^k \in -\partial_{x_j} H_j^k(\tilde{x}^k, \tilde{p}_j^{k+1}). \tag{22}$$

Then

$$\lim_{k\to\infty} \| (\hat{x}^k - \tilde{x}^k, \hat{p}^k - \tilde{p}^k) \| = 0.$$

Proof. The proof of this result is a straightforward adaptation of the proof of Theorem 2.8. □

Remark 2.13. The condition (20) is clearly a strengthening of (13). This condition indeed is analogous to the conditions found in the papers of Cass and Shell, Brock and Scheinkman, and Rockafellar, et. al., which are nicely collected in [11].

2.7. The autonomous case. We now focus our study on the case when the functions L_j^k are independent of time k (i.e., $L_j^k(x, z_j) \equiv L_j(x, z_j)$). In this case, the optimality conditions become

$$\Delta x_j^k \in \partial_{p_j} H_j(x^k, p_j^{k+1}),$$
$$\Delta p_j^k + \alpha_j p_j^{k+1} \in -\partial_{x_j} H_j(x^k, p_j^{k+1}),$$

where

$$H_j(x, p_j) = \sup_{z_j \in \mathbb{R}^{n_j}} \{L_j(x, z_j) + p_j' z_j\}.$$

The above conditions define an autonomous pseudo-Hamiltonian system and the possibility arises that there exists a *steady-state equilibrium.* That is, a pair $(\bar{x}, \bar{p}) \in \mathbb{R}^n \times \mathbb{R}^n$ which satisfies

$$0 \in \partial_{p_j} H_j(\bar{x}, \bar{p}_j),$$
$$\alpha_j \bar{p}_j \in -\partial_{x_j} H_j(\bar{x}, \bar{p}_j).$$

When a unique steady-state equilibrium exists, the turnpike properties discussed above provide conditions when the pair $(\bar{x}, \bar{p})$ becomes an attractor for all bounded (overtaking) equilibria. Moreover, in this case, the support properties and curvature assumptions along a trajectory described on the nonautonomous case become simpler.

DEFINITION 2.14. *Let $(\bar{x}, \bar{p})$ be a steady-state equilibrium. We say that the strong diagonal support property for $(\bar{x}, \bar{p})$ holds if for each $\epsilon > 0$ there exists $\delta > 0$ so that whenever $\|x - \bar{x}\| + \|p - \bar{p}\| > \epsilon$, one has*

$$\sum_{j \in M} \left[(p_j - \bar{p}_j)' \pi_j + (x_j - \bar{x}_j)' (\xi_j - \alpha_j \bar{p}_j) \right]$$
$$> \delta + \sum_{j \in M} \alpha_j (x_j - \bar{x}_j)' (p_j - \bar{p}_j), \tag{23}$$

for all $j \in M$ and pairs (π_j, ξ_j) satisfying

$$\pi_j \in \partial_{x_j} H_j(x, p_j) \text{ and } \xi_j \in -\partial_{p_j} H_j(x, p_j).$$

Remark 2.15. If $\mathbf{x}$ is an M-program with an associated M-price schedule $\mathbf{p}$ such that

$$\|x^k - \bar{x}\| + \|p^k - \bar{p}\| > \epsilon,$$

then making the substitutions $x_j = x_j^k$, $p_j = p_j^{k+1}$, $\pi_j = \Delta x_j^k$, and $\xi_j = \Delta x_j^k$ in (23) immediately shows that the steady state $(\bar{x}, \bar{p})$ satisfies (20). This leads immediately to the following result.

THEOREM 2.16. *Assume that $(\bar{x}, \bar{p})$ is a unique steady-state equilibrium that has the strong diagonal support property given by* (23). *Then for any M-program $\mathbf{x}$ with an associated M-price schedule $\mathbf{p}$ that satisfies*

$$\limsup_{k \to \infty} \| (x^k, p^k) \| < \infty,$$

we have

$$\lim_{k\to\infty} \| \left(x^k - \bar{x}, p^k - \bar{p}\right) \| = 0.$$

Proof. Theorem 2.16 follows immediately from the above remark and Theorem 2.12. □

In a discrete time infinite-horizon optimal control problem with discount factor $\beta \in (0, 1]$, a steady-state equilibrium $\bar{x}$ is assured to be an attractor of overtaking equilibria by requiring the Hamiltonian of the system to be a-concave in x and b-convex in p for values of $a > 0$ and $b > 0$ for which the inequality

$$\left(1 - \frac{1}{\beta}\right)^2 < 4ab$$

holds (see, e.g., Rockafellar [24] or Brock and Scheinkman [6]). We conclude this section by extending this result to the case considered here.

DEFINITION 2.17. *Let $a = (a_1, a_2, \ldots, a_m)$ and $b = (b_1, b_2, \ldots, b_m)$ be two vectors in $\mathbb{R}^m$ with $a_j > 0$ and $b_j > 0$ for all $j \in M$. We say that the combined Hamiltonian $\sum_{j\in M} H_j\left(x, p_j\right)$ is strictly diagonally a-concave in x, b-convex in p if*

$$\sum_{j\in M}\left[H_j(x, p_j) + \frac{1}{2}\left(a_j\|x_j\|^2 - b_j\|p_j\|^2\right)\right]$$

is strictly diagonally concave in x, convex in p.

THEOREM 2.18. *Assume that there exists a unique steady-state equilibrium, $(\bar{x}, \bar{p})$. Let $a = (a_1, a_2, \ldots, a_m)$ and $b = (b_1, b_2, \ldots, b_m)$ be two vectors in $\mathbb{R}^m$ with $a_j > 0$ and $b_j > 0$ for all $j \in M$, and assume that the combined Hamiltonian is strictly diagonally a-concave in x, b-convex in p. Further, let* **x** *be a bounded equilibrium M-program with an associated M-price schedule* **p** *that also remains bounded. Then, if the discount rates β_j, $j \in M$, satisfy the inequalities*

$$\alpha_j^2 = \left(1 - \frac{1}{\beta_j}\right)^2 < 4a_jb_j, \tag{24}$$

the M-program **x** *converges to $\bar{x}$.*

Proof. Let **x** be a bounded M-program and let **p** be its associated M-price schedule (which is also bounded). Then we have the following inclusions for all $k \in \mathbb{N}$ and $j \in M$:

$$\Delta x_j^k - b_j p^{k+1} \in \partial_{p_j}\left[H_j(x^k, p_j^{k+1}) + \frac{1}{2}\left(a_j\|x_j^k\|^2 - b_j\|p_j^{k+1}\|^2\right)\right],$$

$$\Delta p_j^k + \alpha_j p_j^k - a_j x_j^k \in -\partial_{x_j}\left[H_j(x^k, p_j^{k+1}) + \frac{1}{2}\left(a_j\|x_j^k\|^2 - b_j\|p_j^{k+1}\|^2\right)\right],$$

$$-b_j\bar{p} \in \partial_{p_j}\left[H_j(\bar{x}, \bar{p}_j) + \frac{1}{2}\left(a_j\|\bar{x}_j\|^2 - b_j\|\bar{p}_j\|^2\right)\right],$$

$$\alpha_j\bar{p}_j - a_j\bar{x}_j \in -\partial_{x_j}\left[H_j(\bar{x}, \bar{p}_j) + \frac{1}{2}\left(a_j\|\bar{x}_j\|^2 - b_j\|\bar{p}_j\|^2\right)\right].$$

Thus, as a consequence of our SDCCAs, we have for each $k \in \mathbb{N}$ that

$$0 < \sum_{j\in M}\left[\left(p_j^{k+1} - \bar{p}_j\right)'\left(\Delta x_j^k - b_j p^{k+1} + b_j\bar{p}\right)\right.$$

$$\left. + \left(x_j^k - \bar{x}_j\right)'\left(\Delta p_j^k + \alpha_j p_j^{k+1} - a_j x_j^k - \left(\alpha_j\bar{p}_j - a_j\bar{x}_j\right)\right)\right]$$

$$= \sum_{j\in M}\left[\left(p_j^{k+1}-\bar{p}_j\right)'\Delta x_j^k + \left(x_j^k-\bar{x}_j\right)'\Delta p_j^k\right]$$
$$- \sum_{j\in M}\left[b_j\|p_j^{k+1}-\bar{p}_j\|^2 - \alpha_j\left(p_j^{k+1}-\bar{p}_j\right)'\left(x_j^k-\bar{x}_j\right) + a_j\|x_j^k-\bar{x}_j\|^2\right].$$

Summing from $k=0$ to $k=K$ we may write

$$V(K) > \sum_{k=0}^{K}\sum_{j\in M}\left[b_j\|p_j^{k+1}-\bar{p}_j\|^2 - \alpha_j\left(p_j^{k+1}-\bar{p}_j\right)'\left(x_j^k-\bar{x}_j\right) + a_j\|x_j^k-\bar{x}_j\|^2\right],$$

where, as before,

$$V(K) = \sum_{k=0}^{K}\sum_{j\in M}\left[\left(p_j^{k+1}-\bar{p}_j\right)'\Delta x_j^k + \left(x_j^k-\bar{x}_j\right)'\Delta p_j^k\right].$$

This inequality may be equivalently written as

$$V(K) > \sum_{k=0}^{K}\sum_{j\in M} b_j\left\|\left(p_j^{k+1}-\bar{p}_j\right) - \frac{\alpha_j}{2b_j}\left(x_j^k-\bar{x}\right)\right\|^2$$
$$- \left(\frac{\alpha_j^2}{4b_j} - a_j\right)\|x_j^k-\bar{x}\|^2$$
$$> \sum_{k=0}^{K}\sum_{j\in M}\left(a_j - \frac{\alpha_j^2}{4b_j}\right)\|x_j^k-\bar{x}\|^2.$$

To conclude this proof we see, from above, that if the M-program $\mathbf{x}$ does not converge to $\bar{x}$, then $\lim_{K\to\infty} V(K) = \infty$ because of (24). This, however, is a contradiction since we may equivalently write

$$V(K) = \sum_{j\in M}\left(\left(p_j^{K+1}-\bar{p}_j\right)'\left(x_j^{K+1}-\bar{x}_j\right) - \left(p_j^0-\bar{p}_j\right)'\left(x_j^0-\bar{x}_j\right)\right),$$

showing that $V(K)$ must remain bounded for all K. □

Remark 2.19. In the above results we have extended the classical asymptotic turnpike theory to a dynamic game framework with separated dynamics. The fact that each firm controls a distinct dynamical system and that the coupling between the players occurs only through the rewards is essential. An indication of the increased complexities of coupled dynamics is given in Haurie and Leitmann [18].

Remark 2.20. The assumption of the existence and uniqueness of the steady-state equilibrium $(\bar{x}, \bar{p})$ requires the solution of a nonlinear system of inclusions. In the single-player case, the study of solutions of these systems was considered by Rockafellar [24] (see in particular Theorem 5). In particular, he showed that if the Hamiltonian is a-concave in x and b-convex in p, and if there exists a unique steady-state equilibrium for the undiscounted case $\alpha = 0$ (i.e., $\beta = 1$), then for all $\alpha > 0$ sufficiently small there exists a unique steady-state equilibrium pair $(\bar{x}_\alpha, \bar{p}_\alpha)$. This theory has yet to be developed for the multi-player case considered here. In addition, the relationship between the steady-state equilibrium and the solution of a related family of mathematical programming problems was studied, again in the single-player case, by Feinstein and Luenberger [14]. In that work the "implicit programming problem" is defined and the solutions to this optimization problem are shown to correspond to a steady-state equilibrium pair. For the multi-player case, we direct the reader to [19] in which an "implicit static equilibrium" is defined with properties analogous to those presented in [14].

3. Existence of equilibria. In this section we extend Rosen's approach to show existence of equilibria in dynamic autonomous competitive processes, under sufficient smoothness and compactness conditions. Basically we reduce the existence proof to a fixed-point argument for a point-to-set mapping constructed from an associated class of infinite-horizon concave optimization problems.

3.1. Existence of overtaking equilibria in the undiscounted case. Our proof of existence of an overtaking equilibrium for undiscounted dynamic competitive processes uses extensively sufficient overtaking optimality conditions for single-player optimization problems (see [7, Chap. 2]). For this appeal to sufficiency conditions the existence of a bounded attractor to all good programs is important. This is the reason why our existence theory is restricted to autonomous systems for which a steady-state equilibrium provides such an attractor.

Remark 3.1. Existence of overtaking optimal control for autonomous systems (discrete or continuous time) can be established through a reduction to finite cost argument (see, e.g., [7]). There is a difficulty in extending this approach to the case of dynamic open-loop games. It comes from the inherent time-dependency introduced by the other players' decisions. Our approach circumvents this difficulty by implementing a reduction to finite costs for an associated class of infinite-horizon concave optimization problems.

We first make the following assumptions.

Assumption 3.2. The functions $L_j : \mathbb{R}^n \times \mathbb{R}^{n_j} \to \mathbb{R}$ are strictly concave in (x_j, z_j) and additionally we have that $\frac{\partial L_j}{\partial x_j}$ and $\frac{\partial L_j}{\partial z_j}$ are continuous on $\mathcal{R}^n \times \mathbb{R}^{n_j}$ for each $j \in M$.

Assumption 3.3. There exists a unique steady-state equilibrium $\bar{x} \in \mathbb{R}^n$ and a corresponding constant M-price schedule $\bar{p} \in \mathbb{R}^n$ satisfying

$$\begin{aligned} 0 &\in \partial_{p_j} H_j(\bar{x}, \bar{p}_j), \\ 0 &\in \partial_{x_j} H_j(\bar{x}, \bar{p}_j). \end{aligned} \tag{25}$$

To achieve our result we must assure that all admissible M-programs lie in a compact set. Thus we make the following additional assumption.

Assumption 3.4. For each $j \in M$ there exists a closed bounded set $X_j \subset \mathbb{R}^{n_j}$ such that each M-program, $\mathbf{x}$, satisfies $x_j^k \in X_j$ for each $k \in \mathbb{N}$. Additionally we introduce the following notation.

(i) We let Ω denote the set of all M-programs that start at x^o and converge to $\bar{x}$, the unique steady-state equilibrium.

(ii) We define the family of functionals $\rho^K : \Omega \times \Omega \to \mathbb{R}$, $K \in \mathbb{N}$, by the formula

$$\rho^K(\mathbf{x}, \mathbf{y}) \doteq \sum_{k=0}^{K} \sum_{j \in M} \left[L_j \left(\left[\mathbf{x}^j, y_j \right]^k, \Delta y_j^k \right) \right].$$

We view Ω as a subset of all bounded sequences in $\mathbb{R}^n$ endowed with the topology of pointwise convergence.

DEFINITION 3.5. *Let* $\mathbf{x}, \mathbf{y} \in \Omega$. *We say that* $\mathbf{y} \in \Gamma(\mathbf{x})$ *if*

$$\liminf_{K \to \infty} \left(\rho^K(\mathbf{x}, \mathbf{y}) - \rho^K(\mathbf{x}, \mathbf{z}) \right) \geq 0,$$

for all M*-programs* $\mathbf{z}$ *such that* $z_j^0 = x_j^0$. *That is,* $\mathbf{y}$ *is an overtaking optimal solution of the infinite-horizon optimization problem whose objective functional is defined by* $\rho^K(\mathbf{x}, \cdot)$. *Hence* $\Gamma(\mathbf{x})$ *can be viewed as the set of* optimal responses *by all players to an* M*-program* $(\mathbf{x})$.

THEOREM 3.6. *Under the above assumptions, there exists an overtaking equilibrium for the infinite-horizon dynamic game.*

Proof. To prove our result we prove that the set-valued map $\Gamma : \Omega \to 2^{\Omega}$ has a fixed point using the Kakutani fixed-point theorem. To do this we need to show that

1. for each $x \in \Omega$, the set $\Gamma(\mathbf{x})$ is nonempty, convex, and compact.
2. the map $\Gamma(\cdot)$ has a closed graph. That is, if $(\mathbf{y}_l, \mathbf{x}_l)$ is a sequence in $\Omega \times \Omega$ that converges to $(\mathbf{y}, \mathbf{x})$ and additionally satisfies

$$\mathbf{y}_l \in \Gamma(\mathbf{x}_l)$$

for all $l = 1, 2, \ldots$, then $\mathbf{y} \in \Gamma(\mathbf{x})$.

We begin by first showing, for each $\mathbf{x} \in \Omega$, that $\Gamma(\mathbf{x})$ is nonempty. To see this we fix $\mathbf{x} \in \Omega$, let $(\theta_l)_{l=1}^{\infty}$ be an unbounded strictly increasing sequence of positive integers, and define for each $l = 1, 2, \ldots$, the M-programs $\mathbf{x}_l$ as

$$x_l^k = \begin{cases} x^k & \text{if } k < \theta_l, \\ \bar{x} & \text{if } k \geq \theta_l, \end{cases}$$

for $k = 1, 2, \ldots$. Now consider the problem of maximizing the functional $\gamma(\mathbf{x}_l, \cdot)$ defined by the formula

$$\begin{aligned} \gamma(\mathbf{x}_l, \mathbf{z}) &\doteq \sum_{k=0}^{\infty} \sum_{j \in M} \left[L_j \left(\left[\mathbf{x}_l^j, z_j \right]^k, \Delta z_j^k \right) - L_j(\bar{\mathbf{x}}, 0) + \bar{p}_j \Delta z_j^k \right] \\ &= \sum_{k=0}^{\theta_l - 1} \sum_{j \in M} \left[L_j \left(\left[\mathbf{x}_l^j, y_j \right]^k, \Delta y_j^k \right) - L_j(\bar{\mathbf{x}}, 0) + \bar{p}_j \Delta y_j^k \right] \\ &\quad + \sum_{k=\theta_l}^{\infty} \sum_{j \in M} \left[L_j \left(\left[\bar{\mathbf{x}}^j, y_j \right]^k, \Delta y_j^k \right) - L_j(\bar{\mathbf{x}}, 0) + \bar{p}_j \Delta y_j^k \right], \end{aligned}$$

over all M-programs starting at x^o. Observe that for each $k > \theta_l$, the terms of $\gamma(\mathbf{x}_l, \mathbf{z})$ are nonpositive and in fact equal zero whenever we take $z_j^k = \bar{x}_j$. From this it follows easily that there exists an M-program, say $\mathbf{z}$, for which $\gamma(\mathbf{x}_l, \mathbf{z}) > -\infty$ and, moreover, since each term of an M-program lies in a compact subset of $\mathbb{R}^n$, it also follows that $\gamma(\mathbf{x}_l, \mathbf{z})$ is bounded above. Therefore, there exists an M-program $\mathbf{y}_l$ that maximizes $\gamma(\mathbf{x}_l, \mathbf{z})$ as desired. Furthermore, as a result of the strict concavity assumptions given above, it follows that $\mathbf{y}_l \in \Gamma(\mathbf{x}_l)$ and we also have

$$\lim_{k \to \infty} y_l^k = \bar{x}.$$

This asymptotic stability property was proved by Brock in [3] (see also [7]). In this way we generate the sequence $(\mathbf{y}_l)$ in Ω, a compact set. Thus we can assume that the sequence $(\mathbf{y}_l)$ converges to an M-program $\mathbf{y} \in \Omega$. Additionally, we also know that the sequence $(\mathbf{x}_l)$ converges to our original M-program $\mathbf{x}$. We now show that $\mathbf{y} \in \Gamma(\mathbf{x})$. To this end we note that for each $l = 1, 2, \ldots$, there exists an M-price schedule $\mathbf{p}_l$ such that

$$\begin{aligned} \Delta y_{lj}^k &\in \partial_{p_j} H_j \left([x_l^{(j)}, y_{lj}]^k, p_{lj}^{k+1} \right), \\ \Delta p_{lj}^k &\in -\partial_{x_j} H_j \left([x_l^{(j)}, y_{lj}]^k, p_{lj}^{k+1} \right) \end{aligned}$$

or equivalently that

$$\begin{aligned} p_{lj}^{k+1} &= -\frac{\partial}{\partial z_j} L_j \left([x_l^{(j)}, y_{lj}]^k, \Delta y_{lj}^k \right), \\ \Delta p_{lj}^k &= -\frac{\partial}{\partial x_j} L_j \left([x_l^{(j)}, y_{lj}]^k, \Delta y_{lj}^k \right). \end{aligned} \tag{26}$$

Holding $k \in \mathbb{N}$ fixed and letting $l \to \infty$ in the above gives us that

$$p_j^{k+1} \doteq \lim_{l\to\infty} p_{lj}^{k+1} = -\frac{\partial}{\partial z_j} L_j\left([x^{(j)}, y_j]^k, \Delta y_j^k\right) \tag{27}$$

and

$$\Delta p_j^k = -\frac{\partial}{\partial x_j} L_j\left([x^{(j)}, y_j]^k, \Delta y_j^k\right). \tag{28}$$

That is, we can associate with $\mathbf{y} \in \Omega$ an M-price schedule $\mathbf{p}$ and, moreover, since $\mathbf{x}, \mathbf{y} \in \Omega$ we have $\lim_{k\to\infty} p_j^k = \bar{p}_j$ as well. Therefore, by appealing to standard sufficient conditions for overtaking optimality we obtain $\mathbf{y} \in \Gamma(\mathbf{x})$ as desired. Further it is easy to see that, as a result of our concavity assumptions on L_j, $\Gamma(\mathbf{x})$ is a convex set.

Since $\Gamma(\mathbf{x})$ is a subset of Ω, a compact set, the compactness of $\Gamma(\mathbf{x})$ follows immediately once we show that it is closed. To see this let $(\mathbf{y}_l)$ be a sequence in $\Gamma(\mathbf{x})$ that converges to $\mathbf{y}$ and let $(\mathbf{p}_l)$ be the sequence of corresponding M-price schedules. By a direct adaptation of the above argument it can be shown that there is an M-price schedule $\mathbf{p}$ associated with $\mathbf{y}$ that converges to $\bar{p}$ as $k \to \infty$. Hence $\mathbf{y} \in \Gamma(\mathbf{x})$, giving us the desired compactness condition.

It remains to prove that the graph of Γ is closed. To see this let $(\mathbf{y}_l, \mathbf{x}_l) \to (\mathbf{y}, \mathbf{x})$ as $l \to \infty$ be such that $\mathbf{y}_l \in \Gamma(\mathbf{x}_l)$ for each $l \in \mathbb{N}$. Further we let $(\mathbf{p}_l)$ be a sequence of associated M-price schedules, satisfying (26). Then, once again, letting $l \to \infty$ one sees that there exists an M-price schedule $\mathbf{p}$ associated with $\mathbf{x}$ such that $\mathbf{p}_l \to \mathbf{p}$. Thus proceeding as above we see that $\mathbf{y} \in \Gamma(\mathbf{x})$, which gives us the closed graph property.

We have just shown that the conditions needed to apply Kakutani's fixed-point theorem are satisfied. Therefore there exists $\mathbf{x}^* \in \Gamma(\mathbf{x}^*)$. The proof that the M-program $\mathbf{x}^*$ is an overtaking equilibrium now follows as in Rosen [26, Thm. 1]. □

3.2. Existence of equilibria in discounted competitive processes. Under the assumptions made above to establish the existence of an overtaking equilibrium, it is easy to treat the case of autonomous games with discounting. Indeed, by requiring all admissible M-programs to satisfy the compact constraints $x_j^k \in X_j$ for all $j \in M$ and $k \in \mathbb{N}$ and having $\beta_j \in (0, 1)$, it follows that the functionals $\rho^K : \Omega \times \Omega \to \mathbb{R}$, $K \in \mathbb{N}$, now given by

$$\rho^K(\mathbf{x}, \mathbf{y}) \doteq \sum_{k=0}^{K} \sum_{j\in M} \left[(\beta_j)^k L_j\left(\left[\mathbf{x}^j, y_j\right]^k, \Delta y_j^k\right)\right]$$

are bounded both above and below for all $K \in \mathbb{N}$. Thus, the existence of an equilibrium can be assured by an easy modification of the above argument. To do this we enlarge Ω to be the set of all M-programs satisfying the initial condition x^o and now define $\Gamma : \Omega \to 2^\Omega$ as follows:

> For an M-program $\mathbf{x} \in \Omega$ we let $\Gamma(\mathbf{x})$ denote the set of all M-programs $\mathbf{y} \in \Omega$ such that
>
> $$\rho^K(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{z}\in\Omega} \rho^K(\mathbf{x}, \mathbf{z}).$$

Further, the positive discounting eliminates the need to define the associated problems using the functional $\gamma(\cdot, \cdot)$. Therefore we state the following result without proof.

THEOREM 3.7. *Let $\beta_j > 0$ for $j \in M$ and let Assumptions* 3.2–3.4 *hold. Then there exists an overtaking equilibrium for the discounted autonomous infinite-horizon dynamic game.*

Remark 3.8. Both of the existence results given above are for autonomous dynamic games only. We observe however that our arguments would be applicable in the nonautonomous case

if one could assume that the sets $\Gamma(\mathbf{x})$ are nonempty. To make this conclusion would require one to appeal to corresponding nonautonomous existence theorems for optimality (overtaking optimality in the discounted case). This of course would require some growth restrictions on the functions $L_j^k(\cdot,\cdot)$ (e.g., $|L_j^k(x, z_j)| \le \Lambda_j$ for all (x, z_j), $j \in M$, and $k \in \mathbb{N}$).

4. Uniqueness of equilibria. In Rosen's paper [26], the strict diagonal concavity condition was introduced essentially to assure uniqueness of equilibria. In §2 we introduced a similar assumption to get asymptotic stability of equilibrium programs. Indeed this condition also leads to uniqueness.

THEOREM 4.1. *If the assumptions of Theorem* 2.8 *hold and there exists an overtaking equilibrium at* x^o, *then it is unique.*

Proof. Assume that $\hat{\mathbf{x}}$ and $\tilde{\mathbf{x}}$ are two distinct equilibria at x^o, with their respective M-price schedules $\hat{\mathbf{p}}$ and $\tilde{\mathbf{p}}$. According to the necessary conditions, for all $k = 0, 1, \ldots$, there exist

$$\hat{\pi}_j^k \in \partial_{p_j} H_j^k(\hat{x}, \hat{p}_j), \quad \tilde{\pi}_j^k \in \partial_{p_j} H_j^k(\tilde{x}, \tilde{p}_j), \tag{29}$$

$$\hat{\xi}_j^k \in -\partial_{x_j} H_j^k(\hat{x}, \hat{p}_j), \quad \tilde{\xi}_j^k \in -\partial_{x_j} H_j^k(\tilde{x}, \tilde{p}_j) \tag{30}$$

such that

$$\sum_{j\in M}\Big[(\hat{p}_j^{k+1} - \tilde{p}_j^{k+1})'(\hat{\pi}_j^k - \Delta\hat{x}_j^k - \tilde{\pi}_j^k + \Delta\tilde{x}_j^k) + (\hat{x}_j^k - \tilde{x}_j^k)'(\hat{\xi}_j^k - \Delta\hat{p}_j^k - \tilde{\xi}_j^k + \Delta\tilde{p}_j^k)\Big] = 0. \tag{31}$$

Adding over $k = 0, \ldots, K$ we get

$$\sum_{k=0}^{K}\sum_{j\in M}\Big[(\Delta\hat{p}_j^k - \Delta\tilde{p}_j^k)'(\hat{x}_j^k - \tilde{x}_j^k) + (\hat{p}_j^{k+1} - \tilde{p}_j^{k+1})'(\Delta\hat{x}_j^k - \Delta\tilde{x}_j^k)\Big]$$

$$= \sum_{k=0}^{K}\sum_{j\in M}\Big[(\hat{p}_j^{k+1} - \tilde{p}_j^{k+1})'(\hat{\pi}_j^k - \tilde{\pi}_j^k) + (\hat{x}_j^k - \tilde{x}_j^k)'(\hat{\xi}_j^k - \tilde{\xi}_j^k)\Big]. \tag{32}$$

Now, when $K \to \infty$, the right-hand side of (32) tends to a strictly positive number, due to SDCCA, whereas the left-hand side, which collapses to

$$\sum_{j\in M}\Big[(\hat{p}_j^{K+1} - \tilde{p}_j^{K+1})'(\hat{x}_j^{K+1} - \tilde{x}_j^{K+1}) + (\hat{p}_j^0 - \tilde{p}_j^0)'(\hat{x}_j^0 - \tilde{x}_j^0)\Big], \tag{33}$$

goes to zero, due to the turnpike property. This is a contradiction; hence the overtaking equilibrium is unique. □

5. Example. We consider the following class of games as an example where the conditions stated in this paper are easy to check. We assume that $L_j : \mathbb{R}^n \times \mathbb{R}^{n_j} \to \mathbb{R}$ has the form

$$L_j(x, z_j) = g_j(x_j, z_j) + G_j(x),$$

for $j = 1, 2, \ldots m$. Thus the accumulated reward for the jth player over $[0, K]$, $K > 0$ an integer, is given as

$$\phi_j^K(\mathbf{x}) = \sum_{k=0}^{K-1}(\beta_j)^k\left[g_j(x_j^k, \Delta x_j^k) + G_j(x^k)\right].$$

In this formulation there is a separation between the control of each player and the state variables of the other players. We assume that the functions $g_j(\cdot, \cdot)$ are concave and that the combined G's are strictly diagonally concave. In this way it is easy to see that the combined Hamiltonians, defined by

$$\begin{aligned} H_j(x, p_j) &= \sup_{z \in \mathbb{R}^{n_j}} \{L_j(x, z) + p_j \cdot z\} \\ &= \sup_{z \in \mathbb{R}^{n_j}} \{g_j(x_j, z) + p_j \cdot z\} + G_j(x) \\ &= \tilde{H}_j(x_j, p_j) + G_j(x), \end{aligned}$$

are strictly diagonally concave in x and convex in p since each of the Hamiltonians, $\tilde{H}$, are concave in x_j and convex in p_j. We further observe that the Hamiltonian dynamical system that is necessary for an equilibrium becomes, when $\beta_j \equiv 1$,

$$\begin{aligned} \Delta x_j^k &\in \partial_{p_j} \tilde{H}_j(x_j^k, p_j^k), \\ \Delta p_j^k + \nabla_{x_j} G_j(x^k) &\in -\partial_{x_j} \tilde{H}_j(x_j^k, p_j^k). \end{aligned}$$

The conditions for a steady-state equilibrium become

$$\begin{aligned} 0 &\in \partial_{p_j} \tilde{H}_j(\bar{x}_j, \bar{p}_j), \\ -\nabla_{x_j} G_j(\bar{x}) &\in -\partial_{x_j} \tilde{H}_j(\bar{x}_j, \bar{p}). \end{aligned}$$

Thus to insure that the turnpike property holds, we require this steady state $\bar{x}$ to be *strongly diagonally supported* by $\bar{p}$. That is, for each $\epsilon > 0$ there exists $\delta > 0$ so that if $\|x - \bar{x}\| + \|p - \bar{p}\| > \epsilon$, one has

$$\sum_{j \in M} \left[(p_j - \bar{p}_j)' \pi_j + (x_j - \bar{x}_j)' \eta_j\right] + \sum_{j \in M} (x_j - \bar{x}_j)' (\nabla_{x_j} G_j(x) - \nabla_{x_j} G_j(\bar{x})) > \delta$$

for all (π_j, η_j), $j \in M$, satisfying

$$\begin{aligned} \pi_j &\in \partial_{p_j} \tilde{H}_j(x_j, p_j), \\ \eta_j &\in -\partial_{x_j} \tilde{H}_j(x_j, p_j). \end{aligned}$$

This condition is insured if the steady-state equilibrium is unique and the programs remain in a compact set. Indeed this is a consequence of the strict diagonal concavity for the combined G's and of the concavity–convexity property of each $\tilde{H}_j$. The fact that the strong support property is a consequence of compactness and strict diagonal concavity would be an easy adaptation of Atsumi's lemma (see [7]).

6. Conclusion. In this paper we have shown that, for infinite-horizon competitive processes defined over dynamical systems where the players only interact through the payoff functionals but control their own separate dynamics, an assumption very similar to the one made by Rosen in his seminal study of uniqueness of equilibria in static concave games brings together global asymptotic stability of equilibrium trajectories, existence, and uniqueness of overtaking equilibria at each initial state. This extends the so-called turnpike theory of optimal economic processes to a competitive case.

Acknowledgment. We would like to acknowledge the fruitful discussions with and comments from our colleague Professor Arie Leizarowitz throughout the development of this work.

REFERENCES

[1] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory*, Academic Press, London, New York, 1982.
[2] M. Breton, J. A. Filar, A. Haurie, and T. A. Schultz, *On the computation of equilibria in discounted stochastic dynamic games*, in Dynamic Games and Applications in Economics, Lecture Notes Econom. and Math. Systems 265, T. Başar, ed., Springer-Verlag, 1986, pp. 64–87.
[3] W. A. Brock, *On existence of weakly maximal programmes in a multisector economy*, Rev. Econom. Stud., 37 (1970), pp. 275–280.
[4] ———, *Differential games with active and passive variables*, in Mathematical Economics and Game Theory: Essays in Honor of Oskar Morgenstern, R. Henn and O. Moeschlin, eds. Springer-Verlag, Berlin, 1977, pp. 34–52.
[5] W. A. Brock and A. Haurie, *On existence of overtaking optimal trajectories over an infinite time horizon*, Math. Oper. Res., 1 (1976), pp. 337–346.
[6] W. A. Brock and J. A. Scheinkman, *Global asymptotic stability of optimal control systems with application to the theory of economic growth*, J. Econom. Theory, 12 (1976), pp. 164–190.
[7] D. A. Carlson, A. Haurie, and A. Leizarowitz, *Infinite Horizon Optimal Control: Deterministic and Stochastic Systems*, Springer-Verlag, New York, 1991.
[8] ———, *Overtaking Equilibria for Switching Regulator and Tracking Games*, in Advances in Dynamic Games and Applications, Annals of the International Society of Dynamic Games, vol. 1, T. Başar and A. Haurie, eds., Birkhäuser, Boston, 1994, pp. 247–268.
[9] D. Cass, *Optimum growth in aggregative model of capital accumulation: A turnpike theorem*, Econometrica, 34 (1965), pp. 833–850.
[10] ———, *Optimum growth in aggregative model of capital accumulation*, Rev. Econom. Stud., 32 (1965), pp. 233–240.
[11] D. Cass and K. Shell, *The Hamiltonian Approach to Dynamic Economics*, Academic Press, New York, 1976.
[12] ———, *The structure and stability of competitive dynamical systems*, J. Econom. Theory, 12 (1976), pp. 30–70.
[13] A. Cournot, *Recherches sur les principes mathématiques de la théorie des richesses*, Hachette, Paris, 1838.
[14] C. D. Feinstein and D. G. Luenberger, *Analysis of the asymptotic behaviour of optimal control trajectories: The implicit programming problem*, SIAM J. Control Optim., 19 (1981), pp. 561–585.
[15] J. W. Friedman, *Oligopoly and the Theory of Games*, North-Holland, Amsterdam, 1977.
[16] D. Fudenberg and J. Tirole, *Game Theory*, MIT Press, Cambridge, MA, 1992.
[17] A. Haurie, *Existence and global asymptotic stability of optimal trajectories for a class of infinite-horizon, nonconvex systems*, J. Optim. Theory Appl., 31 (1980), pp. 515–533.
[18] A. Haurie and G. Leitmann, *On the global stability of equilibrium solutions for open-loop differential games*, Large Scale Systems, 6 (1984), pp. 107–122.
[19] A. Haurie and M. Roche, *Turnpikes and computation of piecewise open-loop equilibria in stochastic differential games*, J. Econom. Dynamics Control, 18 (1993), pp. 317–344.
[20] A. Haurie and B. Tolwinski, *Definition and properties of cooperative equilibria in a two-player game of infinite duration*, J. Optim. Theory Appl., 46 (1985), pp. 525–534.
[21] L. W. McKenzie, *Turnpike theory*, Econometrica, 44 (1976), pp. 841–866.
[22] J. F. Nash, *Non-cooperative games*, Ann. of Math., 54 (1951), pp. 286–295.
[23] R. T. Rockafellar, *Saddle points of Hamiltonian Systems in Convex Problems of Lagrange*, J. Optim. Theory Appl., 12 (1973), pp. 367–399.
[24] ———, *Saddle points of Hamiltonian systems in convex Lagrange problems having nonzero discount rate*, J. Econom. Theory, 12 (1976), pp. 71–113.
[25] A. Rubinstein, *Equilibrium in supergames with the overtaking criterion*, J. Econom. Theory, 21 (1979), pp. 1–9.
[26] J. B. Rosen, *Existence and uniqueness of equilibrium points for concave N-person games*, Econometrica, 33 (1965), pp. 520–534.
[27] R. Selten, *Reexamination of the perfectness concept for equilibrium points in extensive games*, Internat. J. Game Theory, 4 (1975), pp. 25–55.
[28] C. C. Von Weizäcker, *Existence of optimal programs of accumulation for an infinite time horizon*, Rev. Econom. Stud., 32 (1965), pp. 85–104.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1420–1440, July 1996

017

THE GENERALIZED SOLUTIONS OF NONLINEAR OPTIMIZATION PROBLEMS WITH IMPULSE CONTROL*

BORIS M. MILLER†

Abstract. The optimal control problem described by a nonlinear ordinary differential equation is considered with unbounded controls. The aim of this paper is to provide a representation of generalized (discontinuous) solutions in terms of differential equations with a measure. By using the method of discontinuous time change the problem with nonlinear dependence on the unbounded controls can be reduced to the simpler one with bounded controls. These results are applied to solving the problem of generalized (discontinuous) solution representation in the dynamic system with impulse-type controls.

Key words. nonlinear systems, impulse control, generalized solutions, discontinuous time change

AMS subject classification. 49B05

1. Introduction.

In this paper we discuss the control of a nonlinear dynamic system whose state is governed by the nonlinear differential equation

$$\dot{x}(t) = F(x(t), u(t), w(t), t), \tag{1}$$

where $F(x(t), u(t), w(t), t)$ is a given function, $x(t) \in R^n$, $t \in [0, T]$, $x(t) = x_0$ is an initial condition, and $u(t)$ and $w(t)$ are measurable controls on $[0, T]$: $u(t) \in R^k$ is an ordinary control component, and $w(t) \in R^m$ is a generalized one. The former component corresponds to the bounded control

$$u(t) \in U \subset R^k, \tag{2}$$

where U is a closed and bounded subset of R^k, and the latter corresponds to control which is unbounded in the norm but bounded in the integral sense so that

$$w(t) \in W \subset R^m, \tag{3}$$

$$\int_o^T \|w(t)\| dt \leq M < \infty, \tag{4}$$

where W is a subset of R^m and M is a constant.

The optimization problems with impulse and ordinary controls occur in flight dynamics [5], [7], [8]: for example, the positions of the exhaust vanes of the space vehicle can be considered the bounded control inputs, and the jet flow rate could be impulse (unbounded) controls. Another example of such type of optimization problem is one of observation control [2], [8] for discrete-continuous systems where two different types of observation controls characterize the possibility of controlling both the composition of observations and their timing and density.

Condition (4) means that the admissible controls can be taken as functions that are as close as desired to the impulse or generalized functions. As a result, the solution of the optimal control problem in system (1) with absolutely continuous paths $\{x(\cdot)\}$ under constraints (2)–(4) may not exist within the class of measurable ordinary or relaxed control functions. Thus,

*Received by the editors December 16, 1994; accepted for publication (in revised form) April 25, 1995. This research was supported in part by National Science Foundation grant CMS 94-1447s and International Association for the Promotion of Cooperation with Scientists from the Independent States of the Former Soviet Union (INTAS) grant 94-697.

†Institute for Information Transmission Problems, 19 Bolshoy Karetny per., Moscow 101447, Russia (bmiller@ippi.ac.msk.su).

the concept of a solution should be extended to consider the possible discontinuous behavior of the optimal path under the impulse control action.

If the function $F(x, u, w, t)$ is linear on w, i.e.,

$$F(x, u, w, t) = f(x, u, t) + B(t)w,$$

where the matrix-valued function $B(\cdot)$ is continuous, the concept of solution can be easily extended, interpreting equation (1) as one with measure of the following type [6], [16]:

$$dx(t) = f(x(t), u(t), t)dt + B(t)d\mu(t).$$

However, if the function $B(\cdot)$ is discontinuous or depends on the other control $u(\cdot)$, i.e., $B = B(t, u)$, then the representation of the optimal solution by the above equation can be incorrect without a special "constancy" condition. For the investigation of optimization problems with impulse controls in the system

$$dx(t) = f(x(t), u(t), t)dt + B(t, u(t))d\mu(t),$$

the special method that we refer to as the *method of discontinuous time change* was introduced by Rishel [15]. Rishel's approach is to introduce a new independent variable with respect to which trajectories become absolutely continuous. This leads to the consideration of an auxiliary control problem with bounded controls in which the time variable plays the role of a component of the state variable. This approach enables us to reduce the original optimization problem to the equivalent auxiliary one with bounded controls and to deduce optimality conditions for the original optimal process.

The general approach, based on the idea of replacing time with a new independent variable, was applied to the class of nonlinear systems by Warga [20]–[22], who considered nonlinear systems with an arbitrary dependence on unbounded components of control. Warga's approach provides an effective tool for the investigation of the optimization problems and gives the general representation of optimal paths by the discontinuous time change. We are going to give a further extension of this method, namely, to derive the representation of generalized (discontinuous) solution in optimization problems in terms of differential equations with a measure.

Some results in this area were obtained for special classes of systems, for example, when $B(\cdot)$ depends on the state variable x and t only, i.e., $B = B(x, t)$. If $B(x, t)$ satisfies the Frobenius-type condition [1], [9], [11], [14], [18], it is possible to derive the representation of a generalized solution in the form of a nonlinear differential equation with a measure.

In recent works [10], [12], [13] this representation was obtained for nonlinear systems with the function $F(x, u, w, t)$ in (1) in the form

$$F(x, u, w, t) = f(x, u, v, t) + B(x, u, v, t)w$$

with additional variable v, which satisfies the equation

$$\dot{v} = \|w(s)\|$$

and can be used to describe the dependence of state variables and constraints upon the varying control resource. The complete survey of results concerning the applications of the discontinuous time change method to representation of the generalized solutions and derivation of the optimality conditions can be found in [11].

However, if the function $F(x, u, w, t)$ is nonlinear with respect to w or there are special constraints on variations of state variables, we need a further investigation of this concept to obtain the representation of generalized solutions by differential equation with a measure.

In this paper we consider the concept of a generalized solution as a pointwise limit of some sequence of ordinary solutions on the set of continuity of this generalized solution. The application of Warga's approach [20], [21] to the problem with constraint (4) gives a very natural form of auxiliary system for representation of the generalized solution with an additional variable "new time." In this case the function φ in the equation for the "new time" can be taken in the form $\varphi = 1 + \|w\|$, as in the cases mentioned above [10]–[13]. Then, our next steps in the representation of generalized solution, such as Theorems 3.1, 3.2, and 3.3, follow directly from the results presented in [21] with slight modifications. Section 4 is devoted to the study of representation problem for the right-hand side of the auxiliary system in the "control" form along the lines of [22]. This is the key problem for the representation of the generalized solution by the sampling of some control in the auxiliary system. Here, the main result shows that in the case of the regular behavior of $F(x, u, w, t)$ at infinity, we can obtain this representation of the right-hand side of auxiliary system. Now, in the regular case we can derive the differential equation with a measure for any generalized solution of system (1). Results presented in §5 (Theorems 5.1 and 5.3) yield the representation of any generalized solution in the form of a generalized equation with a measure, that is, similarly to the representation of weak solutions for the controlled systems, described by the ordinary differential equations with a nonconvex right-hand side [6]. Finally, in §6 we return to the original optimization problem to formulate the existence condition for the optimal generalized solution.

2. Problem statement and assumptions. Let the controllable system be described by (1) with the state variable $x(t) \in R^n$ and controls $\{u(t), w(t)\}$ satisfying (2)–(4).

We shall assume that vector function F is continuous with respect to $(x, u, w, t) \in R^{n+k+m+1}$ and that for any $(u, w, t) \in U \times W \times [0, T]$ the function F satisfies a Lipschitz condition, i.e.,

$$\|F(x_1, u, w, t) - F(x_2, u, w, t)\| \leq L_1(1 + \|w\|)(\|x_1 - x_2\|) \tag{5}$$

for any $x_1, x_2 \in R^n$ with constant $L_1 > 0$.

We shall also assume that $F(x, u, w, t)$ has linear growth in both x and w, i.e.,

$$\|F(x, u, w, t)\| \leq L_2(1 + \|x\|)(1 + \|w\|) \tag{6}$$

for any (x, u, w, t) with constant $L_2 > 0$.

Our purpose is to consider the optimal control problem for system (1) with the following performance criterion, which must be minimized:

$$J[x(\cdot), u(\cdot), w(\cdot)] = \varphi_0(x(0), x(t)), \tag{7}$$

where φ_0 is continuous with respect to all the variables, and controls $\{u(\cdot), w(\cdot)\}$ satisfy the constraints (2)–(4) also to the following terminal and phase constraints:

$$h(x(0), x(T)) = 0, \qquad S(x(0), x(T)) \leq 0, \tag{8}$$

$$g(x(t), t) \leq 0 \qquad \text{for} \quad t \in [0, T]. \tag{9}$$

Here h, S, and g are R^{N_1}-, R^{N_2}-, and R^{N_3}-valued continuous functions, respectively, and (8) and (9) can be understood as componentwise relations.

To take into account the possible impulse behavior of the control component $w(\cdot)$ we shall consider the following definition of generalized (discontinuous) solution of system (1).

DEFINITION 2.1. *A right continuous function $x(\cdot)$ of bounded variation on the interval $[0, T]$ is said to be a* generalized solution *of the system* (1) *if there exists a sequence of*

admissible controls $\{u^n(\cdot), w^n(\cdot)\}$ *satisfying the constraints* (2)–(4) *such that the corresponding sequence* $\{x^n(\cdot)\}$ *of solutions of* (1) *with the initial condition* $x^n(0) = x(0-)$ *converges to* $x(\cdot)$ *at all the points of its continuity.*

In other words the generalized solution is a limit of sequence of ordinary solutions in the sense of weak-$\star$ topology in the space of right continuous functions with bounded variation.

A solution of optimization problem will be sought in the class of generalized solutions. We shall require that constraints (8) and (9) hold for a generalized solution, whereas (8) and (9) hold only in the limit for the sequence of ordinary solutions that approximates the former one.

DEFINITION 2.2. *A right continuous function* $x(\cdot)$ *of bounded variation on the interval* $[0, T]$ *is said to be the* admissible generalized solution *of the system* (1) *under the constraints* (8) *and* (9) *if*

(i) $\{x(\cdot)\}$ *satisfies the constraints* (8) *and* (9) *in the sense*

$$h(x(0-), x(T)) = 0, \qquad S(x(0-), x(T)) \le 0,$$
$$g(x(t), t) \le 0 \qquad \text{for} \quad t \in [0, T];$$

(ii) *there exists a sequence of admissible controls* $\{u^n(\cdot), w^n(\cdot)\}$ *satisfying the constraints* (2)–(4) *such that the corresponding sequence* $\{x^n(\cdot)\}$ *of solutions of system* (1) *converges to* $x(\cdot)$ *at all points of continuity and, in addition,*

$$\lim_n x^n(0) = x(0-), \qquad \lim_n x^n(T) = x(T),$$
$$\lim_n \sup_{[0,T]} g_k(x^n(t), t) \le 0 \qquad \textit{for} \quad k = 1, \dots, N_3.$$

DEFINITION 2.3. *An admissible generalized solution* $\{x^0(\cdot)\}$ *is said to be the* optimal generalized solution *if the inequality*

$$\varphi_0(x^0(0-), x^0(T)) \le \varphi_0(x(0-), x(T))$$

takes place for any admissible generalized solution $\{x(\cdot)\}$.

Remark 2.1 These definitions are in accord with the definition of the solution of general optimization problem formulated in [21, Thm. VI.4.2].

Our purpose is to characterize the generalized solution with the aid of auxiliary control system by the method of the discontinuous time change. The next step is to derive the equation for the generalized solution in the form of special differential inclusion with a measure and, finally, to prove the existence theorem for the solution of the optimization problem.

3. Generalized solutions and their representation by discontinuous time change. The application of discontinuous time change for the representation of generalized solutions of system (1) needs the consideration of some auxiliary controlled system whose right-hand side is bounded for arbitrary state variables.

Consider the auxiliary controllable system of differential equations for the pair $y(s) \in R^n$ and $\eta(s) \in R^1$, defined on the interval $[0, T_1]$, where $T_1 \le T + M$:

$$(10) \qquad \begin{pmatrix} \dot{y}(s) \\ \dot{\eta}(s) \end{pmatrix} = \begin{pmatrix} \dfrac{F(y(s), u_1(s), w_1(s), \eta(s))}{1 + \|w_1(s)\|} \\ \dfrac{1}{1 + \|w_1(s)\|} \end{pmatrix} = \tilde{F}(y(s), u_1(s), w_1(s), \eta(s))$$

with the initial conditions

$$y(0) = x(0), \qquad \eta(0) = 0$$

and the controls in the form of functions $u_1(s)$ and $w_1(s)$, which satisfy the constraints

$$u_1(s) \in U, \qquad w_1(s) \in W. \tag{11}$$

In the auxiliary system (10) and (11) the function $F(x, u, w, t)$, sets U and W, and constants T and M are the same as in (1)–(4).

Remark 3.1. This choice of auxiliary system corresponds to the one of function φ [20], [21], which is a "rate" of time change, in the form

$$\varphi(x, u, w, t) = 1 + \|w\|.$$

The right-hand side of (10) satisfies a Lipschitz condition with respect to y, and it has linear growth in the usual sense, as follows from the inequalities (5) and (6).

There exists an interconnection between the systems (1) and (10) and their solutions, specified by the following theorems.

THEOREM 3.1. *Let the triple of functions* $\{x(\cdot), u(\cdot), w(\cdot)\}$ *satisfy* (1) *and the controls* $\{u(\cdot), w(\cdot)\}$ *be measurable and satisfy the constraints* (2)–(4).

Then there exist the set of functions $\{y(\cdot), \eta(\cdot), u_1(\cdot), w_1(\cdot)\}$ *defined on some interval* $[0, T_1]$, *where*

$$\eta(T_1) = T,$$

which satisfy the system (10) *with the initial condition* $y(0) = x(0)$, *and the measurable controls* $\{u_1(\cdot), w_1(\cdot)\}$, *which satisfy the constraints* (11) *such that for any* $t \in [0, T]$

$$x(t) = y(\Gamma(t)), \tag{12}$$

where

$$\Gamma(t) = \inf\{s : \eta(s) > t\}. \tag{13}$$

THEOREM 3.2. *Let* $x(\cdot)$ *be a generalized solution of system* (1). *Then there exists a sequence of admissible controls* $\{u_1^n(\cdot), w_1^n(\cdot)\}$ *satisfying* (11), *on some interval* $[0, T_1]$, $T_1 \leq T + M$, *such that the corresponding sequence of solutions of the system* (10) *with initial conditions*

$$y^n(0) = x(0-), \qquad \eta^n(0) = 0$$

converges uniformly on $[0, T_1]$ *to the pair of functions* $\{y(\cdot), \eta(\cdot)\}$, *and the generalized solution* $x(\cdot)$ *can be represented on* $[0, T]$ *by the relation* (12), *where the function* $\Gamma(\cdot)$ *is defined by* (13), *and* $\Gamma(T) = T_1$ *by definition.*

Remark 3.2. These results follow directly from theorems VI.4.4 and VI.4.5 in Warga's book [21] if the original system is autonomous, i.e., $F(x, u, w, t) = F(x, w, u)$. However, Warga's arguments remain valid in the nonautonomous case without any change. For our problem the assumptions of theorem VI.4.4 [21] follow by choosing, in Warga's notation, $\varphi(v, u, b) = 1 + \|w\|$, where Warga's u represents our (u, w). (Note that all the solutions of original and transformed problems are confined to a compact set, due to assumptions (4) and (5).) The assumptions of theorem VI.4.5 should be slightly modified; namely, the condition

$$\|F(x_1, u, w) - F(x_2, u, w)\| \leq L_1\|x_1 - x_2\| \tag{14}$$

is replaced by

$$\|F(x_1, u, w, t) - F(x_2, u, w, t)\| \leq L_1(1 + \|w\|)\|x_1 - x_2\|. \tag{15}$$

However, the proof of Warga's theorem VI.4.5 [21] remains valid with slight modifications if assumption (14) is replaced by (15). To modify the proof of theorem VI.4.5 [21] we should (using the notation of [21])

(i) replace c_1 in (2) with $2c_1(L+1)$—here L is the upper bound of the variation of the system (1) solution;

(ii) replace c_1 in (5) with $c_1(1+\|w(t)\|)$;

(iii) and replace, in step 2, in the first inequality for $e(s)$

$$\int_0^s e(\tau)d\tau \quad \text{with} \quad \int_0^s (1+\|w(\tau)\|)e(\tau)d\tau,$$

which (using Holder's inequality) yields relation (7) in the form

$$e(s) \le |1-\alpha_j^{-1}|Lc',$$

where (in Warga's notation)

$$c'' = 2c_1\int_0^1 (1+|w(t)|)dt, \quad c' = c''\exp(c'') \le 2(L+1)c_1\exp(2(L+1)c_1).$$

Now, for every generalized solution of the system (1) we have the representation by the sequence of ordinary solutions of the auxiliary system (10). However, for the solution of the optimization problem we should find more appropriate representation of generalized solution. Again, consider the sequence of solutions of the auxiliary system (10), which converges uniformly to some pair of functions $\{y(\cdot), \eta(\cdot)\}$. As follows from the theory of differential inclusions and by virtue of continuity properties of function $\tilde{F}(y, u, w, \eta)$ [6] this pair satisfies the differential inclusion

$$\begin{pmatrix} \dot{y}(s) \\ \dot{\eta}(s) \end{pmatrix} \in \overline{\text{conv}}\,(\tilde{F}(y_1(s), u_1, w_1, \eta(s)) \mid u_1 \in U, w_1 \in W), \tag{16}$$

where the set in the right-hand side of (16) is a closed convex hull of the right-hand sides of the system (10) for every (y, η).

Assume that there exists an appropriate representation of the set in the right-hand side of (16); i.e., there exists the R^{n+1}-valued vector-function $G(y, \omega, \eta)$ and some closed bounded set Ω in a vector space of appropriate dimension such that for any $(y, \eta) \in R^{n+1}$

$$\begin{aligned} \overline{\text{conv}}\,(\tilde{F}(y, u_1, w_1, \eta) \mid u_1 \in U, w_1 \in W) &= \overline{\text{conv}}\,(\tilde{F}(y, U, W, \eta) \\ &= G(y, \Omega, \eta) = \{G(y, \omega, \eta) \mid \omega \in \Omega\}. \end{aligned} \tag{17}$$

In addition, assume that $G(y, \omega, \eta)$ is continuous with respect to (y, ω, η) and for any fixed $\omega \in \Omega$ this function satisfies the Lipschitz condition and has a linear growth with respect to (y, η), i.e.,

$$\|G(y_1, \omega, \eta_1) - G(y_2, \omega, \eta_2)\| \le L\{\|y_1 - y_2\| + |\eta_1 - \eta_2|\} \tag{18}$$

and

$$\|G(y, \omega, \eta)\| \le L(1 + \|y\| + |\eta|) \tag{19}$$

for any $y_1, y_2 \in R^n$, $\eta_1, \eta_2 \in R^1$ with some constant $L > 0$.

Then, due to convexity of the set $G(y, \Omega, \eta)$ and the boundedness of the set Ω for any solution $\{y(\cdot), \eta(\cdot)\}$ satisfying the differential inclusion (16), there exists a measurable control $\omega(\cdot)$ such that $\omega(s) \in \Omega$ almost everywhere on $[0, T_1]$ and $\{y(\cdot), \eta(\cdot)\}$ satisfy the system

$$\begin{pmatrix} \dot{y}(s) \\ \dot{\eta}(s) \end{pmatrix} = G(y(s), \omega(s), \eta(s))$$

on $[0, T_1]$ [6].

Now it is possible to formulate the following result concerning representation of the generalized solution.

THEOREM 3.3. *Let there exist the vector function $G(y, \omega, \eta)$ and the set Ω satisfying the conditions* (17)–(19). *Then for any generalized solution $x(\cdot)$ of the system* (1) *there exists a control $\omega(s) \in \Omega$ defined on some interval $[0, T_1]$ such that*

$$x(t) = y(\Gamma(t)),$$

where

$$\Gamma(t) = \inf\{s : \eta(s) > t\}$$

and the pair of functions $\{y(\cdot), \eta(\cdot)\}$ satisfies the differential equation

$$\begin{pmatrix} \dot{y}(s) \\ \dot{\eta}(s) \end{pmatrix} = G(y(s), \omega(s), \eta(s))$$

with the initial conditions

$$y(0) = x(0-), \qquad \eta(0) = 0$$

and terminal condition

$$\eta(T_1) = T.$$

Assumptions concerning the existence of appropriate function $G(y, \omega, \eta)$ and the set Ω seem to be very artificial; however, it is possible to demonstrate that in the "regular" case they are fulfilled.

4. Existence of the G function in the regular case. We shall say that we consider the "regular" case if the function $F(x, u, w, t)$ demonstrates the regular behavior for large values of variable w.

DEFINITION 4.1. *$F(x, u, w, t)$ is said to be* regular at infinity *if*

(i) *it is continuous with respect to all variables and has a Lipschitz property with respect to x and t, i.e.,*

$$\|F(x', u, w, t') - F(x, u, w, t)\| \leq L(1 + \|w\|)(\|x - x'\| + |t - t'|)$$

for any $x', x \in R^n, t', t \in [0, T], u \in U, w \in W$;

(ii) *for any pair of vectors (u, e): $u \in U, e \in E = \{\|e\| = 1\}$ there exists*

$$\lim_n \frac{F(x, u_n, w_n, t)}{1 + \|w_n\|} = \Phi(x, u, e, t) \tag{20}$$

if

$$u_n \to u, \qquad \frac{w_n}{1 + \|w_n\|} \to e, \qquad \|w_n\| \to \infty, \qquad u_n \in U, \qquad w_n \in W,$$

where $\Phi(x, u, e, t)$ *is continuous with respect to all the variables and the limit in* (20) *is independent of the choice of sequence* $\{u_n, w_n\}$.

In this case, introduce the new control variable ω, defined by the relation

$$\omega = \omega(w) = \frac{w}{1 + \|w\|}. \tag{21}$$

Function (21) maps the unbounded set W in a one-to-one manner to the set $\omega(W)$, which is a subset of interior of the unit ball in R^m. The map that is inverse to (21) is

$$w(\omega) = \frac{\omega}{1 - \|\omega\|}.$$

This inverse map is defined and continuous on the interior of the unit ball in R^m.

Substitution of the control $\omega \in \omega(W)$ instead of $w \in W$ into (10) for $\tilde{F}$ gives

$$\tilde{F}(y, u_1, w_1, \eta) = \left\{ \begin{array}{l} (1 - \|\omega\|) F\left(y, u_1, \dfrac{\omega}{1 - \|\omega\|}, \eta\right) \\ 1 - \|\omega\| \end{array} \right\} = \tilde{G}(y, u_1, \omega, \eta).$$

Now we can demonstrate that in the regular case $\tilde{G}$ is almost the same as desired by Theorem 3.3. First, one can prove that $\tilde{G}(y, u_1, \omega, \eta)$ is continuous with respect to $(y, \eta, u_1, \omega) \in R^n \times R^1 \times U \times \overline{\omega(W)}$. (Here the last set is the closure of the set $\omega(W)$.) The continuity of this function is evident if $\|\omega\| < 1$ due to continuity of superposition of continuous functions. Thus, if we consider an arbitrary point (y, η, u_1, ω), where $\|\omega\| = 1$ and $\omega \in \overline{\omega(W)}$, then there exists a sequence $\{u_1^n, \omega^n\}$, such that $u_1^n \in U$, $\|\omega^n\| < 1$, $\omega^n \in \omega(W)$, and $u_1^n \to u_1$, $\omega^n \to \omega$. Moreover, for every $\omega^n \in \omega(W)$ we can define

$$w^n = \frac{\omega^n}{1 - \|\omega^n\|} \in W$$

such that

$$\|w^n\| \to \infty, \qquad \frac{w^n}{1 + \|w^n\|} \to \omega,$$

and

$$\tilde{G}(y, u^n, \omega^n, \eta) = \tilde{F}(y, u^n, w^n, \eta). \tag{22}$$

By regularity condition, the function in the right-hand side of (22) has a limit if $u^n \to u$ and $w^n/(1 + \|w^n\|) \to \omega$, and this limit is independent of the sequence $\{u^n, w^n\}$ choice. Therefore, $\tilde{G}$ can be extended by continuity to the closure of the set $\omega(W)$. We will use the same symbol $\tilde{G}$ for the extended function.

Consider the relation

$$\begin{aligned} &\|\tilde{G}(y', u_1', \omega', \eta') - \tilde{G}(y, u_1, \omega, \eta)\| \\ &\quad \le \|\tilde{G}(y_1', u_1', \omega', \eta') - \tilde{G}(y, u_1', \omega', \eta)\| + \|\tilde{G}(y, u_1', \omega', \eta) - \tilde{G}(y, u_1, \omega, \eta)\|. \end{aligned} \tag{23}$$

The first term in the right-hand side of (23) can be estimated by the value $L(\|y_m' - y\| + |\eta' - \eta|)$, due to the Lipschitz property of original function $F(x, u, w, t)$, and the last term tends to zero if $(u_1', \omega') \to (u_1, \omega)$ because of the regularity condition. Hence $\tilde{G}(y', u_1', \omega', \eta') \to \tilde{G}(y, u_1, w, \eta)$ if $(y', u_1', \omega', \eta') \to (y, u_1, w, \eta)$, and the function G is continuous. Moreover, it is a Lipschitz one and has a linear growth.

Now for the final construction of the requested function $G(y, \overline{\omega}, \eta)$ we can use an analytical description of convex hall of the bounded set $\tilde{G}(y, U, \overline{\omega(W)}, \eta)$ by the Carathéodory theorem [6]. Hence, the function $G(y, \overline{\omega}, \eta)$ and set Ω can be represented in the form

$$\Omega = \left\{ \overline{\omega} = [u_i, \omega_i, \alpha_i, i = 1, \ldots, n+2] \mid u_i \in U, \omega_i \in \overline{\omega(W)}, \alpha_i \geq 0, \sum_{i=1}^{n+2} \alpha_i = 1 \right\},$$

$$G(y, \overline{\omega}, \eta) = \sum_{i=1}^{n+2} \alpha_i \tilde{G}(y, u_i, \omega_i, \eta).$$

Remark 4.1. Following the results of [20], [21], we note that the homeomorphism $w \to \omega(w) = w/(1 + \|w\|)$ of W onto its image $\omega(W)$ defines a metric compactification $(\bar{Z}, Z, \Phi)$ of W in the terminology of Warga's book [21]. However, we consider it possible to bring in the construction procedure to clarify it in our specific case.

As an example of application of this procedure we can handle control problems for the system, which is sublinear with respect to unbounded control [1], [3], [10], [12]–[14], [17], [18]. Let it be described by the equations

$$\dot{x}(t) = f(x(t), v(t), u(t), t) + B(x(t), v(t), u(t), t)w,$$
$$\dot{v}(t) = \|w(t)\|,$$

where $u \in U$, $w \in K$, U is a compact subset, and K is a convex cone, and the functions f and B are continuous. Then for all $e \in \{\|e\| = 1\}$ there exists a limit

$$\lim \frac{f + Bw}{1 + \|w\|} = Be$$

if $\|w\| \to \infty$, $w/(1 + \|w\|) \to e$, and the right-hand side of auxiliary system is described by the function

$$\tilde{G}(y, z, u, \eta, \omega) = \begin{cases} (1 - \|\omega\|)\left[f + B\dfrac{\omega}{1 - \|\omega\|} \right] = (1 - \|\omega\|)f + B\omega, \\ \|\omega\|, \\ 1 - \|\omega\|, \end{cases}$$

where $0 \leq \|\omega\| \leq 1$. If the set

$$\tilde{G}(y, z, U, \eta, \overline{\omega(K)}) = \{l \in R^{n+2} : l = \tilde{G}(y, z, u, \eta, \omega) \mid u \in U, \omega \in K \cap \{\|\omega\| \leq 1\}\}$$

is convex for any y, z, and η, then one can introduce a new pair of control variables $\{(\alpha, e) : 0 \leq \alpha \leq 1, e \in K \cap \{\|e\| \leq 1\}$, which describes the set $\tilde{G}(y, z, U, \eta, \overline{\omega(K)})$ in the following way:

$$\tilde{G}(y, z, U, \eta, \overline{\omega(K)}) = \left\{ \begin{array}{l|l} \alpha f + (1 - \alpha)Be & \alpha \in [0, 1] \\ (1 - \alpha)\|e\| & e \in K \cap \{\|e\| \leq 1\} \\ \alpha & u \in U \end{array} \right\}.$$

This gives the following description of auxiliary system with the right-hand side:

$$G(y, z, \overline{\omega}, \eta) = \begin{pmatrix} \alpha f(y, z, u, \eta) + (1 - \alpha)B(y, z, u, \eta)e \\ (1 - \alpha)\|e\| \\ \alpha \end{pmatrix},$$

where

$$\overline{\omega} = (\alpha, e, u)$$

such that

$$\overline{\omega} \in \Omega = \{0 \leq \alpha \leq 1,\ e \in K \cap \{\|e\| \leq 1\},\ u \in U\}.$$

If the set $G(y, z, \Omega, \eta)$ is convex for any y, z, and η, we shall obtain the desired representation; otherwise we should use the procedure of convexification using the Carathéodory theorem [6].

5. Representation of the generalized solution via the differential equation with measure. This part of the paper is devoted to the representation of generalized solutions by the special type of differential equations with a measure.

Let us consider the function $x(\cdot)$, which is the generalized solution of the system (1) on the interval $[0, T]$. The function $x(\cdot)$ has the bounded variation and is continuous from the right; thus it admits the representation [4]

$$x(t) = x(0) + x^c(t) + \sum_{\tau \leq t} \Delta x(\tau), \tag{24}$$

where

$$\Delta x(\tau) = x(\tau) - x(\tau -)$$

and $x^c(\cdot)$ is a continuous function. For all components in (24) one can derive the representation in the form of special differential inclusion.

To formulate this result, we introduce the set $G^1(y, \eta) \in R^n$, where $l_1 \in G^1(y, \eta)$ if the extended vector $(l_1, 0) \in G(y, \Omega, \eta)$, i.e.,

$$G^1(y, \eta) = \left\{ l_1 \in R^n : \quad l_1 = \left(\begin{array}{l|l} G_1(y, \omega, \eta) & \\ G_2(y, \omega, \eta) & \omega \in \Omega, \\ \vdots & G_{n+1}(y, \omega, \eta) = 0 \\ G_n(y, \omega, \eta) & \end{array} \right) \right\}.$$

This set is nonempty, because it contains all the partial limits of sequences of the type

$$\left\{ \frac{F(y, u_i, w_i, \eta)}{1 + \|w_i\|} \right\}$$

such that $u_i \in U$, $w_i \in W$, and $\|w_i\| \to \infty$. Any of such sequences is uniformly bounded for fixed (y, η) by virtue of the inequalities (6), and hence, the set of its all partial limits is nonempty. By definition of $G(y, \omega, \eta)$ and Ω, the set $G(y, \Omega, \eta)$ contains all partial limits of the sequences

$$\begin{pmatrix} l_1^i \\ l_{n+1}^i \end{pmatrix} = \left\{ \begin{array}{ll|l} l_1^i \in R^n & l_1^i = \dfrac{F(y, u_i, w_i, \eta)}{1 + \|w_i\|} & u_i \in U \\ l_{n+1}^i \in R^1 \quad : & l_{n+1}^i = \dfrac{1}{1 + \|w_i\|} & w_i \in W \end{array} \right\};$$

hence it contains all partial limits with $\|w_i\| \to \infty$ and $l_{n+1}^i \to 0$.

The set $G^1(y, \eta)$ is convex and closed for all (y, η), because it is the cross section of the convex and closed set $G(y, \Omega, \eta)$.

Let us also introduce the conic hull of the set $G^1(y, \eta)$, denoted by con $G^1(y, \eta)$, where

$$\operatorname{con} G^1(y, \eta) = \left\{ l \in R^n \mid l = \sum_{i=1}^{k} \beta_i l_i, \quad \beta_i \geq 0, \quad k \geq 1, \quad l_i \in G^1(y, \eta) \right\}.$$

The next theorem yields the representation of a generalized solution.

THEOREM 5.1. *Let the set of right-hand sides of the system* (10) *be represented in the form* (17) *with the appropriate function* $G(y, \omega, \eta)$ *and set* Ω. *Then for any generalized solution* $x(\cdot)$ *there exists a scalar nonnegative regular measure* $V(dt)$, *defined on the Borel subsets of the interval* $[0, T]$ *such that the function*

$$v(t) = V\{[0, t]\}$$

has the representation

$$v(t) = \int_0^t \dot{v}(s)ds + v^s(t) + \sum_{\tau \le t} \Delta v(\tau),$$

where $\dot{v}(t)$ *is the derivative of* $v(\cdot)$ *defined almost everywhere in* $[0, T]$; $v^s(t)$ *is a continuous nondecreasing function whose set of growth points has zero Lebesgue measure (i.e., the corresponding measure* $V^s(dt)$ *is singular with respect to the Lebesgue measure);* $\Delta v(\tau) = v(\tau) - v(\tau -)$ *are the jumps of the function* $v(\cdot)$; *and*

$$v(T) = V\{[0, T]\} \le M$$

such that the generalized solution $x(\cdot)$ *can be represented in the form*

$$x(t) = x(0-) + x^a(t) + x^s(t) + \sum_{\tau \le t} \Delta x(\tau),$$

where

(i) $x^a(t)$ *is absolutely continuous with respect to the Lebesgue measure and its derivative satisfies the differential inclusion*

$$\dot{x}^a(t) \in \overline{\text{conv}}\, F(x(t), U, W, t) + \text{con}\, G^1(x(t), t) \quad \text{a.e. in } [0, T]; \tag{25}$$

(ii) $x^s(t)$ *is absolutely continuous with respect to the measure* $V^s(dt)$ *and its derivative with respect to the measure* $V^s(dt)$ *satisfies the differential inclusion*

$$\frac{dx^s(t)}{dv^s(t)} \in G^1(x(t), t) \quad \text{a.e. in } [0, T] \quad \text{with respect to } V^s(dt); \tag{26}$$

(iii) *and*

$$\Delta x(\tau) = x(\tau) - x(\tau -) = y_\tau(\Delta v(\tau)) - x(\tau -),$$

where $y_\tau(\cdot)$ *satisfies, on the interval* $[0, \Delta v(\tau)]$, *the differential inclusion*

$$\dot{y}_\tau(s) \in G^1(y_\tau(s), \tau)$$

with the initial condition

$$y_\tau(0) = x(\tau -).$$

Proof. According to Theorem 3.3, $x(t) = y(\Gamma(t))$, where $\Gamma(t) = \inf\{s \in [0, T_1] : \eta(s) > t\}$ and the pair $(y(s), \eta(s))$ satisfies the differential equation

$$\begin{pmatrix} \dot{y}(s) \\ \dot{\eta}(s) \end{pmatrix} = G(y(s), \omega(s), \eta(s)), \qquad y(0) = x(0-), \qquad \eta(0) = 0,$$

with some measurable control $\omega(s) \in \Omega$ a.e. in $[0, T_1]$. By virtue of results obtained in [9], $\Gamma(\cdot)$ is right continuous and monotonically increasing; therefore, it admits the representation

$$\Gamma(t) = \int_0^t \dot{\Gamma}(s)ds + \Gamma^s(t) + \sum_{\tau \le t} \Delta\Gamma(\tau),$$

where $\dot{\Gamma}(t)$ is a derivative of $\Gamma(\cdot)$, $\Gamma^s(t)$ is a continuous nondecreasing function such that the corresponding measure $\Gamma^s(dt)$ is singular with respect to the Lebesgue measure, and

$$\Delta\Gamma(\tau) = \Gamma(\tau) - \Gamma(\tau-).$$

The interval $[0, T]$ can be represented as a union of three disjoint subsets:

$$[0, T] = D_\Gamma^a \cup D_\Gamma^s \cup D_\Gamma^d,$$

where D_Γ^a is a support of absolutely continuous component of the measure $\Gamma(dt)$, D_Γ^s is a support of the measure $\Gamma^s(dt)$, and D_Γ^d is a union of no more than countable set of points τ such that $\Gamma(\tau) - \Gamma(\tau-) > 0$.

The generalized solution satisfies the integral relation

$$x(t) = y(\Gamma(t)) = x(0) + \int_0^{\Gamma(t)} \dot{y}(s)ds,$$

which can be rewritten

$$\begin{aligned} x(t) &= x(0) + \int_0^{\Gamma(t)} \dot{y}(s) I\{s : \eta(s) \overline{\in} D_\Gamma^d\} ds + \int_0^{\Gamma(t)} \dot{y}(s) I\{s : \eta(s) \in D_\Gamma^d\} ds \\ &= x(0) + \int_0^{\Gamma(t)} \dot{y}(s) I\{s : \eta(s) \overline{\in} D_\Gamma^d\} ds + \sum_{\tau \le t} \int_{\Gamma(\tau-)}^{\Gamma(\tau)} \dot{y}(s) ds \\ &= x(0) + \int_0^{\Gamma(t)} \dot{y}(s) I\{s : \eta(s) \overline{\in} D_\Gamma^d\} ds + \sum_{\tau \le t} \Delta x(\tau), \end{aligned} \tag{27}$$

where $I\{\cdot\}$ means the indicator function.

Substituting the variable $\tau = \eta(s)$ into the first integral in (27) and using the relation $\Gamma(\eta(s)) = s$, which is valid if $\eta(s) \overline{\in} D_\Gamma^d$, we obtain

$$\begin{aligned} &\int_0^{\Gamma(t)} \dot{y}(s) I\{s : \eta(s) \overline{\in} D_\Gamma^d\} ds = \int_0^{\Gamma(t)} \dot{y}(\Gamma(\eta(s))) I\{s : \eta(s) \overline{\in} D_\Gamma^d\} ds \\ &= \int_0^{\eta(\Gamma(t))} \dot{y}(\Gamma(t)) I\{\tau : \tau \overline{\in} D_\Gamma^d\} d\Gamma(\tau) = \int_0^t \dot{y}(\Gamma(\tau)) I\{\tau : \tau \overline{\in} D_\Gamma^d\} d\Gamma(\tau) \\ &= \int_0^t \dot{y}(\Gamma(\tau))(\dot{\Gamma}(\tau)d\tau + d\Gamma^s(\tau)) \\ &= \int_0^t \dot{y}(\Gamma(\tau))\dot{\Gamma}(\tau)d\tau + \int_0^t \dot{y}(\Gamma(\tau))d\Gamma^s(\tau) = x^a(t) + x^s(t). \end{aligned} \tag{28}$$

The union of relations (27) and (28) gives the following representation for $x(t)$:

$$x(t) = x(0-) + x^a(t) + x^s(t) + \sum_{\tau \le t} \Delta x(\tau),$$

where

$$x^a(t) = \int_0^t \dot{y}(\Gamma(\tau))\dot{\Gamma}(\tau)d\tau \tag{29}$$

and $x^a(\cdot)$ is absolutely continuous with respect to the Lebesgue measure;

$$x^s(t) = \int_0^t \dot{y}(\Gamma(\tau))d\Gamma^s(\tau), \tag{30}$$

and $x^s(\cdot)$ is absolutely continuous with respect to measure $\Gamma^s(dt)$; and

$$\Delta x(\tau) = y(\Gamma(\tau)) - y(\Gamma(\tau-)) = \int_{\Gamma(\tau-)}^{\Gamma(\tau)} \dot{y}(s)ds$$

is the jump of the function $x(\cdot)$ at the point $\tau \in D_\Gamma^d$.

We are able to show that $x^a(\cdot)$, $x^s(\cdot)$, and jumps $\Delta x(\tau)$ are the same as claimed in the theorem. First, define the measure $V(dt)$ by the relation

$$V\{(a, b]\} = \Gamma\{(a, b]\} - (b - a). \tag{31}$$

Since

$$\Gamma\{(a, b]\} \geq (b - a),$$

the measure $V(dt)$ is regular and nonnegative. Further,

$$V\{[0, T]\} = \Gamma(T) - T \leq M$$

and

$$V^s(dt) = \Gamma^s(dt), \qquad \Delta v(\tau) = V\{\tau\} = \Delta\Gamma(\tau).$$

Consider (29). As on the subset D_Γ^a the derivative of $\Gamma(\cdot)$ exists a.e. and satisfies inequalities

$$0 < \dot{\Gamma}(t) = \frac{1}{\dot{\eta}(\Gamma(t))} < \infty,$$

then $\dot{\eta}(\Gamma(t)) > 0$ a.e. in $[0, T]$, and the derivative of $x^a(\cdot)$ exists and satisfies the relation

$$\dot{x}^a(t) = \frac{\dot{y}(\Gamma(t))}{\dot{\eta}(\Gamma(t))} \quad \text{a.e. in } [0, T].$$

Now we prove that $\dot{x}^a(t)$ satisfies the inclusion (25). As was proven in [12], the control function $\omega(\Gamma(t))$ is Lebesgue measurable; hence,

$$\begin{pmatrix} \dot{y}(\Gamma(t)) \\ \dot{\eta}(\Gamma(t)) \end{pmatrix} = G(y(\Gamma(t)), \omega(\Gamma(t)), \eta(\Gamma(t))) = G(x(t), \omega(\Gamma(t)), t) \tag{32}$$

a.e. in $[0, T]$.

The set $G(x, \Omega, t)$ is the convex hull of closure of the set $\tilde{F}(x, v, w, t)$; hence by the Carathéodory theorem [6] for all $\omega \in \Omega$ the set $\{(\alpha_i, l_i), i = 1, \ldots, n + 2\}$ exists such that

$$\alpha_i \geq 0, \qquad \sum_{i=1}^{n+2} \alpha_i = 1, \qquad l_i \in \overline{\tilde{F}(x, v, w, t)},$$

and

$$G(x, \omega, t) = \sum_{i=1}^{n+2} \alpha_i l_i = \sum_{i=1}^{n+2} \alpha_i \begin{pmatrix} l_i^1 \\ l_i^2 \end{pmatrix}, \tag{33}$$

where l_i^1 is a vector of the first n components of vector l_i and l_i^2 is the $(n+1)$-th component of this vector.

If $l_i^2 > 0$, then, due to continuity of $\tilde{F}(x, u, w, t)$, there exists a pair (u_i, w_i), $u_i \in U$, $w_i \in W$, such that

$$\begin{pmatrix} l_i^1 \\ l_i^2 \end{pmatrix} = \tilde{F}(x, u_i, w_i, t).$$

If $l_i^2 = 0$, then $l_i^1 \in G^1(x, t)$ by the definition of the set $G^1(x, t)$. Now we can derive the representation

$$G(x(t), \omega(\Gamma(t)), t) = \sum_{i=1}^{n+2} \alpha_i(t) \begin{pmatrix} l_i^1(t) \\ l_i^2(t) \end{pmatrix} I\{i : l_i^2(t) > 0\} + \sum_{i=1}^{n+2} \alpha_i(t) \begin{pmatrix} l_i^1(t) \\ l_i^2(t) \end{pmatrix} I\{i : l_i^2(t) = 0\}$$

with some Lebesgue-measurable function $\{\alpha_i(t), l_i^1(t), l_i^2(t)\}$, $i = 1, \ldots, n+2$. Hence the derivative of $x^a(t)$ is equal to

$$\dot{x}^a(t) = \frac{\sum_{i=1}^{n+2} \alpha_i(t) l_i^1(t)}{\sum_{i=1}^{n+2} \alpha_i(t) l_i^2(t)} \qquad \text{a.e. in} \quad [0, T],$$

and one can derive the following representation for $\dot{x}^a(t)$:

$$\dot{x}^a(t) = \sum_{i=1}^{n+2} \frac{\alpha_i(t) l_i^2(t)}{\sum_{i=1}^{n+2} \alpha_i(t) l_i^2(t)} \frac{l_i^1(t)}{l_i^2(t)} I\{i : l_i^2(t) > 0\} + \sum_{i=1}^{n+2} \frac{\alpha_i(t)}{\sum_{i=1}^{n+2} \alpha_i(t) l_i^2(t)} l_i^1(t) I\{i : l_i^2(t) = 0\}. \tag{34}$$

Denote

$$\beta_i^1(t) = \frac{\alpha_i(t) l_i^2(t)}{\sum_{i=1}^{n+2} \alpha_i(t) l_i^2(t)} I\{i : l_i^2(t) > 0\} \tag{35}$$

and

$$\beta_i^2(t) = \frac{\alpha_i(t)}{\sum_{i=1}^{n+2} \alpha_i(t) l_i^2(t)} I\{i : l_i^2(t) = 0\}. \tag{36}$$

Then

$$\beta_i^1(t) \geq 0, \qquad \sum_{i=1}^{n+2} \beta_i^1(t) = 1, \tag{37}$$

and

$$\beta_i^2(t) \geq 0. \tag{38}$$

Vectors $l_i^1(t)/(l_i^2(t))$ in the first sum of (34) satisfy the inclusion

$$\frac{l_i^1(t)}{l_i^2(t)} \in F(x(t), U, W, t),$$

and vectors $l_i^1(t)$ in the second sum satisfy the inclusion

$$l_i^1(t) \in G^1(x(t), t).$$

Hence the derivative $\dot{x}^a(t)$ has the representation

$$\dot{x}^a(t) = \sum_{i=1}^{n+2} \beta_i^1(t) F(x(t), u_i(t), w_i(t), t) + \sum_{i=1}^{n+2} \beta_i^2(t) l_i^1(t),$$

where $l_i^1(t) \in G^1(x(t), t)$ and functions $\beta_i^1(t)$, $\beta_i^2(t)$ satisfy the relations (37), (38). This proves the theorem's claim for the component $x^a(t)$.

Now we consider the function $x^s(t)$ represented by the equation (30) and note that $\dot{\eta}(\Gamma(t)) = 0$ on the subset $D_\Gamma^s = D_V^s$. Then, similar consideration applied to the component $x^s(t)$ gives the inclusion

$$\dot{y}^s(\Gamma(t)) \in G^1(x(t), t)$$

and the relation

$$x^s(t) = \int_0^t l^1(\tau) dv^s(\tau), \qquad l^1(\tau) \in G^1(x(\tau), \tau),$$

proves the theorem for the component $x^s(t)$.

Since on the interval $[\Gamma(\tau-), \Gamma(\tau)]$, where $\eta(s) = \tau \in D_V^d = D_\Gamma^d$, the derivative of $\eta(s)$ is also equal to zero,

$$\dot{y}(s) \in G^1(y(s), \tau).$$

Therefore, if $s \in [\Gamma(\tau-), \Gamma(\tau)]$, then $y(s)$ satisfies the above differential inclusion with the initial condition $y(\Gamma(\tau-) = x(\tau-)$. Letting

$$y_\tau(s) = y(\Gamma(\tau-) + s)$$

for $s \in [0, \Delta v(\tau)] = [0, \Delta\Gamma(\tau)]$, we obtain the assertion of the theorem for the jumps of $x(\cdot)$. This completes the proof. □

In the case of the regular function $F(x, u, w, t)$ we can give a more precise statement about the representation of the generalized solution. First we prove the theorem concerning the description of the set $G^1(x, t)$.

THEOREM 5.2. *If $F(x, u, w, t)$ is regular at infinity, then*

$$G^1(x, t) = \operatorname{conv} \Phi(x, U, E, t),$$

where

$$\Phi(x, U, E, t) = \{l \in R^n : l = \Phi(x, u, e, t) | \quad u \in U, \quad e \in E = \{\|e\|\} = 1\}.$$

Proof. By the definition of regularity, if vector $l \in \Phi(x, U, E, t)$, i.e., there exist $u \in U$ and $e \in E$ such that $l = \Phi(x, u, e, t)$, then for any sequence $w_n \in W$ such that $\|w_n\| \to \infty$, $w_n/(1 + \|w_n\|) \to e$, we have

$$l = \lim_n \frac{F(x, u, w_n, t)}{1 + \|w_n\|} = \Phi(x, u, e, t).$$

Since $\|w_n\| \to \infty$, $\lim_n 1/(1 + \|w_n\|) = 0$.

Hence, the extended vector

$$\begin{pmatrix} l \\ 0 \end{pmatrix} \in G(x, \Omega, t),$$

where $l \in G^1(x, t)$ by definition of the set $G^1(x, t)$. Therefore, by convexity of $G^1(x, t)$

$$\operatorname{conv} \Phi(x, U, E, t) \subseteq G^1(x, t). \tag{39}$$

To prove the inverse inclusion, note that if $l \in G^1(x, t)$, then

$$\begin{pmatrix} l \\ 0 \end{pmatrix} \in G(x, \Omega, t) = \overline{\operatorname{conv}} \tilde{F}(x, U, W, t).$$

Hence there exists the set of pairs $\{\alpha_i, l_i\}$, $i = 1, \ldots, n + 2$, such that

$$\alpha_i > 0, \quad \sum_{i=1}^{n+2} \alpha_i = 1, \quad \begin{pmatrix} l_i^1 \\ 0 \end{pmatrix} \in \overline{\tilde{F}(x, U, W, t)},$$

and

$$\begin{pmatrix} l \\ 0 \end{pmatrix} = \sum_{i=1}^{n+2} \alpha_i \begin{pmatrix} l_i^1 \\ 0 \end{pmatrix}. \tag{40}$$

Then for every i there exists the sequence $\{u_n^i, w_n^i\}$, $u_n^i \in U$, $w_n^i \in W$, such that

$$\begin{pmatrix} l_i^1 \\ 0 \end{pmatrix} = \lim_n \begin{pmatrix} \dfrac{F(x, u_n^i, w_n^i, t)}{1 + \|w_n^i\|} \\ \dfrac{1}{1 + \|w_n^i\|} \end{pmatrix},$$

and due to uniform boundedness of sequence $\{u_n^i, w_n^i/(1 + \|w_n^i\|)\}$ one can extract the subsequence $\{u_{n_k}^i, w_{n_k}^i\}$ such that

$$u_{n_k}^i \to u^i \in U \quad \text{and} \quad \frac{w_{n_k}^i}{1 + \|w_{n_k}^i\|} \to e^i \in E \quad \text{if} \quad k \to \infty.$$

Then $l_i^1 = \Phi(x, u^i, e^i, t)$, and with (40) it means

$$\begin{pmatrix} l \\ 0 \end{pmatrix} \in \begin{pmatrix} \operatorname{conv} \Phi(x, U, E, t) \\ 0 \end{pmatrix}.$$

Hence

$$G^1(x, t) \subseteq \operatorname{conv} \Phi(x, U, E, t), \tag{41}$$

and combining (39) with (41) we complete the proof. □

We can now formulate the result concerning the representation of generalized solution by the differential equation with a measure.

THEOREM 5.3. *Let the function $F(x, u, w, t)$ be regular and the set of vectors*

$$\mathcal{L} = \left\{ l = \begin{pmatrix} l_1 \\ l_2 \end{pmatrix} \in R^{n+1} \;\middle|\; \begin{array}{l} l_1 = F(x, u, w, t) \\ l_2 = \|w\| \end{array} \;\middle|\; \begin{array}{l} u \in U \\ w \in W \end{array} \right\} \tag{42}$$

be convex for any (x, t).

Then for any generalized solution $x(\cdot)$ of the system (1) *there exist*

(i) *the regular nonnegative measure $V(dt)$ on the interval $[0, T]$, which satisfies the constraint*

$$V\{[0, T]\} \leq M;$$

(ii) *the set of both Lebesgue- and V-measurable controls*

$$\{u_0(t), w_0(t), \beta_i^0(t), \beta_i^s(t), u_i(t), e_i(t)\}, \qquad i = 1, \ldots, n+2,$$

which satisfy the constraints

$$u_0(t) \in U, \quad w_0(t) \in W, \quad \beta_i^0(t) \geq 0,$$

$$u_i(t) \in U, \quad e_i(t) \in E, \quad \beta_i^s(t) \geq 0, \quad \sum_{i=1}^{n+2} \beta_i^s(t) = 1,$$

a.e. in the interval $[0, T]$;

(iii) *the set of Lebesgue-measurable controls*

$$\{u_i^\tau(s), e_i^\tau(s), \beta_i^\tau(s)\}, \qquad i = 1, \ldots, n+2,$$

which are defined on every interval $[0, \Delta v(\tau)]$ for all $\tau \in D_V^d$ and satisfy the constraints

$$u_i^\tau(s) \in U, \qquad e_i^\tau(s) \in E, \qquad \beta_i^\tau(s) \geq 0, \qquad \sum_{i=1}^{n+2} \beta_i^\tau(s) = 1,$$

a.e. in the interval $[0, \Delta v(\tau)]$
such that the generalized solution $x(t)$ and the function $v(t) = V\{[0, t]\}$ satisfy the equations

$$\begin{aligned} x(t) = x(0) &+ \int_0^t F(x(\tau), u_0(\tau), w_0(\tau), \tau) d\tau \\ &+ \int_0^t \sum_{i=1}^{n+2} \beta_i^0(\tau) \Phi(x(\tau), u_i(\tau), e_i(\tau), \tau) d\tau \\ &+ \int_0^t \sum_{i=1}^{n+2} \beta_i^s(\tau) \varphi(x(\tau), u_i(\tau), e_i(\tau), \tau) dv^s(\tau) + \sum_{\tau \leq t} \Delta x(\tau), \end{aligned} \tag{43}$$

where

$$\Delta x(\tau) = y_\tau(\Delta v(\tau)) - x(\tau -)$$

and $y_\tau(s)$ is the solution of the differential equation

$$\dot{y}_\tau(s) = \sum_{i=1}^{n+2} \beta_i^\tau(s) \Phi(y_\tau(s), u_i^\tau(s), e_i^\tau(s), \tau) \tag{44}$$

with the initial condition $y_\tau(0) = x(\tau-)$, *and*

$$v(t) = \int_0^t \|w_0(\tau)\| d\tau + \int_0^t \sum_{i=1}^{n+2} \beta_i^0(\tau) d\tau + v^s(t) + \sum_{\tau \le t} \Delta v(\tau). \tag{45}$$

Proof. By definition of the measure $V(dt)$, according to (31), the derivative of $v(t) = V\{[0, t]\}$ is equal to

$$\dot{v}(t) = \dot{\Gamma}(t) - 1,$$

and, as follows from (32) and (33), the pair of functions $\{x^a(\cdot), v^a(\cdot)\}$ satisfies the equations

$$\begin{aligned} \dot{x}^a(t) &= \sum_{i=1}^{n+2} \beta_i^1(t) \frac{l_i^1(t)}{l_i^2(t)} I\{i : l_i^2(t) > 0\} + \sum_{i=1}^{n+2} \beta_i^2(t) l_i^1(t) I\{i : l_i^2(t) = 0\}, \\ \dot{v}^a(t) &= \sum_{i=1}^{n+2} \beta_i^1(t) \frac{1 - l_i^2(t)}{l_i^2(t)} I\{i : l_i^2(t) > 0\} + \sum_{i=1}^{n+2} \beta_i^2(t), \end{aligned} \tag{46}$$

where $l_i^1(t)$ and $l_i^2(t)$ are the components of the vector function $G(x(t), \omega(\Gamma(t)), t)$, i.e.,

$$\begin{pmatrix} l_i^1(t) \\ l_i^2(t) \end{pmatrix} = G(x(t), \omega(\Gamma(t)), t),$$

and the assembly of coefficients $\beta_i^1(t)$ and $\beta i^2(t)$ is defined by (35), (36).

If $l_i^2(t) > 0$, then there exists a pair of vectors $u_i(t) \in U$, $w_i(t) \in W$, such that

$$\frac{l_i^1(t)}{l_i^2(t)} = F(x(t), u_i(t), w_i(t), t) \tag{47}$$

and

$$\frac{1 - l_i^2(t)}{l_i^2(t)} = \|w_i(t)\|, \tag{48}$$

and if $l_i^2(t) = 0$, then

$$l_i^1(t) \in G^1(x(t), t) = \operatorname{conv} \Phi(x(t), U, E, t). \tag{49}$$

The substitution of relations (47)–(49) into (46) gives

$$\begin{pmatrix} \dot{x}^a(t) \\ \dot{v}^a(t) \end{pmatrix} = \sum_{i=1}^{n+2} \beta_i^1(t) \begin{pmatrix} F(x(t), u_i(t), w_i(t), t) \\ \|w_i(t)\| \end{pmatrix} + \sum_{i=1}^{n+2} \beta_i^2(t) \begin{pmatrix} l_i^1(t) \\ 1 \end{pmatrix}. \tag{50}$$

For the first term in the right-hand side of (50), due to convexity of the set (42) and the relation (37), there exists a pair of functions $\{u_0(t), w_0(t)\}$, such that $u_0(t) \in U$, $w_0(t) \in W$ a.e. in $[0, T]$, and

$$\sum_{i=1}^{n+2} \beta_i^1(t) \begin{pmatrix} F(x(t), u_i(t), w_i(t), t) \\ \|w_i(t)\| \end{pmatrix} = \begin{pmatrix} F(x(t), u_0(t), w_0(t), t) \\ \|w_0(t)\| \end{pmatrix}. \tag{51}$$

Using the standard procedure of measurable selection [23], this functions $\{u_0(t)\}$, $\{w_0(t)\}$ can be shown to be measurable with respect to Lebesgue measure.

The second term in the right-hand side of (51) can be represented in the form

$$\sum_{i=1}^{n+2} \beta_i^2(t) \begin{pmatrix} l_i^1(t) \\ 1 \end{pmatrix} = \left(\sum_{i=1}^{n+2} \beta_i^2(t) \right) \sum_{i=1}^{n+2} \frac{\beta_i^2(t)}{\sum_{i=1}^{n+2} \beta_i^2(t)} \begin{pmatrix} l_i^1(t) \\ 1 \end{pmatrix},$$

where the components in the sum satisfy the inclusion

$$\begin{pmatrix} l_i^1(t) \\ 1 \end{pmatrix} \in \begin{pmatrix} G^1(x(t), t) \\ 1 \end{pmatrix}$$

because of its convexity. Then there exists an assembly of functions

$$\{\gamma_k(t), u_k(t), e_k(t), \quad k = 1, \ldots, n+2\}$$

which satisfy the constraints $\gamma_k(t) \geq 0$, $\sum_{k=1}^{n+2} \gamma_k(t) = 1$, $u_k(t) \in U$, $e_k(t) \in E$ a.e. in $[0, T]$ such that

$$\sum_{i=1}^{n+2} \frac{\beta_i^2(t)}{\sum_{i=1}^{n+2} \beta_i^2(t)} \begin{pmatrix} l_i^1(t) \\ 1 \end{pmatrix} = \sum_{k=1}^{n+2} \gamma_k(t) \begin{pmatrix} \Phi(x(t), u_k(t), e_k(t), t) \\ 1 \end{pmatrix}$$

and

$$\sum_{i=1}^{n+2} \beta_i^2(t) \begin{pmatrix} l_i^1(t) \\ 1 \end{pmatrix} = \sum_{i=1}^{n+2} \beta_i^0(t) \begin{pmatrix} \Phi(x(t), u_i(t), e_i(t), t) \\ 1 \end{pmatrix}, \tag{52}$$

where

$$\beta_i^0(t) = \gamma_i(t) \sum_{k=1}^{n+2} \beta_k^2(t).$$

If the functions $\{\beta_i^0, u_i, e_i\}$ are selected to be Lebesgue measurable, then substitution of relation (50) and (52) into (46) proves the representations (43) and (45) for the absolutely continuous part of functions $\{x(\cdot), v(\cdot)\}$.

By the same arguments we can prove the representation for the singular components of functions $\{x(\cdot), v(\cdot)\}$. From the relations (26) it follows that the derivative of the function $x^s(t)$ with respect to the measure $V^s(dt)$ equals

$$\frac{dx^s(t)}{dv^s(t)} = \sum_{i=1}^{n+2} \beta_i^s(t) \Phi(x(t), u_i(t), e_i(t), t),$$

where the functions $u_i(t)$, $e_i(t)$ are V^s-measurable and

$$\beta_i^s(t) \geq 0, \qquad \sum_{i=1}^{n+2} \beta_i^s(t) = 1.$$

Therefore, if we complete the definition of Lebesgue-measurable functions $\{u_i(t), e_i(t)\}$ in relation (52), which are defined on the set D_V^a, by V^s-measurable functions $\{u_i(t), e_i(t)\}$, we obtain statement (ii) of this theorem. The proof of statement (iii) can be obtained by the same arguments. □

Now we can return to the original optimization problem. The results obtained above give us the opportunity to prove the theorem, which ensures the existence of the generalized solution for the original optimization problem.

6. Existence of a solution for problem of generalized optimization. Here, we shall use the approach based on the auxiliary control problem. Consider the auxiliary control problem for the system

$$\begin{pmatrix} \dot{y}(s) \\ \dot{\eta}(s) \end{pmatrix} = G(y(s), \omega(s), \eta(s)) \tag{53}$$

with control which subjects it to the constraint

$$\omega(s) \in \Omega \tag{54}$$

on the interval $[0, T_1]$, with T_1 satisfying the inequality

$$T_1 \leq T + M \tag{55}$$

with the terminal and phase constraints

$$S(y(0), y(T_1)) \leq 0, \qquad h(y(0), y(T_1)) = 0, \qquad \eta(T_1) = T, \tag{56}$$

$$g(y(s), \eta(s)) \leq 0 \text{ for any } s \in [0, T_1], \tag{57}$$

and the following performance criterion, which should be minimized:

$$J'[y(\cdot), \omega(\cdot), T_1] = \varphi_0(y(0), y(T_1)). \tag{58}$$

If the set of admissible controls in the original optimization problem is nonempty, then the set of controls satisfying the constraints (54)–(57) is also nonempty, and one can prove the equivalence of the original and auxiliary problems.

THEOREM 6.1. *Let the set of generalized solutions of the system* (1) *under constraints* (2), (3) *and* (8), (9) *be bounded and nonempty. Then there exists an optimal generalized solution.*

Both in the regular case and in the case of convexity of the set (42), this optimal solution admits the representation by the differential equation with a measure in the form (43), (44) and appropriate elements (i), (ii), (iii) that could be understood as the generalized controls.

THEOREM 6.2. *The optimization problem* (1)–(3), (7)–(9) *and the auxiliary problem* (53)–(58) *are equivalent; i.e., if the pair* $\{y^0(\cdot), \eta^0(\cdot)\}$ *is the solution of auxiliary problem, then the path* $x^0(t) = y^0(\Gamma^0(t))$ *is the optimal generalized solution of the original problem.*

Conversely, if the assumptions of Theorem 5.3 *hold and if* $x^0(\cdot)$ *is the optimal generalized solution of the original problem, then there exists the optimal solution of the auxiliary problem* $\{y^0(\cdot), \eta^0(\cdot)\}$ *such that the relations* (12) *and* (13) *take place.*

Remark 6.1. These two theorems follow directly from the results of theorems VI.4.4 and VI.4.5 in [21] and Theorems 3.2 and 5.3.

Remark 6.2. This result can be considered the main result of our work because it gives a very useful tool for investigation of the generalized optimization problems. One of the most useful features of this approach is that it gives the opportunity to derive necessary and sufficient conditions of optimality. Examples of its application in this area can be found in [1], [11], [15], [16], [20], [22], where the method of discontinuous time change was used for derivation of necessary conditions of optimality in the maximum principle form.

Acknowledgments. The author is very grateful to the anonymous referees for their valuable comments and suggestions, which strengthened this paper.

REFERENCES

[1] A. BRESSAN AND F. RAMPAZZO, *Impulsive control systems with commutative vector fields*, J. Optim. Theory Appl., 71 (1991), pp. 67–83.

[2] F. L. CHERNOUS'KO AND V. B. KOLMANOVSKII, *Optimal Control with Random Disturbances*, Nauka, Moscow, 1978. (In Russian.)

[3] V. I. GURMAN, *Extension Principle in Control Problems*, Nauka, Moscow, 1985. (In Russian.)

[4] A. N. KOLMOGOROV AND S. V. FOMIN, *Elements of Theory of Function and Functional Analysis*, Nauka, Moscow, 1976. (In Russian.)

[5] V. F. KROTOV, V. Z. BUKREEV, AND V. I. GURMAN, *New Methods of Variational Calculus in Flight Dynamics*, Mashinostroenie, Moscow, 1969. (In Russian.)

[6] E. B. LEE AND L. MARCUS, *Foundation of Optimal Control Theory*, Wiley, New York, 1967.

[7] D. LAWDEN, *Optimal Trajectories for Space Navigation*, Butterworth, London, 1963.

[8] V. V. MALYSHEV, M. N. KRASIL'SHIKOV, AND V. I. KARLOV, *Optimization of Observation and Control in Flight Vehicles*, Mashinostroenie, Moscow, 1989. (In Russian.)

[9] B. M. MILLER, *Sampled-data control of processes described by ordinary differential equation* I, II, Autom. Rem. Control, 39 (1978), pp. 57–67, pp. 338–344.

[10] ———, *Optimization of dynamic systems with a generalized control*, Automat. Rem. Control, 50 (1989), pp. 733–742.

[11] ———, *Method of discontinuous time change in problems of control for impulse and discrete-continuous systems*, Automat. Rem. Control, 54 (1993), pp. 1727–1750.

[12] ———, *The generalized solutions of ordinary differential equations in the impulse control problems, Summary*, J. Math. Systems Estim. Control, 4 (1994), pp. 385–388.

[13] M. MOTTA AND F. RAMPAZZO, *Space-time trajectories of nonlinear systems driven by ordinary and impulsive controls*, Differential and Integral Equations, 8 (1995), pp. 269–288.

[14] YU. V. ORLOV, *Theory of Optimal Systems with Generalized Controls*, Nauka, Moscow, 1988. (In Russian.)

[15] R. W. RISHEL, *An extended Pontryagin principle for control systems, whose control laws contain measures*, SIAM J. Control, 3 (1965), pp. 191–205.

[16] W. W. SCHMAEDEKE, *Optimal control theory for nonlinear vector differential equations containing measures*, SIAM J. Control, 3 (1965), pp. 231–280.

[17] A. N. SESEKIN, *Nonlinear differential equations containing product of discontinuous function by generalized functions*, in Generalized Functions and Differential Equations, Ural. Nauchn. Tsentr Akad. Nauk SSSR, Sverdlovsk, 1985, pp. 48–61. (In Russian.)

[18] S. T. ZAVALISCHIN AND A. N. SESEKIN, *Impulsive Processes. Models and Applications*, Nauka, Moscow, 1991. (In Russian.)

[19] R. V. VINTER AND F. M. F. L. PEREIRA, *A maximum principle for optimal processes with discontinuous trajectories*, SIAM J. Control Optim., 26 (1988), pp. 205–229.

[20] J. WARGA, *Variational problems with unbounded control*, SIAM J. Control, 3 (1965), pp. 424–438.

[21] ———, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[22] J. WARGA AND Q. J. ZHU, *The equivalence of extremals in different representations of unbounded control problems*, SIAM J. Control Optim., 32 (1994), pp. 1151–1169.

[23] L. C. YOUNG, *Lectures on the Calculus of Variation and Optimal Control Theory*, W. B. Saunders, Philadelphia, London, Toronto, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1441–1460, July 1996

018

A DIFFERENTIAL GAME WITH TWO PLAYERS AND ONE TARGET*

PIERRE CARDALIAGUET†

Abstract. We study a two-player differential game in which one of the players wants the state of the system to reach an open target, while the other player wants the state of the system to avoid this target. We show that the victory domains of the players form a partition of the complement of the target. One of them can be characterized by the mean of geometrical conditions (as a "discriminating domain"). Finally we show that the common boundary of the victory domains enjoys a semipermeability property.

Key words. differential games, pursuit-and-evasion games, viability theory

AMS subject classifications. 49J24, 49J52, 90J25, 90J26

Let

$$(1)\qquad \begin{cases} x'(t) = f(x(t), u(t), v(t)), \text{ for almost every } t \geq 0, \\ u(t) \in U,\ v(t) \in V, \\ x(0) = x_0 \end{cases}$$

be a two-controlled dynamical system and Ω be an open subset of $\mathbb{R}^N$. The open set Ω shall be called the *target*. In this paper, we investigate the differential game where the first player—Ursula, playing with u—wants the state of the system to reach Ω in finite time. The second player—Victor, playing with v—wants the state of the system to avoid Ω forever. So Ursula wins if the trajectory $x(\cdot)$ reaches the target Ω in finite time, while Victor wins if the trajectory avoids the target forever. This game is called the target problem.

The target problem has been studied often since Isaacs' pioneer work [21]. It is one of the most interesting *game of kind* (by opposition, in Isaacs' terminology, to the *games of degree*). Many examples of application can be found in [21] and also in [20], [8], and [6]. In their monographs [22], [23], Krasovskii and Subbotin have proved that, from any initial position x_0, either Ursula or Victor wins against any action of her (his) adversary. This result is called the alternative theorem. In Krasovskii and Subbotin's result, the game is played in the framework of the positional strategies. Specifically, in equation (1), the controls $u(t)$ and $v(t)$ are replaced by positional strategies, i.e., maps (without a priori regularity) $\tilde{u} : \mathbb{R}^+ \times \mathbb{R}^N \to U$ and $\tilde{v} : \mathbb{R}^+ \times \mathbb{R}^N \to V$. Then the differential equation (1) has in general no solution and Krasovskii and Subbotin had to provide another definition of solution for the differential equation; these solutions are called the constructive motions. In general, they are not solutions to (1) in the Carathéodory sense, but they are limits of step-by-step motions which are solutions to equation (1). So, for the initial differential system, the alternative theorem gives information only on an *approximated game*.[1]

Our main purpose in this paper is to show that, starting from any initial position, either Ursula or Victor can realize *exactly* her (his) objective, against any action of her (his) opponent, provided that the game is played in the framework of Elliot–Kalton nonanticipative strategies (we recall the definition below). The nonanticipative strategies cannot be played simultaneously by Ursula and Victor because the player who plays the nonanticipative strategy needs knowledge of the control played by his opponent to play. So the game cannot be put in the so-called normal form unlike differential games in the class of positional strategies. But this is

*Received by the editors August 1, 1994; accepted for publication (in revised form) April 26, 1995.

†CEREMADE, Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris cedex 16, France.

[1] In Krasovskii and Subbotin's terminology, this approximated problem is called the game of $(\mathcal{M}^\epsilon, \mathcal{N}^\epsilon)$ approach, ϵ denoting a small positive parameter of approximation.

not surprising because we show below that knowledge of the opponent's strategy is in general necessary for Victor to realize exactly his objective, while it is useless for Ursula. This fact was pointed out in several examples of differential games since Isaacs.

We also intend to characterize, by geometric conditions, the set of initial positions x_0 from which Ursula (resp., Victor) may win against any action of her (resp., his) adversary; this set is called Ursula's (resp., Victor's) victory domain.[2] The characterization of the victory domains enables us to solve numerically the target problem (see the joint work [9]). It is also close to one of the characterizations of the stable bridges (see [23]) for the problem of the $(\mathcal{M}, \mathcal{N})$ approach. Our systematic use of the geometric conditions to study the victory domains distinguishes this investigation from previous studies.

Thanks to this characterization we shall provide a proof of the barrier phenomenon. This phenomenon is the key point in Isaacs' construction of the victory domains of the target problem. It is shown in [21] that, if the boundary of the victory domains is smooth, then each player can prevent the state of the system from crossing this boundary in one direction. Since then, this result has been generalized under less restrictive assumptions on the regularity of the boundary. It has also been proved in the case of control theory, without any assumption on regularity of the boundary (see Quincampoix [25]). Our result is an extension of this last one to the two-player differential games.

We make some remarks about the proofs. The proof of the main theorem (Theorem 2.1) uses some results of viability theory [4] and in particular the measurable viability theorem (see [18]). In the characterization of one of the victory domains, we use the same considerations previously used in the method of program iterations developed by Čencov (see [11]–[13]). The proof of Theorem 2.3 seems to be closely related to the proofs of Krasovskii and Subbotin for the characterization of the stable bridges. Let us finally emphasize that Theorems 2.1 and 2.3 have been discovered, independently and simultaneously, by a Polish mathematician, Plaskacz. The proof of Lemma 4.1 is basically his proof.

This paper is organized as follows. In §1, we define the nonanticipative strategies and some sets that play a central role throughout the paper: the discriminating and leadership domains and the discriminating and leadership kernel of a closed set. The main results of this paper are concerned with the interpretation theorems of the discriminating and leadership domains and of the characterization theorems of the discriminating and leadership kernels for the nonanticipative strategies (§2). These results shall be applied in §3 to the target problem. We show that the victory domains of the players can be characterized by means of the discriminating and leadership kernels. Moreover, we prove that the victory domains form a partition of the complement of the target. We also prove a so-called barrier phenomenon which states that some particular trajectories starting from the boundary of the victory domains remain in a neighbourhood of this boundary.

Finally, §§4 and 5 are devoted, respectively, to the proofs of the interpretation theorems and of the characterization theorems.

1. Definitions, assumptions, and notation.

1.1. The nonanticipative strategies. The first definitions of nonanticipative strategies can be found in [27], [28]. Such strategies have been intensively studied (see for instance [14], [15], [20], [16], [19]), but they were mainly applied to problems of games of degree.

Let us now recall the definition of the nonanticipative strategies. If we denote by

$$\text{(2)} \qquad \begin{cases} \mathcal{U} = \{u(\cdot) : [0, +\infty[\to U, \text{ measurable application } \}, \\ \mathcal{V} = \{v(\cdot) : [0, +\infty[\to V, \text{ measurable application } \} \end{cases}$$

[2] We give a rigorous definition of the victory domains in §3. This set plays a major role in the target problem.

the sets of time-measurable controls, nonanticipative strategies are defined in the following way.

DEFINITION 1.1. *We say that a map $\alpha : \mathcal{V} \to \mathcal{U}$ is a nonanticipative strategy (for Ursula) if it satisfies the following condition. For any $s \geq 0$ and for any $v_1(\cdot)$ and $v_2(\cdot)$ belonging to $\mathcal{V}$, such that $v_1(\cdot)$ and $v_2(\cdot)$ coincide almost everywhere on $[0, s]$, the images $\alpha(v_1(\cdot))$ and $\alpha(v_2(\cdot))$ coincide almost everywhere on $[0, s]$.*

Nonanticipative strategies $\beta : \mathcal{U} \to \mathcal{V}$ (for Victor) are defined in the same way. Namely, for any $s \geq 0$ and for any $u_1(\cdot)$ and $u_2(\cdot)$ belonging to $\mathcal{U}$, such that $u_1(\cdot)$ and $u_2(\cdot)$ coincide almost everywhere on $[0, s]$, the images $\beta(u_1(\cdot))$ and $\beta(u_2(\cdot))$ coincide almost everywhere on $[0, s]$.

1.2. The proximal normals. Proximal normals generalize to the closed sets the usual definition of the outward normals.

DEFINITION 1.2. *Let K be a closed subset of $\mathbb{R}^N$ and x belong to K. A vector $p \in \mathbb{R}^N$ is a* proximal normal *to K at x if*

$$d_K(x + p) = \|p\|,$$

where $d_K(y) := \min_{z \in K} \|z - y\|$.

The set of proximal normals to a closed set K at a point x is denoted by $NP_K(x)$. (Note that $\|\cdot\|$ always denotes the Euclidean norm.)

So a vector p is a proximal normal to a closed set K at a point $x \in K$ if the open ball centered in $x + p$ and of radius $\|p\|$ does not intersect K. Let us point out that the closed ball centered in $x + p$ and of radius $\|p\|$ intersects K at least at the point x. Roughly speaking, this ball is tangent to K at x.

1.3. The discriminating and leadership domains and kernels.

DEFINITION 1.3. *Let $H : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ be a map. A closed set D is a* domain for H *if D satisfies*

$$\forall x \in D, \ \forall p \in NP_D(x), \ H(x, p) \leq 0.$$

We shall study in detail the following two cases:

- If the map H is defined by

$$H(x, p) := \sup_u \inf_v \langle f(x, u, v), p \rangle, \tag{3}$$

then the domains for H are called *discriminating domains* for f. (See [1], [3], [4] for an equivalent definition.)

- If H is defined by

$$H(x, p) := \inf_v \sup_u \langle f(x, u, v), p \rangle, \tag{4}$$

then the domains for H are called *leadership domains* for f (see [10]). Leadership domains are always discriminating domains. The converse is false in general.

In [10], we proved the following theorem.

THEOREM 1.1. *Let $H : \mathbb{R}^N \times \mathbb{R}^N$ be a lower semicontinuous map. Any closed subset K of $\mathbb{R}^N$ contains a largest closed domain for H. This set is called the* kernel of K for H.

So any closed domain for H contained in a closed K is contained in the kernel of K for H. Moreover, this kernel is itself a domain for H. The kernel of K for H may be empty if K does not contain any domain for H.

• If the map H is defined by (3), then the kernel of K for H is called the *discriminating kernel* of K for f and is denoted by $Disc_f(K)$.

• If the map H is defined by (4), then the kernel of K for H is called the *leadership kernel* of K for f and is denoted by $Lead_f(K)$.

Note in particular that for any closed set K, $Lead_f(K) \subset Disc_f(K)$.

1.4. Assumptions and notation. We summarize here the assumptions we shall need on the dynamics throughout this paper. The first ones are concerned with the regularity properties of f:

$$\text{(5)} \qquad \begin{cases} \text{(i)} & \text{U and V are metric compact spaces,} \\ \text{(ii)} & f : \mathbb{R}^N \times U \times V \to \mathbb{R}^N \text{ is continuous,} \\ \text{(iii)} & f(\cdot, u, v) \text{ is an } \ell\text{-Lipschitz map for any } u \text{ and } v. \end{cases}$$

For the study of the discriminating domains and kernels, we also require some convexity properties[3]:

$$\text{(6)} \qquad \begin{cases} \text{(i)} & V \text{ is a convex compact subset of } \mathbb{R}^d \ (d \in \mathbb{N}^*), \\ \text{(ii)} & f \text{ is affine in } v. \end{cases}$$

When (5) is satisfied, we shall denote by $x[x_0, u(\cdot), v(\cdot)]$ the unique solution of the differential equation

$$\text{(7)} \qquad \begin{cases} x'(t) = f(x(t), u(t), v(t)) \text{ for almost every } t \geq 0, \\ x(0) = x_0, \end{cases}$$

where $x_0 \in \mathbb{R}^N$, $u(\cdot) \in \mathcal{U}$, and $v(\cdot) \in \mathcal{V}$.

We shall also denote by B the closed unit ball of the state space $\mathbb{R}^N$, while $\overset{o}{B}$ shall denote the open unit ball. Moreover, the distance map to a closed set K shall be denoted by d_K:

$$d_K(x) := \min_{y \in K} \|y - x\|.$$

For any positive ϵ, $K + \epsilon B$ shall denote the closed set

$$K + \epsilon B := \{y \in \mathbb{R}^N \mid d_K(y) \leq \epsilon\}.$$

2. Statement of the main results.

2.1. Discriminating domains and kernels. We first provide an interpretation of discriminating domains for nonanticipative strategies.

The following theorem states that the discriminating domains are sets in which Victor can ensure that the state of the system remains as soon as he knows what Ursula plays.

THEOREM 2.1 (interpretation theorem). *Assume that f satisfies* (5) *and* (6) *and that D is a closed subset of $\mathbb{R}^N$. Then, the closed set D is a discriminating domain for f if and only if, for any x_0 belonging to D, there exists a nonanticipative strategy β, such that, for any $u(\cdot)$ belonging to $\mathcal{U}$, the solution $x[x_0, u(\cdot), \beta(u(\cdot))]$ remains in D on $[0, +\infty)$. Namely,*

$$\forall t \geq 0, \ x[x_0, u(\cdot), \beta(u(\cdot))](t) \in D.$$

[3]By *affine*, we mean of the form

$$f(x, u, v) = a(x, u) + B(x, u)v.$$

(Recall that $x[x_0, u(\cdot), \beta(u(\cdot))]$ *is the solution to* (7) *with* $v(\cdot) := \beta(u(\cdot))$.*)*

Section 4 is devoted to the proof of Theorem 2.1.

Remark. Theorem 2.1 still holds true even if we only assume that D is a *locally compact set.* But in this case, we have to modify the statement of the theorem: A locally compact set D is a discriminating domain if and only if there exist a nonanticipative strategy β *and a positive time* T, such that, for any control $u(\cdot)$ belonging to $\mathcal{U}$, the solution $x[x_0, u(\cdot), \beta(u(\cdot))]$ remains in D *on* $[0, T]$.

If a discriminating domain D is contained in a closed set K, then Victor can ensure that the state of the system remains in D and so in K. This is in particular the case for the discriminating kernel of K. We prove here a kind of converse: If Victor can ensure that the state of the system remains in K, then the initial position of the system necessarily belongs to the discriminating kernel of K.

THEOREM 2.2 (characterization theorem). *Let* K *be a closed subset of* $\mathbb{R}^N$ *and* f *satisfy* (5) *and* (6). *Then, the discriminating kernel of* K *for* f *is equal to the set of points* x_0 *belonging to* K *for which there exists a nonanticipative strategy* $\beta : \mathcal{U} \to \mathcal{V}$ *such that, for any* $u(\cdot) \in \mathcal{U}$, *the solution* $x[x_0, u(\cdot), \beta(u(\cdot))]$ *remains in* K.

Proof of Theorem 2.2. From Theorem 2.1, for any x_0 belonging to $Disc_f(K)$, there exists a nonanticipative strategy $\beta : \mathcal{U} \to \mathcal{V}$ such that, for any $u(\cdot) \in \mathcal{U}$, the solution $x[x_0, u(\cdot), \beta(u(\cdot))]$ remains in $Disc_f(K)$ on $[0, +\infty)$, and so in K.

To prove the converse, let x_0 belong to K but not to $Disc_f(K)$. Let β be a nonanticipative strategy for Ursula. We have to construct a control $u(\cdot) \in \mathcal{U}$ such that the solution $x[x_0, u(\cdot), \beta(u(\cdot))]$ leaves K in finite time.

For that purpose, let us recall that the viability kernel of a closed set C for a Marchaud set-valued map[4] $F : \mathbb{R}^N \to \mathbb{R}^N$ is the set of initial positions from which starts at least a solution of the differential inclusion for F which remains in K forever. The viability kernel of K for F is a (maybe empty) closed subset of K. It is denoted by $Viab_F(K)$. (Most results on viability theory can be found in Aubin's monograph [4].)

In [10], we have proved that the discriminating kernel of a closed set can be obtained as decreasing intersections of viability kernels. Namely, if we define the sequence of closed sets

$$\begin{cases} K_1 := K, \\ K_{i+1} := \bigcap_{u \in U} Viab_{f(\cdot,u,V)}(K_i), \end{cases}$$

where $f(\cdot, u, V)$ denotes the set-valued map $x \rightsquigarrow \bigcup_{v \in V} f(x, u, v)$, then

$$\bigcap_i K_i = Disc_f(K).$$

Since $x_0 \notin Disc_f(K)$, there is some $i_0 \in \mathbb{N}^*$ such that x_0 belongs to K_{i_0} but not to K_{i_0+1}.

From the very definition of K_{i_0+1}, there exists $u_1 \in U$, such that x_0 does not belong to $Viab_{f(\cdot,u_1,V)}(K_{i_0})$. Thus any solution to the differential inclusion for $f(\cdot, u_1, V)$ starting from x_0 leaves K_{i_0} in finite time. The solution $x[x_0, u_1, \beta(u_1)]$ is a solution to the differential inclusion for $f(\cdot, u_1, V)$, so there is a time t_1 such that $x[x_0, u_1, \beta(u_1)](t_1)$ does not belong to K_{i_0}.

We set $u(\cdot) := u_1$ on $[0, t_1]$. If $x_1 := x[x_0, u(\cdot), \beta(u(\cdot))](t_1)$ does not belong to K_{i_0-1}, then we set $t_2 := t_1$. Otherwise, x_1 belongs to K_{i_0-1}, but not to K_{i_0}. So there exists $u_2 \in U$ such that x_1 does not belong to $Viab_{f(\cdot,u_2,V)}(K)$. We define an intermediate control $\bar{u}(\cdot)$ by $\bar{u}(s) := u(s)$ on $[0, t_1]$ and $\bar{u}(s) := u_2$ on $[t_1, +\infty)$. Since β is nonanticipative, $x[x_0, \bar{u}(\cdot), \beta(\bar{u}(\cdot))](t_1)$ equals x_1. Moreover, $x[x_0, \bar{u}(\cdot), \beta(\bar{u}(\cdot))]$ is a solution to the

[4] I.e., an upper semicontinuous set-valued map, with closed convex values and a linear growth (see [4]).

differential inclusion for $f(\cdot, u_2, V)$ on $[t_1, +\infty)$. Thus there is a time $t_2 \geq t_1$ such that $x[x_1, \bar{u}(\cdot), \beta(\bar{u}(\cdot))](t_2)$ does not belong to K_{i_0-1}. Now set $u(s) = \bar{u}(s)$ on $[t_1, t_2]$. Since $\bar{u}(\cdot)$ and $u(\cdot)$ coincide on $[0, t_2]$ and since β is nonanticipative, $x[x_0, u(\cdot), \beta(u(\cdot))](t_2)$ does not belong to K_{i_0-1}.

Thus it is possible to define by induction a nondecreasing sequence (t_i) and a control $u(\cdot)$ on $[0, t_i]$ such that $x[x_0, u(\cdot), \beta(u(\cdot))](t_i)$ does not belong to K_{i_0+1-i}. In particular, for $i = i_0$, $x[x_0, u(\cdot), \beta(u(\cdot))](t_{i_0})$ does not belong to $K_1 = K$. So we have constructed a control $u(\cdot)$ such that the solution $x[x_0, u(\cdot), \beta(u(\cdot))]$ leaves K in finite time. □

2.2. Leadership domains and kernels. If Ursula has a spy (she plays nonanticipative strategies), then leadership domains are the sets in which Victor can almost ensure that the state of the system remains.

THEOREM 2.3 (interpretation theorem). *Assume that f satisfies* (5) *and that D is a closed subset of $\mathbb{R}^N$. Then, the closed set D is a leadership domain for f if and only if, for any x_0 belonging to D, for any nonanticipative strategy $\alpha : \mathcal{V} \to \mathcal{U}$, for any positive ϵ, and for any time $T \geq 0$, there exists a control $v(\cdot) \in \mathcal{V}$ such that the solution $x[x_0, \alpha(v(\cdot)), v(\cdot)]$ remains*[5] *in $D + \epsilon B$ on $[0, T]$.*

The proof of Theorem 2.3 is provided in §4.

Remark. (1) Unfortunately, it is sometimes impossible to find a time-measurable control $v(\cdot)$ such that $x[x_0, \alpha(v(\cdot)), v(\cdot)]$ remains in the leadership domain D. For example, set $D := \{0\}$ (subset of $\mathbb{R}$) and $f(x, u, v) := \{u + v\}$, where u and v belong to $[-1, 1]$. If the nonanticipative strategy is

$$\begin{cases} \alpha(v(\cdot))(t) = v(t) \text{ if } v(t) \neq 0, \\ \alpha(v(\cdot))(t) = 1 \text{ if } v(t) = 0, \end{cases} \tag{8}$$

then there is no control $v(\cdot)$ such that the solution $x(\cdot) := x[0, \alpha(v(\cdot)), v(\cdot)]$ remains in D (because $x'(t) \neq 0$ for almost every $t \geq 0$).

So, in general, Victor needs knowledge of Ursula's control to prevent the state of the system from leaving K.

(2) Theorem 2.3 still holds true even if we only assume that D is a *locally compact set.* But in this case, we have to modify the statement of the theorem: A locally compact set D is a leadership domain if and only if, for any x_0 belonging to D, *there exists a positive T* such that, for any nonanticipative strategy α and for any $\epsilon > 0$, there is some control $v(\cdot) \in \mathcal{V}$ such that the solution $x[x_0, \alpha(v(\cdot)), v(\cdot)]$ remains in $D + \epsilon B$ on $[0, T]$.

We now characterize the leadership kernel of a closed set.

THEOREM 2.4 (characterization theorem). *Let K be a closed subset of $\mathbb{R}^N$ and f satisfy* (5). *Then, the leadership kernel of K for f is equal to the set of points x_0 belonging to K such that, for any nonanticipative strategy $\alpha : \mathcal{V} \to \mathcal{U}$, for any $\epsilon > 0$, and for any $T \geq 0$, there exists a control $v(\cdot) \in \mathcal{V}$ such that the solution $x[x_0, \alpha(v(\cdot)), v(\cdot)]$ remains in $K + \epsilon B$ on $[0, T]$.*

Theorem 2.4 yields, in particular, Corollary 2.1.

COROLLARY 2.1. *If x_0 belongs to K but not to $Lead_f(K)$, there exist positive T and ϵ and a nonanticipative strategy $\alpha : \mathcal{V} \to \mathcal{U}$, such that for any control $v(\cdot) \in \mathcal{V}$, the solution $x[x_0, \alpha(v(\cdot)), v(\cdot)]$ leaves $K + \epsilon B$ before T (i.e., there is a time $t \leq T$ such that $d_K(x[x_0, \alpha(v), v](t))$ is larger than ϵ).*

The proof of Theorem 2.4 is provided in §5.

3. Applications for the target problem. Let Ω be an open target of $\mathbb{R}^N$ and f satisfy (5), (6). We recall briefly the target problem. Two players, Ursula and Victor, control the

[5] Let us recall that we denote by $D + \epsilon B$ the set of points x such that $d_D(x) \leq \epsilon$.

dynamical system

$$x'(t) = f(x(t), u(t), v(t)).$$

Ursula, acting on u, wants the state of the system to reach the target Ω, while Victor, acting on v, wants the state of the system to avoid Ω.

3.1. The alternative theorem. Let us define now the victory domains of each player.

DEFINITION 3.1 (victory domains).

• Victor's victory domain *is the set of initial positions* $x_0 \notin \Omega$ *for which Victor can find a nonanticipative strategy* $\beta : \mathcal{U} \to \mathcal{V}$ *such that for any time-measurable control* $u(\cdot) \in \mathcal{V}$ *played by Ursula, the solution* $x[x_0, u(\cdot), \beta(u(\cdot))]$ *avoids* Ω *for any* $t \geq 0$, *i.e.,*

$$\forall t \geq 0,\ x[x_0, u(\cdot), \beta(u(\cdot))](t) \notin \Omega.$$

• Ursula's victory domain *is the set of initial positions* $x_0 \notin \Omega$ *for which Ursula can find a nonanticipative strategy* $\alpha : \mathcal{V} \to \mathcal{U}$, *positive* ϵ, *and* T *such that, for any* $v(\cdot) \in \mathcal{U}$ *played by Victor, the solution* $x[x_0, \alpha(v(\cdot)), v(\cdot)]$ *reaches the set* $\Omega_\epsilon := \{x \mid d_{\Omega^c}(x) \geq \epsilon\}$ *before* T. *Namely,*

$$\exists t \leq T,\ d_{\Omega^c}(x[x_0, \alpha(v(\cdot)), v(\cdot)](t)) \geq \epsilon.$$

Remark. Note that $\Omega_\epsilon \subset \Omega$.

Theorems 2.2 and 2.4 yield the following results.

THEOREM 3.1 (alternative theorem). *Assume that* f *satisfies* (5) *and* (6). *Set* $K := \mathbb{R}^N \backslash \Omega$. *Then*

• *Victor's victory domain is equal to* $Disc_f(K)$.

• *Ursula's victory domain is equal to* $K \backslash Lead_f(K)$.

Assume moreover that

$$Lead_f(K) = Disc_f(K). \tag{9}$$

Then the victory domains of the two players form a partition of the closed set K.

Note that equality (9) is fulfilled as soon as Isaacs' condition holds:

$$\forall (x, p) \in \mathbb{R}^{2N},\ \sup_u \inf_v \langle f(x, u, v), p\rangle = \inf_v \sup_u \langle f(x, u, v), p\rangle.$$

If equality (9) does not hold true, there are initial conditions x_0 (where x_0 belongs to $Disc_f(K)$ but not to $Lead_f(K)$) from which Victor wins if he knows Ursula's control and from which Ursula wins if she knows Victor's control.

Let us recall that a similar alternative theorem has been obtained by Krasovskii and Subbotin in the framework of the positional strategies (see [23]).

Here we characterize the victory domains by means of geometric conditions (as discriminating and leadership kernels of a closed set). This characterization is used in the joint work with Quincampoix and Saint-Pierre [9] to compute numerically the victory domains (see also [30] and [26] in the framework of control theory).

3.2. The barrier phenomenon. Since the kernel of a closed set for a map $H : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is the maximal domain for H contained in K, it enjoys some particular properties at its boundary. We have proved in [10] the following theorem.

THEOREM 3.2. *Assume that the map* $H : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ *is continuous and positively homogeneous in the second variable. Let* D *denote the kernel of a closed set* K *for* H. *Assume that a point* x *belongs to* ∂D *but not to* ∂K. *Then*

(1) *if* $v \in NP_D(x)$,

$$H(x, v) \leq 0.$$

(2) *if* $\mu \in NP_{\overline{\mathbb{R}^N \setminus D}}(x)$,

$$H(x, -\mu) \geq 0.$$

Let us point out that (1) follows from the very definition of the domains for H (recall that D is itself a domain for H). So the important part of the theorem is the second assertion.

Let us further point out that, if $\mu \in NP_{\overline{\mathbb{R}^N \setminus D}}(x)$, then the ball $x + \mu + \|\mu\| B$ is contained in D. Roughly speaking, μ is an inward proximal normal. So, if the boundary of D has nonzero outward and inward proximal normals ($\mu = -\nu$ up to a positive multiplicative coefficient), then $H(x, \nu) = 0$. In the case when $H(x, p) := \sup_u \inf_v \langle f(x, u, v), p\rangle$, (1) and (2) can be interpreted as a generalized definition of Isaacs' equation:

$$\sup_u \inf_v \langle f(x, u, v), p\rangle = 0, \tag{10}$$

where p is an outward normal to the smooth boundary of the victory domains.

Isaacs' equation plays a main role in two-player differential game theory. See for instance [21], [7] and the references therein. Smooth surfaces satisfying (10) are called semipermeable surfaces by Isaacs. The reason is that both players can prevent the state of the system from crossing this surface in one direction. In general, the boundary of the victory domains is not smooth, and Isaacs' methods do not work anymore. We prove below that, in any case, the boundary of the victory domains is almost semipermeable. Similar results have been obtained by Quincampoix in the framework of control theory[6] (see [25]).

THEOREM 3.3. *Assume that f satisfies* (5) *and* (6). *Let x_0 belong to $\partial Disc_f(K)$ but not to ∂K. If $\beta : \mathcal{U} \to \mathcal{V}$ is a winning nonanticipative strategy*[7] *for Victor, then there is a time $T > 0$ such that, for any positive ϵ, Ursula can find a control $u(\cdot) \in \mathcal{U}$ such that the solution $x[x_0, u(\cdot), \beta(u(\cdot))]$ remains in $\partial Disc_f(K) + \epsilon B$ on $[0, T]$, i.e.,*

$$\forall t \leq T,\ d_{\partial Disc_f(K)}(x[x_0, u(\cdot), \beta(u(\cdot))](t)) \leq \epsilon.$$

So Ursula can ensure that the state of the system almost remains on the boundary of her victory domain.

Proof of Theorem 3.3. The proof of this result is a direct application of the second remark following Theorem 2.1 applied to the dynamics g defined by

$$\forall x \in \mathbb{R}^N,\ \forall \tilde{u} \in \tilde{U},\ \forall \tilde{v} \in \tilde{V},\ g(x, \tilde{u}, \tilde{v}) := f(x, \tilde{v}, \tilde{u}),$$

where $\tilde{U} := V$ and $\tilde{V} := U$. Indeed, choose $x_0 \in \partial Disc_f(K) \setminus \partial K$ and $R > 0$ such that $x_0 + RB$ is contained in K. From Theorem 3.2, the set $\hat{D} := \overline{Disc_f(K) \setminus K}$ satisfies the following condition on $x_0 + RB$:

$$\forall x \in x_0 + R\overset{o}{B},\ \forall p \in NP_{\hat{D}}(x),\ \sup_u \inf_v \langle f(x, u, v), -p\rangle \geq 0.$$

Since $\sup_u \inf_v \langle f(x, u, v), -p\rangle = -\inf_{\tilde{u}} \sup_{\tilde{v}} \langle g(x, \tilde{u}, \tilde{v}), p\rangle$, we have proved that the locally compact set $\hat{D} \cap (x_0 + R\overset{o}{B})$ is a leadership domain. So the second remark following Theorem 2.1 can be applied. □

[6] Quincampoix's results are concerned with the boundary of the viability kernel and of the invariance kernel (see also the monograph [4]). The reader may refer to [29] for the case where the dynamics are only upper semicontinuous.

[7] A winning nonanticipative strategy for Victor is a strategy which ensures that the state of the system remains in K forever, and so in $Disc_f(K)$.

4. Proof of the interpretation theorems. We prove in this section Theorems 2.1 and 2.3. The proof for the discriminating domains involves results of viability theory. The proof for the leadership domains uses more technical estimates.

4.1. Proof for the discriminating domains.

4.1.1. Necessary condition. Let D be a closed set which enjoys the discriminating property. By definition, for any $x_0 \in D$, there exists a nonanticipative strategy $\beta : \mathcal{U} \to \mathcal{V}$ such that for any $u(\cdot) \in \mathcal{U}$, the solution $x[x_0, u(\cdot), \beta(u(\cdot))]$ remains in D. Let us show that D is a discriminating domain.

For that purpose, fix $x_0 \in D$, β a nonanticipative strategy as above, and $p \in NP_D(x_0)$. We have to prove that $\sup_u \inf_v \langle f(x_0, u, v), p\rangle \leq 0$. Assume on the contrary that

$$\sup_u \inf_v \langle f(x_0, u, v), p\rangle \geq a > 0.$$

There is some $\bar{u} \in U$ such that

$$\forall v \in V, \ \inf_v \langle f(x_0, \bar{u}, v), p\rangle \geq a. \tag{11}$$

If we set $u(t) = \bar{u}$ for any t, the solution $x(\cdot) := x[x_0, u(\cdot), \beta(u(\cdot))]$ remains in D forever from the very definition of β. Since the open ball centered in $x_0 + p$ and of radius $\|p\|$ does not intersect D (p is a proximal normal to D at x_0), $x(\cdot)$ satisfies

$$\forall t \geq 0, \ \|x(t) - (x_0 + p)\| \geq \|p\|. \tag{12}$$

Let us compute now the derivative of $\|x(t) - (x_0 + p)\|^2$:

$$\begin{aligned}
\frac{d}{dt}\frac{1}{2}\|x(t) - (x_0 + p)\|^2 &= \langle x'(t), x(t) - (x_0 + p)\rangle \\
&= \langle f(x(t), \bar{u}, \beta(u)(t)), x(t) - x_0\rangle \\
&\qquad -\langle f(x(t), \bar{u}, \beta(u)(t)), p\rangle \\
&\leq \|x(t) - x_0\|(\|f(x(t), \bar{u}, \beta(u)(t))\| + \ell\|p\|) \\
&\qquad -\langle f(x_0, \bar{u}, \beta(u)(t)), p\rangle \\
&\leq -a/2
\end{aligned}$$

for almost every t small enough (recall that f is ℓ-Lipschitz and satisfies (11)). Thus

$$\|x(t) - (x_0 + p)\|^2 \leq \|p\|^2 - (at)/2 < \|p\|^2$$

for t small enough. This contradicts (12). Thus we have proved that any closed set D satisfying the discriminating property satisfies also $\sup_u \inf_v \langle f(x_0, u, v), p\rangle \leq 0$, for any $x_0 \in D$ and any proximal normal p to D at x. This means that D is a discriminating domain. □

4.1.2. Sufficient condition. Assume now that D is a discriminating domain. Let x_0 belong to D. We want to define a nonanticipative strategy $\beta : \mathcal{U} \to \mathcal{V}$, such that for any $u(\cdot)$ belonging to $\mathcal{U}$, the solution $x[x_0, u(\cdot), \beta(u(\cdot))]$ remains in D.

For that purpose, we introduce the following set-valued map:

$$B(u(\cdot)) = \{v(\cdot) \in \mathcal{V} \mid x[x_0, u(\cdot), v(\cdot)] \text{ remains in } D\}.$$

The set-valued map $B(\cdot) : \mathcal{U} \rightsquigarrow \mathcal{V}$ has nonempty values. Indeed D is a discriminating domain

and the measurable viability theorem[8] [18] states that, for any $x_0 \in D$, for any $u(\cdot) : \mathbb{R}^+ \to U$ measurable, there exists a solution $x(\cdot)$ of the differential inclusion

$$\text{(13)} \qquad \begin{cases} x'(t) \in f(x(t), u(t), V) \text{ for almost all } t \geq 0, \\ x(0) = x_0, \end{cases}$$

which remains in D. Usual selections theorems state (see also Theorem 3.1.1 of [17]) that the solutions to (13) are the same as the solutions to

$$\begin{cases} x'(t) = f(x(t), u(t), v(t)) \text{ for almost all } t \geq 0, \\ v(t) \in V \text{ for almost all } t \geq 0, \\ x(0) = x_0. \end{cases}$$

Thus there exists a control $v(\cdot)$ belonging to $\mathcal{V}$ such that $x(\cdot) = x[x_0, u(\cdot), v(\cdot)]$, i.e., $v(\cdot) \in B(u(\cdot))$.

We claim that $B(\cdot)$ has compact values for the weak topology of the Hilbert space $L^2([0, +\infty[, \mathbb{R}^d, e^{-t})$ (which is $L^2([0, +\infty), \mathbb{R}^d)$ supplied with the Euclidean norm $\|v(\cdot)\|_{L^2} := \int_0^{+\infty} \|v(t)\|^2 e^{-t} dt$) and that $B(\cdot)$ is a nonanticipative set-valued map. Recall that $V \subset \mathbb{R}^d$.

$B(\cdot)$ has weakly compact values. Since V is a convex compact subset of $\mathbb{R}^d$, $\mathcal{V}$ is weakly compact. Let $v_p(\cdot)$ belong to $B(u(\cdot))$. Up to a subsequence (again denoted by $v_p(\cdot)$), $v_p(\cdot)$ converges for the weak topology to some control $v(\cdot)$ belonging to $\mathcal{V}$. Thus, from Alaoglu and Ascoli theorems, the solutions $x[x_0, u(\cdot), v_p(\cdot)]$ converge to the solution $x[x_0, u(\cdot), v(\cdot)]$ for the compact convergence because f is affine in v (for more details, see for instance the proof of the Convergence Theorem 2.4.4 in [4]). Thus $x[x_0, u(\cdot), v(\cdot)]$ remains in D on $[0, +\infty)$, i.e., $v(\cdot)$ belongs to $B(u(\cdot))$. So $B(\cdot)$ has nonempty, compact values.

The set-valued map $B(\cdot)$ is nonanticipative. We say that the set-valued map $B : \mathcal{U} \rightsquigarrow \mathcal{V}$ is nonanticipative if, for any $s \geq 0$, u_1 and u_2 belonging to $\mathcal{U}$ coincide almost everywhere on $[0, s]$, then, for any v_1 belonging to $B(u_1)$, one can find v_2 belonging to $B(u_2)$ which coincides with v_1 almost everywhere on $[0, s]$. We now check that the set-valued map $B(\cdot)$ is nonanticipative.

Indeed, since u_1 and u_2 coincide almost everywhere on $[0, T]$, $x[x_0, u_2, v_1]$ remains in D on $[0, T]$, from the very definition of v_1. Set $v_2(\cdot) := v_1(\cdot)$ on $[0, T]$. The measurable viability theorem yields the existence of a control $v(\cdot)$ such that the solution to the differential equation for $f(\cdot, u_2(t), v(t))$, starting at time T from $x[x_0, u_1, v_1](T)$, remains in D. Set $v_2(\cdot) := v(\cdot)$ on $(T, +\infty)$. Then $v_2(\cdot)$ belongs to $B(u_2(\cdot))$ and coincides almost everywhere on $[0, T]$ with $v_1(\cdot)$.

Thus, to prove Theorem 2.1, it is enough to find a nonanticipative selection of $B(\cdot)$, i.e., a nonanticipative map $\beta : \mathcal{U} \to \mathcal{V}$ such that

$$\forall u(\cdot) \in \mathcal{U}, \ \beta(u(\cdot)) \in B(u(\cdot)).$$

The following lemma provides the existence of a nonanticipative selection of nonanticipative set-valued maps. This lemma is due to Plaskacz [24].

LEMMA 4.1 (Plaskacz). *Let $B(\cdot) : \mathcal{U} \to \mathcal{V}$ be a nonanticipative set-valued map with nonempty, closed values for the weak topology of $L^2(\mathbb{R}^+, V, e^{-t})$. Then there is a nonanticipative selection $\beta(\cdot)$ of $B(\cdot)$.*

---

[8] To apply this result, we have to recall that, under assumptions (5) and (6), the following statements are equivalent (see [10]):

(1) D is a discriminating domain for f.

(2) $\forall x \in D, \ \forall u \in U, \ f(x, u, V) \cap T_D(x) \neq \emptyset$, where $T_D(x) := \{v \in \mathbb{R}^N \mid \liminf_{h \to 0^+} d_D(x + hv)/h = 0\}$.

Thus, if D is a discriminating domain, for any $u(\cdot) \in \mathcal{U}$, the following condition is fulfilled:

$$\forall x \in D, \ f(x, u(t), V) \cap T_D(x) \neq \emptyset \text{ for almost every } t \geq 0.$$

So we can apply Theorem 1 of [18].

Since this result is still unpublished, we provide a proof of Lemma 4.1 in the Appendix for the convenience of the reader.

4.2. Proof for the leadership domains. Let us now prove Theorem 2.3.

4.2.1. Sufficient condition. Suppose that D is not a leadership domain. We have to contradict the conclusion of Theorem 2.3, i.e., to find some x_0 belonging to D and some positive ϵ and T and we have to build a nonanticipative strategy $\alpha(\cdot)$, such that, for any v belonging to $\mathcal{V}$, the solution $x[x_0, \alpha(v), v]$ leaves $D + \epsilon B$ before T.

Since D is not a leadership domain, there are some $x_0 \in D$ and some proximal normal $\nu \in NP_D(x_0)$ such that

$$\inf_v \sup_u \langle f(x_0, u, v), \nu \rangle \;=\; 2a \;>\; 0.$$

Without loss of generality, we can assume that $\|\nu\| \leq 1$. (Recall that $NP_D(x_0)$ is closed and convex and that 0 belongs to $NP_D(x_0)$.)

Since the maps $f(\cdot, u, v)$ are ℓ-Lipschitz for any u and v,

$$\forall y \in \left[x_0 + \frac{a}{\ell} B\right], \;\; \inf_v \sup_u \langle f(y, u, v), \nu \rangle \;\geq\; a.$$

The set-valued map

$$v \rightsquigarrow \left\{ u \in U \mid \inf_{y \in x_0 + \frac{a}{\ell} B} \langle f(y, u, v), \nu \rangle \;\geq\; a \right\}$$

is measurable and has nonempty closed values. (It has a closed graph.) Thus Measurable Selection Theorem 8.1.3 of [5] states that there exists a measurable selection $\pi : V \to U$ of this set-valued map.

Define now the map $\alpha(\cdot) : \mathcal{V} \to \mathcal{U}$ in the following way:

$$\forall v(\cdot) \in \mathcal{V}, \;\; \alpha(v(\cdot))(t) = \pi(v(t)).$$

Note that $\alpha(v(\cdot))$ is measurable because it is the composition of two measurable maps. Moreover, $\alpha(\cdot)$ is obviously a nonanticipative strategy.

Let M be a bound of $\|f(\cdot, \cdot, \cdot)\|$ on $[x_0 + \frac{a}{\ell} B] \times U \times V$. Fix $v(\cdot) \in \mathcal{V}$ and set $x(\cdot) := x[x_0, \alpha(v(\cdot)), v(\cdot)]$. We claim that, for any $t \in [0, \frac{a}{M\ell}]$,

$$\langle x(t) - x_0, \nu \rangle \geq at \;\text{ and }\; \|x(t) - x_0\| \leq Mt. \tag{14}$$

If our claim holds true, we can make the following statement.

LEMMA 4.2. *Let D be a closed subset of $\mathbb{R}^N$ and x belong to D. Assume that ν is a proximal normal to D at x. Let M and a be positive. Then, for any $0 < t < \frac{2a}{M^2}$, one has*

$$\left.\begin{array}{l} \|y - x\| \leq Mt, \\ \langle y - x, \nu \rangle \geq at \end{array}\right\} \Rightarrow \;\; d_D(y) \;\geq at - \frac{M^2 t^2}{2}.$$

Set $\bar{t} := \inf\{\frac{a}{M\ell}, \frac{2a}{M^2}\}$ and $\epsilon := [a - \frac{M^2 \bar{t}}{2}]\bar{t}$. (Note that both $\bar{t}$ and ϵ are positive.) Combining equation (14) with Lemma 4.2 yields $d_D(x(\bar{t})) \geq \epsilon$. Since $\bar{t}$ and ϵ do not depend on $v(\cdot)$, we have defined a nonanticipative strategy $\alpha(\cdot)$ such that, for any $v(\cdot)$ belonging to $\mathcal{U}$, the solution to $x[x_0, \alpha(v(\cdot)), v(\cdot)]$ leaves $D + \epsilon B$ before $\bar{t}$. This is the desired conclusion.

To prove that (14) holds true, note that, for $t \leq \frac{a}{M\ell}$, $x(t)$ belongs to $x_0 + \frac{a}{\ell}$. Thus the following estimates hold:

$$\langle x'(t), v\rangle \geq a \text{ and } \|x'(t)\| \leq M \text{ for almost all } t \in \left[0, \frac{a}{M\ell}\right].$$

We obtain (14) after integration, using the fact that $(\|x(t)\|)' \leq \|x'(t)\|$ for almost every $t \geq 0$. □

Proof of Lemma 4.2. Let us first prove that, if y and t are as in Lemma 4.2, then y belongs to the ball $x + v + \|v\|B$:

$$\|x + v - y\|^2 = \|x - y\|^2 - 2\langle y - x, v\rangle + \|v\|^2 \leq \|v\|^2 + M^2t^2 - 2at.$$

Thus, for $t \in (0, \frac{2a}{M^2})$, $\|x + v - y\|$ is not greater than $\|v\|$, which means that y belongs to $x + v + \|v\|B$.

Since v is a proximal normal, the distance from y to D is less than or equal to the distance from y to the complement of $x + v + \|v\|B$. Thus

$$d_D(y) \geq \|v\| - \|x + v - y\| \geq \|v\| - [\|v\|^2 + M^2t^2 - 2at]^{\frac{1}{2}}.$$

Since, for $s \in [-1, 0]$, one has $1 - [1 + s]^{\frac{1}{2}} \geq -\frac{s}{2}$, the proof is complete. □

4.2.2. Necessary condition. We now suppose that D is a leadership domain. Fix $x_0 \in D$, α a nonanticipative strategy, $\epsilon > 0$, and $T > 0$. We construct a control $v(\cdot)$ such that the solution $x[x_0, \alpha(v(\cdot)), v(\cdot)]$ remains in $D + \epsilon B$ on $[0, T]$.

The proof is divided into three steps. We first state two lemmas. Then we construct the control $v(\cdot)$. Finally, we prove that $v(\cdot)$ satisfies the required condition.

Preliminary lemmas. Throughout the proof, we keep the notation of the following lemma, which shall be proved later.

LEMMA 4.3. *With the notation and the assumptions of Theorem* 2.1, *there is some radius R such that, for any $u(\cdot)$ belonging to $\mathcal{U}$ and any $v(\cdot)$ belonging to $\mathcal{V}$, the solution $x[x_0, u(\cdot), v(\cdot)]$ remains in $x_0 + RB$ on $[0, 2T]$. We denote by M a bound of $\|f(\cdot, \cdot, \cdot)\|$ on $(x_0 + RB) \times U \times V$.*

In particular, Lemma 4.3 states that we can proceed as if f is bounded by M, because we study the solutions only on $[0, T]$.

We now need the following lemma.

LEMMA 4.4. *Assume that the map f satisfies* (5) *and is bounded by M. Let D be a leadership domain. Let $\bar{x} \notin D$ and y belong to the projection of $\bar{x}$ onto D. Choose $\bar{v} \in V$ such that*

$$\sup_u \langle f(\bar{x}, u, \bar{v}), \bar{x} - y\rangle \leq 0.$$

There are positive constants c and τ (which depend only on ℓ and M) such that, for any $u(\cdot) \in \mathcal{U}$, the following estimate holds true:

$$\forall t \leq \tau,\ d_D(x[\bar{x}, u(\cdot), \bar{v}](t))^2 \leq ct^2 + d_D^2(\bar{x})e^{2\ell t}.$$

Lemma 4.4 is proved below.

We are now ready to construct the control $v(\cdot)$. (We keep the notation of Lemmas 4.3 and 4.4.)

Construction of $v(\cdot)$. Fix $a \in (0, \tau)$ small enough such that

$$M^2a^2e^{2T\ell} + ca^2\frac{e^{2T\ell} - 1}{e^{2\ell a} - 1} < \epsilon^2.$$

Let $(t_p = ap)_{p \le T/a}$ be a subdivision of $[0, T]$. We define $v(\cdot)$ recursively, in such a way that $v(\cdot)$ is constant on any interval $[t_p, t_{p+1})$.

On $[0, t_1)$, we set $v(t) = v_1$, where v_1 is any element of V.

Assume we have defined $v(\cdot)$ on $[0, t_p)$. For simplicity, we set $x_p := x[x_0, \alpha(v), v](t_p)$. Since α is nonanticipative, x_p only depends on the restriction of $v(\cdot)$ to $[0, t_p]$.

• If x_p belongs to D, we set $v(t) := v_{p+1}$ on $[t_p, t_{p+1})$, where v_{p+1} is any element of V.

• If x_p does not belong to D, then let y_p be a projection of x_p onto D. Since D is a leadership domain and $x_p - y_p$ is a proximal normal to D at y_p, there exists v_{p+1} such that $\sup_u \langle f(y_p, u, v_{p+1}) , x_p - y_p \rangle \le 0$. Then we set $v(\cdot) := v_{p+1}$ on $[t_p, t_{p+1})$.

From now on, we set $x(t) := x[x_0, \alpha(v), v](t)$.

$d_D(x(t)) \le \epsilon$ *for* $t \in [0, T]$. To show this, we apply Lemma 4.4 with $u(\cdot) := \alpha(v(\cdot))$. Note that in the worst case (i.e., the case when $\sup_{t\in[0,T]} d_D(x(t))$ is maximum), the $x(t_p)$ do not belong to D once $p \ge 1$.

If t belongs to $[t_p, t_{p+1})$, then $v(t) = v_{p+1}$, where v_{p+1} satisfies

$$\sup_u \langle f(y_p, u, v_{p+1}), x_p - y_p \rangle \le 0.$$

Since y_p is a projection of x_p onto D, we can apply Lemma 4.4 to get

$$d_D^2(x(t)) \le c(t - t_p)^2 + d_D^2(x_p) e^{2\ell a}$$

because $a \le \tau$.

Applying Lemma 4.4 again yields

$$d_D^2(x_p) \le ca^2 + d_D^2(x_{p-1}) e^{2\ell a}.$$

By induction we obtain

$$d_D(x_p)^2 \le d_D^2(x_1) e^{2p\ell a} + ca^2 \sum_{j=0}^{i-1} e^{2j\ell a}.$$

Note that, for any $t \le T$, there is $p \le T/a$ such that $t_p \le t < t_{p+1}$. Moreover,

$$d_D^2(x_1) \le \|x_0 - x_1\|^2 \le M^2 a^2$$

because f is bounded by M. So we have finally proved that

$$\forall t \le T, \quad d_D(x(t))^2 \le M^2 a^2 e^{2T\ell} + ca^2 \frac{e^{2T\ell} - 1}{e^{2\ell a} - 1}.$$

From the very definition of a, the right-hand side is smaller than ϵ^2. Thus, we have constructed a control $v(\cdot)$ such that, on $[0, T]$,

$$d_D(x[x_0, \alpha(v(\cdot)), v(\cdot)](t)) \le \epsilon,$$

as desired. □

Proof of Lemma 4.3. For simplicity, we set $x(\cdot) := x[x_0, u(\cdot), v(\cdot)]$, and $\lambda := \sup_u \sup_v \|f(x_0, u, v)\|$. The derivative of $\|x(t) - x_0\|$ is not larger than $\|x'(t)\|$ for almost every t. Thus, for almost every t, one has

$$\begin{aligned} \frac{d}{dt}\|x(t) - x_0\| &\le \|x'(t)\| \\ &\le \|f(x(t), u(t), v(t))\| \\ &\le \lambda + \ell\|x(t) - x_0\| \end{aligned}$$

because f is ℓ-Lipschitz. Gronwall's lemma yields

$$\|x(t) - x_0\| \leq \frac{\lambda}{\ell}[e^{\ell t} - 1].$$

Thus, if we set $R := \frac{\lambda}{\ell}[e^{\ell 2T} - 1]$, Lemma 4.3 follows. □

Proof of Lemma 4.4. For simplicity, set $x(t) := x[x, u(\cdot), \bar{v}](t)$. Let $t \leq T$ for which the derivative of $x(\cdot)$ exists and is equal to $f(x(t), u(t), \bar{v})$. Recall also that f is ℓ-Lipschitz and bounded by M. So we have the following estimate:

$$\begin{aligned}\left(\frac{1}{2}\|x(t) - y\|^2\right)' &= \langle f(x(t), u(t), \bar{v}) \ , \ x(t) - y\rangle \\ &\leq \langle f(y, u(t), \bar{v}), x(t) - y\rangle + \ell\|x(t) - y\|^2 \\ &\leq \langle f(y, u(t), \bar{v}), \bar{x} - y\rangle + M\|x(t) - \bar{x}\| + \ell\|x(t) - y\|^2.\end{aligned}$$

Recall that $\langle f(y, u, \bar{v}) \ , \ \bar{x} - y\rangle \leq 0$ for any u of U. Since $\|f(\cdot, \cdot, \cdot)\|$ is bounded by M, the distance between $x(t)$ and $\bar{x}$ is not larger than Mt. Thus

$$\left(\frac{1}{2}\|x(t) - y\|^2\right)' \leq M^2 t + \ell\|x(t) - y\|^2.$$

This inequality is fulfilled for almost every $t \geq 0$. Gronwall's lemma yields

$$\forall t \geq 0, \ \|x(t) - y\|^2 \leq \|\bar{x} - y\|^2 e^{2\ell t} - \frac{M^2}{\ell}t + \frac{M^2}{2\ell^2}[e^{2\ell t} - 1]. \tag{15}$$

Note that the map $\phi : t \to -\frac{M^2}{\ell}t + \frac{M^2}{2\ell^2}[e^{2\ell t} - 1]$ vanishes at $t = 0$, and $\phi'(0) = 0$. Thus there are positive constants c and τ (which depend only on M and ℓ) such that

$$\forall t \in [0, \tau], \ -\frac{M^2}{\ell}t + \frac{M^2}{2\ell^2}[e^{2\ell t} - 1] \leq ct^2.$$

In particular, (15) yields

$$\forall t \in [0, \tau), \ d_D^2(x(t)) \leq \|x(t) - y\|^2 \leq d_D^2(\bar{x})e^{2\ell t} + ct^2.$$

So we have proved Lemma 4.4. □

5. Proof of the characterization theorem for the leadership kernel. In this section, we prove Theorem 2.4.

From now on, symbols involving α (like $\alpha(\cdot), \bar{\alpha}(\cdot)$), always denote nonanticipative strategies from $\mathcal{V}$ to $\mathcal{U}$.

Set

$$\begin{aligned}L := \{x_0 \in K \mid \forall \alpha(\cdot), \ \forall \epsilon > 0, \ \forall T > 0, \\ \exists v(\cdot) \in \mathcal{V} \text{ such that} \\ \forall t \leq T, \ x[x_0, \alpha(v(\cdot)), v(\cdot)](t) \in K + \epsilon B\}.\end{aligned}$$

We have to show that $L = Lead_f(K)$.

Inclusion $Lead_f(K) \subset L$ follows directly from Theorem 2.1. To prove the opposite inclusion, it is sufficient to show that L is a leadership domain, because $Lead_f(K)$ contains any leadership domain contained in K. For that purpose, we first prove that L is closed and then that L is a leadership domain.

L is closed. This claim is a consequence of the following lemma.

LEMMA 5.1. *If x belongs to K but not to L, there exist positive η, ϵ, and T and a nonanticipative strategy α such that, for any $y \in x + \eta B$, for any control $v(\cdot) \in \mathcal{V}$, the solution $x[y, \alpha(v), v]$ leaves $K + \epsilon B$ before T.*

In particular, the complement of L is open, and thus L is closed.

Proof of Lemma 5.1. From the very definition of L, there are positive ϵ and T and a nonanticipative strategy α, such that for any control v of $\mathcal{V}$, there is some $t \leq T$ with $x[x, \alpha(v), v](t) \notin K + 2\epsilon B$.

Since f is Lipschitz and V is compact, by Gronwall's lemma, there exists some $\eta > 0$ such that, for any $v \in \mathcal{N}$, for any $y \in x + \eta B$,

$$\|x[y, \alpha(v), v](t) - x[x, \alpha(v), v](t)\| \leq \epsilon, \text{ for } t \leq T.$$

Thus

$$\begin{aligned} &d_K(x[y, \alpha(v), v](t)) \\ &\quad \geq d_K(x[x_0, \alpha(v), v](t)) - \|x[y, \alpha(v), v](t) - x[x_0, \alpha(v), v](t)\| \geq \epsilon. \end{aligned}$$

This proves the lemma. □

L is a leadership domain. Assume that L is *not* a leadership domain. Then there is some $x_0 \in L$ such that

$$\text{(16)} \qquad \begin{cases} \exists \alpha(\cdot),\ \epsilon > 0,\ T > 0, \text{ such that } \forall v(\cdot) \in \mathcal{V} \\ \quad \exists t \leq T \text{ with } x[x_0, \alpha(v(\cdot)), v(\cdot)](t) \notin L + \epsilon B. \end{cases}$$

We are going to prove that, if x_0 satisfies (16), then it satisfies

$$\text{(17)} \qquad \begin{cases} \exists \bar{\alpha}(\cdot),\ \bar{\epsilon} > 0,\ \bar{T} > 0, \text{ such that } \forall v(\cdot) \in \mathcal{V} \\ \quad \exists t \leq \bar{T} \text{ with } x[x_0, \bar{\alpha}(v(\cdot)), v(\cdot)](t) \notin K + \bar{\epsilon} B. \end{cases}$$

Since (17) contradicts $x_0 \in L$, L is necessarily a leadership domain.

Let x_0 belong to L, and assume that x_0 satisfies (16). Let E be the closure of

$$\{y \notin L + \epsilon B \mid \exists v \in \mathcal{V} \text{ and } t \leq T \text{ with } y = x[x_0, \alpha(v), v](t)\}.$$

The set E is compact, because f has a linear growth. Moreover, $E \cap L = \emptyset$. For $x \in E$, we define $\eta_x > 0$, $\epsilon_x > 0$, $T_x > 0$, and $\alpha_x(\cdot)$ by the following statements:

- If $x \notin K$, then $\eta_x := d_K(x)/2$, $\epsilon_x := \eta_x$, $T_x := 0$, and $\alpha_x(\cdot) := \alpha(\cdot)$. (Recall that $\alpha(\cdot)$ is defined by (16).)
- If $x \in K$, then η_x, ϵ_x, T_x, and $\alpha_x(\cdot)$ are defined as in Lemma 5.1.

Note that $E \subset \bigcup_{x \in E}(x + \eta_x B)$. Thus there exist $x_1, \ldots, x_p$ belonging to E such that

$$\text{(18)} \qquad E \subset \bigcup_{i=1}^{p}(x_i + \eta_{x_i} B).$$

We are now ready to define

$$\bar{\epsilon} := \min_{i=1,\ldots,p} \epsilon_{x_i} \text{ and } \bar{T} := T + \max_{i=1,\ldots,p} T_{x_i}.$$

To define $\bar{\alpha}$, let us first define the hitting time of E:

$$\forall v(\cdot) \in \mathcal{V},\ \rho(v(\cdot)) := \inf\{t \geq 0 \mid x[x_0, \alpha(v), v](t) \in E\}.$$

Assumption (16) yields that $\rho(v(\cdot)) \leq T$ for any $v(\cdot)$ belonging to $\mathcal{V}$. Moreover, $x[x_0, \alpha(v(\cdot)), v(\cdot)](\rho(v(\cdot)))$ belongs to E, because E is closed.

We are now ready to define $\bar{\alpha}(v(\cdot))$.

• On $[0, \rho(v(\cdot))[$, we set $\bar{\alpha}(v(\cdot))(t) := \alpha(v(\cdot))(t)$.

• On $[\rho(v(\cdot)), +\infty)$, we set $\bar{\alpha}(v(\cdot))(t) := \alpha_i(v(\cdot))(t - \rho(v(\cdot)))$, where i is the smallest integer for which $x[x_0, \alpha(v), v](\rho(v(\cdot)))$ belongs to $x_i + \eta_i B$ (such an i exists from (18)).

We have to prove that $\bar{\alpha}$ is a nonanticipative strategy and that $\bar{\eta}$, $\bar{\epsilon}$, $\bar{T}$, and $\bar{\alpha}(\cdot)$ satisfy the conclusion of (17), i.e.,

$$\forall v(\cdot) \in \mathcal{V}, \ \exists t \leq \bar{T}, \text{ such that } x[x_0, \bar{\alpha}(v(\cdot)), v(\cdot)](t) \notin K + \bar{\epsilon}B.$$

$\bar{\alpha}$ is a nonanticipative strategy. Let $t > 0$ and let v_1 and v_2 belong to $\mathcal{V}$, such that v_1 and v_2 coincide almost everywhere on $[0, t]$. Let

$$t' = \max\{s \leq t \mid \bar{\alpha}(v_1) \equiv \bar{\alpha}(v_2) \text{ on } [0, s]\}.$$

Note that $x[x_0, \bar{\alpha}(v_1), v_1]$ and $x[x_0, \bar{\alpha}(v_2), v_2]$ coincide on $[0, t']$. We have to prove that $t' = t$. For that purpose, we have to discuss two cases.

(i) Assume that $t' < \rho(v_1)$. Then for any $s \leq t'$, $x[x_0, \bar{\alpha}(v_1), v_1](s) = x[x_0, \bar{\alpha}(v_2), v_2](s)$ does not belong to E. Since E is closed, there is $\theta > 0$ such that $\rho(v_1) \geq t' + \theta$ and $\rho(v_2) \geq t' + \theta$.

From the very definition of $\bar{\alpha}$, $\bar{\alpha}(v_1)$ equals $\alpha(v_1)$ on $[0, t' + \theta]$, and $\bar{\alpha}(v_2)$ equals $\alpha(v_2)$ on $[0, t' + \theta]$. Since α is nonanticipative, $\bar{\alpha}(v_1)$ and $\bar{\alpha}(v_2)$ coincide almost everywhere on $[0, \min(t, t' + \theta)]$. So $t' = \min(t, t' + \theta)$, and thus $t' = t$.

(ii) Assume now that $t' \geq \rho(v_1)$. Since $\bar{\alpha}(v_1)$ and $\bar{\alpha}(v_2)$ coincide almost everywhere on $[0, t']$, one has $\rho(v_1) = \rho(v_2)$. We denote by ρ this common value. Let i be the smallest integer such that $x[x_0, \bar{\alpha}(v_1), v_1](\rho) = x[x_0, \bar{\alpha}(v_2), v_2](\rho)$ belongs to $x_i + \eta_i B$.

From the very definition of $\bar{\alpha}$, $\bar{\alpha}(v_1)(t) = \alpha_i(v_1)(t - \rho)$ on $[\rho, \infty)$ and $\bar{\alpha}(v_2)(t) = \alpha_i(v_2)(t - \rho)$ on $[\rho, \infty)$. Since α_i is nonanticipative and since v_1 and v_2 coincide almost everywhere on $[\rho, t]$, $\bar{\alpha}(v_1)$ and $\bar{\alpha}(v_2)$ coincide almost everywhere on $[\rho, t]$. Thus $t = t'$.

In both cases we have proved that, for any $t \geq 0$, if v_1 and v_2 coincide almost everywhere on $[0, t]$, then $\bar{\alpha}(v_1)$ and $\bar{\alpha}(v_2)$ coincide almost everywhere on $[0, t]$. So $\bar{\alpha}$ is nonanticipative.

$\bar{\eta}$, $\bar{\epsilon}$, $\bar{T}$, and $\bar{\alpha}(\cdot)$ satisfy (17). Let v belong to $\mathcal{V}$. We want to prove that $x(\cdot) := x[x_0, \bar{\alpha}(v), v]$ leaves $K + \bar{\epsilon}$ before $\bar{T}$.

From the very definitions of $\bar{\alpha}(v)$ and α, $x(\cdot)$ reaches E at a time $\rho(v) \leq T$. Let i be the smallest integer such that $x(\rho(v))$ belongs to $x_i + \eta_i B$. Suppose either $x_i \notin K$ or $x_i \in K$. If $x_i \notin K$, then

$$d_K(x(\rho(v))) \geq d_K(x_i) - \|x_i - x(\rho(v))\| \geq 2\eta_i - \eta_i \geq \bar{\epsilon}.$$

If $x_i \in K$, then since $\bar{\alpha}(v)(\cdot) = \alpha_i(\cdot - \rho(v))$ on $[\rho(v), \infty)$, $x(\cdot)$ leaves $K + \epsilon_i B$ before $\rho(v) + T_i \leq \bar{T}$; this follows from Lemma 5.1 and from the very definition of $\alpha_i(\cdot)$. In both cases, $x(\cdot)$ leaves $K + \bar{\epsilon}B$ before $\bar{T}$.

So we have proved that (16) implies (17) and so the proof is complete. □

6. Appendix. We now provide a proof of the following lemma.

LEMMA 6.1 (Plaskacz). *Let $B(\cdot) : \mathcal{U} \to \mathcal{V}$ be a nonanticipative set-valued map with nonempty, closed values for the weak topology of $L^2(\mathbb{R}^+, V, e^{-t})$. There is a nonanticipative selection $\beta(\cdot)$ of $B(\cdot)$.*

Proof of Lemma 6.1. Let $\mathcal{B}$ be the set of the set-valued maps $A : \mathcal{U} \rightsquigarrow \mathcal{V}$ with weakly compact nonempty values, which are nonanticipative and which moreover satisfy $A(u(\cdot)) \subset B(u(\cdot))$ for any $u(\cdot)$ belonging to $\mathcal{M}$.

There is a natural partial order on $\mathcal{B}$:

$$[A_1 \preceq A_2] \Leftrightarrow [\forall u(\cdot) \in \mathcal{M},\ A_1(u(\cdot)) \subset A_2(u(\cdot))].$$

The proof is divided into two steps. In the first step, we prove that the order $\preceq$ is inductive. Then Zorn's lemma yields the existence of minimal set-valued maps for $\preceq$. In the second step, we show that minimal set-valued maps are maps, which achieves the proof.

$\preceq$ *is inductive.* For this purpose, let $(A_\lambda)_{\lambda\in\Lambda}$ be a totally ordered subset of $\mathcal{B}$. We denote by A the set-valued map

$$\forall u(\cdot) \in \mathcal{U},\ A(u(\cdot)) := \bigcap_{\lambda\subset\Lambda} A_\lambda(u(\cdot)).$$

Then obviously, $A \preceq A_\lambda$ for any $\lambda \in \Lambda$. We have to prove that A belongs to $\mathcal{B}$. So we have to show that (1) for any $u(\cdot) \in \mathcal{U}$, $A(u(\cdot))$ is nonempty, weakly compact, and contained in $B(u(\cdot))$ and (2) A is a nonanticipative set-valued map.

(1) For any $u(\cdot) \in \mathcal{U}$, $A_\lambda(u(\cdot))$ is a nonempty and weakly compact set, and the family $(A_\lambda(u(\cdot)))$ is totally ordered for the inclusion. Since the decreasing intersection of nonempty compact sets is nonempty and compact, A has nonempty weakly compact values. For any $u(\cdot) \in \mathcal{U}$, for any $\lambda \in \Lambda$, $A_\lambda(u(\cdot))$ is contained in $B(u(\cdot))$. Thus $A(u(\cdot))$ is contained in $B(u(\cdot))$.

(2) It remains to prove that A is nonanticipative. Let $s \geq 0$, and assume that $u_1(\cdot)$ coincides with $u_2(\cdot)$ on $[0, s]$. Let $v_1(\cdot)$ belong to $A(u_1(\cdot))$. We have to find $v_2(\cdot) \in A(u_2(\cdot))$ which coincides with $v_1(\cdot)$ on $[0, s]$.

Since $v_1(\cdot)$ belongs to $A(u_1(\cdot))$, $v_1(\cdot)$ belongs to $A_\lambda(u_1(\cdot))$ for any $\lambda \in \Lambda$. The set-valued maps A_λ are nonanticipative, so, for each λ, there exists $v_2^\lambda(\cdot) \in A_\lambda(u_2(\cdot))$, which coincides with $v_1(\cdot)$ on $[0, s]$.

Fix $\lambda \in \Lambda$, and set

$$P_\lambda := \text{closure}\{v_2^{\lambda'}(\cdot) \mid A_{\lambda'} \preceq A_\lambda\},$$

where the closure denotes the weak closure. Note that, for any $w(\cdot)$ belonging to $\{v_2^{\lambda'}(\cdot) \mid A_{\lambda'} \preceq A_\lambda\}$, $w(\cdot)$ coincides with $v_1(\cdot)$ on $[0, s]$. We use the following lemma to prove that this property also holds true for any $w(\cdot) \in P_\lambda$.

LEMMA 6.2. *Let E be a subset of $\mathcal{V}$ such that there exist $v_1(\cdot) \in \mathcal{V}$ and $s > 0$ with*

$$\forall w \in E,\ w \equiv v_1 \text{ almost everywhere on } [0, s].$$

Then the weak closure $\bar{E}$ of E also enjoys the following property:

$$\forall w \in \bar{E},\ w \equiv v_1 \text{ almost everywhere on } [0, s].$$

Proof. Let w belong to $\bar{E}$. There exists a sequence $w_p \in E$ which converges weakly to w. Thus,

$$\forall t \leq s,\ \int_0^t w_p(\sigma)d\sigma \to \int_0^t w(\sigma)d\sigma.$$

Since, by assumption, w_p and v_1 coincide on $[0, s]$, one has

$$\forall t \in [0, s],\ \int_0^t w_p(\sigma)d\sigma = \int_0^t v_1(\sigma)d\sigma.$$

So,

$$\forall t \in [0, s], \int_0^t w(\sigma)d\sigma = \int_0^t v_1(\sigma)d\sigma.$$

Thus w coincides with v_1 almost everywhere on $[0, s]$. □

Lemma 6.2 states that the elements of P_λ coincide with $v_1(\cdot)$ on $[0, s]$. The family P_λ is a decreasing family of weakly compact sets and P_λ is contained in $A_\lambda(u_2(\cdot))$ for any λ. Thus the set P defined by

$$P := \bigcap_{\lambda \in \Lambda} P_\lambda$$

is nonempty and compact. Moreover,

$$[\forall \lambda \in \Lambda,\ P_\lambda \subset A(u_2(\cdot))] \Rightarrow \left[P \subset \bigcap_{\lambda \in \Lambda} A_\lambda(u_2(\cdot)) = A(u_2(\cdot)) \right].$$

Since, for any $v_2(\cdot)$ belonging to P, $v_2(\cdot)$ belongs to P_λ, $v_2(\cdot)$ coincides with $v_1(\cdot)$ almost everywhere on $[0, s]$. So there exists $v_2(\cdot)$ which belongs to $A(u_2(\cdot))$ and coincides with $v_1(\cdot)$ on $[0, s]$.

For any $s \geq 0$, for any $u_1(\cdot)$ and $u_2(\cdot)$ which coincide almost everywhere on $[0, s]$, for any $v_1(\cdot)$ belonging to $A(u_1(\cdot))$, there exists $v_2(\cdot) \in A(u_2(\cdot))$ which coincides with $v_1(\cdot)$ almost everywhere on $[0, s]$. Thus A is nonanticipative, with nonempty compact values, and with a graph contained in the graph of B, i.e., A belongs to $\mathcal{B}$.

We have proved that the partial order on $\mathcal{B}$ is inductive. Zorn's lemma states that there exist minimal elements for this order. Let $\beta(\cdot) \in \mathcal{B}$ be such a minimal element. We claim that $\beta(\cdot)$ is a map. If our claim holds true, $\beta(\cdot)$ is a nonanticipative selection of B from the very definition of $\mathcal{B}$, and the proof of Lemma 6.1 is complete.

Minimal set-valued maps for $\preceq$ are maps. Assume that, contrary to our claim, the minimal set-valued map β is not a map. There exists $\bar{u}(\cdot) \in \mathcal{U}$, such that the cardinal of $\beta(\bar{u}(\cdot))$ is larger than 1. Let $\bar{v}(\cdot)$ belong to $\beta(\bar{u}(\cdot))$. We are going to construct a set-valued map $A \in \mathcal{B}$, such that $A \preceq \beta$ and $A(\bar{u}(\cdot)) = \{\bar{v}(\cdot)\}$ (in particular, $A \neq \beta$). Thus we obtain a contradiction, because, from the assumption, β is supposed to be minimal.

For any $u(\cdot) \in \mathcal{U}$, we define $A(u(\cdot))$ in the following way.

- If $u(\cdot) = \bar{u}(\cdot)$, we set $A(\bar{u}(\cdot)) := \{\bar{v}(\cdot)\}$.
- If $u(\cdot)$ does not coincide with $\bar{u}(\cdot)$ on any interval $[0, s]$ $(s > 0)$, then we set $A(u(\cdot)) := \beta(u(\cdot))$.
- If $u(\cdot)$ coincides with $\bar{u}(\cdot)$ on some interval $[0, t]$, let $[0, s]$ be the maximal interval on which $\bar{u}(\cdot)$ and $u(\cdot)$ coincide. Then we set

$$A(u(\cdot)) := \{w(\cdot) \in \beta(u(\cdot)) \mid w(\cdot) \equiv \bar{v}(\cdot) \text{ a.e. on } [0, s]\}.$$

Note that $A(u(\cdot)) \subset \beta(u(\cdot))$ for any $u(\cdot)$ belonging to $\mathcal{M}$ and that $A(\bar{u}(\cdot)) = \{\bar{v}(\cdot)\}$. So it remains to prove that A belongs to $\mathcal{B}$.

Since β is nonanticipative and has nonempty values, A has nonempty values. The graph of A is contained in the graph of β, and thus is contained in the graph of B. Moreover, the values of β are weakly compact, so by Lemma 6.2 the values of A are also weakly compact.

Next we prove that A is nonanticipative. For this purpose, let $s > 0$ and assume that $u_1(\cdot)$ and $u_2(\cdot)$ coincide almost everywhere on $[0, s]$. Without loss of generality, we also assume that $[0, s]$ is the maximal interval on which $u_1(\cdot)$ and $u_2(\cdot)$ coincide. Let v_1 belong to $A(u_1(\cdot))$. We have to find $v_2(\cdot) \in A(u_2(\cdot))$ which coincides with $v_1(\cdot)$ on $[0, s]$. There are two cases.

(1) $u_1(\cdot)$ does not coincide with $\bar{u}(\cdot)$ on any interval $[0, t]$ $(t > 0)$. In this case, $v_1(\cdot)$ belongs to $\beta(u_1(\cdot))$. Note moreover that $u_2(\cdot)$ does not coincide with $\bar{u}(\cdot)$ on any interval $[0, t]$ $(t > 0)$, because $u_1(\cdot)$ and $u_2(\cdot)$ coincide on $[0, s]$. Since β is nonanticipative, there is $v_2(\cdot)$ which belongs to $\beta(u_2(\cdot)) = A(u_2(\cdot))$ and which coincides with $v_1(\cdot)$ on $[0, s]$.

(2) $u_1(\cdot)$ coincides with $\bar{u}(\cdot)$ on $[0, t]$. We denote by $[0, t_1]$ the maximal interval on which $u_1(\cdot)$ and $\bar{u}(\cdot)$ coincide and by $[0, t_2]$ the maximal interval on which $u_2(\cdot)$ and $\bar{u}(\cdot)$ coincide. Note that, from the construction of A, $v_1(\cdot)$ and $\bar{v}(\cdot)$ coincide on $[0, t_1]$.

We have to find $v_2(\cdot) \in A(u_2(\cdot))$ which coincides with $v_1(\cdot)$ almost everywhere on $[0, s]$. From the construction of A, this is equivalent to finding $v_2(\cdot) \in \beta(u_2(\cdot))$ which coincides with $v_1(\cdot)$ on $[0, s]$ and with $\bar{v}(\cdot)$ on $[0, t_2]$.

We first prove that, in the set $\{s, t_1, t_2\}$, there are two elements that are equal and smaller than or equal to the third one. Indeed, assume for instance that $s \geq t_1 \geq t_2$. Then

$$\left.\begin{array}{l} u_1(\cdot) \equiv u_2(\cdot) \text{ on } [0, s], \\ u_1(\cdot) \equiv \bar{u}(\cdot) \text{ on } [0, t_1] \end{array}\right\} \Rightarrow u_2(\cdot) \equiv \bar{u}(\cdot) \text{ on } [0, t_1].$$

So $t_2 = t_1$. The other cases (i.e., $t_1 \geq t_2 \geq s$, and so on) can be treated in the same way. To find the desired $v_2(\cdot)$, we have to study three cases.

(1) $s \geq t_1 = t_2$. Since β is nonanticipative, we can find $v_2(\cdot) \in \beta(u_2(\cdot))$ which coincides with $v_1(\cdot)$ on $[0, s]$. Since $v_1(\cdot)$ coincides with $\bar{v}(\cdot)$ on $[0, t_1]$, $v_2(\cdot)$ coincides with $\bar{v}(\cdot)$ on $[0, t_2]$ because $t_1 = t_2$ and $s \geq t_2$. Thus $v_2(\cdot)$ belongs to $A(u_2(\cdot))$ and coincides with $v_1(\cdot)$ on $[0, s]$.

(2) $t_1 \geq s = t_2$. The same argument applies.

(3) $t_2 \geq s = t_1$. Since β is nonanticipative, we can find $v_2(\cdot) \in \beta(u_2(\cdot))$ which coincides with $\bar{v}(\cdot)$ on $[0, t_2]$. In particular, $v_2(\cdot)$ belongs to $A(u_2(\cdot))$ from the construction of A. Since $\bar{v}(\cdot)$ coincides with $v_1(\cdot)$ on $[0, t_1]$, $v_2(\cdot)$ and $v_1(\cdot)$ coincide on $[0, s]$ because $s = t_1$ and $s \leq t_2$.

So in any case, we have found $v_2(\cdot) \in A(u_2(\cdot))$ which coincides with $v_1(\cdot)$ on $[0, s]$. Thus A is nonanticipative, which completes the proof. □

Acknowledgments. The author wishes to express his gratitude to Prof. S. Plaskacz for fruitful discussions. He also wishes to thank the referees for their suggestions and remarks.

REFERENCES

[1] J.-P. AUBIN, *Contingent Isaacs' equation of a differential game*, in Differential Games and Applications, T. Başar and P. Bernhard, eds., Lecture Notes in Control and Inform. Sci., Springer-Verlag, Berlin, New York, 1989, pp. 51–61.

[2] ———, *Victory and defeat in differential games*, in Modeling and Control of Systems, Proceedings of the Bellman Continuum, 1988, A. Blaquière, ed., Lecture Notes in Control and Inform. Sci., Springer-Verlag, Berlin, New York, 1989.

[3] ———, *Differential games: A viability approach*, SIAM J. Control Optim., 28 (1990), pp. 1–27.

[4] ———, *Viability Theory*, Birkhäuser, Boston, MA, 1991.

[5] J.-P. AUBIN AND H. FRANKOWSKA, *Set-valued Analysis*, Birkhäuser, Boston, MA, 1990.

[6] P. BERNHARD, *Contribution à l'étude des jeux différentiels à deux joueurs, somme nulle, et information parfaite*, Thèse de Doctorat d'Etat, Université Pierre et Marie Curie–Paris 6, 1979.

[7] ———, *Differential games: Isaacs equation*, in Systems and Control Encyclopedia, Theory, Technology Applications, M. G. Singh, ed., Pergamon Press, Elmsford, NY, Oxford, 1988.

[8] J. V. BREAKWELL, *Zero-sum differential games with terminal payoff*, in Differential Games and Applications, P. Hagedorn, H. W. Knobloch, and G. H. Olsder, eds., Lecture Notes in Control and Inform. Sci. 3, Springer Verlag, Berlin, New York, 1977.

[9] P. CARDALIAGUET, M. QUINCAMPOIX, AND P. SAINT-PIERRE, *Some algorithms for a game with two-players and one target*, Math. Model. Num. Anal., 28 (1994), pp. 441–461.

[10] P. CARDALIAGUET, *Domaines discriminants en jeux différentiels*, Thèse, Université Paris IX Dauphine, 1994.

[11] A. G. Čencov, *On the structure of a game problem of convergence*, Soviet Math. Dokl., 16 (1975), pp. 1404–1408.
[12] ———, *On a game problem of guidance*, Soviet Math. Dokl., 17 (1976), pp. 73–77.
[13] ———, *On a game problem of converging at a given instant*, Math. USSR Sbornik, 28 (1976), pp. 353–376.
[14] N. J. Elliot and N. J. Kalton, *The existence of value in differential games*, Mem. Amer. Math. Soc., 126 (1972).
[15] ———, *The existence of value in differential games of pursuit and evasion*, J. Differential Equations, 12 (1972), pp. 504–523.
[16] L. C. Evans and P. E. Sougandinis, *Differential games and representation formulas for solutions of Hamilton-Jacobi Equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.
[17] H. Frankowska, *Control of Nonlinear Systems and Differential Inclusions*, Birkhäuser, Boston, MA, to appear.
[18] H. Frankowska, M. Plascasz, and T. Rzezuchowski, *Measurable viability theorem and Hamilton-Jacobi-Bellman equations*, J. Differential Equations, 116 (1995), pp. 265–305.
[19] H. Frankowska and M. Quincampoix, *Isaacs' Equations for Value-Functions of Differential Games*, Internat. Series Numer. Math., 107, Birkhäuser, Verlag, Basel, 1992.
[20] O. Hajek, *Pursuit Games*, Academic Press, San Francisco, New York, 1975.
[21] R. Isaacs, *Differential Games*, Wiley, New York, 1965.
[22] N. N. Krasovskii and A. I. Subbotin, *Positional Differential Games*, Nauka, Moscow, 1974.
[23] ———, *Game-Theoretical Control Problems*, Springer-Verlag, New York, 1988.
[24] S. Plaskacz, Personal communication, 1993.
[25] M. Quincampoix, *Differential inclusion and target problem*, SIAM J. Control Optim., 30 (1992), pp. 324–335.
[26] M. Quincampoix and P. Saint-Pierre, *An algorithm for viability kernels in Hölderian case: Approximation by discrete viability kernels*, J. Math. Systems, Estim. Control, 5 (1995), pp. 115–118.
[27] C. Ryll-Nardzewski, *A theory of pursuit-evasion*, in Advances in Game Theory, Ann. of Math. Stud. 52, Princeton University Press, Princeton, NJ, 1964, pp. 113–126.
[28] E. Roxin, *The axiomatic approach in differential games*, J. Optim. Theory Appl., 3 (1969), pp. 153–163.
[29] P. Saint-Pierre, *Viability of boundary of the viability kernel*, J. Differential Integral Equations, 4 (1991), pp. 1147–1153.
[30] ———, *Approximation of the viability kernel*, Appl. Math. Optim., 29 (1992), 187–209.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 4, pp. 1461–1472, July 1996

019

INFORMATION CAPACITY OF CHANNELS WITH PARTIALLY UNKNOWN NOISE. II. INFINITE-DIMENSIONAL CHANNELS*

C. R. BAKER† AND I.-F. CHAO‡

Abstract. Information capacity is considered for a communication channel in which the noise is the sum of a known Gaussian component and an independent component with unknown statistical distributions. A lower bound on capacity is sought; the unknown noise component is thus assumed to be under the control of an adversary—a jammer. The problem is modeled as a zero-sum two-person game with mutual information as the payoff function. Appropriate constraints are determined on the transmitted signal and the unknown noise component. Although the usual conditions sufficient for application of the general form of the von Neumann minimax theorem are shown not to hold, a solution is obtained for the game: a saddle value, saddle point, and minimax strategy for the jammer are obtained. The essential effect of jamming is to convert the infinite-dimensional channel into a finite-dimensional channel having the same constraints, with the dimensionality depending upon the problem parameters: the covariance of the known Gaussian noise component and the constraints on the transmitted signal and the unknown noise component.

Key words. Shannon theory, channel capacity, information capacity, jamming

AMS subject classifications. 94A15, 94A40, 90D80

1. Introduction.

The subject of this paper is a communication channel in which the noise is the sum of a known Gaussian component and an independent component with unknown statistical distributions. An example is when the interference is the sum of a Gaussian receiver noise and an unknown noise in the transmission medium. One may then ask for the worst-case capacity of the channel; that is, when "nature" chooses the unknown noise component to be that which is least favorable to the channel user, or coder. A natural way to view this problem is that of a jamming channel, where the unknown noise is controlled by a hostile jammer who seeks to minimize the information capacity of the channel.

Information capacity of such a channel is determined in [7] under the assumption that the channel is finite dimensional. By "information capacity" we mean here a saddle point solution to a zero-sum two-person game in which mutual information is the payoff function and the admissible strategies for coder and jammer are determined by constraints on the stochastic signals of the coder and jammer.

In this paper this problem is solved for the infinite-dimensional channel: a channel in which all sample paths belong to a real separable Hilbert space, H, with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. A concrete example is when H is the M-fold product of $L_2[0, T]$; f in H is of the form $f = (f_1, f_2, \ldots, f_M)$, each f_i belonging to $L_2[0, T]$; and $\|f\| = [\sum_{i=1}^{M} \int_0^T f_i^2(t)dt)]^{1/2}$.

Very substantial differences exist between this problem and that for the finite-dimensional channel. In the latter, the initial problem can be reformulated so that a saddle point solution is guaranteed by the von Neumann minimax theorem. Moreover, the solution can be obtained, after some development, by application of constrained optimization and the Kuhn–Tucker conditions. For the infinite-dimensional channel, it will be seen that the von Neumann theorem cannot be applied. Thus, one does not know in advance whether a saddle value exists, much less a saddle point. Moreover, constrained optimization is not available. Nevertheless, we shall prove the existence of a saddle point, obtain the saddle value, and in the process give a minimax strategy for the jammer.

*Received by the editors August 2, 1993; accepted for publication (in revised form) May 1, 1995. This research was supported by NSF grant NCR87113726 and ONR contract N00014-92-C-0094.

†Carolinian Systems Research Corp., Chapel Hill, NC 27514-3002, and Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260.

‡Department of Mathematics, National Central University, Chungli, Taiwan, Republic of China.

The results have potential interest from several viewpoints. One, of course, is that mutual information is a mathematical quantity that accompanies every finite measure on a measurable product space. Game-theoretic formulations with mutual information as payoff function may be of interest to model and analyze various competitions (e.g., in economics), quite aside from communication over noisy channels. For such applications, more general models than those suitable for an additive-interference communication channel may be needed; some results in this direction are given in [1]. Additional generality can be added in various ways; for example, as in [5], finitely additive measures can be included. The problem in a communication channel framework is of long standing, first enunciated by Blachman [10] and Dobrushin [12] as an intrinsic measure of a channel's ability to convey information. A sufficient rationale for the present work and [7] is to give a solution to the game-theoretic problem first described by these authors over 35 years ago.

The results may also have practical use in obtaining coding capacity of communication channels. Information capacity is often used to obtain coding capacity in the sense that the former is typically an upper bound on the latter, and for Gaussian channels, the two are equal, at least for dimension-limited channels [4], [5], where the elements of the code word set must each satisfy an energy constraint. The finite-dimensional results of [7] are sufficient for this purpose in the case of the discrete-time channel. However, for continuous-time channels, one obtains the upper bound on coding capacity in the classical approach by computing the information capacity for each value of T, thus obtaining a quantity C_T, and then taking $\limsup_{T\to\infty} C_T/T$. C_T is computed for the channel having sample functions in $L_2[0, T]$ or its M-fold product.

Another approach to coding capacity of continuous-time channels is given in [4] and [5], where, rather than having the transmission time period increase and included in the constraint, the length of transmission is fixed and the dimensionality of the code word set is incorporated into the constraint. The motivation is that complexity can be as important in cost analysis as the time of transmission—perhaps more so—and that implemented communication channels are necessarily time limited. In this framework, the development involves computation of $\limsup_{n\to\infty} C(n)/n$, where n is the permitted dimensionality of the code word set and $C(n)$ is the capacity of the $L_2[0,T]$ channel when the coder is subject to this constraint.

For solving the first of the two problems just described, knowledge of the information capacity for the infinite-dimensional channel appears necessary; for the second, it seems at least helpful. Thus, the solution of the problem considered here appears to be a useful—perhaps essential—step in order to obtain coding capacity for continuous-time channels subject to jamming. Finally, we note that the basic problem—a channel where the noise is the sum of a known Gaussian component and an independent component with unknown distributions—arises very frequently in applications.

Channels such as those analyzed here are usually characterized as "compound," which refers to a channel known only to the extent that it belongs to a specified family. The family of channels admitted here is defined by energy-type constraints on the jammer and the coder which take into account the properties of the ambient Gaussian noise.

Recent work on the game-theoretic analysis of information capacity has primarily been done by McEliece and his collaborators [11], [18], [19]. Although their work includes analysis of some quantized channels, its intersection with our work is for an essentially one-dimensional Gaussian channel in which the jammer controls all of the noise.

The present work is for the general channel; there is no limitation such as stationarity, memory, or univariate nature. It is anticipated that $\lim_{T\to\infty} C_T/T$ can be given in terms of spectral densities for the stationary Gaussian channel subject to jamming. This can presumably be obtained through limiting arguments based on the results given here, in the same way that

the results of [3] have been used to obtain $\lim_{T\to\infty} C_T/T$ of stationary Gaussian channels without jamming [6].

Previous publications on analysis of channels subject to jamming include papers in which either coding capacity or minimization of mean-square signal distortion is the criterion. Among the former is the Gaussian arbitrarily varying channel (AVC). As analyzed in the existing literature (e.g., [14], [15]), this is an inherently discrete-parameter channel in which the noise is the sum of a known Gaussian process with independent components and (usually) constant variance and a jamming process with unknown statistical properties, with the transmitted signal and the jamming process subject to average or peak power constraints. Comparisons with our approach will thus first require applications of the results of [7] to obtain coding capacity for discrete-time channels. However, extension of the published AVC models and methods to general infinite-dimensional channels is not immediately obvious, especially for continuous-time channels.

An approach to modeling and analysis of jamming channels that can be applied to continuous-time channels has been taken by Basar and Basar [9]. Their work involves the use of mean-square signal distortion as the criterion of optimality and is not directly comparable to our approach.

Additional discussion of related work is contained in [7].

2. Mathematical model. The channel sample paths are described by $Y = X + W + J$, where X is the coder's signal, W is the ambient Gaussian noise, and J is the noise component controlled by the jammer. X, W, and J are mutually independent. These quantities are described by the probabilities μ_X, μ_W, and μ_J on the Borel sets of H; all are assumed countably additive and second order (i.e., $\int_H \|x\|^2 d\mu(x) < \infty$) so that their covariance operators are necessarily trace class. The (average) mutual information of interest is

$$I(X, Y) = I(\mu_{XY}) = \int_{H\times H} \left[\log \frac{d\mu_{XY}}{d\mu_X \otimes \mu_Y}(x, y)\right] d\mu_{XY}(x, y),$$

where $\mu_X \otimes \mu_Y$ denotes product measure and $\mu_{XY}(A) = \mu_X \otimes \mu_W \otimes \mu_J\{(x, w, v) : (x, x + w + v) \in A\}$. $I(\mu_{XY}) \equiv \infty$ when μ_{XY} is not absolutely continuous with respect to $\mu_X \otimes \mu_Y$.

Under these assumptions, μ_X, μ_W, and μ_J have covariance operators R_X, R_W, and R_J: e.g., $\langle R_W u, v\rangle = \int_H \langle x, u\rangle \langle x, v\rangle d\mu_W(x)$. We assume without loss of generality (WLOG) that R_W is strictly positive on H and that all probabilities are zero-mean.

The assumptions imply the existence of a self-adjoint operator S, possibly unbounded, satisfying $R_W + R_J = R_W^{1/2}(I + S)R_W^{1/2}$ or, equivalently, $R_J = R_W^{1/2} S R_W^{1/2}$, where $\text{range}(R_W^{1/2}) \subset \mathcal{D}(S)$, the domain of S [3]. The operator $(I + S)^{-1}$ exists, since S is nonnegative and necessarily bounded, since $\text{range}(R_W^{1/2}) \subset \text{range}(R_W + R_J)^{1/2}$. $R_W = \sum_{n\geq 1} \lambda_n e_n \otimes e_n$, where $\lambda_n > 0$ and $\lambda_n \geq \lambda_{n+1}$ for all $n \geq 1$, $\sum_{n\geq 1} \lambda_n < \infty$, $\{e_n, n \geq 1\}$ is a complete orthonormal set (CONS) for H, and $(e_n \otimes e_n)v \equiv \langle e_n, v\rangle e_n$.

The constraints on the coder are given by $\mu_X[\text{range}(R_W^{1/2})] = 1$ and $E_{\mu_X}\|x\|_W^2 \leq P_1$, where for x in $\text{range}(R_W^{1/2})$, $\|x\|_W^2 \equiv \|R_W^{-1/2}x\|^2$. Such constraints are necessary for the capacity without jamming to be finite [2]. This constraint amounts to a RKHS (reproducing kernel Hilbert space) constraint on the coder's energy in terms of the RKHS of the ambient channel noise covariance. As such, it limits the amount of energy that the coder can place into regions where the ambient noise energy is small.

The jammer's constraint is given by $E_{\mu_J}\|x\|^2 \leq P_2$. The stronger constraints $\mu_J[\text{range}(R_W^{1/2})] = 1$ and $E_{\mu_J}\|x\|_W^2 \leq P_2$ might be thought reasonable, since they are consistent with those imposed on the coder. However, they are too strong; the jammer would not be

able to have any effect on the information capacity under these constraints. This can be seen from the results of [3]; from Theorem 3 of that paper, if S is nonnegative and has zero as the only limit point of its spectrum, then the capacity is equal to $P_1/2$. This situation would hold if one used the constraints $\mu_J[\text{range}(R_W^{1/2})] = 1$ and $E_{\mu_J}\|x\|_W^2 \leq P_2$, since the inequality (using the fact that $R_J = R_W^{1/2} S R_W^{1/2}$) is the same as Trace $S \leq P_2$. Thus, the jammer's constraint must be weaker than that of the coder.

The constraint $E_{\mu_J}\|x\|^2 \leq P_2$ places a constraint on the total jammer energy but not on the relative jammer/noise energy. In terms of frequency, this means that the jammer can use signals that have relatively large energy compared to the noise energy in appropriate frequency ranges. Such a constraint has an immediate intuitive interpretation. The jammer's optimum policy should be that of adding the available energy to the ambient noise energy in such a way that the sum provides maximum interference to the coder. The jammer will thus want to place the available energy in regions available to the coder in which the noise energy is small, and this is provided by the above constraint. The use of the RKHS constraint $E_{\mu_J}\|x\|_W^2 \leq P_2$ would permit the jammer to place the available energy in regions that are available to the coder and with the same flexibility as permitted to the coder. However, it also limits the jammer's energy relative to the ambient noise energy and is evidently too strong for the jammer to have an effect. This may be regarded as somewhat surprising and is one of the first differences that one sees in going from the finite-dimensional to the infinite-dimensional channel. It means that, with the coder selecting the optimum strategy, the jammer cannot reduce the information capacity if the energy limitation on the jammer is subject to the same type of RKHS constraint as that applied to the coder, regardless of the value of P_2, i.e., regardless of the total amount of energy available to the jammer. As already mentioned, the constraint applied to the coder is implied by any constraint that gives finite information capacity in the absence of jamming.

In the form given above, there is no constraint on the probability distribution of the jammer's signal other than $E_{\mu_J}\|x\|^2 \leq P_2$. However, it follows from [16] that for a given R_J, the information capacity (coder's viewpoint) will be minimized by taking μ_J to be Gaussian. *Thus, the jammer should always choose μ_J to be Gaussian, and this will be assumed henceforth.* With this assumption, the jamming channel is a special case of the mismatched Gaussian channel: a Gaussian channel such that the constraint covariance R_W is not the same as the noise covariance [3].

The jammer's strategy is now uniquely determined by the choice of the operator S. For a given strategy, the mutual information can be expressed as [3]

$$I(\mu_{XY}) = F(\mathbf{z}, \alpha) = \frac{1}{2}\sum_n \log\left[1 + z_n(1+\alpha_n)^{-1}\right],$$

where (z_n) and (α_n) are defined as follows. The coder's covariance operator R_X is given by

$$R_X = \sum_n \tau_n (R_W + R_J)^{\frac{1}{2}} u_n \otimes (R_W + R_J)^{\frac{1}{2}} u_n,$$

where $\{u_n, n \geq 1\}$ is a CONS in H, $\tau_n > 0$ for $n \geq 1$, and $\sum_n \tau_n < \infty$. Then

$$\alpha_n \equiv \left\langle SU^* u_n, U^* u_n \right\rangle \qquad \text{and} \qquad z_n \equiv \tau_n \|(I+S)^{\frac{1}{2}} U^* u_n\|^2,$$

where U^* is the unitary operator satisfying $(R_W + R_J)^{1/2} = R_W^{1/2}(I+S)^{1/2}U^*$.

The problem to be considered is to determine if there exists a saddle value for the zero-sum two-person game with $I(\mu_{X,Y})$ as the payoff function; further, if a saddle value exists, then determine whether a saddle point exists; if such a point exists, then give its definition. That

is, we seek to determine whether

$$\sup_{z} \inf_{\alpha} F(\mathbf{z}, \alpha) = \inf_{\alpha} \sup_{z} F(\mathbf{z}, \alpha), \tag{1}$$

where the sup and inf are taken over the admissible signals for the coder and jammer, respectively. If this equality holds, then it defines the saddle value. In that case, one seeks to determine if the saddle value is actually attained by an admissible pair $(\mathbf{z}, \alpha)$, i.e., if a saddle point exists.

The admissible strategies for the coder and the jammer are given by the two constraints defined above. For the coder, the constraint is given by $\sum_n z_n \leq P_1$. Let $\mathcal{C}$ be the set of all densely defined symmetric nonnegative linear operators S on H such that the domain of S contains the range of $R_W^{1/2}$.The constraint on the jammer is that $R_J = R_W^{1/2} S R_W^{1/2}$ for some S in $\mathcal{C}$ such that Trace $R_W^{1/2} S R_W^{1/2} \leq P_2$.

The smallest limit point of the spectrum of S will be denoted by θ. The set of limit points of the spectrum ($\equiv$ the essential spectrum, $\sigma_{\text{ess}}(S)$) of S consists of all eigenvalues of infinite multiplicity, limit points of distinct eigenvalues, and other points of the continuous spectrum. As usual, S is said to have pure point spectrum if it has a CONS of eigenvectors. In this case, the continuous spectrum may not be empty, since any limit point of distinct eigenvalues which is itself not an eigenvalue will belong to the continuous spectrum. The sequence (γ_n) will denote the eigenvalues of S that are strictly less than θ, repeated according to their multiplicity; $\{v_n, n \geq 1\}$ will denote the corresponding eigenvectors. The following result is essential to our development.

LEMMA 1 (see [3, Thm. 3 and Cor. 4]). *Suppose that the jammer's strategy is fixed and (WLOG) order (γ_n) as a nondecreasing sequence.*

(1) *Suppose that $\theta < \infty$. The information capacity $C_W(P_1)$ is then given as follows.*

(a) *If $\{\gamma_n, n \geq 1\}$ is not empty and $\sum_n (\theta - \gamma_n) \leq P_1$, then*

$$C_W(P_1) = \frac{1}{2} \sum_{n=1}^{K} \log \left[\frac{1+\theta}{1+\gamma_n} \right] + \frac{1}{2} \frac{P_1 + \sum_{m=1}^{K} (\gamma_m - \theta)}{1+\theta}$$

(b) *If $\{\gamma_n, n \geq 1\}$ is empty, then $C_W(P_1) = P_1/[2(1+\theta)]$.*

(c) *In* (a), *the capacity can be attained if and only if $P_1 = \sum_{n \geq 1} (\theta - \gamma_n)$. In that case, it is uniquely attained by a Gaussian μ_X with covariance $R_X = \sum_{n \geq 1} \tau_n R_N^{1/2} u_n \otimes R_N^{1/2} u_n$, where $u_n = U v_n$ and $\tau_n = (\theta - \gamma_n)(1+\gamma_n)^{-1}$; for all $n \geq 1$ if (γ_n) is an infinite sequence; for $1 \leq n \leq K$ and $\tau_n = 0$ for $n > K$ when (γ_n) is a finite sequence with K elements. If $P_1 > \sum_{n \geq 1} (\theta - \gamma_n)$, then capacity can be approached as closely as desired by using a covariance R_X of the above form. In (b), the capacity cannot be attained.*

(d) *If $\{\gamma_n, n \geq 1\}$ is not empty, (γ_n) is an infinite sequence, and $P_1 < \sum_n (\theta - \gamma_n)$, then an integer M exists such that $M\gamma_{M+1} > P_1 + \sum_{i=1}^{M} \gamma_i \geq M\gamma_M$ and then*

$$C_W(P_1) = \frac{1}{2} \sum_{i=1}^{M} \left[\frac{P_1 + \sum_{k=1}^{M} \gamma_k + M}{M(1+\gamma_i)} \right].$$

(e) *If $\{\gamma_n, n \geq 1\}$ is not empty, (γ_n) is a finite sequence containing exactly L elements, and $P_1 < \sum_n (\theta - \gamma_n)$, then the following hold:*

(i) *If $P_1 + \sum_{i=1}^{L} \gamma_i \geq L\gamma_L$, then the capacity is given as in* (d), *with $M = L$;*

(ii) *If $M\gamma_{M+1} > P_1 + \sum_{i=1}^{M} \gamma_i \geq M\gamma_M$ for some $M < L$, then the capacity is given as in* (d).

(2) *If $\theta = \infty$, then the information capacity is obtained (and only obtained) by using a Gaussian μ_X having M-dimensional support, with M the smallest integer such that $M\gamma_{M+1} > P_1 + \sum_{i=1}^{M} \gamma_i \geq M\gamma_M$. The capacity $C_W(P_1)$ is then given by*

$$C_W(P_1) = \frac{1}{2} \sum_{i=1}^{M} \left[\frac{P_1 + \sum_{k=1}^{M} \gamma_k + M}{M(1+\gamma_i)} \right].$$

LEMMA 2. *A minimax strategy for the jammer, if it exists, requires an operator S having pure point spectrum with $\theta < \infty$ and all eigenvalues $\leq \theta$ with only a finite number being zero. If $\{\gamma_n, n \geq 1\}$ is not empty, then a minimax strategy requires that $P_1 \geq \sum_{n\geq 1}(\theta - \gamma_n)$ and that $P_1 + \sum_1^M \gamma_n \geq M\gamma_M$ for all γ_M.*

Proof. If the smallest limit point of the spectrum is infinite, then it must occur as a limit point of eigenvalues and then S has pure point spectrum. Suppose that $\gamma_n \nearrow \theta = \infty$. Applying the expression for information capacity given in (2) of Lemma 1, let M be the smallest integer such that $M\gamma_{M+1} > P_1 + \sum_{i=1}^{M} \gamma_i \geq M\gamma_M$. Let $K < M$ be such that $\gamma_K < \gamma_M$. (If no such K exists, then the following argument applies with γ_K replaced by γ_M, since then $P_1 + \sum_{i=1}^{M} \gamma_i > M\gamma_M$.) An increase of ε in the value of γ_K will result in a strict decrease in the value of $C_W(P_1)$. Define a sequence (γ_n') such that $\gamma_n' = \gamma_n$ for $n \notin \{K, M+1\}$, $\gamma_K' = \gamma_K + \varepsilon$, $\gamma_{M+1}' = \gamma_{M+1} - \delta$, with ε and δ satisfying $M(\gamma_{M+1} - \delta) > P_1 + \sum_{i=1}^{M} \gamma_i + \varepsilon \geq \max(M\gamma_M, M\gamma_K + M\varepsilon)$. A jammer covariance defined by the operator S' having eigenvalues (γ_n') will give the same value of the integer M defined in (2) of Lemma 1 as given by the original S. To ensure that the constraint on the jammer remains satisfied by the new sequence (γ_n'), one notes that for $R_W^{1/2} S R_W^{1/2}$ trace-class, its trace equals that of $S^{1/2} R_W S^{1/2}$ [21, p. 31] so that the sequence (γ_n) affects the constraint only through the value of $\sum_{k\geq 1} \gamma_k \langle R_W v_k, v_k \rangle$. Thus, selecting strictly positive ε and δ such that the two preceding inequalities are satisfied and $\delta/\varepsilon \geq \langle R_W v_K, v_K \rangle / \langle R_W v_{M+1}, v_{M+1} \rangle$, the operator S' having eigenvalues (γ_n') will still satisfy the constraint on the jammer and will provide a strictly smaller value of $C_W(P_1)$. This shows that $\theta < \infty$ is necessary for a minimax strategy. A similar argument shows that a jammer strategy as in parts (d) and (e) of (1) of Lemma 1 cannot be minimax. Thus, a minimax strategy requires that $P_1 \geq \sum_n (\theta - \gamma_n)$ if (γ_n) is not empty; from [3, Lem. 4], this implies the inequality $P_1 + \Sigma_1^M \gamma_n \geq M\gamma_M$ for all γ_M.

Attention can now be restricted to jammer strategies such that $\theta < \infty$. Moreover, since only those eigenvalues of S that are $\leq \theta$ affect (by (1) of Lemma 1) the value of $C_W(P_1)$, a minimax strategy requires that all eigenvalues of S not exceed θ; otherwise, eigenvalues that exceed θ will require part of the jammer's energy while not contributing to the reduction in mutual information. Since $\theta \geq 0$, it follows from Lemma 1(1) that θ should be strictly positive and thus that only a finite number of eigenvalues can equal zero.

Finally, to see that S must have pure point spectrum to provide a minimax policy, suppose that θ is in the continuous spectrum of S. If θ is the only point in the continuous spectrum, then it must occur as the limit of distinct eigenvalues, since if θ is in the continuous spectrum, then every interval containing θ must contain a point of the spectrum that is distinct from θ [20, p. 364]. In that case, those eigenvalues must obviously be less than θ if S is to be a minimax strategy.

Now suppose that S is any nonnegative and symmetric operator satisfying the constraint on the jammer and such that θ, the smallest value in $\sigma_{\text{ess}}(S)$, is a point of the continuous spectrum and not a limit point of distinct eigenvalues. The eigenvalues of S that are less than θ must then be finite in number. Let $\{v_i, i \geq 1\}$ be a CONS in H and $\{v_1, \ldots, v_K\}$ the eigenvectors of S corresponding to its eigenvalues $\{\gamma_1, \ldots, \gamma_K\}$. Define the operator S_1 to have pure point spectrum with eigenvectors $\{v_i, i \geq 1\}$ and eigenvalues given by $\{\gamma_1, \ldots, \gamma_K, \theta_1, \ldots \theta_1, \ldots\}$;

that is, θ_1 is an eigenvalue of infinite multiplicity corresponding to the eigenvectors $\{v_n, n \geq K+1\}$. Suppose first that $\theta_1 = \theta$ and $K > 0$. The operator $S - S_1$ is symmetric and nonnegative so that $R_W^{1/2}(S-S_1)R_W^{1/2}$ is self-adjoint, nonnegative, and trace-class. If the latter operator has trace equal to zero, then it is identically zero, and this would then require that $R_W^{1/2}SR_W^{1/2} = R_W^{1/2}S_1R_W^{1/2}$. Since range $(R_W^{1/2})$ is dense in H, this last equality would require that $S = S_1$ on H. As this is false, $P_2 \geq \sum_{j\geq 1}\langle R_W^{1/2}SR_W^{1/2}v_j, v_j\rangle > \sum_{j\geq 1}\langle R_W^{1/2}S_1R_W^{1/2}v_j, v_j\rangle$. From the last two inequalities, the K smallest eigenvalues of S_1 can be increased with S_1 still satisfying the constraint on the jammer and decreasing (from (1(a)) of Lemma 1) the mutual information in comparison with that obtained using the operator S. If $K = 0$, then θ_1 can be taken strictly greater than θ while satisfying the constraint on the jammer and again reducing (from (1(b)) of Lemma 1) the mutual information from that obtained using S. Thus, in order that S define a minimax strategy for the jammer, it is necessary that the continuous spectrum of S be empty except when S has an infinite sequence of eigenvalues and θ is the limit point of those eigenvalues (from below) and not an eigenvalue of S; θ is then the only point in the continuous spectrum. □

Applying Lemma 2 and Lemma 1(1), the coder's optimum strategy (when it exists) requires choosing a Gaussian signal process with the covariance operator R_X defined in terms of the eigenvalues of the ambient noise covariance R_W, the eigenvalues and associated orthonormal eigenvectors of the operator S defining the jamming covariance $R_J = R_W^{1/2}SR_W^{1/2}$, and the unitary operator U satisfying $(R_W + R_J)^{1/2} = R_W^{1/2}(I+S)^{1/2}U^*$. The problem now reduces to determining the jammer's minimax strategy (if it exists), which is to select a Gaussian jamming process with covariance operator R_J such that the value of $C_W(P_1)$ is minimized. Lemma 2 defines properties that the operator S must satisfy if a minimax strategy is to be realized.

An important question at this point is whether the game has a solution. Does a saddle value exist? If so, is there a saddle point? Such questions are usually answered by appealing to the von Neumann minimax theorem, which we now consider.

General forms of the von Neumann minimax theorems are given, for example, in [8]. For a concave/convex function such as F, a saddle value will exist if the set of admissible $\mathbf{z}$ constitutes a compact convex set in a linear topological space H_1, the set of admissible α constitutes a compact convex subset of a linear topological space H_2, and (on the set of admissible $(\mathbf{z}, \alpha)$) F is continuous in $\mathbf{z}$ (resp., α) for each fixed α (resp., $\mathbf{z}$) [8, Thm. 3.5]. For reflexive spaces H_1 and H_2, this criterion can be satisfied by using the weak topology. From [8, Cor. 3.7], both a saddle value and a saddle point will exist if the spaces H_1 and H_2 are reflexive Banach spaces *and* if the set of admissible $\mathbf{z}$ (resp., admissible α) is a closed and bounded subset of H_1 (resp., H_2). As defined, the continuity conditions are satisfied by F and the set of admissible $\mathbf{z}$ constitutes a closed and bounded subset of the nonreflexive space ℓ_1. Of course, the admissible $\mathbf{z}$ constitute a bounded set in the reflexive space ℓ_p for $1 < p < \infty$. However, this set is not closed in any of the reflexive ℓ_p spaces. This can be proven by a construction and argument based on the Baire category theorem. A more general result is given by the following proposition.

PROPOSITION 1. *The set of admissible* $\mathbf{z}$ *is not a closed and bounded subset of any reflexive Banach space.*

Proof. First, for reflexive Banach spaces, convergence in the weak* topology [13, p. 420] is the same as convergence in the weak topology from the definitions of the two topologies. Thus, from [17, Thm. 2.e.7], if a separable Banach space Y is reflexive, then there exists no bounded linear operator T: $\ell_1 \to Y$ having bounded inverse (on the range of T in Y). Let Y be any reflexive Banach space—not necessarily separable—containing the set of admissible $\mathbf{z}$. By linearity, it is seen that Y must contain all of ℓ_1. Let T be the natural injection from

ℓ_1 into Y : Tx is x viewed as an element of Y. *Suppose* that T maps the set of admissible $\mathbf{z}$ into a bounded and closed set in Y. Applying linearity, T must then also map the unit ball $B_1 \equiv \{\mathbf{z} : \sum_{k\geq 1} |z_k| \leq 1\}$ into a bounded and closed set in Y. T is then continuous on ℓ_1; this fact, the separability of ℓ_1, and the fact that $T[B_1]$ is closed imply that $T[B_1]$ is a complete separable metric space under the Y norm metric. By the open mapping theorem [13, p. 57], the operator T^{-1} is continuous on $T[B_1]$ and, by linearity and the definition of B_1, must be continuous on the range of T. The closure of range (T) is separable, since range (T) is separable and T is continuous; reflexive [13, p. 67]; and thus a separable reflexive Banach space under the Y norm, contains ℓ_1 as a subspace, and T^{-1} is continuous on ℓ_1. This contradicts Theorem 2.e.7 of [17]. □

This result shows that the general form of the von Neumann minimax theorem cannot be applied to guarantee the existence of a saddle point (or even a saddle value). We shall prove, however, that a saddle point exists, determine the saddle value, and give a minimax strategy for the jammer.

LEMMA 3 (see [7]). *Define* $(\gamma_{K,i}^m)$, (z_i^m), θ_K^m, g_K^m, *and* T_K^m *as follows, for* $K \geq 0, m \geq K+1$:

$$g_K^m(\theta) = \left[\sum_{i=K+1}^{m} \frac{\lambda_i}{P_2 + \sum_{k=K+1}^{m} \lambda_k + (P_1 - K\theta)\lambda_i}\right]^{-1}, \qquad 0 \leq \theta \leq P_1/K,$$

$$T_K^m = \frac{P_1 \lambda_{K+1}}{P_2 + \sum_{i=K+1}^{m} \lambda_i + K\lambda_{K+1}},$$

$$1 + \gamma_{K,i}^m = \frac{(P_2 + \sum_{k=K+1}^{m} \lambda_k)(1 + \theta_K^m)}{P_2 + \sum_{j=K+1}^{m} \lambda_j + (P_1 - K\theta_K^m)\lambda_i}, \qquad i \geq K+1,$$

$$\gamma_{K,i}^m = 0, \qquad i \leq K,$$

where θ_K^m *is the solution to* $1 + \theta = g_K^m(\theta)$, *and*

$$z_i^m \equiv \theta_K^m - \gamma_{K,i}^m, \qquad i \geq 1.$$

A unique solution exists for θ_K^m. *If* $P_1 < (m-1)P_2/\lambda_m$ *and it is required that the coder's covariance satisfy* $z_n = 0$ *for all* $n \geq m+1$, *then the jamming problem has a saddle point* $(\mathbf{z}^*, \gamma^*)$, *with* $\gamma_i^m = \gamma_{K(m),i}^m$ *for* $i \leq m$ *and* $\gamma_i^* = 0$ *for* $i > m$; $z_i^* = z_i^m$ *for* $i \leq m$ *and* $z_i^* = 0$ *for* $i > m$; *and* $K(m)$ *the smallest integer* k *such that* $\theta_k^m > T_k^m$. *This* $K(m)$ *satisfies* $K(m) \leq m-2$, *and* $P_1 \geq K(m)P_2/\sum_{j=K(m)+1}^{m} \lambda_j$.

The condition $P_1 < (m-1)P_2/\lambda_m$ assumed in Lemma 3 will of course be satisfied for all sufficiently large m; in fact, for all $m \geq k$, where k is such that $P_1 \leq kP_2/(\sum_{j=k+1}^{\infty} \lambda_j)$.

LEMMA 4. *If* K *is any integer* ≥ 0 *such that* $P_1 < (K+1)P_2/\sum_{i\geq K+2} \lambda_i$, *then* $\theta_K^m > T_K^m$ *for all* $m \geq K+1$.

Proof. We will show that $g_K^m(T_K^m) > 1 + T_K^m$ for all $m \geq K+1$. This will prove the statements, since otherwise one would have $\theta_K^m \leq T_K^m$ for some m; since $g_K^m(\theta)$ is a strictly decreasing function of θ for fixed m and K, this would give

$$1 + T_K^m \geq 1 + \theta_K^m = g_K^m(\theta_K^m) \geq g_K^m(T_K^m) > 1 + T_K^m.$$

To see that $g_K(T_K^m) > 1 + T_K^m$, this inequality is equivalent to

$$\sum_{i=K+1}^{m} \frac{\lambda_i[P_2 + \sum_{j=K+1}^{m} \lambda_j + K\lambda_{K+1}]}{(P_2 + \sum_{k=K+1}^{m} \lambda_k)[P_2 + \sum_{n=K+1}^{m} \lambda_n + K\lambda_{K+1} + P_1\lambda_i]} < \frac{P_2 + \sum_{j=K+1}^{m} \lambda_j + K\lambda_{K+1}}{P_2 + \sum_{k=K+1}^{m} \lambda_k + (P_1 + K)\lambda_{K+1}}$$

or

$$\sum_{i=K+1}^{m} \lambda_i \left\{ 1 + \frac{P_1(\lambda_{K+1} - \lambda_i)}{P_2 + \sum_{j=K+1}^{m} \lambda_j + K\lambda_{K+1} + P_1\lambda_i} \right\} < P_2 + \sum_{k=K+1}^{m} \lambda_k.$$

This holds if

$$P_1\lambda_{K+1} \sum_{i=K+2}^{m} \lambda_i < P_2 \left[P_2 + \sum_{j=K+2}^{m} \lambda_j + (K+1)\lambda_{K+1} \right],$$

which is satisfied, since $LHS \leq P_2\lambda_{K+1}(K+1)$. □

Now define T_K and g_K by

$$T_K = \frac{P_1\lambda_{K+1}}{P_2 + \sum_{j=K+1}^{\infty} \lambda_j + K\lambda_{K+1}},$$

$$g_K(\theta) = \left[\sum_{i=K+1}^{\infty} \frac{\lambda_i}{P_2 + \sum_{j=K+1}^{\infty} \lambda_j + (P_1 - K\theta)\lambda_i} \right]^{-1}, \qquad 0 \leq \theta \leq P_1/K,$$

and let θ_K be the solution to $1 + \theta = g_K(\theta)$. It is obvious that $T_K = \lim_m T_K^m$; we now show that the equation $1 + \theta = g_K(\theta)$ has a unique solution, θ_K, and that $\theta_K = \lim_m \theta_K^m$.

LEMMA 5. (a) *$g_K^m(\theta)$ is a strictly decreasing function of m for fixed K and θ.*

(b) *θ_K^m is a strictly decreasing function of m.*

(c) *θ_K is uniquely defined and $\theta_K = \lim_{m\to\infty} \theta_K^m$.*

(d) *$\theta_K \geq T_K$ if and only if*
$\sum_{i=K+1}^{\infty} [P_1(\lambda_{K+1} - \lambda_i)\lambda_i][P_2 + \sum_{j=K+1}^{\infty} \lambda_j + K\lambda_{K+1} + P_1\lambda_i]^{-1} \leq P_2$.

(e) *$\theta_K \geq T_K$ if and only if $\theta_K^m > T_K^m$ for all $m \geq K+1$.*

Proof. (a) It is immediate that $[g_K^{m+1}(\theta)]^{-1} > [g_K^m(\theta)]^{-1}$ if $\theta \leq P_1/K$.

(b) Since $g_K^m(\theta)$ is a strictly decreasing function of m, the solution to $1 + \theta = g_K^{m+1}(\theta)$ must be strictly less than the solution to $1 + \theta = g_K^m(\theta)$.

(c) Let $f(\theta) = 1 + \theta$. f is continuous and strictly monotone increasing. g_K is continuous and strictly monotone decreasing, and $f(0) = 1 < g_K(0)$. If $P_1/K \geq P_2/\sum_{j\geq K+1} \lambda_j$, then $f(P_1/K) = 1 + P_1/K \geq 1 + P_2/\sum_{j\geq K+1} \lambda_j = g_K(P_1/K)$. Thus a unique solution exists to $f = g_K$ if $P_1/K \geq P_2/\sum_{j\geq K+1} \lambda_j$. If $P_1/K < P_2/\sum_{j\geq K+1} \lambda_j$, then the same result holds, since then $f(P_2/\sum_{j\geq K+1} \lambda_j) > g(P_2/\sum_{j\geq K+1} \lambda_j)$.

Since (θ_K^m) is monotone decreasing as m increases and is bounded below by zero, $\lim_m \theta_K^m \equiv \theta_0$ exists. Since $g_K(\theta) = \lim_m g_K^m(\theta)$ for $\theta > 0$, $g_K(\theta_0) = \lim_m g_K^m(\theta_0) \geq \limsup_m g_K^m(\theta_K^m) = \limsup_m (1+\theta_K^m) = 1+\theta_0$. Conversely, $g_K(\theta_0) = \lim_m g_K(\theta_K^m) \leq \limsup_m g_K^m(\theta_K^m) = 1+\theta_0$. Thus, $1 + \theta_0 = g_K(\theta_0)$; since this solution is unique, $\theta_0 = \theta_K$.

(d) $g_K(T_K) \geq 1 + T_K$, since otherwise (proceeding as in the proof of part (c)) the solution to $g_K(\theta) = 1 + \theta$ will occur for $\theta < T_K$. The inequality of (d) then follows from $g_K^{-1}(T_K) \leq (1 + T_K)^{-1}$.

(e) As in the proof of (*d*), $\theta_K^m > T_K^m$ if and only if $g_K^m(T_K^m) > 1 + T_K^m$, and this occurs if and only if

$$\sum_{i=K+1}^{m} \frac{P_1(\lambda_{K+1} - \lambda_i)\lambda_i}{P_2 + \sum_{j=K+1}^{m} \lambda_j + K\lambda_{K+1} + P_1\lambda_i} < P_2.$$

The left-hand side of this inequality is a strictly increasing function of m. Moreover,

$$g_K^{K+1}(\theta_K^{K+1}) = 1 + \theta_K^{K+1} = \left[\frac{\lambda_{K+1}}{P_2 + \left[P_1 - K\theta_K^{K+1} + 1\right]\lambda_{K+1}} \right]^{-1}$$

so that

$$\theta_K^{K+1} = \frac{P_2 + P_1\lambda_{K+1}}{(K+1)\lambda_{K+1}} > \frac{P_1\lambda_{K+1}}{P_2 + (K+1)\lambda_{K+1}} = T_K^{K+1}.$$

Thus, $\lim_{m\to\infty} \theta_K^m \geq \lim_{m\to\infty} T_K^m <=> \theta_K^m \geq T_K^m$ for all $m \geq K+1$. □

The following theorem is the main result of this paper. It will be seen (in Remark 2 below) that the effect of the jammer is essentially to convert the infinite-dimensional channel into a channel of dimension $\leq K+1$ $(0 \leq K < \infty)$. This integer K will be defined in terms of P_1, P_2, and the eigenvalues (λ_n) of R_W.

THEOREM 1. *The jamming problem has a saddle value and a saddle point. The saddle value is given by*

$$I(\mu_{XY}) = \frac{1}{2}\sum_{n\geq K+1} \log\left[1 + \frac{(P_1 - K\theta_K)\lambda_n}{P_2 + \sum_{i\geq K+1}\lambda_i}\right] + \frac{K}{2}\log(1+\theta_K),$$

where θ_K is the unique solution of

$$1 + \theta = \left[\sum_{n\geq K+1} \frac{\lambda_n}{P_2 + \sum_{j\geq K+1}\lambda_j + (P_1 - K\theta)\lambda_n}\right]^{-1}$$

for $0 \leq \theta \leq P_1/K$ and K is the smallest integer $k \geq 0$ such that

$$\sum_{i=k+1}^{\infty} \frac{P_1(\lambda_{k+1} - \lambda_i)\lambda_i}{P_2 + \sum_{j=k+1}^{\infty}\lambda_j + k\lambda_{k+1} + P_1\lambda_i} \leq P_2.$$

A saddle point is given by $(\mathbf{z}^*, \gamma^*)$, where $z_i^* = \theta_K - \gamma_i^*$ for all $i \geq 1$, $\gamma_i^* = 0$ for $i \leq K$, and

$$1 + \gamma_i^* = \frac{[P_2 + \sum_{j\geq K+1}\lambda_j](1+\theta_K)}{P_2 + \sum_{i\geq K+1}\lambda_i + (P_1 - K\theta_K)\lambda_i} \quad \text{for} \quad i \geq K+1.$$

A minimax strategy for the jammer is to choose S to have pure point spectrum with eigenvalues (γ_n^*) and corresponding eigenvectors $\{e_n, n \geq 1\}$. The resulting maximin strategy for the coder is given by (1(c)) *of Lemma* 1.

Proof. Define $Z = \{(z_i) : z_i \geq 0$ for $i \geq 1$ and $\sum_{i\geq 1} z_i \leq P_1\}$. For m such that $P_1 < mP_2/\sum_{i\geq m}\lambda_i$, $Z^m \equiv \{(z_i) \in Z : z_i = 0$ for $i > m\}$. $\Gamma \equiv \{(\gamma_i) : \gamma_i = \langle Sv_i, v_i\rangle$ for a CONS $\{v_i, i \geq 1\}$, $S \in \mathcal{C}$, $\sum_{i\geq 1}\lambda_i\langle Se_i, e_i\rangle \leq P_2\}$, $\Gamma^m = \{(\gamma_i) \in \Gamma : \gamma_j = 0$ for $j > m\}$, where $\mathcal{C}$ is the set of all densely defined symmetric nonnegative linear operators whose domain contains the range of $R_W^{1/2}$. To prove the existence of a saddle value, it is sufficient [8] to show that $\sup_Z \inf_\Gamma F(\mathbf{z}, \gamma) \geq \inf_\Gamma \sup_Z F(\mathbf{z}, \gamma)$. Note that $\sup_Z \inf_\Gamma F(\mathbf{z}, \gamma) \geq \sup_{Z^m} \inf_\Gamma F(\mathbf{z}, \gamma) = \sup_{Z^m} \inf_{\Gamma^m} F(z, \gamma) = F(\mathbf{z}^m, \gamma_{k(m)}^m)$, where $(\mathbf{z}^m, \gamma_{k(m)}^m)$ is defined in Lemma 3.

Thus, $\sup_Z \inf_\Gamma F(\mathbf{z}, \gamma) \geq \liminf_{m\to\infty} F(\mathbf{z}^m, \gamma_{k(m)}^m) = \liminf_{m\to\infty} \frac{1}{2}\Sigma_{i\geq 1}\log[1 + z_i^m/(1+\gamma_{k(m),i}^m)]$. Define $k_0 \equiv \liminf_{m\to\infty} k(m) = \liminf_{m\to\infty}\{k : \theta_k^m > T_k^m\} = K$ (Lemma 5), with K defined as in the theorem. From the definitions of $\gamma_{k(m)}^m$, $\mathbf{z}^m$, γ^*, and $\mathbf{z}^*$, $\lim_n \theta_K^n = \theta_K$ implies $\gamma_{K(m),i}^m \to \gamma_i^*$ and $z_i^m \to z_i^*$ for all $i \geq 1$ as $m \to \infty$. By Fatou's lemma,

$$\liminf_{m\to\infty} \frac{1}{2}\sum_{i=1}^{\infty} \log\left[1 + \frac{z_i^m}{1+\gamma_{K,i}^m}\right] \geq \frac{1}{2}\sum_{i=1}^{\infty}\log\left[1 + \frac{z_i^*}{1+\gamma_i^*}\right]. \tag{2}$$

Thus, $\sup_Z \inf_\Gamma F(\mathbf{z}, \gamma) \geq F(\mathbf{z}^*, \gamma^*) = I(\mu_{XY})$, as given in the theorem.

Conversely, one notes that $\gamma_i^* \nearrow \theta_K$ and that $\sum_{i=1}^{\infty}(\theta_K - \gamma_i^*) = P_1$. Thus, by (1(a)) of Lemma 1, $F(\mathbf{z}^*, \gamma^*) = \sup_Z F(\mathbf{z}, \gamma^*) \geq \inf_\Gamma \sup_Z F(\mathbf{z}, \gamma)$. This shows that

$$\inf_\Gamma \sup_Z F(\mathbf{z}, \gamma) \leq F(\mathbf{z}^*, \gamma^*) \leq \sup_Z \inf_\Gamma F(\mathbf{z}, \gamma)$$

and completes the proof of existence of a saddle value and saddle point and their definitions as given in the theorem. The stated minimax strategy for the jammer is obvious and then determines the coder's optimum strategy from Lemma 1(1). □

Remark 1. The integer K defined in the theorem must be smaller than the smallest integer k satisfying $P_1 < (k+1)P_2/\sum_{j=k+2}^{\infty}\lambda_j$. This follows from Lemma 4 and Theorem 3 of [7] and yields the following result.

COROLLARY 1. *The saddle value of F on Λ has the upper bound*

$$F(\mathbf{z}, \gamma) < \frac{K+1}{2}\log\left[1 + \frac{P_1}{K+1}\right],$$

where K is the smallest integer k satisfying $P_1 < (k+1)P_2/\sum_{j=k+2}^{\infty}\lambda_j$.

Proof. From (1(a)) of Lemma 1, the right side of the first inequality is the value of the channel capacity that would be obtained if the jammer used (γ_n) such that $\gamma_n = 0$ for $n \leq K+1$ and $\gamma_n = P_1/(K+1)$ for $n > K+1$. In that case, $\sum_{n=1}^{\infty}\gamma_n\lambda_n = \sum_{n=K+2}^{\infty}\gamma_n\lambda_n = (P_1/[K+1])\sum_{n\geq K+2}^{\infty}\lambda_n < P_2$. This γ is thus an admissible strategy for the jammer, and the result follows. □

Remark 2. The capacity of the channel in the absence of jamming is $P_1/2$; see [2] or [3]. Since $P_1/2 = \lim_K \frac{1}{2}(K+1)\log[1 + P_1/(K+1)]$, one sees that the minimum effect of jamming can be immediately gauged by determining the value of K, the largest integer such that $P_1 > KP_2/\sum_{n=K+1}^{\infty}\lambda_n$, and applying the above corollary. Since the capacity of the $(K+1)$-dimensional channel in the absence of jamming is $\frac{1}{2}(K+1)\log[1 + P_1/(K+1)]$, one can view the effect of jamming as converting the infinite-dimensional channel into a finite-dimensional channel. The value of this K depends on all the channel parameters: the coder's constraint P_1, the jammer's constraint P_2, and the covariance operator of the ambient Gaussian noise.

COROLLARY 2. *The jammer's minimax strategy satisfies $P_1 = \sum_{n\geq 1}(\theta_K - \gamma_n^*)$.*

Remark 3. The statement of the theorem can be interpreted by considering the sum of the jamming and the ambient noise when the jammer selects the minimax strategy given in the theorem. Define K as in the theorem; K is the largest integer k such that $\gamma_k^* = 0$. Corresponding to the eigenvectors $\{e_n, n \geq 1\}$ of $R_W + R_J$, the eigenvalues are given by λ_j for $j \leq K$, while for $j > K$ the eigenvalues are equal to $\lambda_j(1 + \gamma_j^*)$. Thus, as $j \to \infty$, the eigenvalues converge upward to $\lambda_j(1 + \theta_K)$. Since the coder will typically wish to place the transmitted signal energy, as much as possible, according to smaller eigenvalues of the total channel noise, the effect of the jamming is to increase those smaller eigenvalues by a factor that converges upward to $(1 + \theta_K)$.

3. Concluding comments. The above development and that of [7] provide a complete solution for the information capacity of the Gaussian channel with jamming in which the coder's constraint covariance is matched to the ambient noise covariance. An extension of interest is to obtain similar results when the coder constraint and the ambient noise covariance are mismatched.

In addition to the main result, several of the results may be regarded as unexpected and/or of particular interest. They include the following:

1. The constraint applied to the jammer must be of a weaker type than that applied to the coder. This can be interpreted to mean that the jammer must have more flexibility in choosing

the distribution of the available jamming energy as a function of frequency. If this flexibility is the same as that of the coder, then the jammer cannot decrease the information capacity of the channel, regardless of the amount of available jamming energy.

2. With the constraints applied in this paper, which enable the jammer to reduce the information capacity, the usual sufficient conditions [8] for application of the von Neumann minimax theorem are not satisfied: thus, existence of a saddle point (or even a saddle value) is not guaranteed by applying the von Neumann theorem. Nevertheless, a solution is shown to exist, and the saddle value, a saddle point, and a minimax strategy for the jammer are obtained.

3. The essential effect of jamming is to convert the infinite-dimensional channel into a finite-dimensional channel with the same energy limitations on coder and jammer. The "effective dimensionality" of the channel is determined by the problem parameters: the noise covariance and the constraints.

REFERENCES

[1] C. R. BAKER, *Joint measures and cross-covariance operators*, Trans. Amer. Math. Soc., 186 (1973), pp. 272–289.

[2] ———, *Capacity of the Gaussian channel without feedback*, Inform. and Control, 37 (1978), pp. 70–79.

[3] ———, *Capacity of the mismatched Gaussian channel*, IEEE Trans. Inform. Theory, IT-33 (1987), pp. 802–812.

[4] ———, *Coding capacity for a class of additive channels*, IEEE Trans. Inform. Theory, IT-37 (1991), pp. 233–243.

[5] ———, *Capacity of dimension-limited channels*, J. Multivariate Anal., 37 (1991), pp. 239–258.

[6] C. R. BAKER AND S. IHARA, *Information capacity of the stationary Gaussian channel*, IEEE Trans. Inform. Theory, IT-37 (1991), pp. 1314–1326.

[7] C. R. BAKER AND I.-F. CHAO, *Information capacity of channels with partially unknown noise*, I. *Finite-dimensional channels*, SIAM J. Appl. Math., 56 (1996), pp. 945–962.

[8] V. BARBU AND TH. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Reidel, Boston, 1986.

[9] T. Ü. BASAR AND T. BASAR, *Optimum coding and decoding schemes for the transmission of a stochastic process over a channel with partially unknown statistics*, Stochastics, 8 (1982), pp. 213–237.

[10] N. M. BLACHMAN, *Communication as a game*, IRE WESCON Convention Record, II (1957), pp. 61–66.

[11] J. M. BORDEN, D. M. MASON, AND R. J. MCELIECE, *Some information-theoretic saddlepoints*, SIAM J. Control Optim., 23 (1985), pp. 129–143.

[12] R. L. DOBRUSHIN, *Optimum information transmission through a channel with unknown parameters*, Radiotechnika i Electronika, 4 (1959), pp. 1951–1956.

[13] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part* I, Wiley-Interscience, New York, 1957.

[14] B. HUGHES AND P. NARAYAN, *Gaussian arbitrarily-varying channels*, IEEE Trans. Inform. Theory, IT-34 (1987), pp. 267–284.

[15] ———, *The capacity of a vector Gaussian arbitrarily-varying channel*, IEEE Trans. Inform. Theory, IT-34 (1988), pp. 995–1003.

[16] S. IHARA, *On the capacity of channels with additive non-Gaussian noise*, Inform. and Control, 37 (1978), pp. 34–39.

[17] J. LINDENSTRAUSS AND L. TZAFRIRI, *Classical Banach Spaces* I, Springer-Verlag, Berlin, 1977.

[18] R. J. MCELIECE AND W. E. STARK, *An information theoretic study of communication in the presence of jamming*, Proc. IEEE Int. Conf. Communications, (1981), pp. 45.3.1–45.3.5.

[19] R. J. MCELIECE, *Communications in the presence of jamming—an information theory approach*, in Secure Digital Communications, CISM Courses and Lectures 279, G. Longo, ed., Springer-Verlag, New York, 1983.

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

[21] R. SCHATTEN, *Norm Ideals of Completely Continuous Operators*, Springer-Verlag, Berlin, 1960.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1473–1495, September 1996

A UNIQUENESS RESULT FOR THE LINEAR SYSTEM OF ELASTICITY AND ITS CONTROL THEORETICAL CONSEQUENCES*

ENRIQUE ZUAZUA†

Abstract. We consider the time-dependent linear system of three-dimensional elasticity for an homogeneous and isotropic elastic body Ω with Dirichlet boundary conditions. We prove that when the domain Ω is of class C^2 and is not symmetric with respect to a plane parallel to $x_3 = 0$, any solution whose first two components vanish in some open subset of Ω for a large enough time interval has to be identically zero. The proof of this uniqueness result can be reduced to show that solutions of the scalar wave equation with zero Dirichlet boundary conditions that are of the form $u(x, y, z, t) = \varphi(x, y, t) + \psi(z, t)$ have to be identically zero. This result depends not only on the length of the time interval but also on the geometry of the domain. As a consequence of this uniqueness result we prove that the linear system of elasticity is approximately controllable by means of planar volume forces with zero third component. We show that these results fail for some Lipschitz domains that are piecewise smooth and symmetric with respect to the plane $x_3 = 0$.

Key words. linear system of elasticity, evolution equations, Dirichlet boundary conditions, planar deformations, uniqueness, eigenfunctions, controllability

AMS subject classifications. 73C02, 73C15, 35L10, 93B05

1. Introduction and main results.

Let us consider an homogeneous and isotropic elastic n-dimensional body occupying a bounded domain Ω of $\mathbb{R}^n$ ($n = 2, 3$) of class C^2. ($\Gamma = \partial\Omega$ is of class C^2, and Ω is locally at one side of $\partial\Omega$.)

Let ω be an open and nonempty subset of Ω and $T > 0$. We denote by χ_ω the characteristic function of ω.

We consider the system of linear elasticity with a control (volume force) supported in ω and with homogeneous boundary conditions of Dirichlet type:

$$\begin{cases} u_{tt} - \mu\Delta u - (\lambda + \mu)\nabla \mathrm{div} u = f\chi_\omega & \text{in } \Omega \times (0, T), \\ u = 0 & \text{on } \Gamma \times (0, T), \\ u(0) = u^0, u_t(0) = u^1 & \text{in } \Omega, \end{cases} \tag{1.1}$$

where $\mu, \lambda > 0$ are Lamé's constants.

We assume that $f \in (L^2(\Omega \times (0, T)))^n$ is of the form

$$f = (f_1, \ldots, f_{n-1}, 0), \tag{1.2}$$

i.e., the nth component of f vanishes.

The initial data (u^0, u^1) are supposed to be in $H = (H_0^1(\Omega))^n \times (L^2(\Omega))^n$. Under these conditions, system (1.1) admits a unique solution $(u, u_t) \in C([0, T]; H)$.

We are interested in the controllability properties of this system under restriction (1.2) on the set of controls. More precisely, we investigate its approximate controllability. In other words, we want to know whether, for T large enough, the set

$$R(T) = \{(u(T), u_t(T)) : f \in (L^2(\Omega \times (0, T)))^n \text{ satisfying (1.2)}\}$$

is dense in H for all initial data $(u^0, u^1) \in H$.

Received by the editors December 16, 1993; accepted for publication (in revised form) March 4, 1995. This research was performed while the author was visiting the Department of Computational and Applied Mathematics of Rice University with the support of the Ministerio de Educación y Ciencia (Spain). This research was partially supported by Dirección General de Investigación Cientifica y Tecnológica grants PB90-0245 and PB93-1203 and EEC grant SC1-CT91-0732.

†Matemática Aplicada, Universidad Complutense, 28040 Madrid, Spain (zuazua@sunma4.mat.ucm.es).

Thus, for instance in three space dimensions, we are trying to understand the final configurations that an elastic body may reach by means of the action of arbitrary planar volume forces.

In the absence of restriction (1.2) on the set of controls, the approximate controllability is well known and is a consequence of the following uniqueness result, which turns out to be an immediate corollary of Holmgren's uniqueness theorem: If φ is a solution of

$$\begin{cases} \varphi_{tt} - \mu\Delta\varphi - (\lambda+\mu)\nabla\mathrm{div}\varphi = 0 & \text{in } \Omega \times (0,T), \\ \varphi = 0 & \text{on } \Gamma \times (0,T) \end{cases} \tag{1.3}$$

in the class $C([0,T];(L^2(\Omega))^n) \cap C^1([0,T];(H^{-1}(\Omega))^n)$ such that

$$\varphi_1 \equiv \cdots \equiv \varphi_{n-1} \equiv \varphi_n \equiv 0 \text{ in } \omega \times (0,T)$$

and T is large enough, then $\varphi \equiv 0$ in $\Omega \times (0,T)$.

In that case (i.e., without restriction (1.2)) the control time is

$$T(\Omega) = \frac{2\delta_n(\Omega;\omega)}{\sqrt{\mu}},$$

the quantity δ_n being defined as follows. For any open subset $\mathcal{O}$ of Ω,

$$\delta_n(\Omega;\mathcal{O}) = \sup_{x\in\Omega\setminus\mathcal{O}} \inf_{\gamma\in\xi(x;\mathcal{O})} \ell(\gamma), \tag{1.4}$$

where $\xi(x;\mathcal{O})$ denotes the set of curves in Ω joining x and $\overline{\mathcal{O}}$ and $\ell(.)$ denotes the length of the curve. We set $\delta_n(\Omega;\emptyset) = \infty$.

Since Ω is bounded and smooth, it is clear that $\delta_n(\Omega;\mathcal{O})$ is finite for any open and nonempty subset $\mathcal{O}$ of Ω.

When restriction (1.2) is introduced in the set of controls, the approximate controllability cannot be obtained directly from Holmgren's uniqueness theorem, and we are led to investigate the following uniqueness or unique continuation problem: If φ is a solution of (1.3) in the class $C([0,T];(L^2(\Omega))^n) \cap C^1([0,T];(H^{-1}(\Omega))^n)$ such that

$$\varphi_1 \equiv \cdots \equiv \varphi_{n-1} \equiv 0 \text{ in } \omega \times (0,T) \tag{1.5}$$

and T is large enough, can we deduce that, necessarily, $\varphi \equiv 0$ in $\Omega \times (0,T)$?

In this paper we give the following positive answer to this problem.

THEOREM 1.1. *Suppose that Ω is a bounded domain of class C^2. When $n = 3$, we assume also that either*

(i) *an open subset of $\partial\Omega$ is contained in a plane of the form $x_3 = c$*

or

(ii) *Ω is not symmetric with respect to a plane parallel to $x_3 = 0$.*

Then there exists some $T^*(\Omega) \geq 0$ such that if φ is a solution of (1.3) *in the class $C([0,T];(L^2(\Omega))^n) \cap C^1([0,T];(H^{-1}(\Omega))^n)$ satisfying* (1.5) *with*

$$T > \frac{2\delta_n(\Omega;\omega)}{\sqrt{\mu}} + T^*(\Omega),$$

then $\varphi \equiv 0$ in $\Omega \times (0,T)$.

As a consequence of Holmgren's uniqueness theorem, the proof of Theorem 1.1 can be reduced to the following uniqueness result for scalar wave equations.

THEOREM 1.2. *Let us suppose that Ω satisfies the assumptions of Theorem* 1.1. *Assume that $T > T^*(\Omega)$. Then, if*

$$\psi(x,t) = a(x_1, \dots, x_{n-1}, t) + b(x_n, t)$$

is a solution of the scalar wave equation

$$\begin{cases} \psi_{tt} - \mu\Delta\psi - (\lambda+\mu)\frac{\partial^2\psi}{\partial x_n^2} = 0 \ \text{in } \Omega \times (0,T), \\ \psi = 0 \ \text{on } \partial\Omega \times (0,T), \end{cases} \tag{1.6}$$

necessarily $\psi \equiv 0$ in $\Omega \times (0,T)$.

The proof of Theorem 1.2 can be reduced to the following elliptic version of it.

THEOREM 1.3. *Let us suppose that Ω satisfies the assumptions of Theorem* 1.1. *Then, if*

$$\phi(x) = p(x_1, \dots, x_{n-1}) + \alpha \cos\left(\frac{\kappa}{\sqrt{\lambda+2\mu}} x_n + \beta\right) \tag{1.7}$$

is an eigenfunction of the elliptic eigenvalue problem

$$\begin{cases} -\mu\Delta\phi - (\lambda+\mu)\frac{\partial^2\phi}{\partial x_n^2} = \kappa^2\phi \ \text{in } \Omega, \\ \phi = 0 \ \text{on } \partial\Omega \end{cases} \tag{1.8}$$

for some $\alpha, \beta, \kappa \in \mathbb{R}$ necessarily, $\phi \equiv 0$ in Ω.

Remark 1.4. Observe that in Theorem 1.1 the uniqueness time for the system of elasticity under the weak restriction (1.5) is, in general, larger than the classical uniqueness time that one obtains when all components vanish (i.e., larger than $2\delta_n(\Omega;\omega)/\sqrt{\mu}$). Indeed, under restriction (1.5) the uniqueness time is increased by the quantity $T^*(\Omega)$. We will give below some sharp estimates on this quantity that depend strongly on the geometry of the domain (not only on its size!). In some cases $T^*(\Omega) = 0$ so that the uniqueness time under restriction (1.5) coincides with the classical one. But in general, $T^*(\Omega) > 0$.

Note that when $\omega = \Omega$, the uniqueness time in Theorems 1.1 and 1.2 coincide since $\delta_n(\Omega;\Omega) = 0$.

Remark 1.5. The assumption on the C^2 regularity on the domain Ω is sharp in the sense that the results above fail for some piecewise C^2 domains.

For instance, as we will see in §7, in dimension $n = 2$ when Ω is the polygonal domain

$$\Omega = \Big\{(x_1,x_2) \in \Big(-\sqrt{\mu/(\lambda+2\mu)}, \sqrt{\mu/(\lambda+2\mu)}\Big) \times (-1,1) : \\ |\sqrt{\lambda+2\mu}x_1/\sqrt{\mu}| - 1 < x_2 < 1 - |\sqrt{\lambda+2\mu}x_1/\sqrt{\mu}|\Big\}, \tag{1.9}$$

Theorems 1.1, 1.2, and 1.3 do not hold. Moreover, Theorems 1.1 and 1.2 do not hold for any time $T > 0$.

As we will see below analogous counterexamples may be constructed in three space dimensions (see §7).

However, there is a way of relaxing the C^2 assumption on Ω, as we will see in Theorem 1.8 below.

When $n = 3$, we need the nonsymmetry assumption in order to prove Theorem 1.3. This assumption is not sharp. Indeed, we can construct C^2 domains which are symmetric with respect to the plane $x_3 = 0$ and where the results above hold. Given Ω we define the domain

$$\mathcal{W} = \left\{x \in R^3 : \left(x_1, x_2, \sqrt{\lambda+2\mu}x_3/\sqrt{\mu}\right) \in \Omega\right\}.$$

Let us choose an ellipsoid Ω such that the corresponding $\mathcal{W}$ is a ball of $\mathbb{R}^3$. Observe that if there exists ϕ defined in Ω satisfying (1.7), (1.8), then

$$\varphi(x) = \phi\left(x_1, x_2, \sqrt{\lambda + 2\mu}x_3/\sqrt{\mu}\right) = p(x_1, \ldots, x_{n-1}) + \alpha\cos\left(\frac{\kappa}{\sqrt{\mu}}x_n + \beta\right)$$

is an eigenfunction of the elliptic eigenvalue problem

$$\begin{cases} -\mu\Delta\varphi = \kappa^2\phi \text{ in } \mathcal{W}, \\ \varphi = 0 \text{ on } \partial\mathcal{W}. \end{cases}$$

Such an eigenfunction does not exist since, as is well known, when $\mathcal{W}$ is a ball of $\mathbb{R}^3$, the eigenfunctions are of the form $\varphi(x) = a(|x|)b(\theta)$, where b is an eigenfunction of the Laplacian over the unit sphere.

We have constructed an ellipsoid which is smooth and symmetric with respect to $x_3 = 0$, where the results above hold. In §5 we will discuss with more detail the assumption on the nonsymmetry of the domain. □

In order to give an estimate on the time $T^*(\Omega)$ it is convenient to introduce some notation.

By $\Omega^{n-1} \subset \mathbb{R}^{n-1}$ and $\Omega_1 \subset \mathbb{R}$ we denote respectively the projections of Ω on the hyperplane $x_n = 0$ and on the axis Ox_n. On the other hand, by $\mathcal{U}^{n-1} \subset \mathbb{R}^{n-1}$ (resp., $\mathcal{U}_1 \subset \mathbb{R}$) we denote the union of the projections on the hyperplane $x_n = 0$ (resp., on the axis Ox_n) of all those components of the boundary $\partial\Omega$ that can be written in the form $x_n = h(x_1, \ldots, x_{n-1})$ with h of class C^2 and such that

$$|\nabla' h(x_1, \ldots, x_{n-1})|^2 \neq \frac{\lambda + 2\mu}{\mu}$$

or

$$\Delta' h(x_1, \ldots, x_{n-1}) \neq 0.$$

By ∇' and Δ' we denote the gradient and Laplacian in the variables $(x_1, \ldots, x_{n-1})$.

The following proposition provides an estimate on the value $T^*(\Omega)$.

PROPOSITION 1.6. *Let us suppose that Ω satisfies the assumptions of Theorem* 1.1. *Then the results above hold with*

$$T^*(\Omega) = 2\min\left(\frac{1}{\sqrt{\mu}}\delta_{n-1}(\Omega^{n-1}; \mathcal{U}^{n-1}), \frac{1}{\sqrt{\lambda + 2\mu}}\delta_1(\Omega_1; \mathcal{U}_1)\right).$$

By δ_{n-1} and δ_1 we denote the $(n-1)$- and one-dimensional versions of the quantity δ_n defined in (1.4).

Since Ω is of class C^2 (to be of class C^1 would suffice), it is clear that $\mathcal{U}^{n-1}$ is nonempty and therefore $T^*(\Omega)$ is finite.

Remark 1.7. When dealing with a particular domain Ω the estimate above on $T^*(\Omega)$ might be eventually improved. However, in §6 we will show that the value given to $T^*(\Omega)$ in Proposition 1.6 is, in general, sharp. □

The proof of Theorem 1.1 shows, roughly, that Theorem 1.1 holds if and only if the uniqueness result for the elliptic problem (1.7), (1.8) of Theorem 1.3 holds. A careful analysis of this elliptic problem allows us to relax C^2 regularity assumption on Ω. Indeed, we have the following result.

THEOREM 1.8. *The results above hold if Ω satisfies the following four conditions:*

a) *Ω is a piecewise C^2-bounded domain.*

b) *Some open and nonempty C^2 component of Γ can be written in the form $x_n = h(x_1, \ldots, x_{n-1})$ with $|\nabla' h|^2 \neq (\lambda + 2\mu)/\mu$ or $\Delta' h \neq 0$ everywhere on that component.*

c) *There exists a point of a* C^2 *component of the boundary of* Ω *where the tangent hyperplane to* Ω *exists, and it is parallel to the axis* Ox_n.

d) *When* $n = 3$, *either*

d_1) *an open subset of the boundary of* Ω *is contained on a plane of the form* $x_3 = c$ *or*

d_2) Ω *is not symmetric with respect to a plane of the form* $x_3 = c$.

Moreover, under these conditons, the definition of $T^*(\Omega)$ *given in Proposition* 1.6 *remains valid.*

As a consequence of Theorems 1.1 and 1.8 we have the following approximate controllability result.

THEOREM 1.9. *Let us suppose that* Ω *satisfies the assumptions of Theorem* 1.8. *Then, if*

$$T > 2\frac{\delta(\Omega;\omega)}{\sqrt{\mu}} + T^*(\Omega),$$

system (1.1) *is approximately controllable at time* T *under the constraint* (1.2). *More precisely, for all* (u^0, u^1) *and* (v^0, v^1) *in* H *and* $\varepsilon > 0$ *there exists* $f \in (L^2(\Omega \times (0,T)))^n$ *obeying* (1.2) *such that the solution of* (1.1) *satisfies*

$$\left[\|u(T) - v^0\|^2_{H_0^1(\Omega))^n} + \|u_t(T) - v^1\|^2_{L^2(\Omega))^n}\right]^{1/2} \leq \varepsilon. \tag{1.10}$$

In other words, $R(T)$ *is dense in* H.

The fact that Theorems 1.1 and 1.8 imply Theorem 1.9 can be proven by rather classical arguments in control theory.

Remark 1.10. According to Theorem 1.8, the results above hold when Ω is the n-dimensional cube $\Omega = (0,1)^n$ with $T^*(\Omega) = 0$. In fact, they hold in any cylinder $\Omega = \Theta \times (0, l)$, where Θ is a bounded and piecewise C^2 domain of $\mathbb{R}^{n-1}$.

In this case, when Dirichlet boundary conditions are replaced by periodic ones, it is easy to check explicitly that the exact controllability does not hold with L^2-controls satisfying (1.2), i.e., $R(T) \neq H$ (see §9). In other words, (1.10) does not hold with $\varepsilon = 0$ for all initial and final data. We do not have an explicit counterexample for Dirichlet boundary conditions, but very probably the same happens.

Without the constraint (1.2), exact controllability with $(L^2(\Omega \times (0,T)))^n$-controls holds for a certain class of ω's. For instance, if ω is a neighborhood of the boundary of Ω exact controllability holds with control time, $T(\Omega) = \operatorname{diam}(\Omega \setminus \omega)/\sqrt{\mu}$ (see [4]). In fact, in order to have exact controllability, it is sufficient to take as support of the controls a set ω of the form $\omega = \Omega \cap \mathcal{O}$, where $\mathcal{O}$ is an open neighborhood in $\mathbb{R}^n$ of a subset of the boundary of Ω of the form

$$\Gamma(x^0) = \{x \in \Gamma : (x - x^0) \cdot \nu(x) > 0\}$$

for any $x^0 \in \mathbb{R}^n$. (By $\nu(x)$ we denote the outward unit normal to Ω at $x \in \Gamma$.) □

Remark 1.11. The same type of example shows that Theorem 1.1 and its counterpart in the more general setting of Theorem 1.8 are sharp in the sense that if, instead of (1.5), we impose the weaker condition

$$\varphi_1 \equiv \cdots \equiv \varphi_{n-2} \equiv 0 \quad \text{in } \omega \times (0,T)$$

and periodic boundary conditions, then uniqueness fails (see §9.2). □

The rest of this paper is organized as follows. In §2 we prove the elliptic version of the uniqueness result, i.e., Theorem 1.3. In §3 we prove Theorem 1.2. In §4 we prove the main

uniqueness results for the system of elasticity, Theorems 1.1 and 1.8. In §5 we discuss the assumption on the nonsymmetry of the domain and formulate some open problems that may lead to the characterization of those symmetric domains where the results above hold. In §6 we show that the uniqueness time in Theorem 1.1 and the estimate on $T^*(\Omega)$ given in Proposition 1.6 are sharp. In §7 we discuss how to extend the counterexample of Remark 1.5 to three space dimensions. In §8 we derive the controllability result stated in Theorem 1.9. Finally, in §9 we give two examples of noncontrollability as mentioned in Remarks 1.10 and 1.11.

In the sequel we will denote by ∇', Δ', div′, and curl′ the $(n-1)$-dimensional gradient, Laplacian, divergence, and curl operators in the variables $(x_1, \dots, x_{n-1})$.

Some of the results of this article were announced in [6].

2. Proof of Theorem 1.3. First, we prove Theorem 1.3 in the case $n = 3$ under the additional assumption d). Then, we prove it when $n = 2$ under the more general assumptions.

2.1. The case $n = 3$. Due to the Dirichlet boundary conditions it is obvious that if $\alpha = 0$, then $\phi \equiv 0$. Thus, without loss of generality we can assume that $\alpha = -1$.

On the other hand, if an open subset of the boundary of Ω is contained on a plane of the form $x_3 = c$, then necessarily, p is constant on an open subset of Ω^2, the projection of Ω on $x_3 = 0$. Since p satisfies the elliptic equation

$$-\mu\Delta' p = \kappa^2 p \quad \text{in } \Omega^2$$

by elliptic unique continuation, we deduce that p is constant everywhere in Ω^2. But then $\phi = \phi(x_3)$, and since it satisfies Dirichlet boundary conditions on Γ, we deduce that $\phi \equiv 0$.

Let us consider the case where Ω is not symmetric with respect to a plane parallel to $x_3 = 0$. Let $x^0 = (x_1^0, x_2^0, x_3^0)$ be a point of a C^2 component of $\partial\Omega$ where the tangent plane to $\partial\Omega$ is parallel to the axis Ox_3. Obviously, such a point exists when Ω is bounded and of class C^1. Without loss of generality we may assume that $x_3^0 = 0$.

Let us introduce the cross section of Ω at the level $x_3 = x_3^0 = 0$, i.e., the set

$$\Omega_0 = \{(x_1, x_2) \in \mathbb{R}^2 : (x_1, x_2, 0) \in \Omega\}.$$

Clearly, Ω_0 is a two-dimensional domain of class C^2 in a neighborhood of $(x_1, x_2) = (x_1^0, x_2^0) = y^0$.

Let us now consider a small neighborhood of y^0 in Ω_0,

$$\mathcal{N}_\varepsilon = \{(x_1, x_2) \in \Omega_0 : |(x_1 - x_1^0, x_2 - x_2^0)| < \varepsilon\},$$

and a C^2 trajectory

$$\tau \in [0, 1] \to \theta(\tau) = (x_1(\tau), \dots, x_{n-1}(\tau)) \in \overline{\mathcal{N}_\varepsilon}$$

such that $\theta(0)$ belongs to the interior of $\mathcal{N}_\varepsilon$, $\theta(1) = (x_1^0, x_2^0)$, and $\theta'(1) = \vec{n}$, the outward unit normal vector to $\partial\Omega_0$. This trajectory is the projection on Ω_0 of a C^2-trajectory on $\partial\Omega$ (which we denote by $\sigma = \sigma(\tau) = (x_1(\tau), x_2(\tau), x_3(\tau))$) such that $\sigma(1) = x^0$.

Taking into account that the tangent hyperplane to Ω at x^0 is parallel to the Ox_3 axis, we deduce that, necessarily,

$$\left|\frac{dx_3(\tau)}{d\tau}\right| \to \infty \quad \text{as } \tau \to 1^-. \tag{2.1}$$

In view of the structure of the eigenfunction ϕ and the fact that ϕ vanishes on $\partial\Omega$, along this trajectory $\sigma = \sigma(\tau)$ we have

$$x_3(\tau) = \frac{\sqrt{\lambda + 2\mu}}{\kappa}[\arccos(p(x_1(\tau), x_2(\tau))) - \beta]. \tag{2.2}$$

Then

$$\frac{dx_3(\tau)}{d\tau} = -\frac{\sqrt{\lambda+2\mu}\big(\nabla' p(x_1(\tau), x_2(\tau)) \cdot (x_1'(\tau), x_2'(\tau))\big)}{\kappa\sqrt{1- \mid p(x_1(\tau), x_2(\tau))) \mid^2}}. \tag{2.3}$$

We now observe that p solves the following elliptic equation in Ω_0:

$$-\mu\Delta' p = \kappa^2 p \quad \text{in } \Omega_0, \qquad p = \cos(\beta) \quad \text{on } \partial\Omega_0. \tag{2.4}$$

Since Ω_0 is locally of class C^2 at $y^0 = (x_1^0, x_2^0)$, the classical regularity theory for elliptic equations (see, for instance, [2, §8.11]) guarantees that $p \in C^{1,\alpha}(\overline{\mathcal{N}_\varepsilon})$. Therefore $\nabla' p$ remains bounded as $(x_1, x_2) \to y^0$. Thus, in view of (2.1) and (2.3) we deduce that $|p(y^0)| = 1$. Without loss of generality we may assume that $p(y^0) = 1$. But then $\beta = 2k\pi$ for some $k \in \mathbf{Z}$ and the function ϕ is even with respect to x_3.

On the other hand, over Γ we have

$$x_3 = \frac{\sqrt{\lambda+2\mu}}{\kappa}[\arccos(p(x_1, x_2)) - \beta]. \tag{2.5}$$

This implies that Ω is symmetric with respect to $x_3 = 0$ in the set where $p(x_1, x_2) > -1$. Thus, if $\inf_{\Omega_0} p > -1$, Ω is symmetric with respect to $x_3 = 0$, and this contradicts the nonsymmetry assumption. Then, necessarily, there exists a point y^1 in the interior of Ω_0 where $p(y^1) = -1$ and $\nabla' p(y^1) = 0$. In this case, the symmetry of Ω has to be analyzed with more care. Indeed, in principle, the symmetry of the domain may be broken since, for instance, on the upper half of Γ, x_3 may achieve its maximum at this point y^1 of Ω_0 but at the lower half of Γ, x_3 may keep decreasing in a neighborhood of y^1 by switching from one branch to another of the function arc cos. It is easy to see that this second possibility may not arise without introducing discontinuities on the boundary of Ω.

Thus, Ω is symmetric. This concludes the proof of Theorem 1.3 when $n = 3$. □

Remark 2.1. Note that if $p(y^1) = -1$ at the upper point of the boundary where x_3 achieves its maximum, we have

$$\frac{\partial x_3}{\partial x_\alpha} = -\frac{\sqrt{\lambda+2\mu}}{\kappa}\frac{\partial p(y^1)/\partial x_\alpha}{\sqrt{1- |p(y^1))|^2}}$$

for $\alpha = 1, 2$. But these quantities are not well defined since both numerator and denominator vanish. Applying l'Hôpital's rule we obtain

$$\begin{aligned} l_\alpha = \lim_{(x_1,x_2)\to y^1} \frac{\partial x_3}{\partial x_\alpha} &= -\frac{\sqrt{\lambda+2\mu}}{\kappa}\frac{\partial^2 p(y^1)/\partial x_\alpha^2}{-p(y^1)\partial p(y^1)/\partial x_\alpha/\sqrt{1- |p(y^1))|^2}} \\ &= -\frac{\lambda+2\mu}{\kappa^2}\frac{\partial^2 p(y^1)/\partial x_\alpha^2}{p(y^1) l_\alpha} = \frac{\lambda+2\mu}{\kappa^2}\frac{\partial^2 p(y^1)/\partial x_\alpha^2}{l_\alpha}. \end{aligned}$$

Thus,

$$|\nabla' x_3(y^1)|^2 = l_1^2 + l_2^2 = \frac{\lambda+2\mu}{\kappa^2}\Delta' p(y^1) = -\frac{\lambda+2\mu}{\mu} p(y^1) = \frac{\lambda+2\mu}{\mu} \neq 0.$$

Thus the surface Γ may not be C^1 in a neighborhood of this point in which x_3 achieves its maximal value with $p(y^1) = -1$. Indeed, if it were C^1, we would have $|\nabla' x_3(y^1)| = 0$. The computation above is justified since y^1 is in the interior of Ω_0 and therefore p is C^∞ in a neighborhood of y^1. □

Remark 2.2. Observe that this proof applies under the following assumptions:

a) Ω is piecewise of class $C^{1,\alpha}$.

b) There exists a point on a $C^{1,\alpha}$ component of $\partial\Omega$ where the plane tangent to Ω is parallel to the plane $x_3 = 0$.

c) Ω is not symmetric with respect to a plane of the form $x_3 = c$.

Therefore, in particular, it applies under the assumptions of Theorem 1.8. □

2.2. The case $n = 2$. The proof of the case $n = 3$ applies as well and reduces the analysis to the situation where $|p(y^0)| = 1$. (Note that y^0 reduces to $y^0 = x_1^0$.) Let us assume, without loss of generality, that $p(x_1^0) = 1$. Note that since p solves the one-dimensional version of (2.4), $p(x_1) = \gamma \cos(\kappa x_1/\sqrt{\mu} + \theta)$ for some $\gamma \geq 1$ and real θ. If $\gamma = 1$, then $p'(x_1^0) = 0$, and therefore, by the computation of Remark 2.1, we deduce that

$$|\partial x_2(x_1^0)/\partial x_1|^2 = \frac{\lambda + 2\mu}{\mu},$$

which is finite. This contradicts (2.1).

Suppose now that $\gamma > 1$. Taking into account that $|p| \leq 1$ for all x_1 in Ω_0, we deduce that x_1 has to lie in an interval to one side of x_1^0 where $|\cos(\kappa x_1/\sqrt{\mu} + \theta)| \leq 1/\gamma$. But then

$$x_2 = \frac{\sqrt{\lambda + 2\mu}}{\kappa}[\arccos(p(x_1)) - \beta]$$

is monotonous, and therefore the boundary of Ω is included in a set that does not contain any closed curve. This concludes the proof of Theorem 1.3 when $n = 2$. □

Remark 2.3. Note that when $n = 2$, we do not need any symmetry assumption on Ω because the sole solution of the one-dimensional version of (2.4) such that $p = 1$ on the boundary of a one-dimensional interval and takes values $|p| \leq 1$ everywhere satisfies $p' = 0$ on the boundary. As we will see in §5, the two-dimensional version of this property seems to be much more delicate (it could be even false), and therefore we require the extra nonsymmetry assumption when $n = 3$. □

3. Proof of Theorem 1.2. First we observe that, under the assumptions of Theorem 1.2, ψ may be written as

$$\psi = a(x_1, \ldots, x_{n-1}, t) + b(x_n, t),$$

where a and b are solutions of the following wave equation in space dimensions $n - 1$ and 1, respectively:

$$a_{tt} - \mu\Delta' a = 0 \text{ in } \Omega^{n-1} \times (0, T), \tag{3.1}$$

$$b_{tt} - (\lambda + 2\mu)\frac{\partial^2 b}{\partial x_n^2} = 0 \text{ in } \Omega_1 \times (0, T), \tag{3.2}$$

where Ω^{n-1} and $\Omega_1 = (l_1, l_2)$ are, respectively, the projections of Ω on the hyperplane $x_n = 0$ and the axis Ox_n.

In order to prove Theorem 1.2, in view of Theorem 1.3, it is sufficient to show that ψ is a solution of the wave equation (1.6) in separated variables. More precisely, it is sufficient to prove that ψ is of the form

$$\psi(x, t) = A(t)\phi(x)$$

with A such that

$$-A'' = \kappa^2 A$$

and with ϕ as in (1.7), (1.8).

Let γ be an open and nonempty subset of Γ that can be represented as $x_n = h(x_1, \dots, x_{n-1})$ for $(x_1, \dots, x_{n-1})$ in some open subset $\mathcal{U}$ of Ω^{n-1} with

$$|\nabla' h(x_1, \dots, x_{n-1})|^2 \neq \frac{\lambda + 2\mu}{\mu} \quad \text{or } \Delta' h \neq 0 \quad \text{for all } (x_1, \dots, x_{n-1}) \in \mathcal{U}. \tag{3.3}$$

Obviously, as soon as Ω is bounded and of class C^2, such a subset exists. Note also that, by hypothesis, this set exists if Ω is in the conditions of Theorem 1.8.

Since ψ vanishes on the boundary of Ω, we have

$$a(x_1, \dots, x_{n-1}, t) = -b(h(x_1, \dots, x_{n-1}), t) \quad \forall (x_1, \dots, x_{n-1}) \in \mathcal{U},\ 0 < t < T. \tag{3.4}$$

We distinguish two cases.

Case 1. The image of $h : \mathcal{U} \to \mathbb{R}$ reduces to a point; i.e., h is constant in $\mathcal{U}$, and thus the subset of the boundary that we are considering is a hyperplane parallel to $x_n = 0$.

Case 2. The image of h contains an interval (x_n^0, x_n^1) with $x_n^0 < x_n^1$.

Let us discuss these two cases separately.

Case 1: In view of (3.4) we have

$$a = a(t) \quad \text{in the cylinder } \mathcal{U} \times (0, T).$$

According to (3.1), we have $a_{tt} = 0$, and therefore a is linear in t, i.e.,

$$a = \alpha t + \beta \quad \text{in } \mathcal{U} \times (0, T)$$

for some $\alpha, \beta \in \mathbb{R}$. Thus

$$\psi = \tilde{b}(x_n, t) \quad \text{in } \mathcal{U} \times (0, T),$$

where $\tilde{b} = b + \alpha t + \beta$ is a (new) solution of (3.2). In other words, in the set $\mathcal{U} \times (0, T)$, ψ depends only on x_n and t.

If $\Omega \subset \mathcal{U} \times \mathbb{R}$, since $\psi = 0$ on $\partial\Omega \times (0, T)$, this clearly implies that $\psi \equiv 0$. If not, we observe that by Holmgren's uniqueness theorem (see [3] or [4, Chap. 1]) a has to be of the same form in a larger set. The latter contains

$$\Omega^{n-1} \times \left(\frac{\delta_{n-1}(\Omega^{n-1}; \mathcal{U})}{\sqrt{\mu}}, T - \frac{\delta_{n-1}(\Omega^{n-1}; \mathcal{U})}{\sqrt{\mu}} \right)$$

as soon as $T > 2\delta_{n-1}(\Omega^{n-1}; \mathcal{U})/\sqrt{\mu}$. Since $\Omega \subset \Omega^{n-1} \times \mathbb{R}$, we deduce that

$$\psi = b(x_n, t) \quad \text{everywhere in } \Omega \times \left(\frac{\delta_{n-1}(\Omega^{n-1}; \mathcal{U})}{\sqrt{\mu}}, T - \frac{\delta_{n-1}(\Omega^{n-1}; \mathcal{U})}{\sqrt{\mu}} \right), \tag{3.5}$$

and, in view of the boundary conditions that ψ satisfies, this implies that $\psi \equiv 0$.

Case 2: Since a and b solve the wave equations (3.1) and (3.2), respectively, we have

$$\begin{aligned} &(\lambda + 2\mu - \mu \mid \nabla' h(x_1, \dots, x_{n-1}) \mid^2) \frac{\partial^2 b}{\partial x_n^2}(h(x_1, \dots, x_{n-1}), t) \\ &\quad -\mu \Delta' h(x_1, \dots, x_{n-1}) \frac{\partial b}{\partial x_n}(h(x_1, \dots, x_{n-1}), t) = 0 \end{aligned} \tag{3.6}$$

for all $(x_1, \dots, x_{n-1}) \in \mathcal{U}$ and $t \in (0, T)$.

From (3.3) and (3.6) we deduce immediately that b is of the form

$$b(x_n, t) = A(t)B(x_n) + C(t) \quad \text{in } \mathcal{U} \times (0, T). \tag{3.7}$$

Since b solves the wave equation (3.2) we have

$$A''(t)B(x_n) - (\lambda + 2\mu)A(t)B''(x_n) + C''(t) = 0.$$

Taking one derivative of this equation with respect to x_n we see that

$$A''(t)B'(x_n) - (\lambda + 2\mu)A(t)B'''(x_n) = 0.$$

Thus A and B satisfy

$$A''(t) = -(\lambda + 2\mu)\xi^2 A(t) \tag{3.8}$$

and

$$-B''' = \xi^2 B' \tag{3.9}$$

for some real number ξ^2, and therefore they are necessarily of the form

$$A(t) = \alpha_1 e^{i\xi\sqrt{\lambda+2\mu}t} + \beta_1 e^{-i\xi\sqrt{\lambda+2\mu}t}$$

and

$$B(x_n) = \alpha_2 e^{i\xi x_n} + \beta_2 e^{-i\xi x_n} + \gamma$$

for some complex numbers $\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma$, and ξ.

It is also easy to check that $C = C(t)$ has to be a linear function of t.

On the other hand, (3.4) implies

$$a(x_1, \dots, x_{n-1}, t) = -A(t)D(x_1, \dots, x_{n-1}) - C(t) \quad \text{in } \mathcal{U} \times (0, T), \tag{3.10}$$

where $D(x_1, \dots, x_{n-1}) = B(h(x_1, \dots, x_{n-1}))$.

Since a satisfies the wave equation (3.1), the function $D = D(x_1, \dots, x_{n-1})$ has to satisfy the elliptic equation

$$-\Delta' D = \frac{\lambda + 2\mu}{\mu}\xi^2 D. \tag{3.11}$$

All this implies, in particular, that ψ has the form

$$\begin{cases} \psi(x_1, \dots, x_n, t) = A(t)\phi(x) & \text{in } \mathcal{U} \times (x_n^0, x_n^1) \times (0, T), \\ \phi(x) = D(x_1, \dots, x_{n-1}) + B(x_n) \end{cases} \tag{3.12}$$

as in Theorem 1.3.

By Holmgren's uniqueness theorem it is easy to see that ψ has to preserve this form in a larger set. Indeed, since b satisfies the wave equation (3.2) and has the form (3.7) with A and B satisfying (3.8) and (3.9), by unique continuation of solutions of the one-dimensional wave equation, b has to preserve this structure in the set

$$\mathcal{Q}_1 = \Omega_1 \times \left(\frac{\max(l_2 - x_n^1, x_n^0 - l_1)}{\sqrt{\lambda + 2\mu}}, T - \frac{\max(l_2 - x_n^1, x_n^0 - l_1)}{\sqrt{\lambda + 2\mu}} \right) \tag{3.13}$$

if

$$T > \frac{2}{\sqrt{\lambda+2\mu}}\max(l_2 - x_n^1, x_n^0 - l_1).$$

Note that $\max(l_2 - x_n^1, x_n^0 - l_1) = \delta_1(\Omega_1, (x_n^0, x_n^1))$.

According to (3.10), (3.11) and taking into account that a solves the wave equation (3.1), by Holmgren's uniqueness theorem it is easy to see that a keeps the same structure in the set

$$\mathcal{Q}_2 = \Omega^{n-1} \times \left(\frac{1}{\sqrt{\mu}}\delta_{n-1}(\Omega^{n-1}; \mathcal{U}), T - \frac{1}{\sqrt{\mu}}\delta_{n-1}(\Omega^{n-1}; \mathcal{U})\right) \tag{3.14}$$

as soon as

$$T > \frac{2}{\sqrt{\mu}}\delta_{n-1}(\Omega^{n-1}; \mathcal{U}).$$

Suppose that

$$T > T_0 \quad \text{with } T_0 = 2\min\left(\frac{1}{\sqrt{\mu}}\delta_{n-1}(\Omega^{n-1}; \mathcal{U}), \frac{1}{\sqrt{\lambda+2\mu}}\delta_1(\Omega_1; (x_n^0, x_n^1))\right). \tag{3.15}$$

Then either a is of the form (3.10), (3.11) in the set (3.14) or b is of the form (3.7) in the set (3.13). In any case, as soon as (3.15) holds, ψ is of the form (3.12) in $\Omega \times (T_0, T - T_0)$.

The analysis of the two different cases above shows that the minimal time for uniqueness is the quantity $T^*(\Omega)$ introduced in Proposition 1.6. Indeed, if in (3.15) we take, instead of $\mathcal{U}$ (resp., (x_n^0, x_n^1)) the union of all sets $\mathcal{U}$ (resp., all intervals (x_n^0, x_n^1)) where the above applies, the time T_0 defined in (3.15) coincides with $T^*(\Omega)$.

4. Proof of Theorem 1.1. In view of (1.5) we have

$$\operatorname{div}\varphi = \frac{\partial \varphi_n}{\partial x_n} \quad \text{in } \omega \times (0, T). \tag{4.1}$$

Using the first $n-1$ equations of (1.3) we deduce that

$$\frac{\partial \operatorname{div}\varphi}{\partial x_j} = \frac{\partial^2 \varphi_n}{\partial x_j \partial x_n} = 0 \quad \text{in } \omega \times (0, T) \tag{4.2}$$

for $j = 1, \ldots, n-1$.

Combining (4.2) with the fact that

$$\frac{\partial^2 \varphi_i}{\partial x_j \partial x_n} = 0 \quad \text{in } \omega \times (0, T) \quad \text{for } i, j = 1, \ldots, n-1$$

we see that

$$\frac{\partial^2 \varphi}{\partial x_j \partial x_n} = 0 \quad \text{in } \omega \times (0, T) \quad \text{for } j = 1, \ldots, n-1.$$

For each $j \in \{1, \ldots, n-1\}$, the vector-valued function $\partial^2\varphi/\partial x_j \partial x_n$ solves the system of elasticity (1.3). Thus, applying Holmgren's uniqueness theorem we deduce that

$$\frac{\partial^2 \varphi}{\partial x_j \partial x_n} \equiv 0 \quad \text{in } Q_1 = \Omega \times \left(\frac{\delta_n(\Omega; \omega)}{\sqrt{\mu}}, T - \frac{\delta_n(\Omega; \omega)}{\sqrt{\mu}}\right).$$

Therefore,

$$\varphi = \rho(x_1, \dots, x_{n-1}, t) - \sigma(x_n, t) \text{ in } Q_1 \tag{4.3}$$

with $\rho = (\rho_1, \dots, \rho_n)$ and $\sigma = (\sigma_1, \dots, \sigma_n)$.

From (1.5) we deduce that

$$\rho_j \text{ and } \sigma_j \text{ are independent of } x \text{ in } \omega \times (0, T) \text{ for } j = 1, \dots, n-1. \tag{4.4}$$

On the other hand, $\phi = \operatorname{div}\varphi$ solves the scalar wave equation

$$\phi_{tt} - (\lambda + 2\mu)\Delta\phi = 0 \text{ in } \Omega \times (0, T). \tag{4.5}$$

Since

$$\frac{\partial\phi}{\partial x_j} = \frac{\partial \operatorname{div}\varphi}{\partial x_j} = \frac{\partial^2\varphi}{\partial x_j \partial x_n} = 0 \quad \text{in } \omega \times (0, T) \text{ for } j = 1, \dots, n-1$$

as a consequence of Holmgren's uniqueness theorem (this time applied to (4.5)) we deduce that

$$\frac{\partial\phi}{\partial x_j} = 0 \quad \text{in } Q_2 = \Omega \times \left(\frac{\delta_n(\Omega; \omega)}{\sqrt{\lambda + 2\mu}}, T - \frac{\delta_n(\Omega; \omega)}{\sqrt{\lambda + 2\mu}} \right)$$

for $j = 1, \dots, n-1$, i.e., $\phi = \phi(x_n, t)$ in Q_2.

But, according to (4.3),

$$\phi = \operatorname{div}'(\rho_1, \dots, \rho_{n-1}) - \frac{\partial\sigma_n}{\partial x_n}$$

in $Q_1 = Q_1 \cap Q_2$. Therefore, $\operatorname{div}'(\rho_1, \dots, \rho_{n-1})$ is independent of $(x_1, \dots, x_{n-1})$, i.e.,

$$\frac{\partial\rho_1}{\partial x_1} + \cdots + \frac{\partial\rho_{n-1}}{\partial x_{n-1}} = c(t) \text{ in } Q_1. \tag{4.6}$$

We now distinguish dimensions $n = 2$ and $n = 3$.

When $n = 3$, $w = \operatorname{curl}\varphi$ satisfies the wave equation

$$w_{tt} - \mu\Delta w = 0 \text{ in } \Omega \times (0, T). \tag{4.7}$$

Moreover, (1.5) implies that w_3 vanishes in $\omega \times (0, T)$. Thus, by Holmgren's uniqueness theorem we deduce that

$$w_3 \equiv 0 \text{ in } Q_1. \tag{4.8}$$

But

$$w_3 = \operatorname{curl}'(\rho_1, \rho_2).$$

Therefore, (4.8) implies the existence of a potential $p = p(x_1, x_2, t)$ such that $(\rho_1, \rho_2) = \nabla' p$. In view of (4.4) and (4.6) we see that p obeys

$$\begin{cases} \Delta' p = c(t) & \text{in } Q_1, \\ p = \gamma(t) \cdot (x_1, x_2) + q(t) & \text{in } \{(x, t) \in Q_1 : x \in \omega\}. \end{cases}$$

Thus, by elliptic unique continuation,

$$p = \gamma(t) \cdot (x_1, x_2) + q(t) \quad \text{in } Q_1$$

and then

$$\nabla' p = (\rho_1, \rho_2) \equiv \gamma(t) \qquad \text{in } Q_1.$$

Therefore,

$$\varphi_j = \gamma_j(t) - \sigma_j(x_3, t) \qquad \text{in } Q_1$$

for $j = 1, 2$. But since φ satisfies Dirichlet boundary conditions, we deduce that

$$\varphi_j \equiv 0 \text{ in } Q_1 \text{ for } j = 1, 2.$$

When $n = 2$, (4.6) implies that $\rho_1 = c(t)x_1 + d(t)$ in Q_1 and therefore $\varphi_1(x_1, x_2, t) = c(t)x_1 + d(t) - \sigma_1(x_2, t)$ in Q_1. But then φ_1 cannot be identically zero on the boundary of the bounded domain Ω except if $\varphi_1 \equiv 0$ in Q_1.

Thus, in both cases, it is sufficient to prove that $\varphi_n \equiv 0$ in Q_1. Indeed, if $\varphi_n \equiv 0$ in Q_1, then the initial data for φ at time $t = T/2$ are identically zero, i.e., $\varphi(x, T/2) \equiv \varphi_t(x, T/2) \equiv 0$. The system (1.3) being reversible in time, by uniqueness of solutions, we deduce $\varphi \equiv 0$ everywhere.

We have

$$\varphi_{n,tt} - \mu\Delta\varphi_n - (\lambda + \mu)\partial^2\varphi_n/\partial x_n^2 = 0 \quad \text{in } Q_1,$$

$$\varphi_n = \rho_n(x_1, \ldots, x_{n-1}, t) - \sigma_n(x_n, t) \quad \text{in } Q_1,$$

$$\varphi_n = 0 \quad \text{on } \Gamma \times (0, T).$$

But then, as an immediate consequence of Theorem 1.2, $\varphi_n \equiv 0$ in Q_1.

5. Discussion on the nonsymmetry assumption. In this section we analyze the nonsymmetry assumption that we have introduced in dimension $n = 3$. As was pointed out in Remark 1.5, the nonsymmetry assumption is required only for the proof of Theorem 1.3.

When analyzing the uniqueness property stated in Theorem 1.3 for a smooth domain Ω of $\mathbb{R}^3$ which is symmetric with respect to $x_3 = 0$, the proof of Theorem 1.3 leads naturally to the following question: Is there any smooth domain $\mathcal{O}$ (of class C^2) of $\mathbb{R}^2$ such that for some $\kappa \in \mathbb{R}$ the elliptic problem

$$-\mu\Delta' p = \kappa^2 p \text{ in } \mathcal{O}, \qquad p = 1 \text{ on } \partial\mathcal{O}, \tag{5.1}$$

admits a solution of class C^2 such that

$$\|p\|_\infty = 1, \qquad \nabla' p \neq 0 \text{ everywhere on } \partial\mathcal{O}? \tag{5.2}$$

If such a domain $\mathcal{O}$ and function p exist with $-1 < p < 1$ in the interior of $\mathcal{O}$, we can construct a smooth domain Ω of R^3, symmetric with respect to $x_3 = 0$, $\mathcal{O}$ being its cross section at the level $x_3 = 0$, and where the uniqueness property of Theorem 1.3 fails. Indeed, it is sufficient to define the upper half of the boundary of Ω as being given by the graph of the function

$$x_3 = \frac{\sqrt{\lambda + 2\mu}}{\kappa}[\arccos(p(x_1, x_2)) - \beta] \tag{5.3}$$

with β such that $\cos\beta = 1$. In this case, the function

$$\phi(x) = p(x_1, x_2) - \cos\left(\frac{\kappa}{\sqrt{\lambda + 2\mu}} x_n + \beta\right) \tag{5.4}$$

satisfies (1.8) in this domain Ω without being identically zero.

When p achieves the extremal values 1 and/or -1 in the interior of $\mathcal{O}$, the same can be done. But in this case the surface $\partial\Omega$ is not of class C^1 at those points where (x_1, x_2) are such that p achieves those values, as the computations of Remark 2.1 show.

Therefore, the existence of this type of domain $\mathcal{O}$ and function p satisfying (5.1) and (5.2) is relevant to understand the uniqueness property stated in Theorem 1.3 in the frame of symmetric domains.

In dimension $n = 2$, one has to analyze the one-dimensional version of (5.1), (5.2). As we pointed out in Remark 2.3, there is no open connected domain $\mathcal{O}$ in $\mathbb{R}$ where these properties hold. Thus, Theorem 1.3 holds without nonsymmetry assumptions.

This question may be addressed in a different way. Assume that Ω is a C^2 domain of $\mathbb{R}^3$ and symmetric with respect to the plane $x_3 = 0$. Suppose that the upper surface of $\partial\Omega$ is the graph of a function $h = h(x_1, x_2)$. Then, taking into account that

$$\phi(x) = p(x_1, x_2) - \cos\left(\frac{\kappa}{\sqrt{\lambda + 2\mu}} x_3 + \beta\right)$$

vanishes on the boundary, we have

$$p(x_1, x_2) = \cos\left(\frac{\kappa}{\sqrt{\lambda + 2\mu}} h(x_1, x_2) + \beta\right).$$

Since β has to be such that $\cos\beta = 1$, we can assume that $\beta = 0$. Taking into account that $-\mu\Delta' p = \kappa^2 p$ in $\mathcal{O}$ we deduce that

$$-\mu\sqrt{\lambda + 2\mu}\,\Delta' h + \kappa(\lambda + 2\mu - \mu|\nabla' h|^2)\operatorname{cotg}\left(\frac{\kappa}{\sqrt{\lambda + 2\mu}} h(x_1, x_2)\right) \text{ in } \mathcal{O} = 0. \tag{5.5}$$

In order to understand the uniqueness property of Theorem 1.3 without the symmetry assumption we have to analyze the existence of κ and h such that (5.5) has a smooth solution so that $\partial h/\partial\vec{n} = -\infty$ on $\partial\mathcal{O}$. (By $\vec{n}$ we denote the normal outward unit vector to $\partial\mathcal{O}$.)

In one space dimension (when $\mathcal{O}$ is an interval of $\mathbb{R}$), it is easy to see that all solutions of (5.5) develop singularities so that they are not Lipschitz.

6. Optimality of the time of uniqueness. In this section we prove that, in general, the uniqueness times given in Theorems 1.1 and 1.2 are sharp.

6.1. Scalar wave equations. In this section we prove the optimality of the uniqueness time $T^*(\Omega)$ given in Proposition 1.6. For the sake of simplicity we consider the bidimensional case $n = 2$.

Let us consider the following polygonal domain in $\mathbb{R}^2$ (see Figure 1):

$$\Omega = \left\{(x_1, x_2) \in \left(\frac{\gamma}{2}, \frac{3\gamma}{2}\right) \times \left(-\frac{1}{2}, \frac{1}{2}\right) : \frac{x_1}{2\gamma} - \frac{3}{4} < x_2 < \frac{3}{4} - \frac{x_1}{2\gamma}\right\}$$

$$\cup \left\{(x_1, x_2) \in \left(-\gamma, \frac{\gamma}{2}\right) \times (-1, 1) : \left|\frac{x_1}{\gamma}\right| - 1 < x_2 < 1 - \left|\frac{x_1}{\gamma}\right|\right\}, \tag{6.1}$$

where $\gamma = \sqrt{\mu/(\lambda + 2\mu)}$.

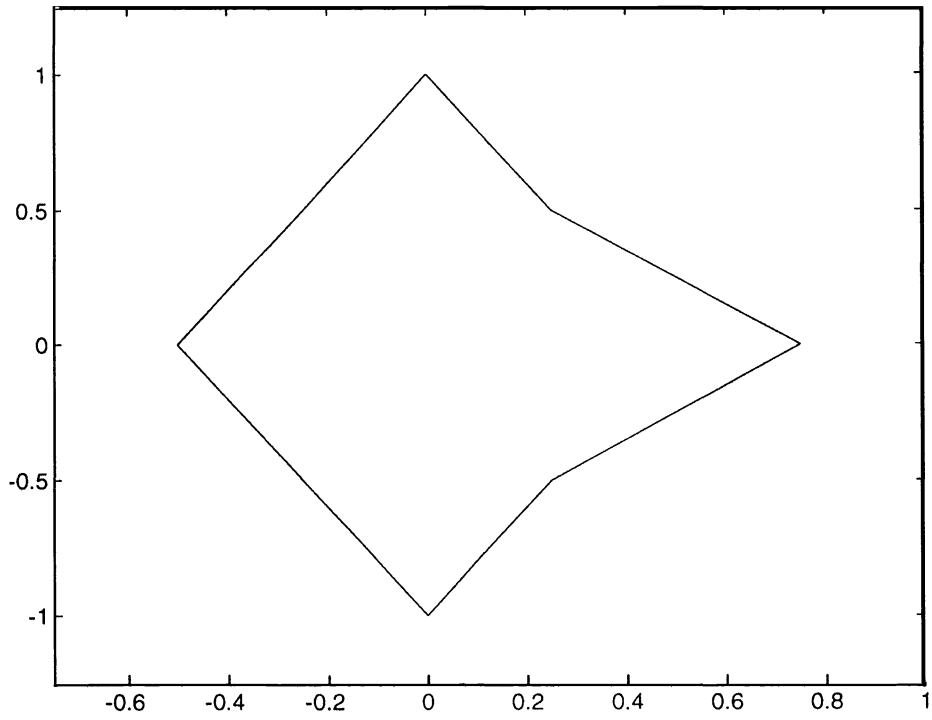


FIG. 1. *Polygonal domain* Ω *as in* (6.1) *with* $\lambda = 2$ *and* $\mu = 1$.

In this case

$$\Omega^{n-1} = \Omega^1 = (-\gamma, 3\gamma/2), \qquad \mathcal{U}^{n-1} = \mathcal{U}^1 = \left(\frac{\gamma}{2}, \frac{3\gamma}{2}\right),$$

$$\Omega_1 = (-1, 1), \qquad \mathcal{U}_1 = \left(-\frac{1}{2}, \frac{1}{2}\right).$$

Then

$$\delta_1(\Omega^1, \mathcal{U}^1) = 3\gamma/2, \qquad \delta_1(\Omega_1, \mathcal{U}_1) = \frac{1}{2},$$

and therefore

$$T^*(\Omega) = \frac{1}{\sqrt{\lambda + 2\mu}}.$$

Let us prove that this uniqueness time $T^*(\Omega)$ is optimal in the context of Theorem 1.2. Let b be an even and x-periodic function with period γ that solves the wave equation

$$b_{tt} - \mu b_{xx} = 0.$$

Let us choose b so that its support in the interval $[-\gamma, \gamma]$ is concentrated in the union of two very small intervals to the right of $-\gamma$ and 0, respectively, i.e.,

$$\text{supp}(b) \subset [-\gamma, -\gamma + \varepsilon] \cup [0, \varepsilon].$$

Then set

$$\psi(x_1, x_2, t) = b(x_1, t) - b(\gamma x_2, t).$$

It is easy to check that ψ solves (1.6) in the cylinder

$$\Omega \times \left(-\frac{(\gamma - \varepsilon)}{2\sqrt{\mu}}, \frac{(\gamma - \varepsilon)}{2\sqrt{\mu}} \right).$$

Since this is true for any $\varepsilon > 0$ and $T^*(\Omega) = \gamma/\sqrt{\mu}$, this implies that the uniqueness time $T^*(\Omega)$ is sharp.

Note that ψ does not satisfy the boundary conditions at the part of the boundary of Ω lying in the strip $\gamma/2 < x_1 < 3\gamma/2$ as soon as the support of $b(\cdot, t)$ enters this region.

As pointed out in Remark 1.4, this shows that the uniqueness time given in Theorem 1.1 is sharp when $\omega = \Omega$.

6.2. System of elasticity. In this section we prove that, in general, the uniqueness time given in Theorem 1.1 is optimal.

Let Ω be the unit ball of $\mathbb{R}^3$. It is easy to check that $T^*(\Omega) = 0$. Then the uniqueness time in Theorem 1.1 reduces to the classical one:

$$T(\Omega) = \frac{2\delta_3(\Omega; \omega)}{\sqrt{\mu}}.$$

Suppose that ω is a neighborhood of the boundary:

$$\omega = \{x \in \Omega : 1 - r < | x | \}$$

with $r < 1$. Then $T(\Omega) = 2(1 - r)/\sqrt{\mu}$.

Let φ be a solution of (1.3) with data at time $t = 0$ with support contained in a small ball around the origin of radius ε. It is easy to see that φ vanishes on ω on the time interval $(-(1 - r - \varepsilon)/\sqrt{\mu}, (1 - r - \varepsilon)/\sqrt{\mu})$. Since ε is arbitrarily small, this implies that $T(\Omega)$ is sharp.

7. Counterexamples to the uniqueness results. In this section we develop the two-dimensional counterexample mentioned in Remark 1.5. We also show how it can be extended to three space dimensions.

7.1. Dimension $n = 2$. Let us consider the following polygonal domain Ω in dimension $n = 2$ (see Figure 2):

$$\Omega = \Big\{ (x_1, x_2) \in \left(-\sqrt{\mu/(\lambda + 2\mu)}, \sqrt{\mu/(\lambda + 2\mu)} \right) \times (-1, 1) : \tag{7.1}$$
$$| \sqrt{\lambda + 2\mu} x_1/\sqrt{\mu} | - 1 < x_2 < 1 - | \sqrt{\lambda + 2\mu} x_1/\sqrt{\mu} | \Big\}.$$

Let b be an even and x-periodic function with period $\sqrt{\mu/(\lambda + 2\mu)}$ that solves the wave equation

$$b_{tt} - \mu b_{xx} = 0.$$

Then

$$\psi(x_1, x_2, t) = b(x_1, t) - b\left(\frac{\sqrt{\mu} x_2}{\sqrt{\lambda + 2\mu}}, t \right)$$

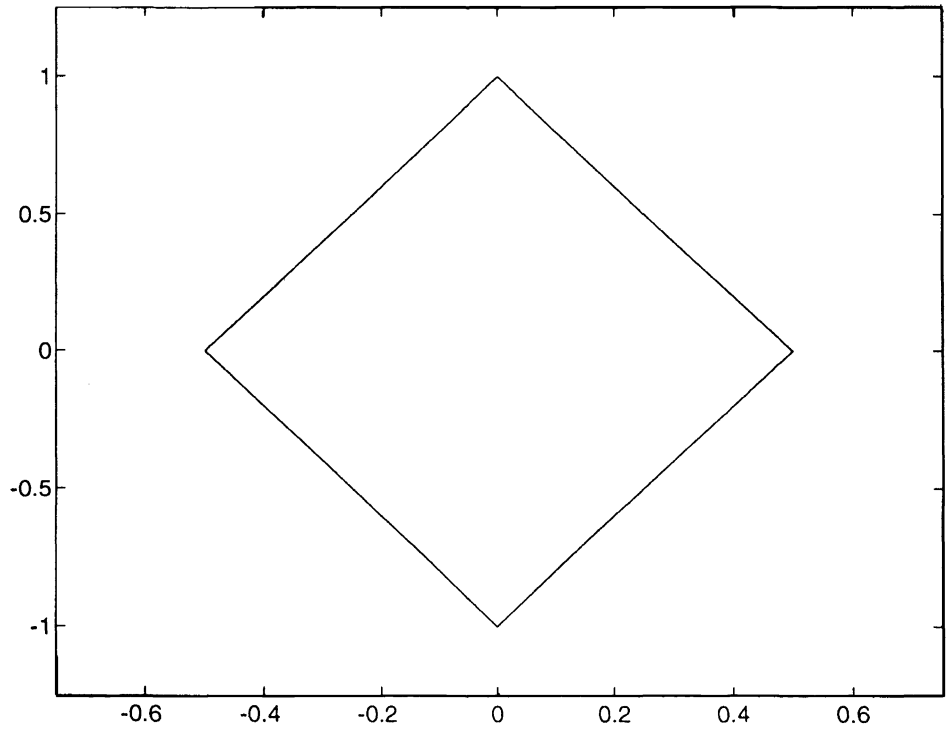


FIG. 2. *Polygonal domain* Ω *as in* (7.1) *with* $\lambda = 2$ *and* $\mu = 1$.

solves (1.6) for all $t > 0$. This shows that Theorem 1.2 does not hold in this domain Ω for any $T > 0$.

On the other hand, $\varphi = (0, \psi)$ is a nontrivial solution of (1.3) that satisfies (1.5) for any $T > 0$. Thus Theorem 1.1 does not hold either.

In a similar way, the function

$$\phi(x) = \cos\left(\frac{2\pi\sqrt{\lambda + 2\mu}}{\sqrt{\mu}} x_1\right) - \cos(2\pi x_2)$$

solves the elliptic eigenvalue problem (1.8) in this domain for

$$\kappa^2 = 4\pi^2(\lambda + 2\mu).$$

This shows that Theorem 1.3 also fails.

7.2. Dimension $n = 3$. In this section we show how the example above can be extended to dimension $n = 3$.

Let us denote by Θ a planar bounded and piecewise smooth domain of $\mathbb{R}^2$. Let $p = p(x_1, x_2)$ be the first eigenfunction of $-\Delta'$ in $H_0^1(\Theta)$, i.e.,

$$\begin{cases} -\mu\Delta' p = \kappa^2 p & \text{in } \Theta, \\ p = 0 & \text{on } \partial\Theta, \end{cases}$$

κ^2 being the first eigenvalue. Since p is bounded, we can assume that $\|p\|_{L^\infty(\Theta)} = 1$. Note that $p > 0$ in Θ.

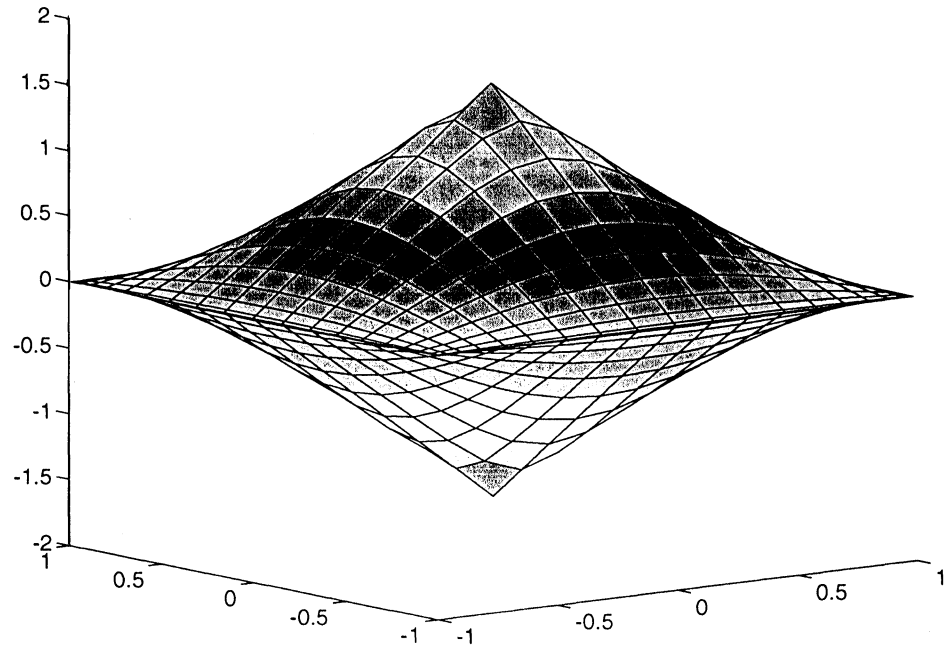


FIG. 3. *Domain* Ω *as in* (7.2) *when* $\Theta = (-1, 1) \times (-1, 1)$ *with* λ *and* μ *such that* $\lambda + 2\mu = \kappa^2$.

Let us consider the function

$$\phi(x_1, x_2, x_3) = p(x_1, x_2) - \sin\Big(\frac{\kappa}{\sqrt{\lambda + 2\mu}} x_3\Big).$$

It is easy to check that ϕ satisfies

$$-\mu\Delta\phi - (\lambda + \mu)\frac{\partial^2\phi}{\partial x_3^2} = \kappa^2\phi \quad \text{in } \Theta \times \mathbb{R}.$$

Let us consider now the surfaces S_1 and S_2 in $\mathbb{R}^3$ corresponding to the graphs of the folowing two functions:

$$x_3 = \frac{\sqrt{\lambda + 2\mu}}{\kappa}\arcsin(p(x_1, x_2)) \quad \text{and} \quad x_3 = -\frac{\sqrt{\lambda + 2\mu}}{\kappa}\arcsin(p(x_1, x_2)). \tag{7.2}$$

These two surfaces define the upper and lower boundaries of a piecewise smooth bounded domain Ω of $\mathbb{R}^3$ which is contained in $\Theta \times (-\pi/2, \pi/2)$. By definition $\phi = 0$ on $\partial\Omega$. Therefore Theorem 1.3 does not hold in this domain.

This type of set Ω reproduces the structure of the polygonal domain given in §7.1.

In these domains Ω the hypothesis c) of Theorem 1.8 fails since there is no point of the boundary of Ω where the tangent plane is orthogonal to the plane $x_3 = 0$. Note that, at the points of the boundary of Ω where $x_3 = 0$, the tangent plane does not exist since the tangent planes of the upper and lower surfaces S_1 and S_2 do not match.

In Figures 3 and 4 we see the domain Ω when Θ is an square and a disk of $\mathbb{R}^2$.

Note that when Θ is smooth, the corresponding domain Ω is Lipschitz. This can be proven easily by using the fact that the normal derivative of p at $\partial\Theta$ is strictly negative. However, as is seen in Figure 3, when Θ is the square, the corresponding domain Ω is not Lipschitz. This is due to the fact that the derivative of the eigenfunction p in the direction of a diagonal of Θ vanishes at the corner.

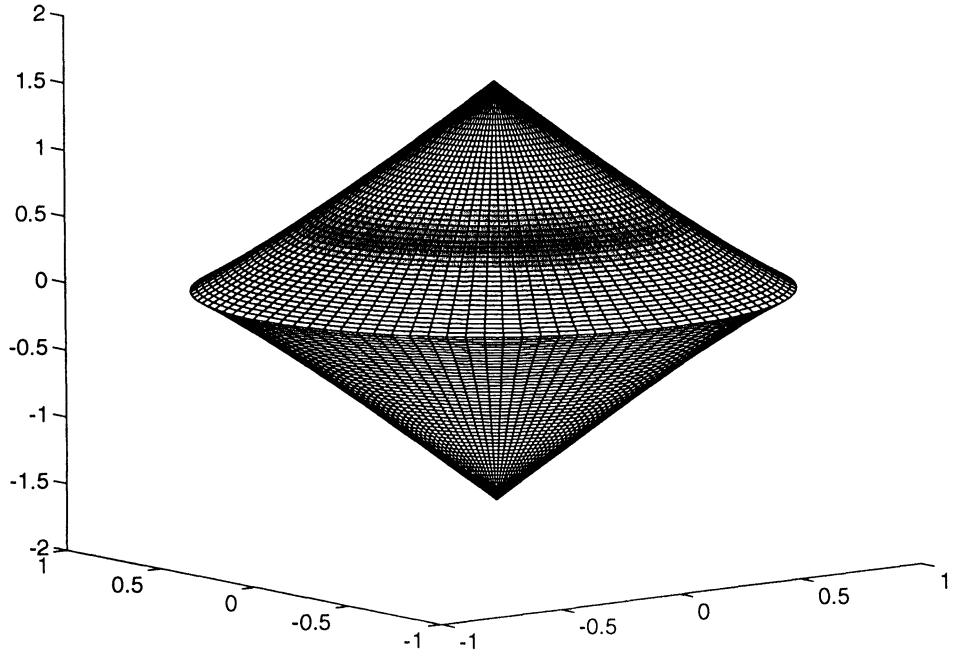


FIG. 4. *Domain Ω as in (7.2) when Θ is the unit disk of $\mathbb{R}^2$ with λ and μ such that $\lambda + 2\mu = \kappa^2$.*

8. Proof of Theorem 1.9. We prove Theorem 1.9 following the approach of Lions [5] and Fabre, Puel, and Zuazua [1].

First we observe that it is sufficient to consider the case where $u^0 \equiv u^1 \equiv 0$. Indeed, given any initial and final data (u^0, u^1), $(v^0, v^1) \in H$ and any $\varepsilon > 0$, let w be the solution of (1.1) with initial data (u^0, u^1) and right-hand side $f = 0$. Then, set $\hat{u} = u - w$. It is easy to check that finding $f \in (L^2(\Omega \times (0, T)))^n$ such that (1.10) holds is equivalent to finding $f \in (L^2(\Omega \times (0, T)))^n$ so that the solution $\hat{u}$ of (1.1) with this control and zero initial data satisfies

$$\left(\|\hat{u}(T) - v^0 + w(T)\|^2_{(H_0^1(\Omega))^n} + \|\hat{u}_t(T) - v^1 + w_t(T)\|^2_{(L^2(\Omega))^n}\right)^{1/2} \leq \varepsilon.$$

Therefore, in the sequel we will assume that $u^0 \equiv u^1 \equiv 0$.

Given any $(v^0, v^1) \in H$ and $\varepsilon > 0$ we introduce the functional $J : \mathcal{H} = (L^2(\Omega))^n \times (H^{-1}(\Omega))^n \to \mathbb{R}$, defined as follows:

$$J(\varphi^0, \varphi^1) = \frac{1}{2}\int_0^T \int_\omega \left(|\varphi_1|^2 + \cdots + |\varphi_{n-1}|^2\right) dx dt - \int_\Omega v^1 \cdot \varphi^0 dx + \langle v^0, \varphi^1 \rangle$$
$$+\varepsilon \|(\varphi^0, \varphi^1)\|_{(L^2(\Omega))^n \times (H^{-1}(\Omega))^n}, \tag{8.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality product between $(H_0^1(\Omega))^n$ and $(H^{-1}(\Omega))^n$ and φ is the solution of

$$\begin{cases} \varphi_{tt} - \mu\Delta\varphi - (\lambda + \mu)\nabla \mathrm{div}\varphi = 0 & \text{in } \Omega \times (0, T), \\ \varphi = 0 & \text{on } \Gamma \times (0, T), \\ \varphi(x, T) = \varphi^0(x), \ \varphi_t(x, T) = \varphi^1(x) & \text{in } \Omega. \end{cases} \tag{8.2}$$

The functional J is coercive in $\mathcal{H}$. More precisely, we have the following result.

LEMMA 8.1. *Under the assumptions of Theorem* 1.8,

$$\liminf_{\|(\varphi^0,\varphi^1)\|_{\mathcal{H}}\to\infty} \frac{J(\varphi^0,\varphi^1)}{\|(\varphi^0,\varphi^1)\|_{\mathcal{H}}} \geq \varepsilon. \tag{8.3}$$

Proof. Let us consider a sequence $(\varphi_j^0, \varphi_j^1)$ in $\mathcal{H}$ such that

$$N_j = \|(\varphi_j^0, \varphi_j^1)\|_{\mathcal{H}} \to \infty \text{ as } j \to \infty.$$

We introduce the normalized initial data

$$(\hat{\varphi}_j^0, \hat{\varphi}_j^1) = (\varphi_j^0, \varphi_j^1)/N_j$$

and the corresponding solutions of (8.2):

$$\hat{\varphi}_j = \varphi_j/N_j.$$

We have

$$\frac{I_j}{N_j} = \frac{J(\varphi_j^0,\varphi_j^1)}{N_j} = \frac{N_j}{2}\int_\omega \left(|\hat{\varphi}_{j,1}|^2 + \cdots + |\hat{\varphi}_{j,n-1}|^2\right) dxdt - \int_\Omega v^1.\hat{\varphi}_j^0 dx + \langle v^0, \hat{\varphi}_j^1\rangle + \varepsilon.$$

We distinguish the following two cases:

(i)

$$\liminf_{j\to\infty} \int_0^T \int_\omega \left(|\hat{\varphi}_{j,1}|^2 + \cdots + |\hat{\varphi}_{j,n-1}|^2\right) dxdt > 0.$$

(ii) there exists a subsequence (denoted by the index j to simplify the notation) such that

$$\int_0^T \int_\omega \left(|\hat{\varphi}_{j,1}|^2 + \cdots + |\hat{\varphi}_{j,n-1}|^2\right) dxdt \to 0 \text{ as } j \to \infty. \tag{8.4}$$

In the first case we clearly have

$$\liminf_{j\to\infty} I_j/N_j = \infty.$$

Let us consider the second case. Since $(\hat{\varphi}_j^0, \hat{\varphi}_j^1)$ is bounded in $\mathcal{H}$, we can extract a subsequence (still denoted by the index j) such that $(\hat{\varphi}_j^0, \hat{\varphi}_j^1) \to (\hat{\varphi}^0, \hat{\varphi}^1)$ weakly in $\mathcal{H}$ as $j \to \infty$. Let us denote by $\hat{\varphi}$ the corresponding solutions of (8.2). In view of (8.4) we have

$$\hat{\varphi}_1 \equiv \cdots \equiv \hat{\varphi}_{n-1} \equiv 0 \text{ in } \omega \times (0, T)$$

and therefore, as a consequence of Theorem 1.8, $(\hat{\varphi}^0, \hat{\varphi}^1) \equiv 0$. Thus

$$(\hat{\varphi}_j^0, \hat{\varphi}_j^1) \to (0, 0) \text{ weakly in } \mathcal{H}. \tag{8.5}$$

From (8.5) we deduce that

$$\liminf_{j\to\infty} \frac{I_j}{N_j} = \liminf_{j\to\infty} \frac{N_j}{2}\int_0^T \int_\omega \left(|\hat{\varphi}_{j,1}|^2 + \cdots + |\hat{\varphi}_{j,n-1}|^2\right) dxdt + \varepsilon \geq \varepsilon. \tag{8.6}$$

Therefore, (8.3) holds. □

As a consequence of the coercivity property (8.3), it is easy to check that the infimum of J over $\mathcal{H}$ is achieved at some $(\hat{\varphi}^0, \hat{\varphi}^1) \in \mathcal{H}$. At this minimizer $(\hat{\varphi}^0, \hat{\varphi}^1)$ the optimality conditions

$$(8.7)\quad \left| \int_0^T \int_\omega \left(\hat{\varphi}_1\rho_1 + \cdots + \hat{\varphi}_{n-1}\rho_{n-1}\right) dxdt - \int_\Omega v^1 \cdot \rho^0 dx + \langle v^0, \rho^1 \rangle \right| \leq \varepsilon \|(\rho^0, \rho^1)\|_{\mathcal{H}}$$

are satisfied for all $(\rho^0, \rho^1) \in \mathcal{H}$, where $\hat{\varphi}$ denotes the solution of (8.2) corresponding to the minimizer $(\hat{\varphi}^0, \hat{\varphi}^1)$ and ρ is the solution of (8.2) with data (ρ^0, ρ^1).

Observe that if u solves (1.1) with $u^0 \equiv u^1 \equiv 0$ and $f = \hat{\varphi}$, then

$$(8.8)\quad \int_0^T \int_\omega \left(\hat{\varphi}_1\rho_1 + \cdots + \hat{\varphi}_{n-1}\rho_{n-1}\right) dxdt = \int_\Omega u_t(T) \cdot \rho^0 dx - \langle u(T), \rho^1 \rangle.$$

Combining (8.7) and (8.8) we obtain

$$\left| -\int_\Omega (v^1 - u_t(T)) \cdot \rho^0 \, dx + \langle v^0 - u(T), \rho^1 \rangle \right| \leq \varepsilon \|(\rho^0, \rho^1)\|_{\mathcal{H}}$$

for all $(\rho^0, \rho^1) \in \mathcal{H}$, and this is equivalent to

$$\|(u(T) - v^0, u_t(T) - v^1)\|_H \leq \varepsilon.$$

9. Two examples of noncontrollability. In this section we develop the examples mentioned in Remarks 1.10 and 1.11 that can be constructed explicitly for periodic boundary conditions. The first one shows that, in general, exact controllability does not hold with L^2-controls obeying (1.2) and periodic boundary conditions. We do not have a counterexample for Dirichlet boundary conditions, but the situation is probably the same in this respect. The second example shows that if we impose further restrictions on the control, the approximate controllability may be lost. However, this result may depend on the type of boundary conditions and cannot be considered a serious indication for Dirichlet boundary conditions.

Throughout this section we assume that the Dirichlet boundary conditions are replaced by the periodic ones.

9.1. An example of nonexact controllability. The exact controllability with controls f in $(L^2(\Omega \times (0, T)))^n$ obeying (1.2) is equivalent to the following estimate for the adjoint system (1.3) (see [4]):

$$(9.1)\quad \|\varphi(0)\|^2_{(L^2(\Omega))^n} + \|\varphi_t(0)\|^2_{(H^{-1}(\Omega))^n} \leq C \int_0^T \int_\omega \left[|\varphi_1|^2 + \cdots + |\varphi_{n-1}|^2\right] dxdt.$$

Let us prove that, in general, (9.1) does not hold even if $\omega = \Omega$.

Let us assume that Ω is the three-dimensional cube $\Omega = (0, \pi)^3$. Let us consider divergence-free solutions of the form

$$\varphi(x, t) = c \sin(k \cdot x) \cos(\sqrt{\mu}\,|k|\,t)$$

with $c \in \mathbb{R}^n$ and $k \in \mathbf{Z}^n$ such that $c \cdot k = 0$. It is easy to check that φ satisfies (1.3). Actually, since $c \cdot k = 0$, we have $\operatorname{div}\varphi = 0$. Therefore, (1.3) reduces to the diagonal system of wave equations

$$\varphi_{tt} - \mu\Delta\varphi = 0.$$

Let us now analyze inequality (9.1) in the case where $\omega = \Omega$. It is a simple computation to check that the energy of the initial data (i.e., the left-hand side of (9.1)) equals $|c|^2/2$. The

integral on the right-hand side, which measures the energy of the first two components of the solution, equals

$$\frac{|c'|^2}{2}\left(T+\frac{\sin(\sqrt{\mu}\,|k|\,T)}{\sqrt{\mu}\,|k|}\right),$$

where $c'=(c_1,c_2)$. For any $T>0$ the quantity

$$\frac{|c|^2}{|c'|^2}\left(T+\frac{\sin(\sqrt{\mu}\,|k|\,T)}{\sqrt{\mu}\,|k|}\right)^{-1}=\left(1+\frac{c_3^2}{|c'|^2}\right)\left(T+\frac{\sin(\sqrt{\mu}\,|k|\,T)}{\sqrt{\mu}\,|k|}\right)^{-1}$$

(which is the quotient between the left- and right-hand sides of (9.1)) may be arbitrarily large. Indeed, since $c\cdot k=0$, we have $c_3=-c'\cdot k'/k_3$ with $k'=(k_1,k_2)$. Thus,

$$1+\frac{c_3^2}{|c'|^2}=1+\frac{|c'\cdot k'|^2}{|c'|^2\,k_3^2}.$$

We see that given $k_3\in\mathbf{Z}$ and $c'\in\mathbb{R}^2$, we may choose $k'\in\mathbf{Z}^2$ so that this quantity becomes arbitrarily large. On the other hand,

$$\left(T+\frac{\sin(\sqrt{\mu}\,|k|\,T)}{\sqrt{\mu}\,|k|}\right)\to T\quad\text{as } |k|\to\infty.$$

Thus (9.1) does not hold. □

Let us recall that if all the components of the control are nonzero, exact controllability with $(L^2(\Omega\times(0,T)))^3$ controls holds for a certain class of ω's. For instance, if ω is a neighborhood of the boundary of Ω, exact controllability holds with control time $T(\Omega)=\operatorname{diam}(\Omega\setminus\omega)/\sqrt{\mu}$ (see [4]).

9.2. An example of nonapproximate controllability under further constraints on the control. In view of Theorem 1.9 it is natural to study if approximate controllability holds when two components of the control are constrained to be identically zero. Thus let us suppose that the control function f is of the form

$$f=(f_1,\dots,f_{n-2},0,0). \tag{9.2}$$

This problem can be formulated in an equivalent way in terms of a uniqueness property for the adjoint system (1.3). This time, the uniqueness property to be understood is the following: If φ is a solution of (1.3) in the class $C([0,T];(L^2(\Omega))^n)\cap C^1([0,T];(H^{-1}(\Omega))^n)$ such that

$$\varphi_1\equiv\dots\equiv\varphi_{n-2}\equiv 0 \text{ in } \omega\times(0,T) \tag{9.3}$$

and T is large enough, can we deduce that, necessarily, $\varphi\equiv 0$ in $\Omega\times(0,T)$? It is easy to check that in dimension $n=3$ when $\Omega=(0,\pi)^3$ this uniqueness result does not hold even when $\omega=\Omega$.

Indeed, we can construct solutions of the form

$$\varphi(x,t)=c\sin(k\cdot x)\cos(\sqrt{\mu}\mid k\mid t)$$

with $c\in\mathbb{R}^3$, $k\in\mathbf{Z}^3$, $k\neq 0$ such that $c\cdot k=0$, $c_1=0$, $c'\neq 0$ so that $\varphi_1\equiv 0$ everywhere but φ is nontrivial. Thus we cannot expect approximate controllability results in three space dimensions with controls of the form $f=(f_1,0,0)$. □

Acknowledgments. The author acknowledges Caroline Fabre for a fruitful discussion on the proof of Theorem 1.3.

REFERENCES

[1] C. FABRE, J. P. PUEL, AND E. ZUAZUA, *Contrôlabilité approchée de l'équation de la chaleur semilinéaire*, C. R. Acad. Sci. Paris, Sér. I Math., 315 (1992), pp. 807–812.

[2] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1983.

[3] F. JOHN, *Partial Differential Equations*, 4th ed., Appl. Math. Sci. 1, Springer-Verlag, New York, 1986.

[4] J. L. LIONS, *Contrôlabilité exacte, stabilisation et perturbations de systèmes distribués*, Tome 1, *Contrôlabilité exacte*, Rech. Math. Appl. 8, Masson, Paris, 1988.

[5] ———, *Remarks on approximate controllability*, J. Anal. Math., 59 (1992), pp. 103–116.

[6] E. ZUAZUA, *Contrôlabilité approchée du système de l'élasticité avec contraintes sur les contrôles*, C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 61–66.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1496–1511, September 1996

002

NEW NECESSARY CONDITIONS FOR THE GENERALIZED PROBLEM OF BOLZA*

P. D. LOEWEN† AND R. T. ROCKAFELLAR‡

Abstract. Problems of optimal control are considered in the neoclassical Bolza format, which centers on states and velocities and relies on nonsmooth analysis. Subgradient versions of the Euler–Lagrange equation and the Hamiltonian equation are shown to be necessary for the optimality of a trajectory, moreover in a newly sharpened form that makes these conditions equivalent to each other. At the same time, the assumptions on the Lagrangian integrand are weakened substantially over what has been required previously in obtaining such conditions.

Key words. optimal control, calculus of variations, nonsmooth analysis, problem of Bolza, Euler–Lagrange condition, Hamiltonian condition, transversality condition

AMS subject classifications. 49K15, 49K05, 49K24

1. Introduction. Among the classical problems in the calculus of variations, that of Bolza marked a high point of complication, involving all the kinds of side conditions then viewed as important. With deceptive simplicity, the *generalized* problem of Bolza can be stated in one line:

$$(\mathcal{P}) \qquad \text{minimize } \Lambda[x] := l(x(a), x(b)) + \int_a^b L(t, x(t), \dot{x}(t))\, dt,$$

where the minimization takes place over all absolutely continuous functions ("arcs") $x\colon [a, b] \to \mathbb{R}^n$. Its generality rests on allowing l and L to be extended real valued, hence not necessarily differentiable or even continuous.

The tactic of admitting such a broad range of choices for l and L, first adopted in Rockafellar [21], enables $(\mathcal{P})$ to encompass a vast array of dynamic optimization problems, including those governed by controlled differential equations, differential inclusions, and incorporating endpoint constraints of every conceivable form. For example, $(\mathcal{P})$ subsumes the problem

$$(\mathcal{P}_1) \qquad \begin{aligned} &\text{minimize } \Lambda_1[x] := l_1(x(a), x(b)) + \int_a^b L_1(t, x(t), \dot{x}(t))\, dt \\ &\text{subject to } (x(a), x(b)) \in S \text{ and } \dot{x}(t) \in F(t, x(t)) \text{ a.e. } t \in [a, b] \end{aligned}$$

for a set $S \subset \mathbb{R}^n \times \mathbb{R}^n$ and a multifunction $F : [a, b] \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. Indeed, it suffices to take $l = l_1 + \Psi_S$ and $L = L_1 + \Psi_{\text{gph}\, F}$, where Ψ_S and $\Psi_{\text{gph}\, F}$ are the indicators of S and the graph of F (having the value 0 on these sets but ∞ outside). In the classical problem of Bolza, S and the graph of F were specified by side conditions of the kind $l_i(x(a), x(b)) = 0$ and $L_j(t, x(t), \dot{x}(t)) = 0$, with i and j in given finite index sets, all functions being assumed smooth (cf. Bliss [2, p. 189]); eventually the equations were supplemented by inequalities, and "isoperimetric" constraints were listed too (cf. Hestenes [8, p. 348]). Isoperimetric constraints fit into $(\mathcal{P}_1)$ by the trick of adding more state variables and modifying S and F accordingly. (In these classical formulations the interval $[a, b]$ was permitted to vary, and this could be built into $(\mathcal{P}_1)$ and $(\mathcal{P})$ as well, but we focus on the fixed-interval case here, reserving the variable-interval extension for elsewhere.)

*Received by the editors October 24, 1994; accepted for publication (in revised form) March 28, 1995.

†Department of Mathematics, University of British Columbia, Vancouver, BC V6T 1Z2, Canada (loew@math.ubc.ca). The research of this author was supported by the Natural Science and Engineering Research Council of Canada.

‡Department of Mathematics, University of Washington, Seattle, WA 98195 (rtr@math.uwashington.edu). The research of this author was supported by National Science Foundation grant DMS-9200303.

On the other hand, problems in optimal control of the wide form below can also be fitted into the pattern of $(\mathcal{P})$:

$$(\mathcal{P}_C)\qquad \begin{aligned} &\text{minimize } \Phi[x,u] := \phi(x(a),x(b)) + \int_a^b f(t,x(t),u(t))\,dt \\ &\text{subject to } \dot{x}(t) \in F(t,x(t),u(t)),\ u(t) \in U(t,x(t)) \text{ a.e. } t \in [a,b], \\ &\qquad \text{and } (x(a),x(b)) \in S. \end{aligned}$$

To arrange this, simply take $l = \phi + \Psi_S$ as before and

$$L(t,x,v) = \inf_u \{f(t,x,u) \,:\, u \in U(t,x),\ v \in F(t,x,u)\},$$

interpreting the right side as ∞ when there is no u in $U(t,x)$ for which $F(t,x,u)$ contains v. Note that the dynamics here involve a controlled differential inclusion and that the set of admissible controls displays explicit state dependence—two features beyond the scope of the classical theory. It is more difficult to force $(\mathcal{P}_C)$ into the framework of $(\mathcal{P}_1)$, which underscores the importance of $(\mathcal{P})$ as the model of choice when a full spectrum of control applications is envisioned. For more on this approach to optimal control, see [24] and [29].

Our aim is to establish necessary conditions for optimality in $(\mathcal{P})$ that retain both the form and the power of their classical precursors, the equations of Euler–Lagrange and Hamilton, despite the nonsmooth, extended-real-valued setting. This program for the generalized problem of Bolza is not new: it began with Rockafellar's work in the case where both functions l and $L(t,\cdot,\cdot)$ are convex [21–23, 25], and it was greatly advanced beyond such full convexity by Clarke [3–6] and others. Most recently there have been contributions by Loewen and Rockafellar [13, 14], Mordukhovich [19], and Ioffe and Rockafellar [10].

The current work has two especially distinguishing features. First, it provides a sharpened version of the Hamiltonian optimality condition that is *equivalent* to the sharpened form of the Euler–Lagrange condition we introduced in [14]. Second, it assumes significantly less than before about the Lagrangian L; it does not demand that L have the form $L_1 + \Psi_{\text{gph}\,F}$ in which Lipschitz properties are expected of L_1 and F, as, for instance, in [13]. It does ask for the convexity of L in the velocity argument, in contrast to the recent papers [19] and [10], but those works are more restrictive in other respects and anyway concern the Euler–Lagrange condition only.

The convexity of L in the velocity argument is essential for the equivalence between the Euler–Lagrange condition and the Hamiltonian condition, whatever their versions. Indeed, aside from the classical case of a smooth function L, or the fully convex case where L is convex in the state and velocity arguments together and some other special cases covered by [30], results asserting the simultaneous necessity of both conditions were elusive. The best that could be claimed, in [14], was the existence of at least one adjoint arc for which both conditions in a certain form were satisfied. (Other adjoint arcs might fulfill just one of the two.)

The sharpened Euler–Lagrange condition that we use in relating an extremal arc $\overline{x}$ to an adjoint arc p asserts that

$$\dot{p}(t) \in \operatorname{co}\left\{v \,:\, (v,p(t)) \in \partial L(t,\overline{x}(t),\dot{\overline{x}}(t))\right\} \text{ a.e. } t \in [a,b]. \tag{1.1}$$

Here ∂ refers to the possibly nonconvex *limiting subgradient* set (see Loewen [12] for notation and terminology), known also under various other names: limiting proximal subgradient set in Clarke [6], approximate subdifferential in Ioffe [9], subdifferential in Mordukhovich [19], subgradient set in the general sense in Rockafellar [31]. (The subgradients are those of $L(t,\cdot,\cdot)$ with t fixed.) Under the hypotheses of this paper (see §2), the inclusion (1.1) implies that for almost all t the vector $\dot{\overline{x}}(t)$ maximizes the function $v \mapsto \langle p(t)\,,\, v\rangle - L(t,\overline{x}(t),v)$.

The sharpened Hamiltonian condition that we establish for the first time as necessary for optimality, by virtue of its equivalence to (1.1), is

$$\dot{p}(t) \in \operatorname{co}\left\{w \,:\, (-w, \dot{\overline{x}}(t)) \in \partial H(t, \overline{x}(t), p(t))\right\} \text{ a.e. } t \in [a, b]. \tag{1.2}$$

The Hamiltonian H is, as usual, the Legendre–Fenchel transform of the Lagrangian L in its velocity variable:

$$H(t, x, p) := \sup\left\{\langle p\,,\, v\rangle - L(t, x, v) \,:\, v \in \mathbb{R}^n\right\}.$$

Clearly, (1.2) is a strict improvement on the form $(-\dot{p}(t), \dot{\overline{x}}(t)) \in \operatorname{co} \partial H(t, \overline{x}(t), p(t))$, taken as standard until now, since it convexifies only in the first argument. It implies, in particular, that for almost all t, the vector $p(t)$ maximizes the function $q \mapsto \langle q\,,\, \dot{\overline{x}}(t)\rangle - H(t, \overline{x}(t), q)$.

The weakened assumptions on L that suffice for these developments are set out in hypotheses (H4) and (H5) of §2. The first of these is a very mild "epi-continuity" assumption. Geometrically it amounts to insisting that, for each fixed t, the set epi $L(t, x, \cdot)$ should vary continuously with x. The second is a growth condition on subgradients, reducing when $L(t, x, v)$ is smooth to a local inequality of the form $|\nabla_x L| \le \kappa\big(1 + |\nabla_v L|\big)$. It implies, through a result of Mordukhovich [18], the Aubin ("pseudo-Lipschitz") continuity of the multifunction $x \rightrightarrows$ epi $L(t, x, \cdot)$ near the optimal arc. Our need for Aubin continuity on a tube of uniform size around the minimizing trajectory makes it necessary to formulate a quantitative generalization of Mordukhovich's result in §4.

We give special attention in §7 to the Lipschitz-plus-indicator case where $L = L_1 + \Psi_{\text{gph}\, F}$, showing for that version of the problem that the present results yield a full suite of (sharpened) Lagrangian and Hamiltonian necessary conditions for optimality in both normal and abnormal forms beyond what we had previously obtained in [13] and [14]. This recalls the work of Smirnov [32], who proposed the version of (1.1) for $L = \Psi_{\text{gph}\, F}$ as a necessary condition in 1991 but whose requirements that F be bounded and autonomous are significantly relaxed here. (Smirnov's result and proof are linked to prior work of Mordukhovich [15–17], who has recently given conditions [19] under which the necessity of (1.1) can be established in the absence of convexity hypotheses.) The main thrust of our effort, however, goes the other way: we demonstrate how to transform $(\mathcal{P})$ in its full generality into an instance of the differential inclusion problem in [14], and with some new machinery we then apply the results in that paper in combination with the Lagrangian–Hamiltonian equivalence theorem in [31].

State constraints requiring $x(t)$ to belong to a set $X(t) \subset \mathbb{R}^n$ can in principle be incorporated into problem $(\mathcal{P})$ by adding an indicator term in the specification of L, but for technical reasons it is better, at least in the theory as it now stands, to keep them explicit. The treatment of such constraints is taken up in §6.

2. The main result. Our main result is Theorem 2.1, a set of necessary conditions for an arc $\overline{x}$ to provide a local minimum in problem $(\mathcal{P})$. So let $\overline{x}$ be given, and fix some $\varepsilon > 0$ in order to define a suitable neighbourhood of $\overline{x}$:

$$\Omega := \{(t, x) \,:\, t \in [a, b],\ |x - \overline{x}(t)| < \varepsilon\},$$
$$\Omega_t := \{x \,:\, |x - \overline{x}(t)| < \varepsilon\}, \quad a \le t \le b.$$

We impose five conditions on $\overline{x}$ and the functions l and L relative to the set Ω; they are described below as (H1)–(H5). For simplicity in dealing with subgradients of $L(t, x, v)$ and $H(t, x, p)$ we use the notation ∂L and ∂H instead of the more cumbersome (but precise) $\partial_{(x,v)} L$ and $\partial_{(x,p)} H$. In general, as already mentioned, we write $\partial f(z)$ for the set of limiting subgradients associated with a lower semicontinuous function f at the point z; the singular counterpart to this set is $\partial^\infty f(z)$. See Loewen [12] for details.

THEOREM 2.1. *Assume* (H1)–(H5). *Suppose that for every arc x with graph in Ω, one has $\Lambda[x] \geq \Lambda[\overline{x}]$. Then either the normal conditions or the degenerate conditions given below are valid.*

Normal conditions: For some arc p on $[a, b]$,

(a) $\dot{p}(t) \in \operatorname{co}\{v : (v, p(t)) \in \partial L(t, \overline{x}(t), \dot{\overline{x}}(t))\}$ *a.e.* $t \in [a, b]$,

(b) $(p(a), -p(b)) \in \partial l(\overline{x}(a), \overline{x}(b))$.

Degenerate conditions: For some nonzero arc p on $[a, b]$,

(a^∞) $\dot{p}(t) \in \operatorname{co}\{v : (v, p(t)) \in \partial^\infty L(t, \overline{x}(t), \dot{\overline{x}}(t))\}$ *a.e.* $t \in [a, b]$,

(b^∞) $(p(a), -p(b)) \in \partial^\infty l(\overline{x}(a), \overline{x}(b))$.

(In particular, if the only arc p on $[a, b]$ satisfying conditions (a^∞)–(b^∞) *is the zero arc, then the normal conditions are satisfied.) In the normal conditions, assertion* (a) *is equivalent to*

(a′) $\dot{p}(t) \in \operatorname{co}\{w : (-w, \dot{\overline{x}}(t)) \in \partial H(t, \overline{x}(t), p(t))\}$ *a.e.* $t \in [a, b]$.

Also, conditions (a) *and* (a′) *imply that for almost all t in $[a, b]$,*

(c) $p(t) \in \partial_v L(t, \overline{x}(t), \dot{\overline{x}}(t)) = \operatorname{argmax}_{q \in \mathbb{R}^n} \{\langle q, \dot{\overline{x}}(t)\rangle - H(t, \overline{x}(t), q)\}$, *and*
$\dot{\overline{x}}(t) \in \partial_p H(t, \overline{x}(t), p(t)) = \operatorname{argmax}_{v \in \mathbb{R}^n} \{\langle p(t), v\rangle - L(t, \overline{x}(t), v)\}$.

Hypotheses. The terms in the Bolza functional Λ are required to have the following properties, expressed in terms of the constant $\varepsilon > 0$ in the definition of Ω and two positive-valued integrable functions δ and κ on $[a, b]$.

(H1) The endpoint cost function $l(x_a, x_b)$: $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is lower semicontinuous on $\Omega_a \times \Omega_b$.

(H2) The integrand $L(t, x, v)$: $\Omega \times \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is measurable with respect to the σ-field $\mathcal{L} \times \mathcal{B}$ generated by products of Lebesgue subsets of $[a, b]$ with Borel subsets of $\mathbb{R}^n \times \mathbb{R}^n$.

(H3) For each fixed pair (t, x) in Ω, the function $v \mapsto L(t, x, v)$ is convex.

(H4) For almost every t in $[a, b]$, the function $(x, v) \mapsto L(t, x, v)$ is lower semicontinuous on $\Omega_t \times \mathbb{R}^n$ and has the following epi-continuity property: for any point $(\widehat{x}, \widehat{v})$ where $|\widehat{x} - \overline{x}(t)| < \varepsilon$ and $L(t, \widehat{x}, \widehat{v})$ is finite, and for any sequence $x_k \to \widehat{x}$ in Ω_t, there exists a sequence $v_k \to \widehat{v}$ along which $L(t, x_k, v_k) \to L(t, \widehat{x}, \widehat{v})$.

(H5) The ratio $\kappa(t)/\delta(t)$ is essentially bounded. For almost all t in $[a, b]$, one has

$$|w| \leq \kappa(t)\,(1 + |p|) \text{ for all } (w, p) \in \partial L(t, x, v),$$

whenever $|x - \overline{x}(t)| < \varepsilon$, $\left|(v, L(t, x, v)) - (\dot{\overline{x}}(t), L(t, \overline{x}(t), \dot{\overline{x}}(t)))\right| < \delta(t)$.

As the absence of measures in the statement of Theorem 2.1 signals, state constraints cannot be implicit in the instance of $(\mathcal{P})$ under consideration at this stage. To see how such restrictions are ruled out, note that (H4) makes the set

$$G_t = \{x \in \Omega_t : L(t, x, v) < \infty \text{ for some } v \text{ with } |v - \dot{\overline{x}}(t)| < \delta(t)\}$$

be *open* for almost all t. Indeed, (H4) says that no point $\widehat{x}$ in G_t can be a boundary point. Of course, $\overline{x}(t)$ lies in G_t. Thus our paradigm allows L to impose certain velocity constraints through the use of infinite penalties but does not allow unilateral state constraints to be covered in the same way. State constraints in the explicit form $x(t) \in X(t)$ $\forall t$ can nonetheless be handled by our methods, as will be explained in §6.

Hypothesis (H5) can be viewed as a combined growth condition and Lipschitz condition. As a growth condition, it resembles the "condition of Morrey type" underlying Clarke and Vinter's Proposition 3.2 in [7], a result which establishes the validity of the Euler–Lagrange equation in the calculus of variations without the a priori assumption that the minimizing arc is Lipschitzian. Indeed, in the special case where $L(t, \cdot, \cdot)$ is continuously differentiable on Ω_t, (H5) reduces to

$$|\nabla_x L(t, x, v)| \leq \kappa(t)\big(1 + |\nabla_v L(t, x, v)|\big) \qquad \forall (x, v) \in \Omega_t.$$

As a Lipschitz condition, (H5) is a subgradient characterization of the Aubin continuity of the epigraphical multifunction associated with L, as we shall see in §4. A Hamiltonian formulation of this assumption is derived in §5, where its relationship to Clarke's "strong Lipschitz condition" [4] is easiest to discern.

Note that the seemingly weaker form of (H5) obtained by substituting the proximal subgradient set $\widehat{\partial} L$ for ∂L is actually equivalent to the form stated here, because ∂L is defined by taking limits of proximal subgradients.

The method of proof. There is a well-known equivalence between Bolza problems and Mayer problems, mediated by the technique of state augmentation. Indeed, consider the domain $\widetilde{\Omega} = \Omega \times \mathbb{R}$ in one more state dimension, the extra state variable being denoted by y, and define the epigraphical multifunction $E : \widetilde{\Omega} \rightrightarrows \mathbb{R}^{n+1}$ by

$$E(t, x, y) := \operatorname{epi} L(t, x, \cdot).$$

(The set $E(t, x, y)$ does not actually depend on y.) If the arc $\overline{x}$ figuring in our hypotheses solves $(\mathcal{P})$, then the arc $(\overline{x}, \overline{y})$, with

$$\overline{y}(t) := \int_a^t L(r, \overline{x}(r), \dot{\overline{x}}(r))\, dr,$$

solves the following differential inclusion problem:

$$(\mathcal{P}') \qquad \begin{array}{l} \text{minimize } k\big(x(a), y(a), x(b), y(b)\big) := l(x(a), x(b)) + y(b) + \Psi_{\{0\}}(y(a)) \\ \text{subject to } (\dot{x}(t), \dot{y}(t)) \in E(t, x(t), y(t)) \text{ a.e. } t \in [a, b]. \end{array}$$

The right side in the dynamic constraint here is unbounded. Necessary conditions for optimality in problems of this sort were the subject of a previous paper [14]. Our procedure in the current paper is basically to check the hypotheses in [14] and then to translate the conclusions of that work into the context of $(\mathcal{P})$. Checking the hypotheses takes a certain amount of work, since the transition from the subgradient hypothesis (H5) to the Lipschitz conditions required by [14] is not completely straightforward (see §4). Likewise, an additional state-augmentation argument is necessary to reduce the case of a general lower semicontinuous endpoint cost l to the Lipschitz-plus-indicator form treated in [14] (see the proof of Theorem 3.1). Finally, it is insufficient simply to transcribe the conclusions of [14]: the sharpened Hamiltonian inclusion featured here relies on a careful analysis of the relationship between the Hamiltonian and Eulerian forms of the necessary conditions, as carried out by Rockafellar [31].

3. Proof of the main result. To prove Theorem 2.1, we shall apply an intermediate result for unbounded differential inclusions which is readily derived from [14, Thm. 4.3]. The reformulated problem $(\mathcal{P}')$ under consideration fits the general pattern:

$$(\overline{\mathcal{P}}) \qquad \begin{array}{l} \text{minimize } k(z(a), z(b)) \\ \text{subject to } \dot{z}(t) \in E(t, z(t)) \text{ a.e. } t \in [a, b]. \end{array}$$

The hypotheses of [14] for this kind of problem refer to a distinguished arc $\overline{z}$ and, for some fixed $\eta > 0$, its "graphical neighbourhood"

$$U = \{(t, z) : t \in [a, b],\ |z - \overline{z}(t)| < \eta\},$$
$$U_t = \{z : |z - \overline{z}(t)| < \eta\}, \quad a \le t \le b.$$

They read as follows:

(h1) The endpoint cost function $k: U_a \times U_b \to \mathbb{R} \cup \{\infty\}$ is lower semicontinuous.

(h2) The sets $E(t, z)$ are nonempty, closed, and convex for each (t, z) in U and empty for each (t, z) outside U.

(h3) The multifunction E is measurable with respect to the σ-field $\mathcal{L} \times \mathcal{B}$ generated by products of Lebesgue subsets of $[a, b]$ with Borel subsets of $\mathbb{R}^m$.

(h4) There are integrable functions δ and K on $[a, b]$, with K/δ essentially bounded, such that almost every t in $[a, b]$ obeys

$$E(t, y) \cap (\dot{\overline{z}}(t) + \delta(t)\mathbb{B}) \subseteq E(t, z) + K(t)|y - z|\mathbb{B} \quad \forall y, z \in U_t.$$

Here and elsewhere in this paper, $\mathbb{B}$ denotes the closed unit ball in the Euclidean space of appropriate dimension.

THEOREM 3.1 (see [14]). *If hypotheses* (h1)–(h4) *hold and* $\overline{z}$ *solves problem* $(\overline{\mathcal{P}})$, *then there exists an arc* q *on* $[a, b]$ *such that*

(a) $\dot{q}(t) \in \operatorname{co}\left\{w : (w, q(t)) \in N_{\operatorname{gph} E(t,\cdot)}\left(\overline{z}(t), \dot{\overline{z}}(t)\right)\right\}$ *a.e.* $t \in [a, b]$; *and*

(b) *one of the following transversality conditions holds:*

(i) $(q(a), -q(b)) \in \partial k(\overline{z}(a), \overline{z}(b))$, *or*

(ii) $(q(a), -q(b)) \in \partial^\infty k(\overline{z}(a), \overline{z}(b))$, *with* q *not identically zero.*

Proof. A simple trick reduces the general lower semicontinuous endpoint cost function k to one in the Lipschitz-plus-indicator form analyzed in [14]. Indeed, it suffices to define the constant arc $\overline{r} = k(\overline{z}(a), \overline{z}(b))$ and then to observe that the pair $(\overline{z}, \overline{r})$ solves the problem

$$\begin{aligned} &\text{minimize } r(b) \\ &\text{subject to } (z(a), z(b), r(a)) \in \operatorname{epi} k, \quad r(b) \in \mathbb{R}, \\ &\qquad (\dot{z}(t), \dot{r}(t)) \in E(t, z(t)) \times \{0\} \text{ a.e. } t \in [a, b]. \end{aligned}$$

The stated result follows from conditions (b) and (d) of [14, Thm. 4.3] by elementary subgradient calculus.

Note that although the statement of [14, Thm. 4.3] involves stronger Lipschitz conditions on the multifunction E specifically tailored to the modulus of integrability of the function $\dot{\overline{z}}$, they are present only to facilitate a statement free of explicit references to the quantity $\dot{\overline{z}}(t)$. In fact, the conditions of [14, Prop. 2.2] are sufficient for the conclusions of [14, Thm. 4.3], and it is these we have applied here—taking $R = \delta$ and $m = K$. □

Leaving aside the verification of hypotheses (h1)–(h4) for now, let us derive the conclusions of Theorem 2.1 from those of Theorem 3.1. To apply the latter result, we take $m = n+1$, with $z = (x, y)$ as a pattern and $\overline{z} = (\overline{x}, \overline{y})$ as the optimal arc. Of course, $E(t, z) = E(t, x, y)$ and $k(x_a, y_a, x_b, y_b) = l(x_a, x_b) + y_b + \Psi_{\{0\}}(y_a)$. Observe that

$$\begin{aligned} \partial k(\overline{z}(a), \overline{z}(b)) &= \{(\alpha, \zeta, \beta, 1) : (\alpha, \beta) \in \partial l(\overline{x}(a), \overline{x}(b)),\ \zeta \in \mathbb{R}\}, \\ \partial^\infty k(\overline{z}(a), \overline{z}(b)) &= \left\{(\alpha, \zeta, \beta, 0) : (\alpha, \beta) \in \partial^\infty l(\overline{x}(a), \overline{x}(b)),\ \zeta \in \mathbb{R}\right\}. \end{aligned}$$

In terms of these data, Theorem 3.1 provides an adjoint arc (p, q): $[a, b] \to \mathbb{R}^n \times \mathbb{R}$ satisfying two conditions. First, the transversality condition 3.1(b) implies that either $q(b) = -1$ and $(p(a), -p(b)) \in \partial l(\overline{x}(a), \overline{x}(b))$ or $q(b) = 0$ and $(p(a), -p(b)) \in \partial^\infty l(\overline{x}(a), \overline{x}(b))$ with (p, q) not identically zero. Second, the Euler–Lagrange condition 3.1(a) asserts that for almost all t in $[a, b]$,

$$(\dot{p}(t), \dot{q}(t)) \in \operatorname{co}\left\{(v, w) : (v, w, p(t), q(t)) \in N_{\operatorname{gph} E(t,\cdot)}\left(\overline{x}(t), \overline{y}(t), \dot{\overline{x}}(t), \dot{\overline{y}}(t)\right)\right\}. \tag{3.1}$$

Now for fixed t, one has

$$\operatorname{gph} E(t, \cdot) = \{(x, y, v, r) : (x, v, r) \in \operatorname{epi} L(t, \cdot, \cdot),\ y \in \mathbb{R}\}.$$

In terms of $\overline{L}(t) = L(t, \overline{x}(t), \dot{\overline{x}}(t))$, this implies

$$\begin{aligned} &N_{\operatorname{gph} E(t,\cdot)}\left(\overline{x}(t), \overline{y}(t), \dot{\overline{x}}(t), \dot{\overline{y}}(t)\right) \\ &\qquad = \left\{(v, 0, \pi, -\rho) : (v, \pi, -\rho) \in N_{\operatorname{epi} L(t,\cdot,\cdot)}\left(\overline{x}(t), \dot{\overline{x}}(t), \overline{L}(t)\right)\right\}. \end{aligned}$$

Using this relation in (3.1), we get

$$(\dot{p}(t), \dot{q}(t)) \in \operatorname{co}\left\{v : (v, p(t), q(t)) \in N_{\operatorname{epi} L(t,\cdot,\cdot)}\left(\overline{x}(t), \dot{\overline{x}}(t), \overline{L}(t)\right)\right\} \times \{0\}. \tag{3.2}$$

In particular, the second component of this inclusion implies that $\dot{q}(t) = 0$ almost everywhere. Thus q is a constant function whose value is either 0 or -1. In the case where $q = -1$, one has the normal conditions of Theorem 2.1:

(a) $\dot{p}(t) \in \operatorname{co}\left\{v : (v, p(t)) \in \partial L(t, \overline{x}(t), \dot{\overline{x}}(t))\right\}$ a.e. $t \in [a, b]$, and

(b) $(p(a), -p(b)) \in \partial l(\overline{x}(a), \overline{x}(b))$.

In the case where $q = 0$, the degenerate conditions of Theorem 2.1 follow instead:

(a$^\infty$) $\dot{p}(t) \in \operatorname{co}\left\{v : (v, p(t)) \in \partial^\infty L(t, \overline{x}(t), \dot{\overline{x}}(t))\right\}$ a.e. $t \in [a, b]$, and

(b$^\infty$) $(p(a), -p(b)) \in \partial^\infty l(\overline{x}(a), \overline{x}(b))$, and p is not the zero arc.

In [31, Thm. 1.1], Rockafellar proves that the inclusions (a) and (a′) in Theorem 2.1 are equivalent for each fixed t with the properties described in (H4), provided that every such t also satisfies the condition

$$(w, 0) \in \partial^\infty L(t, \overline{x}(t), \dot{\overline{x}}(t)) \implies w = 0. \tag{3.3}$$

Note that (3.3) holds for almost all t, by (H5). Indeed, any point $(w, 0)$ in the cone $\partial^\infty L(t, \overline{x}(t), \dot{\overline{x}}(t))$ must have the form $(w, 0) = \lim_{\nu\to\infty} r_\nu(w_\nu, p_\nu)$ for sequences $r_\nu \to 0^+$ and $(w_\nu, p_\nu) \in \hat{\partial} L(t, x_\nu, v_\nu)$ along which $(x_\nu, v_\nu) \underset{L}{\to} (\overline{x}(t), \dot{\overline{x}}(t))$. Since $\hat{\partial} L \subseteq \partial L$ always, (H5) implies that for all ν sufficiently large,

$$|r_\nu w_\nu| \le \kappa\,(|r_\nu p_\nu| + r_\nu).$$

In the limit as $\nu \to \infty$, it follows that $|w| \le 0$, so (3.3) holds. Under the same hypotheses, Rockafellar [31] shows that the equivalent conditions (a) and (a′) both imply the argmax conditions in (c).

To complete the proof of Theorem 2.1, we must demonstrate that Theorem 3.1 is truly applicable—by checking hypotheses (h1)–(h4). Conditions (h1)–(h3) hold for any $\eta \in (0, \varepsilon]$ as obvious consequences of the corresponding hypotheses (H1)–(H3) on l and L. The real issue is hypothesis (h4), which calls for the Aubin continuity of the multifunction E (see Aubin [1]) with respect to a certain restricted tube around $\overline{z}$. This condition follows from hypothesis (H5) and Theorem 4.3 in the next section.

To see this, fix a time t in $[a, b]$ at which the conditions in (H4)–(H5) hold. Since t will be fixed throughout this argument, and since E does not actually depend on y, we suppress both the t- and y-dependence of E and L, writing simply $E(x) = \operatorname{epi} L(x, \cdot)$ for $|x - \overline{x}| < \varepsilon$. (We also write $\dot{\overline{x}}$ instead of $\dot{\overline{x}}(t)$ and use the shorthand $\overline{L} = L(\overline{x}, \dot{\overline{x}})$.) As noted above, $\operatorname{gph} E = \operatorname{epi} L$; thus (H4) implies that $\operatorname{gph} E$ is closed and that condition (i) of Theorem 4.3 holds. Condition (ii), on the other hand, requires that

$$|w| \le R|(p, -\lambda)| \text{ for all } (w, p, -\lambda) \in N_{\operatorname{epi} L}(x, v, r), \quad \text{whenever } |x - \overline{x}| < \varepsilon \text{ and } \left|(v, r) - (\dot{\overline{x}}, \overline{L})\right| < \delta.$$

Now the "proximal subgradient formula" [26, 11] asserts that every nonzero vector $(w, p, -\lambda)$ in $N_{\operatorname{epi} L}(x, v, r)$ can be realized as the limit of a sequence of proximal normals $(w_\nu, p_\nu, -\lambda_\nu)$ in $\hat{N}_{\operatorname{epi} L}(x_\nu, v_\nu, r_\nu)$ for which $\lambda_\nu > 0$ and the corresponding base points obey $(x_\nu, v_\nu, r_\nu) \underset{\operatorname{epi} L}{\to} (x, v, r)$. For every term in such a sequence, one has $(w_\nu/\lambda_\nu, p_\nu/\lambda_\nu) \in \hat{\partial} L(x_\nu, v_\nu)$, so (H5) gives

$$|w_\nu/\lambda_\nu| \le \kappa\,(1 + |p_\nu/\lambda_\nu|).$$

Multiplying through by $\lambda_\nu > 0$ and letting $\nu \to \infty$, we obtain

$$|w| \leq \kappa\,(|p| + |\lambda|) \leq 2\kappa|(p, -\lambda)|.$$

This argument applies to every triple $(w, p, -\lambda)$ in $N_{\text{epi}\,L}\,(x, v, r)$, so condition (ii) of Theorem 4.3 holds with $R = 2\kappa$. The conclusion of Theorem 4.3 establishes (h4), with $K(t) = \sqrt{1 + 4\kappa(t)^2}$, any constant $0 < \eta \leq \varepsilon_0(t) := \min\{\varepsilon, \delta(t)/(9K(t))\}$, and $\delta(t)$ equal to one sixth of the value of the $\delta(t)$ provided in (H5). (The function ε_0 is bounded away from zero because $\kappa/\delta \in L^\infty$ by (H5).) This completes the proof of Theorem 2.1.

4. On uniform Aubin continuity. This section and the next furnish technical support for the proof and interpretation of Theorem 2.1 above. Both are intended, though, to stand alone as having independent interest: they involve functions and constants whose names are deliberately suggestive but are logically distinct from those identified in §§1–3 and 6–7.

DEFINITION 4.1. *Let* $\Gamma\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ *be a multifunction and* $(\overline{x}, \overline{y})$ *be a point in* $\operatorname{gph}\Gamma$. *To say that* Γ *is Aubin continuous at* $(\overline{x}, \overline{y})$ *with parameters* $\varepsilon > 0$, $\delta > 0$, *and* $K > 0$ *means that one has*

$$\Gamma(x) \cap (\overline{y} + \delta\mathbb{B}) \subseteq \Gamma(y) + K|y - x|\mathbb{B} \qquad \forall x, y \in \overline{x} + \varepsilon\mathbb{B}. \tag{4.1}$$

The modulus of Aubin continuity for a given multifunction Γ at a point $(\overline{x}, \overline{y})$ in $\operatorname{gph}\Gamma$ is the number $\kappa_\Gamma(\overline{x}, \overline{y})$, defined as the infimum of all $K > 0$ satisfying (4.1) for some $\varepsilon > 0$ and $\delta > 0$. Mordukhovich [18, Thm. 5.7] has shown that if $\operatorname{gph}\Gamma$ is closed, then

$$\kappa_\Gamma(\overline{z}, \overline{y}) = \sup\left\{|\alpha| \,:\, (\alpha, \beta) \in N_{\operatorname{gph}\Gamma}\,(\overline{x}, \overline{y})\,,\ |\beta| \leq 1\right\}.$$

For the purposes of this paper, Aubin continuity with some fixed $\delta > 0$ is required at every point in some ε-neighbourhood of a given arc, and knowing only that κ_Γ is finite at every point along the arc is not sufficient. We need quantitative estimates of the constants ε, δ, and K in terms of a neighbourhood of $(\overline{x}, \overline{y})$ in which the (generalized) slope of vectors normal to the graph of Γ is bounded.

Our approach to this problem is patterned on that introduced in Rockafellar [28, Rem. 3.14]: we prove that Γ has the desired Aubin continuity properties by showing that the function d_Γ defined in the following lemma satisfies a corresponding Lipschitz continuity condition. This in turn is accomplished by using Rockafellar's 1985 results [27] for estimating the subgradients of marginal functions. (Although the facts we appropriate from these earlier papers were phrased in terms of Clarke subgradients, they apply equally well to the limiting subgradients we are working with here.)

LEMMA 4.2. *Given a multifunction* Γ *with closed graph* G, *consider* $d_\Gamma(x, v) := d_{\Gamma(x)}\,(v)$. *If* d_Γ *is Lipschitz of rank* K *on the set* $(\overline{x} + \varepsilon\mathbb{B}) \times (\overline{y} + \delta\mathbb{B})$ *for some constants* $\varepsilon > 0$, $\delta > 0$, *then condition* (4.1) *holds, with the same constants.*

Proof. This is elementary—see Rockafellar [28, Thm. 2.3, (b)⇒(a)]. □

THEOREM 4.3. *Let* $\Gamma\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ *be a multifunction; write* $G = \operatorname{gph}\Gamma$, *and assume that* G *is a closed set. Let* $(\overline{x}, \overline{y})$ *in* G *be a point for which some constants* $\varepsilon > 0$, $\delta > 0$, *and* $R > 0$ *satisfy two conditions*:

(i) *For any point* (x, y) *in* G *with* $|x - \overline{x}| < \varepsilon$ *and* $|y - \overline{y}| < \delta$, *and for any sequence* $x_k \to x$, *there is a sequence* $y_k \in \Gamma(x_k)$ *such that* $y_k \to y$.

(ii) $|\alpha| \leq R|\beta|$ *for all* $(\alpha, \beta) \in N_{\operatorname{gph}\Gamma}\,(x, y)$ *whenever* $|x - \overline{x}| < \varepsilon$ *and* $|y - \overline{y}| < \delta$.

Then Γ *is Aubin continuous at* $(\overline{x}, \overline{y})$ *with parameters* $K = \sqrt{1 + R^2}$, $\delta_0 = \delta/6$, *and* $\varepsilon_0 = \min\{\varepsilon, \delta/(9K)\}$, *i.e.,*

$$\Gamma(y) \cap (\overline{x} + \delta_0\mathbb{B}) \subseteq \Gamma(x) + K|y - x|\mathbb{B} \qquad \forall x, y \in \overline{x} + \varepsilon_0\mathbb{B}. \tag{4.2}$$

Proof. According to Lemma 4.2, we only have to show that d_Γ is Lipschitz continuous of rank K on the set $(\overline{x} + \varepsilon_0 \mathbb{B}) \times (\overline{\gamma} + \delta_0 \mathbb{B})$. To accomplish this, it suffices to show that $\partial d_\Gamma(x, v) \subseteq K\mathbb{B}$ for all (x, v) in this set. We therefore set out to estimate ∂d_Γ, relying on Rockafellar [27].

A convenient characterization of d_Γ, valid for all x and v without restriction, is

$$\begin{aligned} d_\Gamma(x, v) &= \min\{|v - \gamma| \,:\, (x, \gamma) \in G\} \\ &= \min\left\{ f(x, v, \gamma) \,:\, (x, v, \gamma) \in S \right\}, \end{aligned} \tag{4.3}$$

where $f(x, v, \gamma) = |x - \gamma|$ and $S = \{(x, v, \gamma) \,:\, (x, \gamma) \in G,\ v \in \mathbb{R}^m\}$. In the latter form, d_Γ is revealed as the marginal function associated with an optimization problem depending on parameters (x, v). Such functions have been studied extensively, in particular by Rockafellar [27], whose results we shall employ here. Let us denote by $\Sigma(x, v)$ the set of minimizing vectors γ in (4.3) above. Theorem 8.3 of [27] implies that any point (x, v) has a neighbourhood in which d_Γ is Lipschitzian and satisfies the following estimate for limiting subgradients:

$$\partial d_\Gamma(x, v) \ \subseteq\ \{(\xi, \eta) \,:\, (\xi, \eta, 0) \in \partial f(x, v, \gamma) + N_S(x, v, \gamma)\ \exists \gamma \in \Sigma(x, v)\}.$$

(The hypotheses of Rockafellar's result are easy to verify, because the function f here is Lipschitzian, so $\partial^\infty f \equiv \{0\}$, and because f grows rapidly enough to make the inf-compactness condition obvious.) We note that whenever $\gamma \in \Sigma(x, v)$,

$$\begin{aligned} \partial f(x, v, \gamma) &\subseteq \{(0, u, -u) \,:\, |u| \le 1\}, \\ N_S(x, v, \gamma) &= \left\{ (\alpha, 0, \beta) \,:\, (\alpha, \beta) \in N_G(x, \gamma) \right\}. \end{aligned}$$

(A sharper version of the first inclusion is possible, but this one is adequate for our purposes.) It follows that any point (ξ, η) in $\partial d_\Gamma(x, v)$ obeys

$$(\xi, \eta, 0) = (0, u, -u) + (\alpha, 0, \beta) \text{ for some } \gamma \in \Sigma(x, v),\ (\alpha, \beta) \in N_G(x, \gamma),\ u \in \mathbb{B}.$$

Thus

$$\partial d_\Gamma(x, v) \subseteq \{(\alpha, \beta) \in N_G(x, \gamma) \,:\, \gamma \in \Sigma(x, v),\ |\beta| \le 1\}.$$

For those points (x, v) where all the pairs (x, γ) with $\gamma \in \Sigma(x, v)$ lie in the set specified by hypothesis (ii), that condition implies that every pair (α, β) on the right side has $|\alpha| \le R$ and $|\beta| \le 1$, so $|(\alpha, \beta)|^2 \le (R^2 + 1) = K^2$. Thus

$$\{x\} \times \Sigma(x, v) \subseteq (\overline{x} + \varepsilon\mathbb{B}) \times (\overline{\gamma} + \delta\mathbb{B}) \implies \partial d_\Gamma(x, v) \subseteq K\mathbb{B}. \tag{4.4}$$

This reveals the key to the result: the location of the set $\Sigma(x, v)$.

CLAIM. *Fix $(\widehat{x}, \widehat{\gamma})$ in G with $|\widehat{x} - \overline{x}| < \varepsilon$, $|\widehat{\gamma} - \overline{\gamma}| < \delta$. Then for some $\mu > 0$, one has $\Sigma(x, v) \subseteq \overline{\gamma} + \delta\mathbb{B}$ whenever $|x - \widehat{x}| < \mu$, $|v - \widehat{\gamma}| < (\delta - |\widehat{\gamma} - \overline{\gamma}|)/3$.*

To prove this, suppose not: then there are sequences $x_k \to \widehat{x}$ and v_k for which

$$|v_k - \widehat{\gamma}| < \frac{\delta - |\widehat{\gamma} - \overline{\gamma}|}{3} \tag{4.5}$$

and yet $\Sigma(x_k, v_k)$ contains some point outside $\overline{\gamma} + \delta\mathbb{B}$. Call this point π_k. Then $|\pi_k - \overline{\gamma}| > \delta$ and consequently

$$\begin{aligned} d_{\Gamma(x_k)}(v_k) &= |(\pi_k - \overline{\gamma}) - (v_k - \overline{\gamma})| \\ &\ge |\pi_k - \overline{\gamma}| - |v_k - \overline{\gamma}| \\ &> \delta - |v_k - \widehat{\gamma}| - |\widehat{\gamma} - \overline{\gamma}|. \end{aligned} \tag{4.6}$$

But the semicontinuity property (i) provides a sequence $\gamma_k \in \Gamma(x_k)$ such that $\gamma_k \to \widehat{\gamma}$. For this sequence,

$$d_{\Gamma(x_k)}(v_k) \le |v_k - \gamma_k| \le |v_k - \widehat{\gamma}| + |\widehat{\gamma} - \gamma_k|. \tag{4.7}$$

Concatenating inequalities (4.6) and (4.7) and applying condition (4.5) yield

$$\begin{aligned} 2|v_k - \widehat{\gamma}| &\ge \delta - |\widehat{\gamma} - \overline{\gamma}| - |\widehat{\gamma} - \gamma_k| \\ &> 3|v_k - \widehat{\gamma}| - |\widehat{\gamma} - \gamma_k|, \end{aligned}$$

whence $|v_k - \widehat{\gamma}| < |\gamma_k - \widehat{\gamma}|$. In particular, $v_k \to \widehat{\gamma}$. Taking the limit in the previous inequality then gives $0 \ge \delta - |\widehat{\gamma} - \overline{\gamma}|$. This is a contradiction, since the right side here is positive by construction. The claim holds.

We apply the claim first to the point $(\widehat{x}, \widehat{\gamma}) = (\overline{x}, \overline{\gamma})$. In view of (4.4), this shows that ∂d_Γ is bounded by K throughout the interior of some set $(\overline{x} + \mu\mathbb{B}) \times (\overline{\gamma} + (\delta/3)\mathbb{B})$ but provides no information about the size of $\mu > 0$. To balance the need for quantitative information in both directions, consider

$$\widehat{\mu} = \sup\{\mu \in (0, \varepsilon) \,:\, \partial d_\Gamma(x, v) \subseteq K\mathbb{B} \;\forall x \in \overline{x} + \mu\mathbb{B},\; v \in \overline{\gamma} + \delta_0\mathbb{B}\},$$

where δ_0 is defined in the theorem statement. Note that for every (x, v) where $|x - \overline{x}| < \widehat{\mu}$ and $|v - \overline{\gamma}| \le \delta_0$, one has $\partial d_\Gamma(x, v) \subseteq K\mathbb{B}$. Thus d_Γ is Lipschitz continuous of rank K on the set just described. In particular, Γ is Aubin continuous there, and consequently every y with $|y - \overline{x}| < \widehat{\mu}$ obeys

$$\begin{aligned} \overline{\gamma} \in \Gamma(\overline{x}) \cap (\overline{\gamma} + \delta_0\mathbb{B}) \subseteq \Gamma(y) + K|y - \overline{x}|\mathbb{B}, \text{ i.e.,} \\ \Gamma(y) \cap (\overline{\gamma} + K|y - \overline{x}|\mathbb{B}) \ne \emptyset. \end{aligned} \tag{4.8}$$

The closed-graph property of Γ makes it elementary to extend (4.8) to all y in the closed set $\overline{x} + \widehat{\mu}\mathbb{B}$.

Let us prove that $\widehat{\mu} \ge \varepsilon_0$. Suppose this statement is false, i.e., $\widehat{\mu} < \varepsilon_0$: then every sufficiently large integer k admits a corresponding point (x_k, v_k) with $|x_k - \overline{x}| < \widehat{\mu} + 1/k < \varepsilon$, $|v_k - \overline{\gamma}| \le \delta_0$, but $\partial d_\Gamma(x, v) \not\subseteq K\mathbb{B}$. By passing to a subsequence if necessary we can assume that (x_k, v_k) converges to some point $(\widehat{x}, \widehat{v})$ satisfying $|\widehat{x} - \overline{x}| \le \widehat{\mu} < \varepsilon_0$ and $|\widehat{v} - \overline{\gamma}| \le \delta_0 = \delta/6$. From the strong form of (4.8) described in the previous paragraph there exists some point $\widehat{\gamma}$ in $\Gamma(\widehat{x})$ such that $|\widehat{\gamma} - \overline{\gamma}| \le K|\widehat{x} - \overline{x}|$. Our claim applies to the point $(\widehat{x}, \widehat{\gamma})$: it says that the estimate $\partial d_\Gamma \subseteq K\mathbb{B}$ holds throughout some set of the form

$$(\widehat{x} + \mu\mathbb{B}) \times \left(\widehat{\gamma} + \frac{\delta - |\widehat{\gamma} - \overline{\gamma}|}{3}\mathbb{B}\right),$$

where $\mu > 0$ and (by choice of ε_0)

$$\frac{\delta - |\widehat{\gamma} - \overline{\gamma}|}{3} \ge \frac{\delta - K|\widehat{x} - \overline{x}|}{3} \ge \frac{\delta - K\left(\frac{\delta}{9K}\right)}{3} = \frac{8\delta}{27}.$$

But the point $(\widehat{x}, \widehat{v})$ satisfies

$$\begin{aligned} |\widehat{v} - \widehat{\gamma}| \le |\widehat{v} - \overline{\gamma}| + |\overline{\gamma} - \widehat{\gamma}| &\le \frac{\delta}{6} + K|\widehat{x} - \overline{x}| \\ &\le \frac{\delta}{6} + K\left(\frac{\delta}{9K}\right) = \frac{5\delta}{18}. \end{aligned}$$

Now $\frac{8\delta}{27} > \frac{5\delta}{18}$, so these two estimates show that $(\widehat{x}, \widehat{v})$ lies in the *interior* of a set in which ∂d_Γ is bounded by K. This contradicts the stated properties of the sequence (x_k, v_k) and completes our justification that $\widehat{\mu} \geq \varepsilon_0$.

These arguments show that d_Γ is Lipschitz with rank K on a set containing $(\overline{x} + \varepsilon_0\mathbb{B}) \times (\overline{\gamma} + \delta_0\mathbb{B})$. The desired result now follows from Lemma 4.2. □

5. Aubin continuity in Lagrangian and Hamiltonian terms. Like §4, this section is logically independent of the others in the paper, although the similarity of the notation is deliberate.

Given $\overline{x}$ in $\mathbb{R}^n$ and $\varepsilon > 0$, write $\Omega = \overline{x} + \varepsilon\mathbb{B}$ and consider a Lagrangian $L: \Omega \times \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Assume that for every $x \in \Omega$, the function $v \mapsto L(x, v)$ is closed, proper, and convex. Use L to define the multifunction $E(x) := \operatorname{epi} L(x, \cdot)$. Note that for every $x \in \Omega$, this multifunction has nonempty closed convex values. In this section we characterize the Aubin continuity of E near $(\overline{x}, \overline{v})$ in terms of conditions on L and its associated Hamiltonian $H(x, p) := \sup\{\langle p, v\rangle - L(x, v) : v \in \mathbb{R}^n\}$.

THEOREM 5.1. *Fix any point $(\overline{v}, \overline{L})$ in $E(\overline{x})$, along with scalars $\varepsilon > 0$, $K \geq 0$. Write $\mathbb{B}' = \mathbb{B} \times [-1, 1]$ for the unit ball in in $\mathbb{R}^n \times \mathbb{R}$. Then for any $x, y \in \overline{x} + \varepsilon\mathbb{B}$, conditions* (a)–(c) *below are equivalent*:

(a) $E(x) \cap ((\overline{v}, \overline{L}) + \delta\mathbb{B}') \subseteq E(y) + K|y - x|\mathbb{B}'$.

(b) *For any $u \in \overline{v} + \delta\mathbb{B}$ obeying $L(x, u) \leq \overline{L} + \delta$, there exists $v \in \mathbb{R}^n$ satisfying*

(i) $|v - u| \leq K|y - x|$, *and*

(ii) $L(y, v) \leq \max\{\overline{L} - \delta, L(x, u)\} + K|y - x|$.

(c) *For any $p \in \mathbb{R}^n$,*

$$\inf_{\substack{p' \in \mathbb{R}^n \\ \theta > 0}} \{\theta H(x, p'/\theta) + \delta|p' - p| + \delta|\theta - 1| + \langle p - p', \overline{v}\rangle + (\theta - 1)\overline{L}\} \leq H(x, p) + K(1 + |p|)|y - x|. \tag{5.1}$$

Proof. (a⇒b) Suppose that (a) holds. If $u \in \mathbb{R}^n$ satisfies $L(x, u) \leq \overline{L} + \delta$, then the point $(u, \max\{\overline{L} - \delta, L(x, u)\})$ lies in the left side shown in condition (a); thus

$$(u, \max\{\overline{L} - \delta, L(x, u)\}) \in E(y) + K|y - x|\mathbb{B}'.$$

In particular, there has to be a point (h, r) with $\max\{|h|, |r|\} \leq K|y - x|$ such that

$$(u, \max\{\overline{L} - \delta, L(x, u)\}) + (h, r) \in E(y).$$

The special shape of the epigraph set $E(y)$ allows us to replace r with the larger value $K|y - x|$ on the left side: in this case, defining $v = u + h$ gives $|v - u| \leq K|y - x|$ and

$$(v, \max\{\overline{L} - \delta, L(x, u)\} + K|y - x|) \in E(y), \quad \text{i.e., } L(y, v) \leq \max\{\overline{L} - \delta, L(x, u)\} + K|y - x|.$$

(b⇒a) Suppose that (b) holds. Let (u, r) be a vector on the left side in (a). Then $L(x, u) \leq r$ and $\overline{L} - \delta \leq r \leq \overline{L} + \delta$, so (b) provides a vector v such that

(i′) $|v - u| \leq K|y - x|$,

(ii′) $L(y, v) \leq \max\{\overline{L} - \delta, L(x, u)\} + K|y - x| \leq r + K|y - x|$.

Thus $r \geq L(y, v) - K|y - x|$: the special shape of the epigraph set $E(y)$ ensures that $(u, r) \in E(y) + K|y - x|\mathbb{B}$, as required.

(c⇔a) The right side in (a) is a nonempty, closed convex set, since it arises as the sum of a closed convex set and a compact convex set. A separation theorem customized for epigraphs implies that an equivalent formulation of (a) is

$$\sigma_{\mathrm{LS}}(p, -1) \leq \sigma_{\mathrm{RS}}(p, -1) \quad \forall p \in \mathbb{R}^n. \tag{5.2}$$

We calculate

$$\begin{aligned}\sigma_{\mathrm{RS}}(p,-1) &= \sigma_{E(y)}(p,-1) + K|y-x|\sigma_{\mathbb{B}'}(p,-1)\\ &= H(t,y,p) + K\|(p,-1)\|_*|y-x|,\end{aligned} \tag{5.3}$$

where $\|(v,r)\|_* = |v| + |r|$ is the norm on $\mathbb{R}^n \times \mathbb{R}$ dual to the one defining $\mathbb{B}'$ there.

Basic convex analysis (Rockafellar [20, Chap. 16]) affirms that for any nonempty convex sets C and D, with D compact, one has

$$\sigma_{C\cap D} = \mathrm{cl}(\sigma_C \,\square\, \sigma_D) = \sigma_C \,\square\, \sigma_D.$$

(The second equation here holds because the convex function $\sigma_C \,\square\, \sigma_D$ is finite, and hence continuous, on the whole space.) It follows that

$$\begin{aligned}&\sigma_{\mathrm{LS}}(p,-1)\\ &\quad= \left(\sigma_{E(x)} \,\square\, \sigma_{(\overline{v},\overline{L})+\delta\mathbb{B}'}\right)(p,-1)\\ &\quad= \inf_{(p',q')} \left\{\sigma_{E(x)}(p',q') + \delta\left\|(p,-1)-(p',q')\right\|_* + \left\langle p-p',\,\overline{v}\right\rangle + (-1-q')\overline{L}\right\}.\end{aligned}$$

Now $\sigma_{E(x)}(p',q') = \infty$ whenever $q' > 0$, so the latter infimum can be restricted to those points where $q' \le 0$. Furthermore, at any point (p',q') where $q' = 0$, the special features of epigraph sets imply that the quantity $\sigma_{E(x)}(p',0) + \delta\left\|(p,-1)-(p',0)\right\|_*$ can be realized as a limit of some sequence $\sigma_{E(x)}(p'_k,q'_k) + \delta\left\|(p,-1)-(p'_k,q'_k)\right\|_*$ with $q'_k < 0$. Thus the infimum can be restricted to points where $q' < 0$. So we write $\theta = -q' > 0$ and use the observation that

$$\sigma_{E(x)}(p',-\theta) = \theta\sigma_{E(x)}(p'/\theta,-1) = \theta H(x,p'/\theta)$$

to obtain

$$\sigma_{\mathrm{LS}}(p,-1) = \inf_{\substack{p'\in\mathbb{R}^n\\ \theta>0}} \left\{\theta H(x,p'/\theta) + \delta\left\|(p'-p,\theta-1)\right\|_* + \left\langle p-p',\,\overline{v}\right\rangle + (\theta-1)\overline{L}\right\}. \tag{5.4}$$

Equations (5.3) and (5.4) reveal that condition (5.2) is equivalent to (c), whereas (5.2) is equivalent to (a) by construction. □

Clarke's strong Lipschitz condition. In treating the generalized problem of Bolza in [4, Chap. 4], Clarke imposes a Hamiltonian requirement called the "strong Lipschitz condition," which asks that for all x and y in some large enough ball,

$$H(y,p) \le H(x,p) + K\,(1+|p|)\,|y-x| \qquad \forall p \in \mathbb{R}^n. \tag{5.5}$$

Our next corollary shows that Aubin continuity of the sort utilized here is a less demanding hypothesis.

COROLLARY 5.2. *If H satisfies the strong Lipschitz condition* (5.5)*, then H satisfies each of the equivalent conditions* (a)–(c) *in Theorem* 5.1 *for every $\delta > 0$.*

Proof. Choose $p' = p$, $\theta = 1$ in (5.1) to see that the right side of (5.1) is majorized by the right side of (5.5). Thus (5.5) implies condition (c) of Theorem 5.1. □

To see that (5.1) can be strictly weaker than (5.5), consider the example of $L(x,v) = \frac{1}{2}[v^2 + x^2v^2]$. It is easy to compute that $H(x,p) = \frac{1}{2}p^2/(1+x^2)$. For any $\varepsilon > 0$, then, there is a constant $\sigma_\varepsilon > 0$ such that $|H_x(x,p)| \ge \sigma_\varepsilon|p|^2$ for some x in $[-\varepsilon,\varepsilon]$. In particular,

the strong Lipschitz condition (5.5) fails. However, for any fixed x, y in $[-\varepsilon, \varepsilon]$, the choices $\theta = 1$ and $p' = p\sqrt{1+y^2}/\sqrt{1+x^2}$ in (5.1) give

$$\text{LS}(5.1) \le H(x,p) + \delta\left|p\frac{\sqrt{1+y^2}}{\sqrt{1+x^2}} - p\right| \le H(x,p) + (\delta\sqrt{1+x^2})|p||y-x|.$$

Thus inequality (5.1) holds for any $\delta > 0$, with $K = \delta\sqrt{1+\varepsilon^2}$.

In his later treatment of the generalized problem of Bolza in [5], Clarke replaces his "strong Lipschitz condition" with a "weak Lipschitz condition." Although the latter condition is difficult to compare to our Aubin continuity assumption, it does hold for the simple example introduced above.

6. Problems with explicit state constraints. In deriving Theorem 3.1 from Loewen and Rockafellar [14], we have transcribed only the conclusions that pertain in the absence of explicit state constraints. Such constraints are handled in [14], however, and a proof perfectly analogous to the one given in §3 allows us to incorporate them into the main result of this paper as well. In this section we summarize the new ideas required in this broader context and develop the associated enlargement of Theorem 2.1. Fuller explanations of the new ingredients and ideas appear in [13, 14].

Consider the following extension of problem $(\mathcal{P})$ in which state constraints now explicitly enter:

$$(\mathcal{P}^*) \qquad \begin{aligned} &\text{minimize } \Lambda[x] := l(x(a), x(b)) + \int_a^b L(t, x(t), \dot{x}(t))\,dt \\ &\text{subject to } x(t) \in X(t) \ \forall t \in [a,b]. \end{aligned}$$

We retain hypotheses (H1)–(H5) of §2 and impose the following conditions on the state constraint multifunction X:

(H6) Each set $X(t)$ is closed, and the multifunction $t \rightrightarrows X(t)$ is lower semicontinuous, which means that for every point $(t_0, x_0) \in \Omega \cap (\text{gph } X)$ and for every sequence $t_k \to t_0$ in $[a,b]$, there exists a sequence $x_k \to x_0$ with $x_k \in X(t_k)$ for all k.

For each (t,x) in $\Omega \cap (\text{gph } X)$ let

$$\begin{aligned} \overline{N}_X(t,x) = \text{cl co}\big\{v \in \mathbb{R}^n \ : v = \lim_{k\to\infty} v_k \text{ for some sequences} \\ v_k \in \hat{\hat{N}}_{X(t_k)}(x_k),\ (t_k,x_k) \xrightarrow[\text{gph } X]{} (t,x)\big\}. \end{aligned} \tag{6.1}$$

This closed convex cone specifies the directions in which the adjoint function p is allowed to jump when the state constraint is active. Recall that a vector-valued measure dp is called $\overline{N}_X(t, \bar{x}(t))$-valued when dp can be written as $v(t)\,d\mu(t)$ for some nonnegative measure μ on $[a,b]$ with $dp \ll \mu$ and some measurable selection $v(t) \in \overline{N}_X(t, \bar{x}(t))$ μ-a.e.

With these additional ingredients, our main result takes the form stated below. This version differs from the original one, Theorem 2.1, primarily in that its adjoint function p is only of bounded variation, not absolutely continuous as it must be when $X \equiv \mathbb{R}^n$. In particular, the endpoints $p(a)$ and $p(b)$ may differ from the one-sided limits $p(a+)$ and $p(b-)$ in cases where the measure dp has an atom at one or both ends of the interval $[a,b]$.

THEOREM 6.1. *Assume* (H1)–(H6). *Suppose that the arc* $\bar{x}$ *solves problem* $(\mathcal{P}^*)$ *and that the constraint qualification below is satisfied*:

$$\text{(CQ)} \qquad \textit{the cone } \overline{N}_X(t, \bar{x}(t)) \textit{ is pointed for all } t \textit{ in } [a,b].$$

Then either the normal conditions or the degenerate conditions below are satisfied by some function $p \in BV([a,b];\mathbb{R}^n)$ *for which the singular part of the measure* dp *is* $\overline{N}_X(t,\overline{x}(t))$*-valued, and hence in particular is supported on the set*

$$\left\{t \,:\, \overline{N}_X(t,\overline{x}(t)) \neq \{0\}\right\} = \{t \in [a,b] \,:\, (t,\overline{x}(t)) \in \operatorname{bdy} \operatorname{gph} X\}\,.$$

Normal conditions:

(a) $\dot{p}(t) \in \operatorname{co}\left\{w \,:\, (w,p(t)) \in \partial L(t,\overline{x}(t),\dot{\overline{x}}(t))\right\} + \overline{N}_X(t,\overline{x}(t))$ *a.e.* $t \in [a,b]$,

(b) $(p(a),-p(b)) \in \partial l(\overline{x}(a),\overline{x}(b))$.

Singular conditions: The function p *is nonzero, and*

(a^∞) $\dot{p}(t) \in \operatorname{co}\left\{w \,:\, (w,p(t)) \in \partial^\infty L(t,\overline{x}(t),\dot{\overline{x}}(t))\right\} + \overline{N}_X(t,\overline{x}(t))$ *a.e.* $t \in [a,b]$,

(b^∞) $(p(a),-p(b)) \in \partial^\infty l(\overline{x}(a),\overline{x}(b))$.

(*In particular, if the only such function* p *satisfying conditions* (a^∞)–(b^∞) *is identically zero, then the normal conditions are satisfied.*) *In the normal conditions, assertion* (a) *is equivalent to*

(a′) $\dot{p}(t) \in \operatorname{co}\left\{w \,:\, (-w,\dot{\overline{x}}(t)) \in \partial H(t,\overline{x}(t),p(t))\right\} + \overline{N}_X(t,\overline{x}(t))$ *a.e.* $t \in [a,b]$.

Also, conditions (a) *and* (a′) *imply that for almost all* t *in* $[a,b]$,

(c) $p(t) \in \partial_v L(t,\overline{x}(t),\dot{\overline{x}}(t)) = \operatorname{argmax}_{q\in\mathbb{R}^n}\left\{\langle q\,,\,\dot{\overline{x}}(t)\rangle - H(t,\overline{x}(t),q)\right\}$, *and*

$\dot{\overline{x}}(t) \in \partial_p H(t,\overline{x}(t),p(t)) = \operatorname{argmax}_{v\in\mathbb{R}^n}\{\langle p(t)\,,\,v\rangle - L(t,\overline{x}(t),v)\}$.

7. Application: The Lipschitz-plus-indicator case. Many practical problems permit a clear distinction to be drawn between the constraints and the costs. They can thus be expressed in the form of ($\mathcal{P}_1$) in §1. To this model we can now add the possibility of explicit state constraints. We focus then on the problem

$$(\mathcal{P}_1^*)\qquad \begin{aligned} &\text{minimize } \Lambda_1[x] := l_1(x(a),x(b)) + \int_a^b L_1(t,x(t),\dot{x}(t))\,dt \\ &\text{subject to } (x(a),x(b)) \in S \text{ and } \dot{x}(t) \in F(t,x(t)) \text{ a.e. } t \in [a,b] \\ &\text{along with } x(t) \in X(t)\ \forall t \in [a,b], \end{aligned}$$

in which the endpoint cost function l_1 and, for each t, the running cost function $L_1(t,\cdot,\cdot)$ are assumed to be locally Lipschitz continuous. To display this problem as an instance of the general problem's state-constrained version ($\mathcal{P}^*$) treated in §6, it suffices, as we have noted above, to take $l = l_1 + \Psi_S$ and $L = L_1 + \Psi_{\operatorname{gph} F}$, where Ψ_S and $\Psi_{\operatorname{gph} F}$ are the indicators of S and the graph of the multifunction F.

Suitable hypotheses on l_1 and S, as well as L_1 and F, are as follows. Again, they refer to the constant $\varepsilon > 0$ appearing in the definition of Ω and to two positive-valued integrable functions δ and κ.

($H1^+$) The endpoint cost function $l_1\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitzian on $\Omega_a \times \Omega_b$; the target set $S \subseteq \mathbb{R}^n \times \mathbb{R}^n$ is closed.

($H2^+$) The integrand $L_1\colon \Omega \times \mathbb{R}^n \to \mathbb{R}$ and the multifunction $F\colon \Omega \rightrightarrows \mathbb{R}^n$ are $\mathcal{L} \times \mathcal{B}$-measurable.

($H3^+$) For each fixed pair (t,x) in Ω, the function $v \mapsto L_1(t,x,v)$ is convex on $\mathbb{R}^n$, while the set $F(t,x)$ is convex.

($H4^+$) For almost every t in $[a,b]$, the function $(x,v) \mapsto L_1(t,x,v)$ is finite-valued and lower semicontinuous on $\Omega_t \times \mathbb{R}^n$, while the multifunction $x \rightrightarrows F(t,x)$ has closed graph. Furthermore, one has the following epi-continuity property: for any point $(\widehat{x},\widehat{v})$ in $\operatorname{gph} F$ where $|\widehat{x} - \overline{x}(t)| < \varepsilon$ and for any sequence $x_k \to \widehat{x}$ in Ω_t, there exists a sequence $v_k \to \widehat{v}$ satisfying both $v_k \in F(t,x_k)$ and $L_1(t,x_k,v_k) \to L_1(t,\widehat{x},\widehat{v})$.

($H5^+$) The ratio $\kappa(t)/\delta(t)$ is essentially bounded. For almost all t in $[a,b]$, the function $(x,v) \mapsto L_1(t,x,v)$ is Lipschitz of rank $\kappa(t)$ on the set $(\overline{x}(t) + \varepsilon\mathbb{B}) \times (\dot{\overline{x}}(t) + \delta(t)\mathbb{B})$, while

the multifunction F satisfies

$$|w| \leq \kappa(t)\,(1 + |p|) \text{ for all } (w, p) \in N_{\operatorname{gph} F(t,\cdot)}(x, v)$$

whenever $|x - \overline{x}(t)| < \varepsilon$, $\left|v - \dot{\overline{x}}(t)\right| < \delta(t)$.

This case is especially interesting because the Lipschitz continuity of l_1 and L_1 ensures that the singular subgradients of l and L coincide with the usual subgradients of the reduced functions $l_0 = \Psi_S$ and $L_0(t, x, v) = \Psi_{\operatorname{gph} F}(t, x, v)$. This makes it possible to expand the degenerate conditions of Theorem 6.1, which now take the form

(a^∞) $\dot{p}(t) \in \operatorname{co}\left\{v : (v, p(t)) \in \partial L_0(t, \overline{x}(t), \dot{\overline{x}}(t))\right\} + \overline{N}_X(t, \overline{x}(t))$ a.e.,

(b^∞) $(p(a), -p(b)) \in \partial l_0(\overline{x}(a), \overline{x}(b)) = N_S(\overline{x}(a), \overline{x}(b))$.

Rockafellar's equivalence results in [31] certainly apply to L_0 as well as to L, and consequently condition (a^∞) has the equivalent Hamiltonian form

$$\dot{p}(t) \in \operatorname{co}\left\{w : (-w, \dot{\overline{x}}(t)) \in \partial H_0(t, \overline{x}(t), p(t))\right\} \text{ a.e. } t \in [a, b],$$

where, of course, $H_0(t, x, p) := \sup\{\langle p, v\rangle : v \in F(t, x)\}$ is the Hamiltonian corresponding to L_0. Either of the equivalent forms of (a^∞) implies a corresponding argmax condition analogous to (c) in Theorem 6.1.

To summarize these developments, define the Lagrangian and Hamiltonian of index λ, for any $\lambda \geq 0$, by

$$L_\lambda(t, x, v) := \lambda L_1(t, x, v) + \Psi_{\operatorname{gph} F}(t, x, v),$$
$$H_\lambda(t, x, p) := \sup\{\langle p, v\rangle - \lambda L_1(t, x, v) : v \in F(t, x)\}.$$

Then the following result holds.

THEOREM 7.1. *Assume* ($H1^+$)–($H5^+$) *and* (H6). *Suppose that the arc* $\overline{x}$ *solves problem* $(\mathcal{P}_1^*)$ *and that the constraint qualification* (CQ) *of Theorem* 6.1 *is satisfied. Then there exist* $p \in BV([a, b]; \mathbb{R}^n)$ *and a constant* $\lambda \in \{0, 1\}$, *not both zero, such that for almost all* t *in* $[a, b]$,

(a) $\dot{p}(t) \in \operatorname{co}\left\{w : (-w, \dot{\overline{x}}(t)) \in \partial H_\lambda(t, \overline{x}(t), p(t))\right\} + \overline{N}_X(t, \overline{x}(t))$,

(b) $\dot{p}(t) \in \operatorname{co}\left\{w : (w, p(t)) \in \partial L_\lambda(t, \overline{x}(t), \dot{\overline{x}}(t))\right\} + \overline{N}_X(t, \overline{x}(t))$,

(c) $p(t) \in \partial_v L_\lambda(t, \overline{x}(t), \dot{\overline{x}}(t)) = \operatorname{argmax}_{q \in \mathbb{R}^n}\left\{\left\langle q, \dot{\overline{x}}(t)\right\rangle - H_\lambda(t, \overline{x}(t), q)\right\}$,

$\dot{\overline{x}}(t) \in \partial_p H_\lambda(t, \overline{x}(t), p(t)) = \operatorname{argmax}_{v \in F(t, \overline{x}(t))}\{\langle p(t), v\rangle - \lambda L_1(t, \overline{x}(t), v)\}$.

Furthermore,

(d) $(p(a), -p(b)) \in \partial\,(\lambda l_1 + \Psi_S)\,(\overline{x}(a), \overline{x}(b)) \subseteq \lambda \partial l_1(\overline{x}(a), \overline{x}(b)) + N_S(\overline{x}(a), \overline{x}(b))$,

(e) *the singular part of the measure* dp *is* $\overline{N}_X(t, \overline{x}(t))$*-valued and thus is supported on the set* $\left\{t : \overline{N}_X(t, \overline{x}(t)) \neq \{0\}\right\} = \{t : (t, \overline{x}(t)) \in \operatorname{bdy} \operatorname{gph} X\}$.

Note that when $L_1 \equiv 0$, problem $(\mathcal{P}_1^*)$ reproduces the unbounded differential inclusion control problem of Loewen and Rockafellar [14]. The conclusions of Theorem 7.1 then correspond exactly to those of [14, Thm. 4.3] but with three major improvements: they allow for a nonzero integrand L_1, offer the alternative formulation of the Aubin continuity hypothesis in ($H5^+$), and present a sharper Hamiltonian inclusion in (a).

REFERENCES

[1] J.-P. AUBIN, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.

[2] G. A. BLISS, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, 1946.

[3] F. H. CLARKE, *The generalized problem of Bolza*, SIAM J. Control Optim., 14 (1976), pp. 682–699.

[4] ———, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983. (Republished in Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 1991.)

[5] F. H. CLARKE, *Hamiltonian analysis of the generalized problem of Bolza*, Trans. Amer. Math. Soc., 301 (1987), pp. 385–400.
[6] ———, *Methods of Dynamic and Nonsmooth Optimization*, CBMS-NSF Regional Conference Series 57, Society for Industrial and Applied Mathematics, Philadelphia, 1989.
[7] F. H. CLARKE AND R. B. VINTER, *Regularity properties of solutions to the basic problem in the calculus of variations*, Trans. Amer. Math. Soc., 289 (1985), pp. 73–98.
[8] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, Wiley, New York, 1966.
[9] A. D. IOFFE, *Approximate subdifferentials and applications* I: *The finite dimensional theory*, Trans. Amer. Math. Soc., 281 (1984), pp. 389–416.
[10] A. D. IOFFE AND R. T. ROCKAFELLAR, *The Euler and Weierstrass conditions for nonsmooth variational problems*, Calculus of Variations and Partial Differential Equations, 4 (1996), pp. 59–87.
[11] P. D. LOEWEN, *The proximal subgradient formula in Banach space*, Canad. Math. Bull., 31 (1988), pp. 353–361.
[12] ———, *Optimal Control via Nonsmooth Analysis*, CRM-AMS Lecture Notes in Mathematics 2, American Mathematical Society, Providence, RI, 1993.
[13] P. D. LOEWEN AND R. T. ROCKAFELLAR, *The adjoint arc in nonsmooth optimization*, Trans. Amer. Math. Soc., 325 (1991), pp. 39–72.
[14] ———, *Optimal control of unbounded differential inclusions*, SIAM J. Control Optim., 32 (1994), pp. 442–470.
[15] B. S. MORDUKHOVICH, *Metric approximations and necessary optimality conditions for general classes of nonsmooth extremal problems*, Dokl. Acad. Nauk SSSR 254 (1980), pp. 1072–1076. (In Russian; English translation in Soviet Math. Dokl., 22 (1980), pp. 526–530.)
[16] ———, *Optimization and approximation of differential inclusions*, Kibernetika 6 (1988), pp. 83–89. (In Russian; English translation in Cybernetics, 24 (1988), pp. 781–788.)
[17] ———, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988, Chapter 3.
[18] ———, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.
[19] ———, *Discrete approximations and refined Euler-Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 33 (1995), pp. 882–915.
[20] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[21] ———, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Analysis Appl., 23 (1970), pp. 174–222.
[22] ———, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–40.
[23] ———, *State constraints in convex problems of Bolza*, SIAM J. Control, 10 (1972), pp. 691–715.
[24] ———, *Existence theorems for general control problems of Bolza and Lagrange*, Adv. in Math., 15 (1975), pp. 312–333.
[25] ———, *Dual problems of Lagrange for arcs of bounded variation*, in Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976, pp. 155–192.
[26] ———, *Proximal subgradients, marginal values, and augmented Lagrangians in nonconvex optimization*, Math. Oper. Res., 6 (1981), pp. 424–436.
[27] ———, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Anal., 9 (1985), pp. 665–698.
[28] ———, *Lipschitzian properties of multifunctions*, Nonlinear Anal., 9 (1985), pp. 867–885.
[29] ———, *Hamiltonian trajectories and duality in the optimal control of linear systems with convex costs*, SIAM J. Control Optim., 27 (1989), pp. 1007–1025.
[30] ———, *Dualization of subgradient conditions for optimality*, Nonlinear Anal., 20 (1993), pp. 627–646.
[31] ———, *Equivalent subgradient versions of Hamiltonian and Euler–Lagrange equations in variational analysis*, SIAM J. Control Optim., 34 (1996), pp. 1300–1314.
[32] G. V. SMIRNOV, *Discrete approximations and optimal colutions to differential inclusions*, Kibernetika, 10 (1991), pp. 76–79. (In Russian; English translation in Cybernetics, 27 (1991), pp. 101–107.)
[33] R. B. VINTER AND H. ZHENG, *The extended Euler–Lagrange condition for nonconvex variational problems*, SIAM J. Control Optim., 35 (1997), to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1512–1532, September 1996

003

BOUNDARY EXACT CONTROLLABILITY OF INTERFACE PROBLEMS WITH SINGULARITIES I: ADDITION OF THE COEFFICIENTS OF SINGULARITIES*

SERGE NICAISE†

Abstract. We prove the exact controllability by boundary action of hyperbolic interface problems with singularities. The idea is to replace the classical space of controls with the space of their regular parts, which is augmented with the coefficents of singularities. This leads to a classical boundary control but with an internal control, which is a distribution with a support equal to the singular vertices.

Key words. interface problems, singularities, control

AMS subject classifications. 93C20, 35B37, 35L67

1. Introduction.

The description of various models of multiple-link flexible structures consisting of finitely many interconnected flexible elements, such as strings, beams, plates, shells, or combinations of them, recently has been the subject of great interest [10, 3, 12, 13]. The problem of controllability, or even stabilizability, of such structures has been considered very little. Let us quote the works of Lagnese, Leugering, and Schmidt [25, 11, 12] for one-dimensional (1-d) networks; the books of Lagnese, Leugering, and Schmidt [12, Chap. VII] and J.-L. Lions [15, Chap. VI] about transmission problems excluding singularities; and finally, the papers of Puel and Zuazua [24] and the author [20, 21] for multidimensional structures. In all these works, either no singularity occurs or, if there are some singularities, they are cancelled by an appropriate choice of the multiplier. In the present papers, we show with simple examples of multilink structures—namely, the wave equation in two-dimensional (2-d) networks (see below for a precise definition; also [4, 22])—how to manage the presence of singularities and the controllability problem using the Hilbert uniqueness method of J.-L. Lions [15]. Indeed, this method usually needs the regularity $H^{3/2+\varepsilon}$ for some $\varepsilon > 0$ (in our case, it means $H^{3/2+\varepsilon}(P_i)$ for all faces P_i), which is, in general, not satisfied for 2-d networks. For the classical wave equation in the plane (or in the space), i.e., without transmission, different strategies were proposed to overcome this difficulty: Grisvard [7] imposed strong geometrical assumptions in order to adapt the multipliers method; to avoid these geometrical conditions, Niane and Seck [18, 19] and Heibig and Moussaoui [8] proposed using classical boundary controls whose support stays far from the singular points and adding internal controls located in a small neighbourhood of the singular vertices. Here, for interface problems, we proposed the following slightly different method: we replace the boundary control with its regular part and add to the space of controls the coefficients of the singularities. This leads to a classical boundary control but with an internal control, which is a distribution with a support equal to the singular vertices.

The second part will propose two other methods consisting of acting by a classical boundary control whose support does not contain a neighbourhood of the singular points and adding internal controls located near these singular points (more in the spirit of [18, 8, 19], where such a method was introduced).

The structure of this paper is as follows: In §2, we recall some notations and definitions concerning 2-d networks. We also give the decomposition into a regular part and a singular one for elements from the domain of the Laplace operator. Section 3 is devoted to the solution of the wave equation and the regularity of its coefficients of the singularities. This is made

*Received by the editors February 27, 1995; accepted for publication April 12, 1995.

†Institut des Sciences et Techniques de Valenciennes, Université de Valenciennes et du Hainaut Cambrésis, B.P. 311, F-59304 Valenciennes cedex, France.

by using a localization procedure and Bessel functions as Moussaoui and Sadallah [16] and Moussaoui and Tran [17] made in the non-transmission case. In §4, we establish an identity with multiplier, different from the usual one since a remainder depending on the coefficients of singularities appears. This identity yields an estimate of the energy. The weak solution of the wave equation is considered in §5, as is its interpretation in terms of partial differential equations. Section 6 is devoted to the setting of the Hilbert uniqueness method. Finally, we give in §7 an example of a 2-d network for which we do not have exact controllability by exterior boundary control. As for 1-d networks [12, § II.5.2], this network contains circuits.

2. Preliminaries. Let us fix Ω, a 2-d polygonal topological network (for short, a 2-d network; see [4, 22] for more details) which is a subset of $\mathbf{R}^n$ (n fixed ≥ 2) formed by a finite union of disjoint nonempty subsets P_i, $i \in \mathcal{I}$ such that

(i) each P_i is a simply connected open subset of a plane Π_i of $\mathbf{R}^n$, P_i being a polygonal domain of Π_i;

(ii) $\cup_{i\in\mathcal{I}} \overline{P_i}$ is connected;

(iii) for all $i, j \in \mathcal{I}$, $i \neq j$, $\overline{P_i} \cap \overline{P_j}$ either is empty or is a common vertex or is a whole common side.

For all $i \in \mathcal{I}$, we can fix once and for all a system of Cartesian coordinates in the plane Π_i. We assume that the boundary of P_i is the union of a finite number of linear segments $\overline{\Gamma}_{ij}$, $j \in \{1, \ldots, N_i\}$, numbered according to the trigonometric orientation. S_{ij} denotes the common vertex between $\overline{\Gamma}_{ij}$ and $\overline{\Gamma}_{ij+1}$, and ω_{ij} is the interior angle at S_{ij} between Γ_{ij} and Γ_{ij+1}. Now, ν_{ij} denotes the unitary outer normal vector on Γ_{ij} and τ_{ij} is the unitary tangent vector to Γ_{ij} so that (ν_{ij}, τ_{ij}) is a direct orthonormal basis.

The vertices of Ω will denote the vertices of all P_i's, and in the same way, the edges will be the sides of all P_i's. The faces will be the P_i's. We shall denote by $\mathcal{A}$ (resp., $\mathcal{S}$) the set of edges (resp., vertices) of Ω. For a fixed $S \in \mathcal{S}$, $\mathcal{I}(S)$ will denote the set of adjacent faces to S, i.e., $\mathcal{I}(S) = \{i \in \mathcal{I} : S \in \overline{P_i}\}$. Similarly, $\mathcal{A}(S)$ will be the set of adjacent edges to S, i.e., $\mathcal{A}(S) = \{A \in \mathcal{A} : S \in \overline{A}\}$. Finally, for $A \in \mathcal{A}$, $\mathcal{I}_A$ will be the set of adjacent faces to A, i.e., $\mathcal{I}_A = \{i \in \mathcal{I} : \exists j \in \{1, \ldots, N_i\}$ such that $\Gamma_{ij} = A\}$; moreover, for any $i \in \mathcal{I}_A$, (ν_i, τ_i) will denote (ν_{ij}, τ_{ij}) for the unique $j \in \{1, \ldots, N_i\}$ such that $\Gamma_{ij} = A$, when no confusion is possible.

Later, γ_{ij} will denote on P_i the trace operator on the edge Γ_{ij}. For a function u defined on Ω, u_i will be the restriction of u on the face P_i:

$$u_i : P_i \to \mathbf{C} : x \mapsto u(x).$$

For the sake of simplicity, we shall write

$$\int_\Omega u \, dx = \sum_{i\in\mathcal{I}} \int_{P_i} u_i(x) \, dx.$$

We now recall the definition of Sobolev spaces on Ω.

DEFINITION 2.1. *Let s be a nonnegative real number. Then*

$$\mathcal{H}^s(\Omega) = \prod_{i\in\mathcal{I}} H^s(P_i), \tag{2.1}$$

where $H^s(P_i)$ is the usual Sobolev space on P_i [6] with norm denoted by $\|\cdot\|_{s,P_i}$. $\mathcal{H}^s(\Omega)$ is a Hilbert space for the product norm:

$$\|u\|_{s,\Omega} = \left(\sum_{i\in\mathcal{I}} \|u_i\|^2_{s,P_i} \right)^{1/2}. \tag{2.2}$$

In the following discussion, for the sake of simplicity, we shall write H for $L^2(\Omega) := \mathcal{H}^0(\Omega)$.

Let us fix a partition of the set of edges $\mathcal{A}$ into two subsets $\mathcal{N}$ and $\mathcal{D}$. ($\mathcal{D}$ will be the part of the "boundary" where we shall consider Dirichlet boundary conditions, while $\mathcal{N}$ will be the part where we shall consider Neumann or transmission conditions.) For convenience, we suppose that $\mathcal{D}$ is not empty. We also fix a sequence of positive real numbers $\{\alpha_i\}_{i\in\mathcal{I}}$ and denote by α the function defined on Ω by $\alpha_i(x) = \alpha_i$ for all $x \in P_i$.

We consider the following boundary value problem: given $f \in L^2(\Omega)$, let u be a solution of

$$-\Delta u_i = f_i \text{ in } P_i \quad \forall i \in \mathcal{I}, \tag{2.3}$$

$$\gamma_{ij} u_i = 0 \text{ on } \Gamma_{ij} \quad \forall\, \Gamma_{ij} \in \mathcal{D}, \tag{2.4}$$

$$\gamma_{ij} u_i = \gamma_{kl} u_k \quad \text{when } \Gamma_{ij} = \Gamma_{kl}, \tag{2.5}$$

$$\sum_{i\in\mathcal{I}:\exists j:\Gamma_{ij}=A} \alpha_i \gamma_{ij} \frac{\partial u_i}{\partial \nu_{ij}} = 0 \text{ on } A \quad \text{when } A \in \mathcal{N}. \tag{2.6}$$

When Ω is a polygon of the plane, the union of several polygons P_i, problem (2.3)–(2.6) is called a transmission problem, which was studied by Kellogg [9], Lemrabet [14], and Dobrowolski [5]. The extension to 2-d networks was considered in [4, 22].

The problem (2.3)–(2.6) admits the next variational formulation, (2.7): introduce the Hilbert space

$$V = \{u \in \mathcal{H}^1(\Omega) \text{ fulfilling (2.4) and (2.5)}\}$$

and the continuous bilinear form on V:

$$a(u, v) = \sum_{i\in\mathcal{I}} \alpha_i \int_{P_i} \nabla u_i \cdot \nabla v_i \, dx \quad \forall u, v \in V.$$

Therefore, we shall say that u is a weak solution of problem (2.3)–(2.6) if $u \in V$ satisfies

$$a(u, v) = \int_\Omega f v \, dx \quad \forall v \in V. \tag{2.7}$$

Since $\mathcal{D}$ is nonempty, the form a is V-coercive. Consequently, the existence and uniqueness of the solution of (2.7) follow from the Lax–Milgram lemma. Moreover, since V is dense and compactly imbedded into $L^2(\Omega)$ and the form a is symmetric, it also induces a positive self-adjoint operator A from $L^2(\Omega)$ into $L^2(\Omega)$, with a compact inverse, defined by

$$\begin{cases} D(A) = \{u \in V : \exists f \in H : a(u, v) = \int_\Omega f v \, dx, \forall v \in V\}, \\ Au = f \quad \forall u \in D(A). \end{cases} \tag{2.8}$$

According to [9, 14, 22], $u \in D(A)$ does not have, in general, the optimal regularity $\mathcal{H}^2(\Omega)$ and admits a decomposition into a regular part and a singular part. To describe this decomposition, we recall the following notation (see [22, §1.6]): For a fixed vertex $S \in \mathcal{S}$ of Ω and any $i \in \mathcal{I}(S)$, let us denote by $\omega_i(S)$ the interior opening of P_i at S. We now introduce polar coordinates (r_i, θ_i) on Π_i centred at S such that the half-lines $\theta_i = 0$ and $\theta_i = \omega_i(S)$ contain the two edges of P_i of extremity S. The 1-d network $\mathcal{R}(S)$ associated with S is then defined by

$$\mathcal{R}(S) = \bigcup_{i\in\mathcal{I}(S)} \{(\cos\theta_i, \sin\theta_i) \in \Pi_i : 0 < \theta_i < \omega_i(S)\}.$$

Roughly speaking, we can identify the nodes of $\mathcal{R}(S)$ with the edges of Ω adjacent to S. We are now able to introduce the Laplace–Beltrami operator Δ_S on $\mathcal{R}(S)$: $D(\Delta_S) = \{(u_i)_{i\in\mathcal{I}(S)} : u_i \in H^2(]0, \omega_i(S)[)$ satisfying (2.9), (2.10) and (2.11) below$\}$, and for any $u = (u_i)_{i\in\mathcal{I}(S)} \in D(\Delta_S)$, we set $\Delta_S u = (-u_i'')_{i\in\mathcal{I}(S)}$:

$$u_i(A) = 0 \quad \forall A \in \mathcal{A}(S) \cap \mathcal{D}, \tag{2.9}$$

$$u_i(A) = u_j(A) \quad \forall A \in \mathcal{A}(S), \quad i, j \in \mathcal{I}_A, \tag{2.10}$$

$$\sum_{i\in\mathcal{I}_A} \alpha_i \frac{\partial u_i}{\partial \nu_i}(A) = 0 \quad \forall A \in \mathcal{A}(S) \cap \mathcal{N}. \tag{2.11}$$

According to the results of §1.6 of [22], Δ_S is a nonnegative self-adjoint operator on $L^2(\mathcal{R}(S))$ (equipped with the inner product

$$(u, v)_{\mathcal{R}(S),\alpha} := \sum_{i\in\mathcal{I}(S)} \alpha_i \int_0^{w_i(S)} u_i(\theta_i) v_i(\theta_i)\, d\theta_i)$$

with a discrete set of eigenvalues $\{\lambda_{S,n}^2\}_{n\in\mathbf{N}^*}$ repeated according to their multiplicity (for convenience, $\lambda_{S,n}$ is always supposed to be nonnegative) and of associated eigenvectors $\{\varphi^{S,n}\}_{n\in\mathbf{N}^*}$ satisfying the orthogonality conditions

$$(\varphi^{S,n}, \varphi^{S,m})_{\mathcal{R}(S),\alpha} = \delta_{nm} \quad \forall n, m \in \mathbf{N}^*. \tag{2.12}$$

Theorem 2.27 of [22] yields the following theorem.

THEOREM 2.2. *Let $u \in D(A)$. Then it admits the decomposition*

$$u = u_R + \sum_{S\in\mathcal{S}} \sum_{0<\lambda_{S,n}\le 1/2} c_{S,n} S^{S,n}, \tag{2.13}$$

where $u_R \in \mathcal{H}^{3/2+\varepsilon}(\Omega)$, for $\varepsilon > 0$ small enough, is the regular part of u; $S^{S,n} = \eta_S r^{\lambda_{S,n}} \varphi^{S,n}(\theta)$ are the so-called singular functions of the operator A; η_S is a cutoff function such that $\eta_S = 1$ (resp., 0) near S (resp., near the other vertices); and finally $c_{S,n} \in \mathbf{C}$ is the coefficient of the singularity $S^{S,n}$, which admits the expression

$$c_{S,n} = -\int_\Omega Au \cdot K^{S,n}\, dx \tag{2.14}$$

for some function $K^{S,n} \in L^2(\Omega)$, which behaves like $r^{-\lambda_{S,n}}$ near S.

In the following discussion, for the sake of brevity, we shall write $\sum_{S,n}$ for $\sum_{S\in\mathcal{S}} \sum_{0<\lambda_{S,n}\le 1/2}$.

3. The wave equation. Since H, V, and the form a fulfill the hypotheses of Remark 4.4 of [20], Theorems 4.1–4.3 of [20] may be applied to A. In particular, we have the following theorem.

THEOREM 3.1. *Let $\varphi_0 \in D(A^s)$, $\varphi_1 \in D(A^{s-1/2})$ and $f \in L^1(0, T; D(A^{s-1/2}))$, with $s \ge 1/2$. Then the problem*

$$\begin{cases} \varphi''(t) + A\varphi(t) = f(t), & t \in [0, T], \\ \varphi(0) = \varphi_0, \\ \varphi'(0) = \varphi_1, \end{cases} \tag{3.1}$$

has a unique solution $\varphi \in C([0, T], D(A^s)) \cap C^1([0, T], D(A^{s-1/2}))$ *fulfilling*

$$\begin{aligned}(3.2)\qquad &\|\varphi\|_{C([0,T],D(A^s))} + \|\varphi\|_{C^1([0,T],D(A^{s-1/2}))}\\ &\le C\{\|\varphi_0\|_{D(A^s)} + \|\varphi_1\|_{D(A^{s-1/2})} + \|f\|_{L^1(0,T;D(A^{s-1/2}))}\}\end{aligned}$$

for some constant $C > 0$ *independent of* φ.

In the particular case $s = 1$, the solution φ satisfies $\varphi \in C([0, T], D(A))$; consequently, by Theorem 2.2, $\varphi(t)$ admits the decomposition (2.13) for all t, with coefficients $c_{S,n}(t)$ depending on t. Our next aim is to give regularity results for these coefficients $c_{S,n}(t)$. Following [16, 17], we first work in a neighbourhood of a vertex and use an explicit basis of the eigenfunctions of the Laplace operator. More precisely, for a fixed vertex $S \in \mathcal{S}$, let us introduce the 2-d network

$$\Omega_S = \bigcup_{i \in \mathcal{I}(S)} \Omega_{Si},$$

where $\Omega_{Si} = \{(r_i \cos\theta_i, r_i \sin\theta_i) \in \Pi_i : 0 < r_i < 1, 0 < \theta_i < \omega_i(S)\}$. Remark that Ω_S coincides with Ω in a neighbourhood of S. Similarly, we shall denote by Γ_{Sij} the segment $\{(r_i \cos\eta_i, r_i \sin\eta_i) \in \Pi_i : 0 < r_i < 1\}$, with $\eta_i = 0$ or $\omega_i(S)$, which coincides with Γ_{ij} near S. As before, the Laplace operator A_S on Ω_S corresponds to the following data: $V_S = \{u \in \mathcal{H}^1(\Omega_S)$ fulfilling (2.4), (2.5) on Γ_{Sij} and $u_i(1, \theta_i) = 0$, for a.e. $\theta_i \in]0, \omega_i(S)[$, and all $i \in \mathcal{I}(S)\}$ and

$$a_S(u, v) = \sum_{i \in \mathcal{I}(S)} \alpha_i \int_{\Omega_{Si}} \nabla u_i \nabla v_i \, dx \quad \forall u, v \in V_S.$$

Using the method of separation of variables, one can prove the following lemma.

LEMMA 3.2. *For all* $n, k \in \mathbb{N}^*$, $\lambda_{n,k} := (j_{\lambda_{S,n};k})^2$ *is an eigenvalue of* $\alpha_S^{-1} A_S$ *(obviously* α_S *is the function defined on* Ω_S *by* $\alpha_i(x) = \alpha_i$ *for all* $x \in \Omega_{Si}$*) of associated eigenvector*

$$(3.3)\qquad w^{n,k} = \alpha_{n,k} J_{\lambda_{S,n}}(j_{\lambda_{S,n};k} r)\varphi^{S,n}(\theta),$$

where $J_\gamma(s)$ *is the Bessel function of index* γ *and* $j_{\gamma;k}$ *is its* k*th positive zero. The real number* $\alpha_{n,k}$ *is a normalization factor chosen so that*

$$(3.4)\qquad \sum_{i \in \mathcal{I}(S)} \int_{\Omega_{Si}} \alpha_i |w_i^{n,k}(x)|^2 \, dx = 1.$$

Moreover $\{\sqrt{\alpha_S} w^{n,k}\}_{n,k \in \mathbb{N}^*}$ *is an orthonormal basis of* $L^2(\Omega_S)$ *equipped with the inner product* $(\cdot, \cdot)_{S,\alpha}$ *defined by*

$$(u, v)_{S,\alpha} = \sum_{i \in \mathcal{I}(S)} \alpha_i \int_{\Omega_{Si}} u_i(x) v_i(x) \, dx.$$

Since $J_\gamma(s) \cong s^\gamma$ near $s = 0$, we may see that $w^{n,k} \in \mathcal{H}^2(\Omega_S)$ iff $\lambda_{S,n} > 1$. Consequently, only the eigenvectors $w^{n,k}$ with $0 \le \lambda_{S,n} \le 1$ will arise in the computation of the coefficient $c_{s,n}$ in the decomposition (2.13). As in [16, 17], we need the following technical step.

LEMMA 3.3. *If* $0 \le \lambda_{S,n} \le 1$, *there exist* $C_1, C_2 > 0$ *(depending on* n *but not on* k*) such that*

$$(3.5)\qquad C_1 k^2 \le \lambda_{n,k} \le C_2 k^2,$$

$$(3.6)\qquad C_1 \sqrt{k} \le |\alpha_{n,k}| \le C_2 \sqrt{k} \quad \forall k \in \mathbb{N}^*.$$

Proof. Equation (3.5) is a direct consequence of MacMahon formula [1, p. 371], which says that

$$j_{\gamma;k} = \left(k + \frac{\gamma}{2} - \frac{1}{4}\right)\pi + \bigcirc\left(\frac{1}{k}\right) \tag{3.7}$$

for a fixed γ. The estimates (3.6) follow from simple properties of Bessel functions and remarking that (3.4) is equivalent to

$$|\alpha_{n,k}|^2 \int_0^1 |J_{\lambda_{S,n}}(j_{\lambda_{S,n};k} r)|^2 \, r dr = 1. \qquad \square$$

We are now able to clarify the regularity of the eigenvectors $w^{n,k}$.

PROPOSITION 3.4. *For all $n \in \mathbf{N}^*$, we have*

$$w^{n,k} = w_R^{n,k} + a_{n,k} S^{S,n}, \tag{3.8}$$

where $w_R^{n,k} \in \mathcal{H}^{3/2+\varepsilon}(\Omega_S)$ for some $\varepsilon > 0$; $a_{n,k} = 0$ if $\lambda_{S,n} > 1/2$ or if $\lambda_{S,n} = 0$. Otherwise there exist $C_1, C_2 > 0$ (depending on n but not on k) such that

$$C_1 (\lambda_{n,k})^{\frac{\lambda_{S,n}}{2}+\frac{1}{4}} \le |a_{n,k}| \le C_2 (\lambda_{n,k})^{\frac{\lambda_{S,n}}{2}+\frac{1}{4}} \quad \forall k \in \mathbf{N}^*. \tag{3.9}$$

Proof. As $w^{n,k} \in V_S$ is solution of $A_S w^{n,k} = \alpha_S \lambda_{n,k} w^{n,k} \in L^2(\Omega_S)$, Theorem 2.2 implies

$$w^{n,k} = w_R^{n,k} + \sum_{0<\lambda_{S,m}\le 1/2} a_{n,k,m} S^{S,m}, \tag{3.10}$$

where $w_R^{n,k} \in \mathcal{H}^{3/2+\varepsilon}(\Omega_S)$ for $\varepsilon > 0$ small enough. Moreover, with the help of Theorem 2.27 of [22], we can show that (see also Theorem 2.2)

$$a_{n,k,m} = -\lambda_{n,k} \sum_{i \in \mathcal{I}(S)} \alpha_i \int_{\Omega_{Si}} w_i^{n,k} K_i^{S,m} \, dx,$$

where $K^{S,m} = \frac{1}{\lambda_{S,m}}(r^{-\lambda_{S,m}} - r^{\lambda_{S,m}})\varphi^{S,m}$. Using the expression of $w^{n,k}$ and the orthogonality properties (2.12) of the $\varphi^{S,m}$'s, we arrive at

$$a_{n,k,m} = -\delta_{nm} \frac{\alpha_{n,k}\lambda_{n,k}}{\lambda_{S,m}} \int_0^1 J_{\lambda_{S,n}}(j_{\lambda_{S,n};k} r)(r^{-\lambda_{S,m}} - r^{\lambda_{S,m}}) r dr. \tag{3.11}$$

This leads to (3.8). The asymptotic behaviour of $a_{n,k}$ comes from (3.11), the properties of the Bessel functions, and Lemma 3.3. $\square$

Let us return to the wave equation (3.1).

THEOREM 3.5. *Let $\varphi_0 \in D(A)$, $\varphi_1 \in V$, and $f \in L^2(0, T; V)$. If $T > 2$, then the solution $\varphi \in C([0, T], D(A)) \cap C^1([0, T], V)$ of (3.1) admits the decomposition*

$$\varphi = \varphi_R + \sum_{S,n} c_{S,n} S^{S,n}, \tag{3.12}$$

where $\varphi_R \in C([0, T], \mathcal{H}^{3/2+\varepsilon}(\Omega)) \cap H^1(0, T; V)$ for some $\varepsilon > 0$ and $c_{S,n} \in H^s(0, T)$, for any $s \le \frac{3}{2} - \lambda_{S,n}$ with norms depending continuously on $\|\varphi_0\|_{D(A)} + \|\varphi_1\|_V + \|f\|_{L^2(0,T;V)}$.

Proof. Multiplying φ by the cutoff function η_S, we are reduced to the case $\Omega = \Omega_S$. Writing φ in the basis $\{w^{n,k}\}_{n,k\in\mathbb{N}^*}$ and using (3.8), we arrive at (3.12) with

$$c_{S,n}(t) = \sum_{k\in\mathbb{N}^*} a_{n,k} \left\{ \cos(t\sqrt{\lambda_{n,k}})(\varphi_0, w^{n,k})_{S,\alpha} + \frac{\sin(t\sqrt{\lambda_{n,k}})}{\sqrt{\lambda_{n,k}}}(\varphi_1, w^{n,k})_{S,\alpha} + \int_0^t \frac{\sin((t-s)\sqrt{\lambda_{n,k}})}{\sqrt{\lambda_{n,k}}}(f(s), w^{n,k})_{S,\alpha}\, ds \right\}, \tag{3.13}$$

where φ_R is the remainder. To prove the inclusion $c_{S,n} \in H^s(0,T)$, we use Ingham's inequalities [2] in each of the three series in the right-hand side of (3.13). Indeed applying Theorem 2.1 of [2] to

$$w := \sum_{k\in\mathbb{N}^*} \lambda_{n,k}^{s/2} a_{n,k} \cos(t\sqrt{\lambda_{n,k}})(\varphi_0, w^{n,k})_{S,\alpha},$$

we get the existence of a constant $C > 0$ such that

$$\int_0^T |w(t)|^2\, dt \le C\|\varphi_0\|^2_{D(A)}$$

if $T > 2$ because (3.7) yields assumption (2.1) of [2] with $\gamma = \pi$. Other terms are treated similarly. □

In the above theorem, the assumption $f \in L^2(0,T;V)$ is not satisfactory for our next applications. We prefer $f \in L^1(0,T;V)$, which is unfortunately unavailable; therefore, we replace it with $f \in L^1(0,T;D(A))$.

THEOREM 3.6. *Let $\varphi \in C([0,T], D(A^{3/2})) \cap C^1([0,T], D(A))$ be the solution of* (3.1) *with $\varphi_0 = \varphi_1 = 0$ and $f \in L^1(0,T;D(A))$. Then it admits the decomposition* (3.12)*, where $\varphi_R \in C^1([0,T], \mathcal{H}^{3/2+\varepsilon}(\Omega))$ for some $\varepsilon > 0$ and $c_{S,n} \in C^1([0,T])$, with norms depending continuously on $\|f\|_{L^1(0,T;D(A))}$.*

Proof. Since $D(A^{3/2}) \hookrightarrow D(A)$, $\varphi(t)$ clearly has the expansion (3.12). The conclusion follows from the expression (2.14) of $c_{S,n}(t)$, the regularity $\varphi \in C^1([0,T], D(A))$, and Lebesgue's bounded convergence theorem. □

4. Estimate of the energy. As usual [15, 7], the estimate of the energy is based on an identity with multiplier, which, in our case, takes a slightly different form since we multiply the equation $\varphi'' + A\varphi = 0$ by $m \cdot \nabla\varphi_R$ instead of $m \cdot \nabla\varphi$. (Due to the singular behaviour of φ, the factor $m \cdot \nabla\varphi$ is too singular.) This will be made in several technical steps.

It is quite natural that the boundary control is applied only on the exterior boundary of the domain. This means that we have to distinguish between "interior" and "exterior" Neumann edges: more precisely, we set

$$\begin{aligned} \mathcal{N}_{\text{int}} &= \{A \in \mathcal{N} : \#\mathcal{I}_A \ge 2\}, \\ \mathcal{N}_{\text{ext}} &= \{A \in \mathcal{N} : \#\mathcal{I}_A = 1\}. \end{aligned}$$

Dirichlet edges, on the contrary, may always be seen as exterior edges. For $A \in \mathcal{N}_{\text{ext}} \cup \mathcal{D}$, we then denote by i_A the unique element of $\mathcal{I}_A$.

For fixed points $x_{0i} \in \Pi_i$, $i \in \mathcal{I}$, we define the function m on Ω by $m_i(x) = x - x_{0i}$.

From now on, we suppose that the following geometrical conditions are satisfied (see the end of this section for some examples):

(H1) $\sum_{i\in\mathcal{I}_A} m_i \cdot \nu_i = 0$ on $A\ \forall A \in \mathcal{N}_{\text{int}}$.

(H2) $m_i \cdot \tau_i = m_j \cdot \tau_i$ on $A\ \forall i, j \in \mathcal{I}_A$, $A \in \mathcal{N}_{\text{int}}$.

(H3) $\sum_{i\in\mathcal{I}_A} \alpha_i m_i \cdot \nu_i \geq 0$ on $A\ \forall A \in \mathcal{N}_{\text{int}}$.

(H4) For all $A \in \mathcal{N}_{\text{int}}$ and any $w \in \mathbf{R}^{\#\mathcal{I}_A}$ such that $\sum_{i\in\mathcal{I}_A} \alpha_i w_i = 0$, we have

$$\sum_{i\in\mathcal{I}_A} \alpha_i m_i \cdot \nu_i (w_i)^2 \leq 0 \text{ on } A.$$

LEMMA 4.1. *Let $\varphi \in \mathcal{H}^{3/2+\varepsilon}(\Omega) \cap D(A)$ for some $\varepsilon > 0$. Then the following inequality holds:*

$$\begin{aligned}\int_\Omega A\varphi m \cdot \nabla\varphi dx \geq & -\frac{1}{2}\sum_{A\in\mathcal{D}}\int_A \alpha_{i_A} m_{i_A}\cdot\nu_{i_A}\left(\frac{\partial\varphi_{i_A}}{\partial\nu_{i_A}}\right)^2 d\sigma \\ & + \frac{1}{2}\sum_{A\in\mathcal{N}_{\text{ext}}}\int_A \alpha_{i_A} m_{i_A}\cdot\nu_{i_A}\left(\frac{\partial\varphi_{i_A}}{\partial\tau_{i_A}}\right)^2 d\sigma.\end{aligned} \tag{4.1}$$

Proof. The regularity of φ allows us to apply Green's formula on each P_i, which leads to (see Lemma 3.3 of [7])

$$\begin{aligned}\int_{P_i} \Delta\varphi_i m_i \cdot \nabla\varphi_i\, dx = \sum_{j=1}^{N_i}\Bigg\{ & -\frac{1}{2}\int_{\Gamma_{ij}} m_i\cdot\nu_{ij}|\nabla\varphi_i|^2\, d\sigma \\ & + \int_{\Gamma_{ij}} \frac{\partial\varphi_i}{\partial\nu_{ij}} m_i\cdot\nabla\varphi_i\, d\sigma\Bigg\}.\end{aligned}$$

Multiplying this identity by α_i and summing on $i \in \mathcal{I}$, we get

$$\begin{aligned}\int_\Omega A\varphi m\cdot\nabla\varphi dx = \sum_{i\in\mathcal{I}}\sum_{j=1}^{N_i}\Bigg\{ & \frac{1}{2}\int_{\Gamma_{ij}} \alpha_i m_i\cdot\nu_{ij}|\nabla\varphi_i|^2\, d\sigma \\ & - \int_{\Gamma_{ij}} \alpha_i \frac{\partial\varphi_i}{\partial\nu_{ij}}\left(m_i\cdot\nu_{ij}\frac{\partial\varphi_i}{\partial\nu_{ij}} + m_i\cdot\tau_{ij}\frac{\partial\varphi_i}{\partial\tau_{ij}}\right) d\sigma\Bigg\}.\end{aligned} \tag{4.2}$$

Using the condition (H2) and the fact that φ satisfies (2.4), (2.5), and (2.6), we show that

$$\sum_{i\in\mathcal{I}}\sum_{j=1}^{N_i}\int_{\Gamma_{ij}} \alpha_i \frac{\partial\varphi_i}{\partial\nu_{ij}}\frac{\partial\varphi_i}{\partial\tau_{ij}} m_i\cdot\tau_{ij}\, d\sigma = 0.$$

Consequently, the right-hand side of (4.2) is reduced to

$$\frac{1}{2}\sum_{A\in\mathcal{A}}\int_A\sum_{i\in\mathcal{I}_A}\alpha_i m_i\cdot\nu_i\left\{-\left(\frac{\partial\varphi_i}{\partial\nu_i}\right)^2 + \left(\frac{\partial\varphi_i}{\partial\tau_i}\right)^2\right\} d\sigma.$$

The terms corresponding to $A \in \mathcal{D}\cup\mathcal{N}_{\text{ext}}$ are those of the right-hand side of (4.1) due to (2.4) and (2.6). For $A \in \mathcal{N}_{\text{int}}$, since φ satisfies (2.6), assumptions (H3) and (H4) yield

$$\int_A \sum_{i\in\mathcal{I}_A}\alpha_i m_i\cdot\nu_i\left\{-\left(\frac{\partial\varphi_i}{\partial\nu_i}\right)^2 + \left(\frac{\partial\varphi_i}{\partial\tau_i}\right)^2\right\} d\sigma \geq 0.$$

This proves (4.1). □

The inequality (4.1) can be applied to the solution of the wave equation in the following way.

PROPOSITION 4.2. *Let* $\varphi \in C([0,T], D(A)) \cap C^1([0,T], V) \cap C^2([0,T], H)$ *be a solution of* (3.1) *with* $f = 0$. *Then setting* $Q = \Omega \times (0,T)$, $\Sigma_A = A \times (0,T)$ *for all* $A \in \mathcal{A}$, *we have*

$$\int_Q (D_t\varphi_R)^2\,dxdt \le \frac{1}{2}\sum_{A\in\mathcal{D}}\int_{\Sigma_A}\alpha_{i_A}m_{i_A}\cdot\nu_{i_A}\left(\frac{\partial\varphi_{Ri_A}}{\partial\nu_{i_A}}\right)^2 d\sigma dt \tag{4.3}$$
$$-\frac{1}{2}\sum_{A\in\mathcal{N}_{\text{ext}}}\int_{\Sigma_A}\alpha_{i_A}m_{i_A}\cdot\nu_{i_A}\left(\frac{\partial\varphi_{Ri_A}}{\partial\tau_{i_A}}\right)^2 d\sigma dt$$
$$+\frac{1}{2}\sum_{A\in\mathcal{N}_{\text{ext}}}\int_{\Sigma_A}m_{i_A}\cdot\nu_{i_A}\left(D_t\varphi_{Ri_A}\right)^2 d\sigma dt + R,$$

where φ_R *is the regular part of* φ *appearing in its decomposition* (3.12) *and* R *is a remainder given by*

$$R = -\int_\Omega D_t\varphi m\cdot\nabla\varphi_R\,dx\,|_0^T \tag{4.4}$$
$$+\sum_{S,n}\int_Q [D_t c_{S,n}\{2S^{S,n} + m\cdot\nabla S^{S,n}\}D_t\varphi_R - c_{S,n}AS^{S,n}m\cdot\nabla\varphi_R]dxdt.$$

Proof. For such a solution φ, an integration by parts with respect to the variable t in $(0,T)$ and the expansion (3.12) of φ lead to

$$\int_Q D_t^2\varphi m\cdot\nabla\varphi_R\,dxdt = \int_\Omega D_t\varphi m\cdot\nabla\varphi_R\,dx\,|_0^T \tag{4.5}$$
$$-\int_Q D_t\varphi m\cdot\nabla D_t\varphi_R\,dxdt = \int_\Omega D_t\varphi m\cdot\nabla\varphi_R\,dx\,|_0^T$$
$$-\int_Q D_t\varphi_R m\cdot\nabla D_t\varphi_R\,dxdt - \sum_{S,n}\int_Q D_t c_{S,n}S^{S,n}m\cdot\nabla D_t\varphi_R\,dxdt.$$

The second term of this right-hand side is transformed, using Green's formula in P_i for all $i \in \mathcal{I}$ (allowed thanks to Theorem 3.5). This leads to

$$\int_Q D_t\varphi_R m\cdot\nabla D_t\varphi_R\,dxdt = -\int_Q (D_t\varphi_R)^2 dxdt$$
$$+\frac{1}{2}\sum_{A\in\mathcal{A}}\int_{\Sigma_A}(D_t\varphi_R)^2\left(\sum_{i\in\mathcal{I}_A}m_i\cdot\nu_i\right)d\sigma dt.$$

Since φ_R satisfies (2.4) and taking into account assumption (H1), we arrive at

$$\int_Q D_t\varphi_R m\cdot\nabla D_t\varphi_R\,dxdt = -\int_Q (D_t\varphi_R)^2 dxdt \tag{4.6}$$
$$+\frac{1}{2}\sum_{A\in\mathcal{N}_{\text{ext}}}\int_{\Sigma_A}(D_t\varphi_R)^2 m_{i_A}\cdot\nu_{i_A}d\sigma dt.$$

The third term of the right-hand side of (4.5) is treated similarly; i.e., we first apply Green's formula on each P_i, use the fact that $S^{S,n}$ and φ_R satisfy (2.5), and take into account hypothesis (H1). This yields

$$\int_Q D_t c_{S,n}S^{S,n}m\cdot\nabla D_t\varphi_R\,dxdt = -R^1_{S,n}, \tag{4.7}$$

where we set

$$R^1_{S,n} := \int_Q D_t c_{S,n}\{2S^{S,n} + m \cdot \nabla S^{S,n}\} D_t \varphi_R \, dxdt.$$

The two identities (4.6) and (4.7) into (4.5) give

$$\int_Q D_t^2 \varphi m \cdot \nabla \varphi_R \, dxdt = \int_\Omega D_t \varphi m \cdot \nabla \varphi_R \, dx \,|_0^T + \int_Q (D_t \varphi_R)^2 dxdt - \frac{1}{2} \sum_{A \in \mathcal{N}_{\text{ext}}} \int_{\Sigma_A} (D_t \varphi_R)^2 m_{i_A} \cdot \nu_{i_A} d\sigma dt + \sum_{S,n} R^1_{S,n}. \tag{4.8}$$

Using the expansion (3.12) of φ, we may write

$$\int_Q A\varphi m \cdot \nabla \varphi_R \, dxdt = \int_Q A\varphi_R m \cdot \nabla \varphi_R \, dxdt + \sum_{S,n} R^2_{S,n},$$

where $R^2_{S,n} = \int_Q c_{S,n} A S^{S,n} m \cdot \nabla \varphi_R \, dxdt$. The application of Lemma 4.1 to φ_R yields

$$\int_Q A\varphi m \cdot \nabla \varphi_R \, dxdt \geq \sum_{S,n} R^2_{S,n} - \frac{1}{2} \sum_{A \in \mathcal{D}} \int_{\Sigma_A} \alpha_{i_A} m_{i_A} \cdot \nu_{i_A} \left(\frac{\partial \varphi_{Ri_A}}{\partial \nu_{i_A}} \right)^2 d\sigma dt + \frac{1}{2} \sum_{A \in \mathcal{N}_{\text{ext}}} \int_{\Sigma_A} \alpha_{i_A} m_{i_A} \cdot \nu_{i_A} \left(\frac{\partial \varphi_{Ri_A}}{\partial \tau_{i_A}} \right)^2 d\sigma dt. \tag{4.9}$$

The sum of (4.8) and (4.9) is the inequality (4.3). □

Let us now set

$$\mathcal{D}^+ = \{A \in \mathcal{D} : m_{i_A} \cdot \nu_{i_A} > 0 \text{ on } A\},$$
$$\mathcal{N}^+_{\text{ext}}(\text{resp.}, \mathcal{N}^-_{\text{ext}}) = \{A \in \mathcal{N}_{\text{ext}} : m_{i_A} \cdot \nu_{i_A} > 0 \text{ (resp., } < 0) \text{ on } A\}.$$

For any $\{\varphi_0, \varphi_1\} \in D(A) \times V$, let $\varphi \in C([0, T], D(A)) \cap C^1([0, T], V)$ be the solution of (3.1) with $f = 0$, which admits the decomposition (3.12), and define

$$|||\{\varphi_0, \varphi_1\}|||^2 = \sum_{A \in \mathcal{D}^+} \int_{\Sigma_A} \left(\frac{\partial \varphi_{Ri_A}}{\partial \nu_{i_A}} \right)^2 d\sigma dt + \sum_{A \in \mathcal{N}^-_{\text{ext}}} \int_{\Sigma_A} \left(\frac{\partial \varphi_{Ri_A}}{\partial \tau_{i_A}} \right)^2 d\sigma dt + \sum_{A \in \mathcal{N}^+_{\text{ext}}} \int_{\Sigma_A} \left(D_t \varphi_{i_A} \right)^2 d\sigma dt + \sum_{S,n} \|c_{S,n}\|^2_{1,(0,T)}. \tag{4.10}$$

We are now ready to establish the main result of this section.

PROPOSITION 4.3. *Let $\varphi \in C([0, T], D(A)) \cap C^1([0, T], V) \cap C^2([0, T], H)$ be a solution of* (3.1) *with $f = 0$. Then there exists a minimal time $T_0 > 0$ such that for all $T > T_0$, there exists a constant $C > 0$ (depending on T but not on φ_0, φ_1) such that*

$$(T - T_0) E_0 \leq C |||\{\varphi_0, \varphi_1\}|||^2, \tag{4.11}$$

where E_0 denotes the energy of φ at time $t = 0$, namely,

$$E_0 = \frac{1}{2}\{\|\varphi_1\|_H^2 + a(\varphi_0, \varphi_0)\}.$$

Proof. According to Theorem 4.3, Remark 4.4, and identity (4.24) of [20], we may write

$$\int_Q |D_t\varphi|^2 dxdt = TE_0 + \frac{1}{2}\int_\Omega D_t\varphi\varphi \, dx \,|_0^T . \tag{4.12}$$

Using the expansion (3.12) of φ, (4.12) becomes

$$\begin{aligned} TE_0 &\le -\frac{1}{2}\int_\Omega D_t\varphi\varphi \, dx \,|_0^T \\ &+ C_1\int_Q |D_t\varphi_R|^2 dxdt + C_1\sum_{S,n}\int_Q |D_t c_{S,n} S^{S,n}|^2 dxdt, \end{aligned}$$

where $C_1 > 0$ depends only on the number of eigenvalues $\lambda_{S,n}$ in $]0, 1/2]$. In view of the inequality (4.3), it remains to be shown that

$$R - \frac{1}{2}\int_\Omega D_t\varphi\varphi \, dx \,|_0^T \le C\{|||\{\varphi_0, \varphi_1\}|||^2 + E_0\}. \tag{4.13}$$

The estimate of the term $\int_\Omega D_t\varphi\varphi \, dx \,|_0^T$ with respect to E_0 is classical (see [15, 7]) and is thus omitted. Let us proceed to the estimation of R. Applying Cauchy–Schwarz's inequality, the first term is estimated as follows:

$$\begin{aligned} &\left|\int_\Omega D_t\varphi m\cdot\nabla\varphi_R \, dx \,|_0^T\right| \le C\sum_{t=0,T}\left(\int_\Omega |D_t\varphi(t)|^2 dx\right)^{1/2}\left(\int_\Omega |\nabla\varphi_R(t)|^2 dx\right)^{1/2} \\ &\le CE_0^{1/2}\sum_{t=0,T}\left\{\left(\int_\Omega |D_t\varphi(t)|^2 dx\right)^{1/2} + \sum_{S,n}|c_{S,n}(t)|\left(\int_\Omega |\nabla S^{S,n}|^2 dx\right)^{1/2}\right\}. \end{aligned}$$

Accordingly, the continuous imbedding $H^1(0, T) \hookrightarrow C([0, T])$ (Sobolev imbedding theorem) and the usual estimate

$$ab \le \frac{1}{2}(a^2 + b^2) \quad \forall a, b \in \mathbf{R} \tag{4.14}$$

lead to

$$\begin{aligned} \left|\int_\Omega D_t\varphi m\cdot\nabla\varphi_R \, dx \,|_0^T\right| &\le C\left\{E_0 + \sum_{S,n}\|c_{S,n}\|_{1,(0,T)}^2\right\} \\ &\le C\{E_0 + |||\{\varphi_0, \varphi_1\}|||^2\}. \end{aligned}$$

Let us estimate $R_{S,n}^1$. The treatment of $R_{S,n}^2$ is similar and then omitted. Cauchy–Schwarz's inequality and the expansion (3.12) of φ yield

$$\begin{aligned} |R_{S,n}^1| &\le C\|c_{S,n}\|_{1,(0,T)}\left\{\left(\int_Q |D_t\varphi|^2 dxdt\right)^{1/2} + \sum_{T,m}\|c_{T,m}\|_{1,(0,T)}\right\} \\ &\le C\left\{\sum_{T,m}\|c_{T,m}\|_{1,(0,T)}^2 + \|c_{S,n}\|_{1,(0,T)}T^{1/2}E_0^{1/2}\right\}, \end{aligned}$$

due to the conservation of energy. Applying the estimate (4.14) with $a = \|c_{S,n}\|_{1,(0,T)}T^{1/2}$ and $b = E_0^{1/2}$, one gets

$$|R^1_{S,n}| \le C\left\{(1+T)\sum_{T,m}\|c_{T,m}\|^2_{1,(0,T)} + E_0\right\}.$$

This completes the proof. □

Let us now fix $T > T_0$ such that the inequality (4.11) holds. Then the application

$$D(A) \times V \to \mathbf{R}^+ : \{\varphi_0, \varphi_1\} \to |||\{\varphi_0, \varphi_1\}|||$$

is a norm stronger than the norm induced by $V \times H$, due to Proposition 4.3. As in [15, 7, 20], we define F as the closure of $D(A) \times V$ for this new norm (obviously, F depends on the points x_{0i} and T), and we have the algebraic and topological inclusions

$$D(A) \times V \hookrightarrow F \hookrightarrow V \times H. \tag{4.15}$$

For the inhomogeneous wave equation (3.1) (i.e., f not necessarily equal to 0), we can now state the following proposition.

PROPOSITION 4.4. *Let $\{\varphi_0, \varphi_1\} \in F$ and $f \in L^1(0, T; D(A))$. Then the unique solution φ of* (3.1) *admits the decomposition* (3.12) *and fulfills*

$$\frac{\partial \varphi_{Ri_A}}{\partial \nu_{i_A}} \in L^2(\Sigma_A) \quad \forall A \in \mathcal{D}^+, \tag{4.16}$$

$$\frac{\partial \varphi_{Ri_A}}{\partial \tau_{i_A}} \in L^2(\Sigma_A) \quad \forall A \in \mathcal{N}^-_{\text{ext}}, \tag{4.17}$$

$$D_t\varphi_{i_A} \in L^2(\Sigma_A) \quad \forall A \in \mathcal{N}^+_{\text{ext}}, \tag{4.18}$$

$$c_{S,n} \in H^1(0,T) \quad \forall S \in \mathcal{S},\ 0 < \lambda_{S,n} \le 1/2. \tag{4.19}$$

Moreover, there exists a constant $C > 0$ (independent of $\{\varphi_0, \varphi_1\}$ and f) such that

$$\begin{aligned}&\left\{\sum_{A\in\mathcal{D}^+}\int_{\Sigma_A}\left(\frac{\partial\varphi_{Ri_A}}{\partial\nu_{i_A}}\right)^2 d\sigma dt + \sum_{A\in\mathcal{N}^-_{\text{ext}}}\int_{\Sigma_A}\left(\frac{\partial\varphi_{Ri_A}}{\partial\tau_{i_A}}\right)^2 d\sigma dt\right.\\ &\left.+\sum_{A\in\mathcal{N}^+_{\text{ext}}}\int_{\Sigma_A}\left(D_t\varphi_{i_A}\right)^2 d\sigma dt + \sum_{S,n}\|c_{S,n}\|^2_{1,(0,T)}\right\}^{1/2}\\ &\le C\{|||\{\varphi_0,\varphi_1\}||| + \|f\|_{L^1(0,T;D(A))}\}.\end{aligned} \tag{4.20}$$

Proof. We argue as in Theorem 5.6 of [7]: Using the uniqueness of the solution of (3.1) and the linearity of the wave equation, we may write $\varphi = \varphi^{(1)} + \varphi^{(2)}$, where $\varphi^{(1)}$ (resp., $\varphi^{(2)}$) corresponds to the Cauchy data $\{\varphi_0, \varphi_1\}$ (resp., the right-hand side f). The conclusion for $\varphi^{(1)}$ follows from the definition of the space F. On the other hand, Theorem 3.6 and the usual trace theorems yield the desired properties of $\varphi^{(2)}$. □

We finish this section by giving two examples of 2-d networks for which the conditions (H1)–(H4) are satisfied.

Example 1. First we recall the example considered in [23]. Let $\Omega = P_1 \cup P_2$ be a bounded polygonal domain of $\mathbf{R}^2$, divided into two connected parts, P_1 and P_2, separated by a piecewise polygonal line I (one can readily check that Ω is then a 2-d network). If $m_1(x) = m_2(x) = x - x_0$ for some $x_0 \in \mathbf{R}^2$, one can check that the conditions (H1)–(H4) are reduced to

$$m_1 \cdot \nu_1(\alpha_1 - \alpha_2) \ge 0 \text{ on } I.$$

This condition means that $m_1 \cdot \nu_1$ has to keep the same sign as $(\alpha_1 - \alpha_2)$ along I. Examples of domains P_1, P_2 for which this condition is fulfilled are then easily built.

Example 2. Let $\Omega = \cup_{i=1}^5 P_i$ be included in $\mathbf{R}^3$, where the P_i's are defined by

$$P_i = \{(x, y, 0) : 0 < y < 1,\ i - 2 < x < i - 1\}, \qquad i = 1, 2, 3,$$
$$P_i = \{(i - 4, y, z) : 0 < y < 1,\ 0 < z < 1\}, \qquad i = 4, 5.$$

Take $X_0 = (x_0, y_0, 0) \in \mathbf{R}^3$ and $m_i(x) = \text{proj}_{\Pi_i}(x - X_0)$ (proj_{Π_i} means the orthogonal projection on the plane Π_i). If $\alpha_3 \leq \alpha_2 \leq \alpha_1$, then one can check that (H1)–(H4) are satisfied if $x_0 \leq 0$.

5. Weak solutions of the wave equation. We transpose Proposition 4.4 to get the next theorem.

THEOREM 5.1. *For all $u_0 \in H$, $u_1 \in V'$, $w_A \in L^2(\Sigma_A)$, $A \in \mathcal{D}^+ \cup \mathcal{N}_{\text{ext}}^+ \cup \mathcal{N}_{\text{ext}}^-$, and all $w_{S,n} \in H^1(0, T)$, $S \in \mathcal{S}$, $0 < \lambda_{S,n} \leq 1/2$, there exist unique $u \in L^\infty(0, T; D(A)')$, $\{\psi_1, \psi_0\} \in F'$, which are solutions of*

$$\begin{aligned}(5.1)\qquad &\int_0^T \langle u(t), f(t)\rangle_{D(A)'-D(A)} dt + \langle\{\psi_1, \psi_0\}, \{\varphi_0, -\varphi_1\}\rangle_{F'-F} \\ &= \langle u_1, \varphi(0)\rangle_{V'-V} - \langle u_0, \varphi'(0)\rangle_{H'-H} - \sum_{A\in\mathcal{D}^+} \int_{\Sigma_A} w_A \frac{\partial \varphi_{Ri_A}}{\partial \nu_{i_A}}\, d\sigma dt \\ &- \sum_{A\in\mathcal{N}_{\text{ext}}^-} \int_{\Sigma_A} w_A \frac{\partial \varphi_{Ri_A}}{\partial \tau_{i_A}}\, d\sigma dt - \sum_{A\in\mathcal{N}_{\text{ext}}^+} \int_{\Sigma_A} w_A D_t \varphi_{i_A}\, d\sigma dt \\ &- \sum_{S,n} \int_0^T \{w_{S,n} c_{S,n} + w'_{S,n} c'_{S,n}\} dt\end{aligned}$$

for all $f \in L^1(0, T; D(A))$, $\{\varphi_0, -\varphi_1\} \in F$, where $\varphi = \varphi_R + \sum_{S,n} c_{S,n} S^{S,n}$ is the unique solution of

$$(5.2)\qquad \begin{cases} \varphi \in C([0, T], V) \cap C^1([0, T], H), \\ \varphi''(t) + A\varphi(t) = f(t), t \in [0, T], \\ \varphi(T) = \varphi_0, \varphi'(T) = \varphi_1. \end{cases}$$

The interpretation of (5.1) in terms of partial differential equations is delicate and will be given for more regular data. First we need a density result.

LEMMA 5.2. *Let us denote*

$$K = \{w \in C^\infty([0, T]) : w'''(0) = w'(0) = w'(T) = 0 \text{ and } w''(0) = w(0)\}.$$

Then K is dense in $H^1(0, T)$.

Proof. Let $u \in K^\perp$. Then it fulfills

$$u'' - u = 0 \text{ in } \mathcal{D}'(0, T).$$

Consequently, $u \in C^\infty([0, T])$ and is then a linear combination of e^t and e^{-t}. Since the mapping

$$K \to \mathbf{R}^2 : w \longmapsto (w(0), w(T))$$

is onto, we conclude that $u \equiv 0$. □

The next trace lifting result is easily proved by using the simple geometry of Ω and is left to the reader.

LEMMA 5.3. *Let* $w_A \in \mathcal{D}(\Sigma_A)$, $A \in \mathcal{D}^+ \cup \mathcal{N}_{\text{ext}}^+ \cup \mathcal{N}_{\text{ext}}^-$. *Then there exists* $v \in \mathcal{D}(0, T, \prod_{i\in\mathcal{I}} C^\infty(\overline{P}_i))$ *fulfilling* (2.5) *and* (2.6) *for all* $A \in \mathcal{N}_{\text{int}}$, *and*

$$v_{i_A} = \begin{cases} \alpha_{i_A}^{-1} w_A \text{ on } A \quad \forall A \in \mathcal{D}^+, \\ 0 \text{ on } A \quad \forall A \in \mathcal{D} \setminus \mathcal{D}^+. \end{cases} \tag{5.3}$$

$$\frac{\partial v_{i_A}}{\partial \nu_{i_A}} = \begin{cases} \alpha_{i_A}^{-1} D_t w_A \text{ on } A \quad \forall A \in \mathcal{N}_{\text{ext}}^+, \\ \alpha_{i_A}^{-1} \frac{\partial w_A}{\partial \tau_{i_A}} \text{ on } A \quad \forall A \in \mathcal{N}_{\text{ext}}^-, \\ 0 \text{ on } A \quad \forall A \in \mathcal{N}_{\text{ext}} \setminus (\mathcal{N}_{\text{ext}}^+ \cup \mathcal{N}_{\text{ext}}^-). \end{cases} \tag{5.4}$$

THEOREM 5.4. *Let* $u \in L^\infty(0, T; D(A)')$, $\{\psi_1, \psi_0\} \in F'$ *be the unique solutions of* (5.1) *with data* $u_0 \in V$, $u_1 \in H$, $w_A \in \mathcal{D}(\Sigma_A)$, $A \in \mathcal{D}^+ \cup \mathcal{N}_{\text{ext}}^+ \cup \mathcal{N}_{\text{ext}}^-$, *and* $w_{S,n} \in K$, $S \in \mathcal{S}$, $0 < \lambda_{S,n} \le 1/2$. *Then* $u \in C^1([0, T], H)$ *satisfies* (5.3) *and* (5.5)–(5.7):

$$u_i'' - \alpha_i \Delta u_i = 0 \text{ in } \mathcal{D}'(P_i \times (0, T)) \quad \forall i \in \mathcal{I}, \tag{5.5}$$

$$u(0) = u_0, \qquad u'(0) = u_1, \tag{5.6}$$

$$u(T) = \psi_0, \qquad u'(T) = \psi_1. \tag{5.7}$$

Proof. We proceed as in Theorem 5.3 of [20] with the necessary adaptations. Let us fix $v \in \mathcal{D}(0, T, \prod_{i\in\mathcal{I}} C^\infty(\overline{P}_i))$ obtained in Lemma 5.3 and set

$$\begin{aligned} \hat{w}_{S,n} &= w_{S,n} - w_{S,n}'' \\ &\quad - \sum_{A\in\mathcal{D}^+} \int_A w_A \frac{\partial S^{S,n}}{\partial \nu_{i_A}} d\sigma - \sum_{A\in\mathcal{N}_{\text{ext}}^-} \int_A w_A \frac{\partial S^{S,n}}{\partial \tau_{i_A}} d\sigma. \end{aligned} \tag{5.8}$$

Since the w_A's are regular and $w_{S,n} \in C^\infty([0, T])$, $\hat{w}_{S,n}$ also belongs to $C^\infty([0, T])$. Let $K^{S,n} \in L^2(\Omega)$ be the dual singular function defined in [22] and recalled in Theorem 2.2, and define

$$f_i = v_i'' - \alpha_i \Delta v_i + \sum_{S,n} \hat{w}_{S,n}'' K_i^{S,n} \quad \forall i \in \mathcal{I}.$$

Since $f \in L^2(0, T; H)$, Lemma I.3.4 of [15] guarantees the existence of a unique solution $\psi \in C([0, T], V) \cap C^1([0, T], H) \cap H^2(0, T; V')$ of

$$\begin{cases} \langle \psi''(t), w \rangle + a(\psi(t), w) \\ \qquad = -\int_\Omega f(t) w dx, \text{ a.e. } t \in [0, T], \quad \forall w \in V, \\ \psi(0) = u_0, \quad \psi'(0) = u_1. \end{cases} \tag{5.9}$$

Let us now show that

$$u = \psi + v + \sum_{S,n} \hat{w}_{S,n} K^{S,n} \tag{5.10}$$

is the unique solution of (5.1) when $\psi_0 = u(T)$, $\psi_1 = u'(T)$. Note that the inclusion $w_{S,n} \in K$ leads to the initial conditions

$$u(0) = \psi(0) = u_0, \qquad u'(0) = \psi'(0) = u_1,$$

which is (5.6).

From the density of $\mathcal{D}(0, T; D(A^{3/2}))$ into $L^1(0, T; D(A))$, it suffices to check (5.1) for $\varphi \in C([0, T], D(A^2)) \cap C^1([0, T], D(A^{3/2})) \cap C^2([0, T], D(A))$. Since $\psi \in H^2(0, T; V')$, the integrations by parts over $(0, T)$ are allowed. Taking into account the initial conditions satisfied by φ and u, we get

$$\begin{aligned}(5.11)\qquad &\int_0^T \langle u(t), \varphi''(t) + A\varphi(t)\rangle dt - \langle u(T), \varphi_1\rangle + \langle u'(T), \varphi_0\rangle \\ &= \langle u_1, \varphi(0)\rangle - \langle u_0, \varphi'(0)\rangle \\ &+ \int_0^T \{\langle u''(t), \varphi(t)\rangle + \langle u(t), A\varphi(t)\rangle_{D(A)'-D(A)}\} dt.\end{aligned}$$

Let us now transform the term $\langle u(t), A\varphi(t)\rangle_{D(A)'-D(A)}$: from the expansion (5.10) of u, we may write

$$\begin{aligned}(5.12)\qquad \langle u, A\varphi\rangle_{D(A)'-D(A)} &= a(\psi, \varphi) + \int_\Omega vA\varphi \, dx \\ &+ \sum_{S,n} \hat{w}_{S,n} \int_\Omega K^{S,n} A\varphi \, dx.\end{aligned}$$

Since v vanishes in a neighbourhood of the vertices, Green's formula and Lemma 5.3 yield

$$\begin{aligned}\int_\Omega vA\varphi \, dx &= \int_\Omega Av\varphi \, dx + \sum_{A\in\mathcal{N}_{\text{ext}}^+} \int_A D_t w_A \varphi_{i_A} d\sigma \\ &+ \sum_{A\in\mathcal{N}_{\text{ext}}^-} \int_A \frac{\partial w_A}{\partial \tau_{i_A}} \varphi_{i_A} d\sigma - \sum_{A\in\mathcal{D}^+} \int_A w_A \frac{\partial \varphi_{i_A}}{\partial \nu_{i_A}} d\sigma.\end{aligned}$$

Integrating this identity with respect to t and applying some integrations by parts on Σ_A, we arrive at

$$\begin{aligned}(5.13)\qquad \int_Q vA\varphi \, dxdt &= \int_Q Av\varphi \, dxdt - \sum_{A\in\mathcal{N}_{\text{ext}}^+} \int_{\Sigma_A} w_A D_t \varphi_{i_A} d\sigma dt \\ &- \sum_{A\in\mathcal{N}_{\text{ext}}^-} \int_{\Sigma_A} w_A \frac{\partial \varphi_{i_A}}{\partial \tau_{i_A}} d\sigma dt - \sum_{A\in\mathcal{D}^+} \int_{\Sigma_A} w_A \frac{\partial \varphi_{i_A}}{\partial \nu_{i_A}} d\sigma dt.\end{aligned}$$

Moreover, Theorem 2.2 implies that

$$(5.14)\qquad \int_\Omega A\varphi K^{S,n} dx = -c_{S,n},$$

when φ admits the decomposition (3.12).

Taking into account (5.12)–(5.14), the identity (5.11) becomes

$$\begin{aligned}(5.15)\qquad &\int_0^T \langle u(t), \varphi'' + A\varphi\rangle dt - \langle u(T), \varphi_1\rangle + \langle u'(T), \varphi_0\rangle \\ &= \langle u_1, \varphi(0)\rangle - \langle u_0, \varphi'(0)\rangle - \sum_{A\in\mathcal{N}_{\text{ext}}^+} \int_{\Sigma_A} w_A D_t \varphi_{i_A} d\sigma dt \\ &- \sum_{A\in\mathcal{N}_{\text{ext}}^-} \int_{\Sigma_A} w_A \frac{\partial \varphi_{i_A}}{\partial \tau_{i_A}} d\sigma dt - \sum_{A\in\mathcal{D}^+} \int_{\Sigma_A} w_A \frac{\partial \varphi_{i_A}}{\partial \nu_{i_A}} d\sigma dt \\ &- \sum_{S,n} \int_0^T \hat{w}_{S,n}(t) c_{S,n}(t) \, dt.\end{aligned}$$

This proves that u is the solution of (5.1) with $u(T) = \psi_0$, $u'(T) = \psi_1$ if we show that

$$\sum_{A\in\mathcal{N}_{\text{ext}}^-}\int_{\Sigma_A} w_A\frac{\partial\varphi_{i_A}}{\partial\tau_{i_A}}d\sigma dt + \sum_{A\in\mathcal{D}^+}\int_{\Sigma_A} w_A\frac{\partial\varphi_{i_A}}{\partial\nu_{i_A}}d\sigma dt$$
$$+\sum_{S,n}\int_0^T \hat{w}_{S,n}(t)c_{S,n}(t)\,dt = \sum_{A\in\mathcal{D}^+}\int_{\Sigma_A} w_A\frac{\partial\varphi_{Ri_A}}{\partial\nu_{i_A}}d\sigma dt$$
$$+\sum_{A\in\mathcal{N}_{\text{ext}}^-}\int_{\Sigma_A} w_A\frac{\partial\varphi_{Ri_A}}{\partial\tau_{i_A}}d\sigma dt + \sum_{S,n}\int_0^T (w_{S,n}(t)c_{S,n}(t) + w'_{S,n}(t)c'_{S,n}(t))\,dt.$$

This last identity is proved using the expansion (3.12) of φ, an integration by parts in time (taking into account the boundary conditions $w'_{S,n}(0) = w'_{S,n}(T) = 0$, due to the inclusion $w_{S,n} \in K$) and the expression (5.8) of $\hat{w}_{S,n}$.

Let us finally note that the boundary conditions (5.3) come from (5.10); the boundary conditions (5.3) which are fulfilled by v and the fact that ψ and $K^{S,n}$ both satisfy (2.4). □

Roughly speaking, ψ introduced in the above proof satisfies (2.6); therefore, one can say that u satisfies (2.6) for all $A \in \mathcal{N}_{\text{int}}$ and (5.4) in a weak sense.

Up to now, the functions $w_{S,n}$ appear neither in the partial differential equation (5.5) nor in the boundary conditions (5.3) and (5.4) fulfilled by u. This is because we were working in the setting of $\mathcal{D}'(P_i \times (0, T))$.

DEFINITION 5.5. *Let $D := \{\varphi = \Phi_{|\Omega} : \Phi \in \mathcal{D}(\mathbf{R}^n)\}$, and for all $S \in \mathcal{S}$, $0 < \lambda_{S,n} \le 1/2$, introduce the distribution $T^{S,n}$ in D' by*

$$\langle T^{S,n}, \varphi\rangle = \int_\Omega \alpha K^{S,n}\triangle\varphi\,dx - \sum_{A\in\mathcal{N}}\int_A \sum_{i\in\mathcal{I}_A}\alpha_i\frac{\partial\varphi_i}{\partial\nu_i}K_i^{S,n}\,d\sigma$$
$$+\sum_{A\in\mathcal{D}}\int_A \alpha_{i_A}\frac{\partial K_{i_A}^{S,n}}{\partial\nu_{i_A}}\left(\varphi_{i_A} - \eta_S\varphi_{i_A}(S)\right)d\sigma. \tag{5.16}$$

Note that $T^{S,n}$ is well defined due to the decomposition (2.73) in [22] of $K^{S,n}$.

LEMMA 5.6. *For all $S \in \mathcal{S}$, $0 < \lambda_{S,n} \le 1/2$, we have*

$$\langle T^{S,n}, \varphi\rangle = c_{S,n} \tag{5.17}$$

for any $\varphi \in D(A)$, where $c_{S,n}$ is the coefficient of the singularity $S^{S,n}$ of φ appearing in its decomposition (2.13).

The support of $T^{S,n}$ is equal to S in the sense that for any $\varphi \in D$ such that $\varphi \equiv 0$ in a neighbourhood of S, we have

$$\langle T^{S,n}, \varphi\rangle = 0. \tag{5.18}$$

Proof. Since $\varphi \in D(A)$ satisfies (2.4), (2.5), (2.6), and $\varphi_{i_A}(S) = 0$, for all $S \in \bar{A}$ such that $A \in \mathcal{D}$, we get

$$\langle T^{S,n}, \varphi\rangle = \int_\Omega \alpha K^{S,n}\triangle\varphi\,dx = c_{S,n}$$

due to Theorem 2.2.

On the other hand, for any $\varphi \in D$ such that $\varphi = 0$ on $B(S, \delta) \cap \Omega$ for some $\delta > 0$, applying Green's formula on $P_i \setminus B(S, \delta/2)$ for all $i \in \mathcal{I}$, one gets (5.18). □

According to Lemma 5.6, the identity (5.15) implies that

$$\begin{aligned}(5.19)\qquad &\int_0^T \langle u(t), \varphi'' + A\varphi\rangle dt - \langle u(T), \varphi_1\rangle + \langle u'(T), \varphi_0\rangle \\ &= \langle u_1, \varphi(0)\rangle - \langle u_0, \varphi'(0)\rangle - \sum_{A\in\mathcal{N}_{\text{ext}}^+}\int_{\Sigma_A} w_A D_t\varphi_{i_A} d\sigma dt \\ &- \sum_{A\in\mathcal{N}_{\text{ext}}^-}\int_{\Sigma_A} w_A \frac{\partial\varphi_{i_A}}{\partial\tau_{i_A}} d\sigma dt - \sum_{A\in\mathcal{D}^+}\int_{\Sigma_A} w_A \frac{\partial\varphi_{i_A}}{\partial\nu_{i_A}} d\sigma dt \\ &- \sum_{S,n}\int_0^T \hat{w}_{S,n}(t)\langle T^{S,n}, \varphi\rangle\, dt \quad \forall\varphi\in C^\infty([0,T], D).\end{aligned}$$

This means that u satisfies

$$(5.20)\qquad u'' + Au = -\sum_{S,n}\hat{w}_{S,n}T^{S,n} \text{ in } C^\infty([0,T], D)'$$

if we admit that u satisfies (5.3) and (5.4).

Let us remark that (5.20) covers (5.5) since

$$\langle T^{S,n}, \varphi\rangle = 0 \quad \forall\varphi\in\prod_{i\in\mathcal{I}}\mathcal{D}(P_i\times(0,T)).$$

All these considerations lead us to call the solution u of (5.1) the weak solution of (5.20), (5.6), (5.3), and (5.4). In order to give a meaning to the final conditions (5.7), we need the next regularity result.

THEOREM 5.7. *Under the assumption of Theorem* 5.1, *let* u, $\{\psi_1, \psi_0\}$ *be the solutions of* (5.1). *Then* $u\in C([0,T], D(A)')\cap C^1([0,T], D(A^{3/2})')$ *and* u *satisfies the final conditions* (5.7).

Proof. We argue as at the end of paragraph 5 of [20]: first we reduce the wave equation to the first-order equation

$$(5.21)\qquad \begin{cases}\Phi' + B\Phi = g,\\ \Phi(0) = \Phi_0,\end{cases}$$

where B is an operator from $\mathcal{H} = V\times H$ into itself defined by $D(B) = D(A)\times V$ and for all $\Phi = (\varphi, \xi)\in D(B)$, $B\Phi = (-\xi, A\varphi)$. Using Lemma 5.4 of [20] and Theorems 3.5 and 3.6, we can show that if $\Phi = (\varphi, \xi)\in C([0,T], \mathcal{H})$ is the unique solution of (5.21) with $\Phi_0\in F$ and $g\in L^1(0,T; D(A^{3/2})\times D(A))$, then φ admits the decomposition (3.12) and satisfies

$$\begin{aligned}&\left\{\sum_{A\in\mathcal{D}^+}\int_{\Sigma_A}\left(\frac{\partial\varphi_{Ri_A}}{\partial\nu_{i_A}}\right)^2 d\sigma dt + \sum_{A\in\mathcal{N}_{\text{ext}}^-}\int_{\Sigma_A}\left(\frac{\partial\varphi_{Ri_A}}{\partial\tau_{i_A}}\right)^2 d\sigma dt\right.\\ &\left.+\sum_{A\in\mathcal{N}_{\text{ext}}^+}\int_{\Sigma_A}\left(D_t\varphi_{i_A}\right)^2 d\sigma dt + \sum_{S,n}\|c_{S,n}\|^2_{1,(0,T)}\right\}^{1/2}\\ &\le C\{|||\Phi_0||| + \|g\|_{L^1(0,T;D(A^{3/2})\times D(A))}\}.\end{aligned}$$

By transposition and density, we arrive at the conclusion. □

6. The Hilbert uniqueness method. The application of the Hilbert uniqueness method of J.-L. Lions [15] is now standard: first, by Proposition 4.4, for $\{\varphi_0, \varphi_1\} \in F$, there exists a unique solution $\varphi \in C([0,T], V) \cap C^1([0,T], H)$ of (3.1) with $f = 0$ satisfying (4.20). Second, consider $\psi \in L^\infty(0,T, D(A)')$, $\{\chi_1, -\chi_0\} \in F'$, the unique solutions of

$$
\begin{aligned}
(6.1)\qquad & \int_0^T \langle \psi(t), g(t)\rangle dt - \langle \{\chi_1, -\chi_0\}, \{\eta_0, \eta_1\}\rangle \\
&= -\sum_{A\in\mathcal{D}^+} \int_{\Sigma_A} \frac{\partial \varphi_{Ri_A}}{\partial \nu_{i_A}} \frac{\partial \eta_{Ri_A}}{\partial \nu_{i_A}}\, d\sigma dt - \sum_{A\in\mathcal{N}^-_{\text{ext}}} \int_{\Sigma_A} \frac{\partial \varphi_{Ri_A}}{\partial \tau_{i_A}} \frac{\partial \eta_{Ri_A}}{\partial \tau_{i_A}}\, d\sigma dt \\
&\quad - \sum_{A\in\mathcal{N}^+_{\text{ext}}} \int_{\Sigma_A} D_t\varphi_{i_A} D_t\eta_{i_A}\, d\sigma dt - \sum_{S,n} \int_0^T (c_{S,n} d_{S,n} + c'_{S,n} d'_{S,n})\, dt
\end{aligned}
$$

for all $g \in L^1(0,T; D(A))$, $\{\eta_0, \eta_1\} \in F$, where $\eta = \eta_R + \sum_{S,n} d_{S,n} S^{S,n}$ is the unique solution of

$$
(6.2)\qquad \begin{cases} \eta \in C([0,T], V) \cap C^1([0,T], H), \\ \eta''(t) + A\eta(t) = g(t), t \in [0,T], \\ \eta(0) = \eta_0, \qquad \eta'(0) = \eta_1. \end{cases}
$$

Its existence comes from Theorem 5.1, inverting the order of time; moreover, Theorem 5.7 gives a meaning to the initial conditions

$$
\psi(0) = \chi_0, \qquad \psi'(0) = \chi_1.
$$

Accordingly, the next operator,

$$
\Lambda : F \to F' : \{\varphi_0, \varphi_1\} \to \{\chi_1, -\chi_0\},
$$

is well defined and is an isomorphism, because the identity (6.1) with $\eta = \varphi$ yields

$$
\langle \Lambda\{\varphi_0, \varphi_1\}, \{\varphi_0, \varphi_1\}\rangle = |||\{\varphi_0, \varphi_1\}|||^2 \quad \forall \{\varphi_0, \varphi_1\} \in F.
$$

This leads to the main result of this paper.

THEOREM 6.1. *For all $u_0 \in H$, $u_1 \in V'$, there exist $w_A \in L^2(\Sigma_A)$, $A \in \mathcal{D}^+ \cup \mathcal{N}^+_{\text{ext}} \cup \mathcal{N}^-_{\text{ext}}$, and $w_{S,n} \in H^1(0,T)$, $S \in \mathcal{S}$, $0 < \lambda_{S,n} \le 1/2$ such that the weak solution $u \in C([0,T], D(A)') \cap C^1([0,T], D(A^{3/2})')$ of the wave equation* (6.3) (*in the sense of* (5.1)) *satisfies $u(T) = u'(T) = 0$:*

$$
(6.3)\qquad \begin{cases} u''(t) + Au(t) = -\sum_{S,n} \hat{w}_{S,n}(t) T^{S,n}, \qquad t \in [0,T], \\ u(0) = u_0, \qquad u'(0) = u_1, \\ u_{i_A} = \begin{cases} \alpha_{i_A}^{-1} w_A \text{ on } A \quad \forall A \in \mathcal{D}^+, \\ 0 \text{ on } A \quad \forall A \in \mathcal{D} \setminus \mathcal{D}^+, \end{cases} \\ \dfrac{\partial u_{i_A}}{\partial \nu_{i_A}} = \begin{cases} \alpha_{i_A}^{-1} D_t w_A \text{ on } A \quad \forall A \in \mathcal{N}^+_{\text{ext}}, \\ \alpha_{i_A}^{-1} \dfrac{\partial w_A}{\partial \tau_{i_A}} \text{ on } A \quad \forall A \in \mathcal{N}^-_{\text{ext}}, \\ 0 \text{ on } A \quad \forall A \in \mathcal{N}_{\text{ext}} \setminus (\mathcal{N}^+_{\text{ext}} \cup \mathcal{N}^-_{\text{ext}}). \end{cases} \end{cases}
$$

Recall that in the case of smoother controls w_A and $w_{S,n}$, $\hat{w}_{S,n}$ is given by (5.8) (*see Theorem* 5.4).

Proof. Since $\{u_1, -u_0\} \in V' \times H \subset F'$, there exists a unique solution $\{\varphi_0, \varphi_1\} \in F$ of

$$\Lambda\{\varphi_0, \varphi_1\} = \{u_1, -u_0\}.$$

We take the solution φ of (3.1) with $f = 0$, which admits the decomposition (3.12) and then the solution ψ of (6.1). The conclusion follows with $u = \psi$, $w_A = \frac{\partial \varphi_{Ri_A}}{\partial \nu_{i_A}}$ for all $A \in \mathcal{D}^+$, $w_A = D_t \varphi_{i_A}$ for all $A \in \mathcal{N}^+_{\text{ext}}$, $w_A = \frac{\partial \varphi_{Ri_A}}{\partial \tau_{i_A}}$ for all $A \in \mathcal{N}^-_{\text{ext}}$, and $w_{S,n} = c_{S,n}$ for all $S \in \mathcal{S}, 0 < \lambda_{S,n} \leq 1/2$, because of the reversibility of the wave equation and Proposition 4.4. □

Remark. In the above theorem, the boundary controls are classical. The influence of the singularities is translated through the terms $\hat{c}_{S,n}(t)T^{S,n}$, and each of them can be seen as a distributional internal control with a support concentrated at the singular vertex S (due to Lemma 5.6). The introduction of these terms is the price to pay to avoid the regularity hypothesis $D(A) \hookrightarrow \mathcal{H}^{3/2+\varepsilon}(\Omega)$ for some $\varepsilon > 0$ leading to strong geometrical conditions on the domains Ω [12, Chaps. 4 and 7]. On the other hand, the conditions (H1)–(H4) that we imposed are not related to the singularities but linked to the multiplier method, since they were introduced to avoid internal control. Consequently, they also appear for transmission problems without singularities [15, Chap. VI].

7. Lack of controllability for 2-d networks with circuits. In this section, we shall give an example of a 2-d network for which we do not have exact controllability by boundary control on the exterior edges with the help of the Hilbert uniqueness method. Inspired by the results of Lagnese, Leugering, and Schmidt for 1-d networks [12, § II.5.2], we choose a network with circuits. More precisely, let us define $\Omega \subset \mathbf{R}^3$ by

$$\Omega = \bigcup_{i=-5}^{5} P_i,$$

where

$$\begin{aligned}
&P_0 = (0,1) \times (0,1) \times \{0\}, \qquad P_5 = (-1,0) \times (0,1) \times \{1\},\\
&P_1 = \{0\} \times (0,1) \times (0,1), \qquad P_3 = \{-1\} \times (0,1) \times (0,1),\\
&P_2 = (-1,0) \times \{1\} \times (0,1), \qquad P_4 = (-1,0) \times \{0\} \times (0,1),\\
&P_{-k} = \{(x_1, x_2, -x_3) : (x_1, x_2, x_3) \in P_k\}.
\end{aligned}$$

The exterior edges of Ω are the exterior edges of P_0, i.e.,

$$\begin{aligned}
\Gamma_{01} &= (0,1) \times \{0\} \times \{0\},\\
\Gamma_{02} &= \{1\} \times (0,1) \times \{0\},\\
\Gamma_{03} &= (0,1) \times \{1\} \times \{0\}.
\end{aligned}$$

We take $\mathcal{D} = \{\Gamma_{0i}\}_{i=1,2,3}$, which means that we consider Dirichlet boundary conditions on the exterior boundary of Ω. Finally, for simplicity, we fix $\alpha_i \equiv 1$ for all $i \in \mathcal{I} := \{-5, \ldots, 5\}$.

As in [12], the lack of controllability comes from the existence of a special eigenvector $w \neq 0$ of the Laplace operator A on Ω of eigenvalue $\lambda > 0$ then fulfilling

$$-\triangle w_i = \lambda w_i \text{ in } P_i \quad \forall i \in \mathcal{I} \tag{7.1}$$

and the boundary and interface conditions (2.4)–(2.6) and also the supplementary conditions

$$\frac{\partial w_0}{\partial \nu} = 0 \text{ on } \Gamma_{0i} \quad \forall i = 1, 2, 3. \tag{7.2}$$

Indeed, let us consider the 2-d network $\tilde{\Omega} = \cup_{i=1}^{5} P_i$ (with the same P_i as in the definition of Ω). Consider the Laplace operator $\tilde{A}$ on $\tilde{\Omega}$ with Dirichlet boundary conditions on its exterior boundary (corresponding to the edges of the P_i's included in the plane $x_3 = 0$). Take $\tilde{w} = (w_i)_{i=1,\dots,5}$, an eigenvector $\neq 0$ of $\tilde{A}$ of eigenvalue $\lambda > 0$. (In other words, $\tilde{w}$ satisfies (7.1) for $i = 1, \dots, 5$ and Dirichlet boundary conditions.) From Theorem 2.27 of [22], one can show that $\tilde{w} \in \mathcal{H}^{3/2+\varepsilon}(\tilde{\Omega})$ for some $\varepsilon > 0$. We now define w on the whole of Ω by antisymmetry:

$$w_0 \equiv 0,\ w_{-k}(x_1, x_2, -x_3) = -w_k(x_1, x_2, x_3) \quad \forall (x_1, x_2, x_3) \in P_k, \qquad k = 1, \dots, 5.$$

From the inclusion $\tilde{w} \in D(\tilde{A})$, we readily check that $w \in D(A)$ and satisfies (7.1). The property (7.2) is immediate since $w_0 \equiv 0$ in P_0. Finally, w is inherited from $\tilde{w}$ of the regularity $\mathcal{H}^{3/2+\varepsilon}(\Omega)$ for some $\varepsilon > 0$.

We now say that we have exact controllability at time $T > 0$ by Dirichlet control on $\mathcal{D}$ with the help of the Hilbert uniqueness method if there exists a Hilbert space F such that (4.15), Proposition 4.4 and Theorem 5.1 hold (with $\mathcal{D}^+ = \mathcal{D}, \mathcal{N}_{\text{ext}} = \emptyset$) and if moreover the continuous mapping

$$C_T : \prod_{A \in \mathcal{D}} (L^2(\Sigma_A)) \times \prod_{S,n} H^1(0, T) \longrightarrow F' : (w_A)_{A \in \mathcal{D}} \times (w_{S,n})_{S,n} \longmapsto \{\psi_1, \psi_0\},$$

where u, $\{\psi_1, \psi_0\}$ are the unique solutions of (5.1) with $u_0 = u_1 = 0$, is surjective.

From (5.1), we directly see that

$$C_T^*(\{\varphi_0, -\varphi_1\}) = \left(\frac{\partial \varphi_{Ri_A}}{\partial \nu_{i_A}} \right)_{A \in \mathcal{D}} \times (c_{S,n})_{S,n}$$

when $\varphi = \varphi_R + \sum_{S,n} c_{S,n} S^{S,n}$ is the unique solution of (5.2) with $f = 0$.

If C_T is surjective, then $\ker C_T^* = \{0\}$, which, in our case, is impossible because the pair $\{w, 0\} \in D(A) \times V \hookrightarrow F$ belongs to $\ker C_T^*$. Indeed the unique solution η of

$$\begin{cases} \eta''(t) + A\eta(t) = 0, & t \in [0, T], \\ \eta(T) = w, \qquad \eta'(T) = 0 \end{cases}$$

is given by $\eta(t) = w \cos(\sqrt{\lambda}(T - t))$. Consequently, the property (7.2) satisfied by w and its regularity leads to

$$\frac{\partial \eta_{i_A}}{\partial \nu_{i_A}} = 0 \text{ on } A \quad \forall A \in \mathcal{D},$$

$$c_{S,n} = 0 \quad \forall S \in \mathcal{S},\ 0 < \lambda_{S,n} \leq 1/2.$$

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions,* Dover, New York, 1965.

[2] J. M. Ball and M. Slemrod, *Nonharmonic Fourier series and the stabilization of distributed semi-linear control systems,* Comm. Pure Appl. Math., 32 (1979), pp. 555–587.

[3] P. G. Ciarlet, H. Le Dret, and R. Nzengwa, *Junctions between three-dimension and two-dimensional linearly elastic structures*, J. Math. Pures Appl., 68 (1989), pp. 261–295.

[4] M. Dauge and S. Nicaise, *Oblique derivative and interface problems on polygonal domains and networks,* Comm. Partial Differential Equations, 14 (1989), pp. 1147–1192.

[5] M. Dobrowolski, *Numerical Approximation of Elliptic Interface and Corner Problems,* Habilitationsschrift, Bonn, 1981.

[6] P. Grisvard, *Elliptic Problems in Nonsmooth Domains,* Monographs and Studies in Mathematics 21, Pitman, Boston, 1985.

[7] P. GRISVARD, *Contrôlabilité exacte des solutions de l'équation des ondes en présence de singularités,* J. Math. Pures Appl., 68 (1989), pp. 215–259.
[8] A. HEIBIG AND M. MOUSSAOUI, *Exact Controllability of the Wave Equation for Domains with Slits and for Mixed Boundary Conditions*, Preprint, ENS Lyon, 1993.
[9] R. B. KELLOGG, *Singularities in interface problems,* in Synspade 70, B. Hubbard, ed., Academic Press, New York, 1971, pp. 351–400.
[10] J. E. LAGNESE, G. LEUGERING, AND E. J. P. G. SCHMIDT, *Modeling of dynamic networks of thin thermoelastic beams*, Math. Methods Appl. Sci., 16 (1993), pp. 327–358.
[11] ———, *Control of planar networks of Timoshenko beams*, SIAM J. Control Optim., 31 (1993), pp. 780–811.
[12] ———, *Modeling, Analysis and Control of Dynamic Elastic Multi-link Structures*, Birkhäuser, Boston, 1994.
[13] H. LE DRET, *Problèmes variationnels dans les multi-domaines. Modélisation des jonctions et applications*, RMA 19, Masson, Paris, 1991.
[14] K. LEMRABET, *Régularité de la solution d'un problème de transmission,* J. Math. Pures Appl., 56 (1977), pp. 1–38.
[15] J.-L. LIONS, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués,* tome 1, RMA 8, Masson, Paris, 1988.
[16] M. MOUSSAOUI AND B. SADALLAH, *Régularité des coefficients de propagation des singularités pour l'équation de la chaleur dans un polygone plan,* C. R. Acad. Sci. Paris Sér. I Math., 293 (1981), pp. 297–300.
[17] M. MOUSSAOUI AND V. H. TRAN, *Sur les coefficients de singularité des solutions de l'équation des ondes dans un polygone plan,* C. R. Acad. Sci. Paris Sér. I Math., 316 (1993), pp. 257–260.
[18] M.-T. NIANE AND O. SECK, *Contrôlabilité exacte de l'équation des ondes avec conditions mêlées,* C. R. Acad. Sci. Paris Sér. I Math., 318 (1994), pp. 945–948.
[19] ———, *Contrôlabilité exacte frontière de l'équation des ondes en présence de fissures,* C. R. Acad. Sci. Paris Sér. I Math., 316 (1993), pp. 695–700.
[20] S. NICAISE, *Exact controllability of a pluridimensional coupled problem*, Rev. Mate. Univ. Complut. Madrid, 5 (1992), pp. 91–135.
[21] ———, *About the Lamé system in a polygonal or a polyhedral domain and a coupled problem between the Lamé system and the plate equation* II*: Exact controllability*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 20 (1993), pp. 163–191.
[22] ———, *Polygonal Interface Problems,* Series "Methoden und Verfahren der Mathematischen Physik" 39, Peter Lang Verlag, Frankfurt am Main, 1993.
[23] ———, *Contrôlabilité exacte frontière des problèmes de transmission avec singularités,* C. R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 663–668.
[24] J. P. PUEL AND E. ZUAZUA, *Exact controllability for a model of multidimensional flexible structure*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 323–344.
[25] E. J. P. G. SCHMIDT, *On the modelling and exact controllability of networks of vibrating strings*, SIAM J. Control Optim., 30 (1992), pp. 229–245.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1533–1554, September 1996

004

APPROXIMATE FEEDBACK LINEARIZATION: A HOMOTOPY OPERATOR APPROACH*

ANDRZEJ BANASZUK† AND JOHN HAUSER‡

Abstract. In this paper, we present an approach for finding feedback linearizable systems that approximate a given single-input nonlinear system on a given compact region of the state space. First, we show that if the system is close to being involutive, then it is also close to being linearizable. Rather than working directly with the characteristic distribution of the system, we work with characteristic one-forms, i.e., with the one-forms annihilating the characteristic distribution. We show that homotopy operators can be used to decompose a given characteristic one-form into an exact and an antiexact part. The exact part is used to define a change of coordinates to a normal form that looks like a linearizable part plus nonlinear perturbation terms. The nonlinear terms in this normal form depend continuously on the antiexact part, and they vanish whenever the antiexact part does. Thus, the antiexact part of a given characteristic one-form is a measure of nonlinearizability of the system. If the nonlinear terms are small, by neglecting them we obtain a linearizable system approximating the original system. One can design control for the original system by designing it for the approximating linearizable system and applying it to the original one. We apply this approach for design of locally stabilizing feedback laws for nonlinear systems that are close to being linearizable.

Key words. nonlinear systems, feedback linearization, differential-geometric methods, differential forms

AMS subject classifications. 93C10, 93B29, 93B18, 53C65

1. Introduction.

Consider a single-input system

$$\dot{x} = f(x) + g(x)u, \tag{1}$$

where f and g are smooth vector fields defined on a compact contractible region $\mathcal{M}$ of R^n containing the origin. (Typically, $\mathcal{M}$ is a closed ball in R^n.) We assume that $f(0) = 0$, i.e., that the origin is an equilibrium for $\dot{x} = f(x)$. The classical problem of feedback linearization can be stated as follows: find in a neighborhood of the origin a smooth change of coordinates $z = \Phi(x)$ (a local diffeomorphism) and a smooth feedback law $u = k(x) + l(x)u_{\text{new}}$ such that the closed-loop system in the new coordinates with new control is linear,

$$\dot{z} = Az + Bu_{\text{new}}, \tag{2}$$

and controllable. We usually require that $\Phi(0) = 0$.

We assume that the system (1) has *the linear controllability* property

$$\dim \operatorname{span} \{g, ad_f g, \ldots, ad_f^{n-1} g\} = n \ \forall x \in \mathcal{M} \tag{3}$$

(where $ad_f^i g$ are iterated Lie brackets of f and g). We define the *characteristic distribution* for (1):

$$D := \operatorname{span} \{g, ad_f g, \ldots, ad_f^{n-2} g\}. \tag{4}$$

(It is an $n-1$-dimensional smooth distribution by assumption of linear controllability (3).) We shall call any nowhere vanishing one-form ω annihilating D a *characteristic one-form for* (1). All the characteristic one-forms for (1) can be represented as multiples of some fixed

*Received by the editors January 10, 1994; accepted for publication (in revised form) April 26, 1995. This research was sponsored in part by NSF grants PYI ECS-9157835 and DMS-9207703.

†Department of Mathematics, University of California, Davis, CA 95616-8633 (banaszuk@math.ucdavis.edu).

‡Electrical and Computer Engineering, University of Colorado, Boulder, CO 80309-0425 (hauser@boulder.colorado.edu).

characteristic one-form ω_0 by a smooth nowhere vanishing function (zero-form) β. Suppose that there is a nonvanishing β so that $\beta\omega_0$ is exact, i.e., $\beta\omega_0 = d\alpha$ for some smooth function α. (d denotes the exterior derivative.) Then ω_0 is called *integrable* and β is called an *integrating factor for* ω_0. The following result is standard [14, 15].

THEOREM 1.1. *Suppose that the system* (1) *has the linear controllability property* (3) *on* $\mathcal{M}$. *Let* D *be the characteristic distribution and* ω_0 *be a characteristic one-form for* (1). *The following statements are equivalent:*

1. *Equation* (1) *is feedback linearizable in a neighborhood of the origin in* $\mathcal{M}$.
2. D *is involutive in a neighborhood of the origin in* $\mathcal{M}$.
3. ω_0 *is integrable in a neighborhood of the origin in* $\mathcal{M}$.

As is well known, a generic nonlinear system is not feedback linearizable for $n > 2$. However, in some cases, it may make sense to consider *approximate* feedback linearization. Namely, if one can find a feedback linearizable system *close* to (1), there is hope that a control designed for the feedback linearizable system and applied to (1) will give satisfactory performance if the feedback linearizable system is *close* enough to (1). The first attempt in this direction goes back to [16], where it was proposed to apply to (1) a change of variables and feedback that yield a system of the form

$$\dot{z} = Az + Bu_{\text{new}} + \mathcal{O}(z, u_{\text{new}}), \tag{5}$$

where the term $\mathcal{O}(z, u_{\text{new}})$ contains higher-order terms. The aim was to make $\mathcal{O}(z, u_{\text{new}})$ of as high order as possible. Then we can say that the system (1) is approximately feedback linearized in a small neighborhood of the origin. Reference [13] introduced a new algorithm to achieve the same goal with fewer steps.

Another idea has been investigated in [11]. Roughly speaking, the idea was to neglect nonlinearities in (1) responsible for the failure of the involutivity condition in Theorem 1.1. This approach happened to be successful in the *ball-and-beam* system, when neglect of centrifugal force acting on ball yielded a feedback linearizable system. Application of a control scheme designed for the system with centrifugal force neglected to the original system gave much better results than applying a control scheme based on classical Jacobian linearization. This approach has been further investigated in [10, 18, 19] for the purpose of approximate feedback linearization about the manifold of constant operating points. However, a general approach to deciding which nonlinearities should be neglected to get the best approximation has not been set forth.

In [17] a design of a control law for the systems with cubic nonlinearities is proposed which uses a change of variables that directly minimizes the terms P and Q in L_2.

All of the above-mentioned work (except [17]) dealt with applying a change of coordinates and a preliminary feedback so that the resulting system looks like linearizable part plus nonlinear terms of highest possible order around an equilibrium point or an equilibrium manifold. However, in many applications one requires a large region of operation for the nonlinearizable system. In such a case, demanding the nonlinear terms to be neglected to be of highest possible order may, in fact, be quite undesirable. One might prefer that the nonlinear terms to be neglected be small in a uniform sense over the region of operation. In the present paper we propose an approach to approximate feedback linearization that uses a change of coordinates and a preliminary feedback to put a system (1) in *a perturbed Brunovsky form*,

$$\dot{z} = Az + Bu_{\text{new}} + P(z) + Q(z)u_{\text{new}}, \tag{6}$$

where $P(z)$ and $Q(z)$ vanish at $z = 0$ and are "small" on $\mathcal{M}$. We obtain upper bounds on uniform norms of P and Q (depending on some measures of noninvolutivity of D) on any compact, contractible $\mathcal{M}$.

Our approach is an indirect one. We begin with approximating characteristic one-forms by exact forms using *homotopy operators*. Namely, on any contractible region $\mathcal{M}$ one can define a linear operator H that satisfies

$$\omega = d(H\omega) + Hd\omega \tag{7}$$

for any form ω.

Note that the homotopy identity (7) allows to decompose any given one-form into the *exact part* $d(H\omega)$ and an "error" part $\epsilon := Hd\omega$, which we will call the *antiexact part of* ω. For given ω_0 annihilating D and a scaling factor β we define $\alpha_\beta := H\beta\omega_0$ and $\epsilon_\beta := Hd\beta\omega_0$. Note that the one-form ϵ_β measures how exact $\omega_\beta := \beta\omega_0$ is. If it is zero, then ω_β is exact and the system (1) is linearizable, and the zero-form α_β and its first $n-1$ Lie derivatives along f are the new coordinates. In the case that ω_0 is not exactly integrable, i.e., when no exact integrating factor β exists, we choose β so that $d\beta\omega_0$ is *smallest* in some sense (because this also makes ϵ_β small). We will call this β an *approximate integrating factor for* ω_0. We will use the zero-form α_β and its first $n-1$ Lie derivatives along f as the new coordinates as in the linearizable case. In those new coordinates the system (1) is in the form

$$\dot{z} = Az + Bru + Bp + Eu, \tag{8}$$

where r and p are smooth functions, $r \neq 0$ around the origin, and the term E (the obstruction to linearizablity) depends *linearly* on ϵ_β and some of its derivatives. (In particular, E vanishes whenever ϵ_β does.) We choose $u = r^{-1}(u_{\text{new}} - p)$, where u_{new} is a new control variable. After this change of coordinates and control variable the system is of the form (6) with $Q := r^{-1}E$, $P := -r^{-1}pE$. We obtain estimates on the uniform norm of Q and P (via estimates on r, p, and E) in terms of the error one-form ϵ_β, for any fixed β, on any compact, contractible region $\mathcal{M}$. Most important is that Q and P depend in a continuous way on ϵ_β and some of its derivatives, and they vanish whenever ϵ does.

From another point of view our approach can be viewed as a robustness analysis of exact feedback linearization. It is of obvious interest to analyze what happens to an exactly linearizable system subject to a small perturbation that destroys the property of being linearizable. One can expect that if linearization was used as an intermediate tool to achieve stabilization, tracking, disturbance rejection, and so on, a small perturbation, yielding a system "close" to being linearizable, still allows one to apply the control designed for the original linearizable system, guaranteeing satisfactory performance. In the present paper we propose some tools to measure a distance of a nonlinearizable perturbed system from a linearizable one, thus allowing us to measure how small the small perturbation is. In particular, we provide analysis of robustness of stabilizing feedback design based on feedback linearization.

We anticipate many applications of transforming (1) into (6). The idea behind making Q and P small is to neglect them in design. Intuitively speaking, we can neglect them if they are "small enough." What "small enough" means will depend on the particular application.

We should warn that one cannot expect that an exact or approximate feedback linearization will always help improve performance. The point is that the idea of linearization is to get rid of nonlinearities because we don't know how to deal with them. It may happen, though, that removing of some nonlinear terms may negatively affect the performance of the system. For instance, consider the problem of stabilization of the feedback linearizable system $\dot{x} = -x^3 + u$. One can remove the nonlinear term $-x^3$ using feedback, but this doesn't help stabilization at all. The term $-x^3$ actually helps to stabilize the system, especially for large initial conditions. Still, there are enough examples of systems in which nonlinear terms cause problems to justify the present study.

The paper is organized as follows. In §2 we introduce notation and some auxiliary results. We also explain construction of characteristic one-forms. In §3 we discuss the problem of optimal scaling of the characteristic one-forms. We review the construction of exact integrating factors and introduce and study best approximate L^p integrating factors. In §4 we show how homotopy operators can be use to decompose characteristic one-forms into exact and antiexact parts. In §5 we prove that a change of coordinates based on the exact part of any characteristic one-form obtained with a homotopy operator having its center at the origin defines a local diffeomorphism that takes the system (1) to a normal form that looks like a linearizable part perturbed by some nonlinear terms. The nonlinear perturbation terms depend linearly on the antiexact part of the characteristic one-form. In §6 we obtain some upper bounds on nonlinear perturbation terms using the antiexact part of a characteristic one-form and thus establish a continuity relationships between some measures of noninvolutivity and nonlinearizability. In §7 we apply the results of the paper to study locally linearizing feedback laws for the system (1).

2. Notation and auxiliary results. In the present paper we apply the theory of differential alternating forms. We refer to standard texts such as [1, 7, 8, 9, 12] for all the notions not defined here.

We denote by $T\mathcal{M}$ the tangent bundle to $\mathcal{M}$ and by $\Omega^k(\mathcal{M})$ the set of all alternating k-forms on $\mathcal{M}$, i.e., the space of all k-linear antisymmetric functionals on $T\mathcal{M}$. $\Omega(\mathcal{M})$ will denote the direct sum of all $\Omega^k(\mathcal{M})$. Let $\zeta \in \Omega(\mathcal{M})$ and $v \in T\mathcal{M}$. Then $d\zeta$ will denote the exterior derivative of ζ, and $L_v\zeta$ will denote the Lie derivative of ζ along v.

By $i_v(\xi)$ we will mean the interior product (contraction) of a vector field v with a k-form ξ, which is a $k-1$ form defined by

$$i_v(\xi)(v_1, v_2, \ldots, v_{k-1}) := \xi(v, v_1, v_2, \ldots, v_{k-1}).$$

Note that if ξ is a one-form, then $i_v(\xi) = \xi(v)$. Below we summarize some properties of interior and exterior (wedge) products and exterior and Lie derivatives.

PROPOSITION 2.1. *Let $\xi_1 \in \Omega^k(\mathcal{M})$, $\xi_2 \in \Omega(\mathcal{M})$, and $v \in T\mathcal{M}$ be arbitrary. Then*

1. $i_v(\xi_1 \wedge \xi_2) = (i_v\xi_1) \wedge \xi_2 + (-1)^k \xi_1 \wedge (i_v\xi_2)$.
2. $i_v i_v \xi_2 = 0$.
3. $d(\xi_1 \wedge \xi_2) = (d\xi_1) \wedge \xi_2 + (-1)^k \xi_1 \wedge (d\xi_2)$.
4. $d(L_v\xi_2) = L_v(d\xi_2)$.

While we did not need any additional structure except the differential one to study the problem of finding exact integrating factors, in the case of approximate integrating factors we need some means of measuring the distance between k-forms (for instance, ω_0 from an exact form $d\alpha$, or $d\beta\omega_0$ from 0), both at a point and globally (on $\mathcal{M}$). For this, we use a Riemannian metric, i.e., a positive definite (pointwise) inner product $\langle , \rangle$ on the tangent space to $\mathcal{M}$. This inner product induces an inner product on p-forms (see [1, §6.2]), which we will denote by the same symbol. Namely, let $\{e^i\}$, $i = 1, \ldots, n$, be an orthonormal basis for $\Omega^1(\mathcal{M})$. Then the inner product on p-forms is uniquely defined by requiring $\{e^{i_1} \wedge \cdots \wedge e^{i_p} | i_1 < \cdots < i_p\}$ to be an orthonormal basis for $\Omega^p(\mathcal{M})$. The corresponding pointwise norm will be denoted by $|\cdot|$. We obtain a global inner product $\langle\langle , \rangle\rangle$ of p-forms on $\mathcal{M}$ by integrating the pointwise one over $\mathcal{M}$. A standard metric associated with coordinate system $x_1, x_2, \ldots, x_n$ is the one in which the vector fields $\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n}$ and thus the one-forms $dx_1, dx_2, \ldots, dx_n$ are orthonormal. The standard (in coordinates x_i) volume element on $\mathcal{M}$ is

$$\mu := dx_1 \wedge dx_2 \wedge \cdots \wedge dx_n.$$

For any one-form η we will denote by $\eta^\#$ the dual vector field to η, i.e., the unique smooth vector field satisfying $\langle \xi, \eta \rangle = i_{\eta^\#}(\xi) = \xi(\eta^\#)$ for any one-form ξ. For instance, if we use the

standard metric, we have $(\sum_{i=1}^{n} \eta_i dx_i)^{\#} = \sum_{i=1}^{n} \eta_i \frac{\partial}{\partial x_i}$. We have the following elementary result.

PROPOSITION 2.2. *Let* $\xi_1, \xi_2 \in \Omega^1(\mathcal{M})$ *and* $\zeta \in \Omega^2(\mathcal{M})$. *Then*

$$\langle \xi_1 \wedge \xi_2, \zeta \rangle = \langle \xi_2, i_{\xi_1^{\#}} \zeta \rangle = \zeta(\xi_1^{\#}, \xi_2^{\#}).$$

Let $\zeta \in \Omega^p(\mathcal{M})$ and $\xi \in \Omega^{m-p}(\mathcal{M})$. In the following discussion we will deal with the operator $W_\xi : \Omega^p(\mathcal{M}) \mapsto \Omega^m(\mathcal{M})$ defined by $W_\xi \zeta := \zeta \wedge \xi$.

To obtain a one-form annihilating an $n-1$-dimensional distribution D in $\mathcal{M}$, we contract any volume element of $\mathcal{M}$ by any basis of D. For instance, we may choose

$$\omega_0 := i_g i_{ad_f g} \ldots i_{ad_f^{n-2} g} \mu, \tag{9}$$

where μ is the standard volume element in coordinates x_i.

3. Approximate integrating factors. Before we discuss the approximate integrating factors for nonintegrable characteristic one-form ω_0, let us remind the reader how, given an integrable characteristic one-form ω_0, one constructs an exact integrating factor for ω_0. The construction will suggest what can be done in the case of nonintegrable characteristic one-form ω_0. Let us begin with following standard result (see, e.g., [1, §6.4; 8, §4.2]).

PROPOSITION 3.1. *Let* ω_0 *be a nonvanishing one-form on* $\mathcal{M}$. *The following statements are equivalent:*

1. ω_0 *is integrable.*
2. *There is a one-form* γ *such that*

$$d\omega_0 = \gamma \wedge \omega_0. \tag{10}$$

3. *There is a zero-form* θ *such that*

$$d\omega_0 = d\theta \wedge \omega_0. \tag{11}$$

4. ω_0 *satisfies*

$$d\omega_0 \wedge \omega_0 = 0. \tag{12}$$

Note that statement 4 provides a test for integrability of ω_0 and thus for linearizability of the system (1). Let us present one possible way of proving $4 \Rightarrow 2$.

Let X be any smooth vector field on $\mathcal{M}$ satisfying $i_X(\omega_0) = \omega_0(X) = 1$. Then (see Proposition 2.1, statement 1)

$$0 = i_X(d\omega_0 \wedge \omega_0) = i_X(d\omega_0) \wedge \omega_0 + d\omega_0 \wedge i_X(\omega_0) = i_X(d\omega_0) \wedge \omega_0 + d\omega_0.$$

Choosing $\gamma := -i_X(d\omega_0)$, we see that $4 \Rightarrow 2$. Even though it is not immediately seen, one can choose X so that $i_X(\omega_0) = 1$ and $-i_X(d\omega_0) = d\theta$ for some zero-form θ, thus proving (3). The condition (3) is most important in construction of the integrating factor for ω_0. Namely, once we know the zero-form θ, choosing $\beta := e^{-\theta}$, we obtain $d\beta\omega_0 = 0$ and $\beta > 0$ as required. Note that an integrating factor β obtained in that way is not unique. Namely, if $\beta\omega_0 = d\alpha$ for some zero-form α, one can replace β with $h(\alpha)\beta$, where $h(\alpha)$ is smooth and positive. It is now easily checked that $dh(\alpha)\beta\omega_0 = 0$ so that $h(\alpha)\beta$ is an integrating factor for ω_0 whenever β is.

Given a nonintegrable characteristic one-form ω_0, we try to find a *best possible* integrating factor for it. Let us recall that the goal is to make $d\beta\omega_0$ "as small as possible." Below we define some precise meaning for making $d\beta\omega_0$ "small" by an appropriate choice of β. We

want to avoid the trivial solution $\beta = 0$. To the contrary, we want to formulate the problem of construction of best approximate integrating factor so that the solution yields $\beta > 0$ for all $x \in \mathcal{M}$. In coordinates, minimization of $d\beta\omega_0$ can be understood as making the differences of the mixed partial derivatives $\frac{\partial \beta\omega_{0i}}{\partial x_j} - \frac{\partial \beta\omega_{0j}}{\partial x_i}$ as small as possible.

We will first establish a pointwise measure of exactness for $\beta\omega_0$ and then construct a global one from the pointwise one. Let

$$\kappa(\omega_0)(x) := \frac{|(d\omega_0)(x)|}{|(\omega_0)(x)|}, \tag{13}$$

where $|\cdot|$ is the pointwise norm of a form given by the Riemannian metric. Note that such a measure of exactness of ω_0 is invariant under scaling of the one-form ω_0 by a constant, nonzero function. Now we define global measures of exactness of ω_0. A uniform measure can be obtained by taking the supremum of $\kappa(\omega_0)(x)$ over $\mathcal{M}$,

$$\chi_\infty(\omega_0) := \sup\{\kappa(\omega_0)(x), x \in \mathcal{M}\}. \tag{14}$$

Note that the supremum exists since $\mathcal{M}$ is compact (by assumption).

An average measure of integrability is obtained by integrating $\kappa(\omega_0)(x)$ over $\mathcal{M}$. Let $p > 0$. Then

$$\chi_p(\omega_0) := \left(\int_{\mathcal{M}} (\kappa(\omega_0)(x))^p \mu\right)^{1/p}, \tag{15}$$

where μ is the volume element associated with the Riemannian metric. Now, for $1 \leq p \leq \infty$, we can define the best approximate L^p integrating factor β for ω_0 as the zero-form that minimizes $\chi_p(\beta\omega_0)$.

The best situation that one might hope for when facing the problem of construction of the best approximate integrating factor is that there is a single function β that is, in fact, the best L^p approximate integrating factor for every $1 \leq p \leq \infty$. This will be the case if we can find a function β that minimizes $\kappa(\beta\omega_0)(x)$ at every point $x \in \mathcal{M}$. In certain special cases there is, in fact, an easy solution to this problem.

Let ω_0 be a given one-form on $\mathcal{M}$, and consider the decomposition

$$d\omega_0 = \gamma \wedge \omega_0 + \tau. \tag{16}$$

This equation should be interpreted as an "approximation" to (10) with the two-form τ playing the role of an error term. There are infinitely many ways of decomposing $d\omega_0$ as above, because for any γ, one can simply choose $\tau := d\omega_0 - \gamma \wedge \omega_0$. We know from Proposition 3.1 that we can choose $\tau = 0$ in the case of integrable ω_0. If ω_0 fails to be integrable, we will try to choose γ and τ in (16) so that the two-form τ is smallest possible in a least-squares sense, i.e., with respect to the (global) L^2 norm of forms on $\mathcal{M}$,

$$\|\xi\| := \left(\int_{\mathcal{M}} |\xi|^2 \mu\right)^{1/2} \quad \text{for } \xi \in \Omega(\mathcal{M}). \tag{17}$$

Note that this smallest possible τ measures how far ω_0 is from being closed (and thus exact).

It happens that the problem of finding γ and τ satisfying (16) with $\|\tau\|$ minimal can be solved pointwise.

As in the case of integrable ω_0, we will use an interior product of a vector field X satisfying $i_X(\omega_0) = \omega_0(X) = 1$ with $d\omega_0 \wedge \omega_0$ to obtain a decomposition of type (16). We have (see Proposition 2.1, statement 1)

$$i_X(d\omega_0 \wedge \omega_0) = i_X(d\omega_0) \wedge \omega_0 + d\omega_0 \wedge i_X(\omega_0) = i_X(d\omega_0) \wedge \omega_0 + d\omega_0$$

so that

$$d\omega_0 = (-i_X(d\omega_0)) \wedge \omega_0 + i_X(d\omega_0 \wedge \omega_0).$$

This relation has the required form (16) with $\gamma := -i_X(d\omega_0)$, $\tau := i_X(d\omega_0 \wedge \omega_0)$.

We will denote by $\tau_{\min}$ the two-form τ satisfying (16) for some γ with a minimum pointwise norm $|\tau(x)|$ at every $x \in \mathcal{M}$. (It is clear that this will also be the two-form with minimal global norm $\|\cdot\|$ among all two-forms τ satisfying (16).) Below we give an explicit formula for $\tau_{\min}$.

PROPOSITION 3.2. *Put* $X_0 := |\omega_0|^{-2}\omega_0^{\#}$. *Then* $i_{X_0}(\omega_0) = 1$ *and*

$$\tau_{\min} = i_{X_0}(d\omega_0 \wedge \omega_0). \tag{18}$$

Among all γ *satisfying* (16) *for* $\tau = \tau_{\min}$, *the one with a minimal norm is*

$$\gamma_{\min} := -i_{X_0}(d\omega_0). \tag{19}$$

$\gamma_{\min}$ *is pointwise orthogonal to* ω_0. *All other one-forms* γ *satisfying* (16) *can be represented as* $\gamma_\eta = \gamma_{\min} + \eta\omega_0$ *for some zero-form* η.

Proof. First, note that $\tau = \tau_{\min}$ if and only if $\langle \xi \wedge \omega_0, \tau\rangle = 0$ pointwise on $\mathcal{M}$ for any one-form ξ. (This follows from the fact that by a standard least-squares argument $\tau_{\min}$ must be orthogonal to the space $R(W_{\omega_0}) = \{\xi \wedge \omega_0 | \xi \in \Omega^1(\mathcal{M})\}$.) Note also that $i_{X_0}(\omega_0) = |\omega_0|^{-2}\omega_0(\omega_0^{\#}) = |\omega_0|^{-2}\langle\omega_0, \omega_0\rangle = 1$. Thus, it is immediately seen that (16) is satisfied for τ and γ given by (18) and (19). To see that τ defined by (18) is actually $\tau_{\min}$ we will show that $\langle \xi \wedge \omega_0, \tau\rangle = 0$ pointwise on $\mathcal{M}$ for any one-form ξ. Using Propositions 2.2 and 2.1, statement 2, we have

$$\langle \xi \wedge \omega_0, \tau\rangle = -\langle \omega_0 \wedge \xi, i_{X_0}(d\omega_0 \wedge \omega_0)\rangle = -|\omega_0|^{-2}\langle \xi, i_{\omega_0^{\#}} i_{\omega_0^{\#}}(d\omega_0 \wedge \omega_0)\rangle = 0.$$

We have shown that $\tau = \tau_{\min}$.

To prove that $\gamma_{\min}$ is pointwise orthogonal to ω_0, note that $\langle \omega_0, \gamma_{\min}\rangle = i_{\omega_0^{\#}}\gamma_{\min} = -|\omega_0|^{-2} i_{\omega_0^{\#}} i_{\omega_0^{\#}}(d\omega_0) = 0$. □

Note that $\tau_{\min} = 0$ is zero on $\mathcal{M}$ if and only if ω_0 is integrable on $\mathcal{M}$, and $\tau_{\min} = 0$ and $\gamma_{\min} = 0$ on $\mathcal{M}$ if and only if ω_0 is exact on $\mathcal{M}$. Let $\beta \in \Omega^0(\mathcal{M})$. We have

$$d\beta\omega_0 = (d\beta + \beta\gamma_{\min}) \wedge \omega_0 + \beta\tau_{\min}. \tag{20}$$

Now we can obtain the following pointwise lower bound for $\kappa(\beta\omega_0)$.

PROPOSITION 3.3. *For any* $x \in \mathcal{M}$ *and* $\beta \in \Omega^0(\mathcal{M})$ *we have*

$$\kappa(\beta\omega_0)(x) \geq \frac{|\tau_{\min}(x)|}{|\omega_0(x)|}. \tag{21}$$

Proof. Since the two-forms $(d\ln\beta + \gamma_{\min}) \wedge \omega_0$ and $\tau_{\min}$ are pointwise orthogonal for every β (see the proof of Proposition 3.2), we have

$$\begin{aligned}
\kappa(\beta\omega_0)(x) &= \frac{|d\beta \wedge \omega_0 + \beta d\omega_0|}{|\beta\omega_0|} = \frac{|d\beta \wedge \omega_0 + \beta(\gamma_{\min} \wedge \omega_0) + \beta\tau_{\min}|}{|\beta\omega_0|} \\
&= \frac{|(d\beta + \beta\gamma_{\min}) \wedge \omega_0 + \beta\tau_{\min}|}{|\beta\omega_0|} = \frac{|(d\ln\beta + \gamma_{\min}) \wedge \omega_0 + \tau_{\min}|}{|\omega_0|} \\
&= \left(\left(\frac{|(d\ln(\beta) + \gamma_{\min}) \wedge \omega_0|}{|\omega_0|}\right)^2 + \left(\frac{|\tau_{\min}|}{|\omega_0|}\right)^2\right)^{1/2} \geq \frac{|\tau_{\min}|}{|\omega_0|}.
\end{aligned}$$

□

The best that we can hope for is that the lower bound for $\kappa(\beta\omega_0)(x)$ obtained above is sharp; i.e., there is a zero-form β such that $\kappa(\beta\omega_0)(x) = \frac{|\tau_{\min}(x)|}{|\omega_0(x)|}$ for every $x \in \mathcal{M}$. This will be the case if $(d\ln(\beta) + \gamma_{\min}) \wedge \omega_0 = 0$ for some choice of β. A necessary and sufficient condition for this is $\gamma_\eta := \gamma_{\min} + \eta\omega_0 = d\theta$ for some zero-forms η and θ. Then we choose $\beta := e^{-\theta}$ and obtain $(d\ln\beta + \gamma_\eta) \wedge \omega_0 = 0$. Note that the zero-form β is everywhere strictly positive, as required.

Example 3.1. Consider

$$\begin{aligned} \dot{x}_1 &= x_2 + h_1(x_3) + h_2(x_1, x_2), \\ \dot{x}_2 &= x_3 + h_3(x_3) + h_4(x_1, x_2), \\ \dot{x}_3 &= u, \end{aligned} \tag{22}$$

where $h_i(\cdot)$ are any smooth functions with $h_i(0) = \frac{\partial h_i}{\partial x_j}(0) = 0$. We have $g = \frac{\partial}{\partial x_3}$, $f = (x_2 + h_1(x_3) + h_2(x_1, x_2))\frac{\partial}{\partial x_1} + (x_3 + h_3(x_3) + h_4(x_1, x_2))\frac{\partial}{\partial x_2}$, $ad_f g := [f, g] = (-h_1'(x_3))\frac{\partial}{\partial x_1} - (1 + h_3'(x_3))\frac{\partial}{\partial x_2}$, $\omega_0 = (1 + h_3'(x_3))dx_1 - h_1'(x_3)dx_2$, $d\omega_0 = h_1''(x_3)dx_2 \wedge dx_3 - h_3''(x_3)dx_1 \wedge dx_3$, and $d\omega_0 \wedge \omega_0 = (h_1''(x_3)dx_2 \wedge dx_3 - h_3''(x_3)dx_1 \wedge dx_3) \wedge ((1 + h_3'(x_3))dx_1 - h_1'(x_3)dx_2) = ((1 + h_3'(x_3))h_1''(x_3) - h_1'(x_3)h_3''(x_3))dx_1 \wedge dx_2 \wedge dx_3$. We see that the system is exactly feedback linearizable in a neighborhood of the origin if and only if $(1 + h_3'(x_3))h_1''(x_3) - h_1'(x_3)h_3''(x_3) = 0$, which is the case if $h_1(\cdot) = 0$.

It happens that for this system we can actually construct the best approximate L^p integrating factor β in above-mentioned sense, i.e., the one that works for every $1 \le p \le \infty$. Suppose that we use the standard metric in coordinates x_1, x_2, x_3. We have

$$\gamma_{\min} = \frac{h_1'(x_3)h_1''(x_3) + (1 + h_3'(x_3))h_3''(x_3)}{h_1'(x_3)^2 + (1 + h_3'(x_3))^2} dx_3,$$

$$\tau_{\min} = \frac{(h_1''(x_3) + h_3'(x_3)h_1''(x_3) - h_1'(x_3)h_3''(x_3))}{h_1'(x_3)^2 + (1 + h_3'(x_3))^2}((1 + h_3'(x_3))dx_2 \wedge dx_3 + h_1'(x_3)dx_1 \wedge dx_3).$$

Note that $\gamma_{\min}$ depends only on x_3, and thus it is exact. One can check that $\gamma_{\min} = d\theta$ for $\theta = \ln(h_1'(x_3)^2 + (1 + h_3'(x_3))^2)^{1/2}$. The best approximate integrating factor in above-mentioned sense is $\beta_0 = e^{-\theta} = (h_1'(x_3)^2 + (1 + h_3'(x_3))^2)^{-1/2} = |\omega_0|^{-1}$. (Note that this choice makes the pointwise length of $\beta_0\omega_0$ equal to 1 everywhere.) □

In [2] we show that the lower bound for $\kappa(\beta\omega_0)(x)$ is always sharp if the metric is the standard metric in some special coordinates. There are, however, examples of systems for which the lower bound for $\kappa(\beta\omega_0)(x)$ is not sharp. In this case a more sophisticated analysis is required (cf. [6, 4]), which leads to some variational problems whose solutions for β are given as solutions of elliptic eigenvalue problems. A simple alternative would be an approximation of γ_η by an exact form $dH\gamma_\eta$ using a homotopy operator H (see the next section).

Note that even though the minimal $\chi_p(\beta\omega_0)$ seems to be a natural measure of integrability of ω_0 and thus also a measure of noninvolutivity of the characteristic distribution D, it may not be a sufficient measure for the problem of approximate linearization. There are some indications that one should actually minimize $d(\beta\omega_0)$ *together* with its first $n-1$ Lie derivatives along f. This problem is addressed in [6, 5].

4. Homotopy operator. On any contractible region $\mathcal{M}$ one can define a linear operator $H : \Omega^k(\mathcal{M}) \mapsto \Omega^{k-1}(\mathcal{M})$ that partially inverts the exterior derivative, i.e.,

$$\omega = (dH + Hd)\omega \quad \forall\omega \in \Omega^k(\mathcal{M}) \tag{23}$$

or, in other words,

$$\omega = d(H\omega) + Hd\omega. \tag{24}$$

Any operator with such property will be called a *homotopy* operator. Following [9], we present a construction of such an operator. Consider the *cylinder* $I \times \mathcal{M}$ where $I := [0, 1]$, and define a family of maps $j_\lambda : \mathcal{M} \mapsto I \times \mathcal{M}$ by $j_\lambda(x) = (\lambda, x)$ for $\lambda \in I$. Note that k-forms on the cylinder can be represented in coordinates $\lambda, x_1, \ldots, x_n$ as sums of monomials of two types: $a(\lambda, x)dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k}$ and $a(\lambda, x)d\lambda \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_{k-1}}$. We now define a linear operator $K : \Omega^k(I \times \mathcal{M}) \mapsto \Omega^{k-1}(\mathcal{M})$ such that its action on these two types of monomials is given by

$$K(a(\lambda, x)dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k}) = 0, \tag{25}$$

$$K(a(\lambda, x)d\lambda \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_{k-1}}) = \left(\int_0^1 a(\lambda, x)d\lambda\right) dx_{i_1} \wedge \cdots \wedge dx_{i_{k-1}}. \tag{26}$$

The operator K satisfies ([9, §3.6])

$$K(d\omega) + d(K\omega) = j_1^*\omega - j_0^*\omega, \tag{27}$$

where $j_\lambda^* : \Omega^k(I \times \mathcal{M}) \mapsto \Omega^k(\mathcal{M})$ is the pullback induced by j_λ. Note that the above result doesn't require $\mathcal{M}$ to be contractible. Now, by definition, $\mathcal{M}$ is contractible iff there is a smooth mapping $\phi : I \times \mathcal{M} \mapsto \mathcal{M}$ such that $\phi(1, x) = x, \phi(0, x) = x^0$, where x^0 is a distingushed point in $\mathcal{M}$. Such a mapping ϕ is called a *homotopy* or *contraction* (of $\mathcal{M}$ to x^0). The point x^0 is called the homotopy center. Since we have $(\phi \circ j_1)(x) = x$ and $(\phi \circ j_0)(x) = x^0$ $\forall x \in \mathcal{M}$, the pullback $\phi^* : \Omega^k(\mathcal{M}) \mapsto \Omega^k(I \times \mathcal{M})$ induced by the mapping ϕ satisfies

$$j_1^*(\phi^*\omega) = \omega, \ j_0^*(\phi^*\omega) = 0. \tag{28}$$

Therefore, (27), (28), and the fact that the exterior derivative commutes with a pullback together imply

$$K\phi^*(d\omega) + d(K\phi^*\omega) = \omega. \tag{29}$$

Thus, the operator $H := K \circ \phi^*$ satisfies (24) so that it is a homotopy operator.

Note that different choices of homotopy centers x^0 and homotopies ϕ yield different homotopy operators. The one we will use is probably the simplest one: it will act on one-forms by integrating them along straight lines (in coordinates x_i) from a distingushed point x^0 in $\mathcal{M}$ (usually the origin). Such a homotopy operator will be called *radial* (see, e.g., [8, §5.3] and [9, §3.7]). If $\mathcal{M}$ is star shaped in coordinates x_i with respect to x^0 (i.e., $\mathcal{M}$ can be contracted to x^0 by straight lines lying entirely in $\mathcal{M}$), a simple choice for a homotopy ϕ is $\phi(\lambda, x) = x^0 + \lambda(x - x^0)$. Let $\omega = \sum_{i_1 \ldots i_k} \omega^{i_1 \ldots i_k}(x)dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k}$. Now, one can explicitly calculate the pullback ϕ^*,

$$\begin{aligned}
\phi^*\omega &= \sum_{i_1,\ldots,i_k} \omega^{i_1,\ldots,i_k}(x^0 + \lambda(x - x^0))d(x_{i_1}^0 + \lambda(x_{i_1} - x_{i_1}^0)) \\
&\quad \wedge d(x_{i_2}^0 + \lambda(x_{i_2} - x_{i_2}^0)) \wedge \cdots \wedge d(x_{i_k}^0 + \lambda(x_{i_k} - x_{i_k}^0)) \\
&= \sum_{i_1,\ldots,i_k} \omega^{i_1,\ldots,i_k}(x^0 + \lambda(x - x^0))((x_{i_1} - x_{i_1}^0)d\lambda + \lambda dx_{i_1}) \\
&\quad \wedge ((x_{i_2} - x_{i_2}^0)d\lambda + \lambda dx_{i_2}) \wedge \cdots \wedge ((x_{i_k} - x_{i_k}^0)d\lambda + \lambda dx_{i_k}),
\end{aligned} \tag{30}$$

and we can express the action of the radial homotopy operator as

$$(H\omega)(x) = \sum_{i_1,\dots,i_k} \left(\sum_{i_j} ((-1)^{j+1}(x_{i_j} - x^0_{i_j}) \right.$$
$$\left. \int_0^1 \lambda^{k-1} \omega^{i_1,\dots,i_k}(x^0 + \lambda(x - x^0))d\lambda) \right) dx_{i_1} \wedge \cdots \wedge \widehat{dx_{i_j}} \cdots \wedge dx_{i_k},$$

(31)

where the symbol $\widehat{\ }$ over dx_{i_j} indicates that it is omitted.

Once again consider the system of Example 3.1. Choose $\omega_0 = (1 + h_3'(x_3))dx_1 - h_1'(x_3)dx_2$ and $\beta = 1$. Then, for $x_0 = 0$, we see that

$$\alpha = \int_0^1 (x_1(1 + h_3'(\lambda x_3)) - x_2 h_1'(\lambda x_3))d\lambda = x_1 - \frac{x_2 h_1(x_3)}{x_3} + \frac{x_1 h_3(x_3)}{x_3},$$

and the error one-form ϵ is given by

$$\epsilon = -(h_3(x_3) - x_3 h_3'(x_3))/x_3^2 dx_1 + (h_1(x_3) - x_3 h_1'(x_3))/x_3^2 dx_2$$
$$+ (x_2 h_1(x_3) - x_1 h_3(x_3) - x_2 x_3 h_1'(x_3) + x_1 x_3 h_3'(x_3))/x_3^2 dx_3.$$

(Alternatively, once α is known, one can use the formula $\epsilon = \beta\omega_0 - d\alpha$ instead of $\epsilon = Hd\beta\omega_0$ to obtain the error one-form ϵ.)

Let $|\cdot|$ denote the pointwise norm of a form induced by the standard metric in x_i coordinates. We have shown in the previous section that the best approximate integrating factor for the above system in the sense of minimizing $\frac{|(d\beta\omega_0)(x)|}{|(\beta\omega_0)(x)|}$ pointwise everywhere is $\beta_0 = |\omega_0|^{-1} = (h_1'(x_3)^2 + (1 + h_3'(x_3))^2)^{-1/2}$. Contrary to the case $\beta = 1$ it is now impossible to evaluate $H\beta_0\omega_0$ and $Hd\beta_0\omega_0$ explicitly. The integration must be performed case by case for specific functions $h_1(\cdot)$ and $h_3(\cdot)$. We don't expect to be always able to perform the integration symbolically, because the result might not be an elementary function.

The homotopy operator that we will use in subsequent discussion uses the origin in coordinates x_i as the homotopy center. One can obtain other homotopy operators choosing different homotopy centers x_0. Moreover, one doesn't have to integrate over straight lines from the center. Note that the notion of a straight line is associated with specific choice of coordinates. Hence, if we change coordinates, we immediately obtain a homotopy operator, namely, the radial homotopy operator in the new coordinates. Moreover, an arithmetic mean of homotopy operators is again a homotopy operator. For exactly linearizable systems, once we have found an exact integrating factor for a characteristic one-form ω, any homotopy operator will give the same exact part and zero antiexact part. However, for nonlinearizable systems the choice of a homotopy operator will make a difference. Apparently, the choice of a particular homotopy operator will influence the exact and antiexact parts of a characteristic one-form ω. It is not clear to us yet what should be the best choice for approximate feedback linearization. This issue is currently under investigation. One may expect that the optimal homotopy operator might be rather complicated. Hence, even though there is no reason to believe that the radial homotopy operator in the original coordinates will be the best one (i.e., yielding the smallest antiexact part of ω), there is a good chance that it will be the simplest one to apply. Moreover, any homotopy operator with the center at the origin that satisfies $\phi(\lambda, 0) \equiv 0$ (in particular the radial one) will always yield $\epsilon(0) = 0$. Since ϵ is a smooth one-form vanishing at the origin, it will be small in a neighborhood of the origin.

The following result shows some that there are some limitations to what can be achieved in approximating nonexact characteristic forms by exact ones.

THEOREM 4.1. *Let H be any homotopy operator on $\mathcal{M}$ and ω be any characteristic one-form for the system* (1). *Let $\alpha_H := H\omega$ and $\epsilon_H := Hd\omega$. Then for any closed curve c in $\mathcal{M}$ we have*

$$\int_c L_f^i \epsilon_H = \int_c L_f^i \omega$$

for $i = 0, 1, \ldots$.

Proof. Note that $L_f^i\omega = L_f^i(d\alpha_H + \epsilon_H) = dL_f^i\alpha_H + L_f^i\epsilon_H$. Now, the proof follows from the fact that the integral of an exact form over any closed curve is zero. □

The above result is a law of preservation of hassle. No matter how one chooses a homotopy operator H, the average value of ("a component along c" of) $\epsilon_H := Hd\omega$ on any closed curve c is constant. Different homotopy operators may only distribute ϵ along c in a different way. A similar result holds true for any Lie derivative (of any order) of ϵ along any vector field.

This result can be used to obtain lower bounds for the uniform norm of the error one-form ϵ_H and its Lie derivatives along f on $\mathcal{M}$, independent of the choice of homotopy.

Let us conclude the section by an example of a homotopy operator that is optimal in some precise sense. Let ω be a one-form. The so-called Hodge decomposition of ω is a decomposition of the form $\omega = d\alpha + \epsilon$, where $d\alpha$ is the best L^2 approximation of ω among exact one-forms (cf. [1, §7.5; 3, 4, 5, 6]). Let δ denote the formal adjoint operator to the exterior derivative d and $\Delta := \delta d + d\delta$ denote the Laplace–De Rham operator ([1, §7.5]). One can show that $\Delta\alpha = \delta\omega$ and $\Delta\epsilon = \delta d\omega$. These equations (together with some boundary conditions) allow us to find α and ϵ appearing in the Hodge decomposition of ω. Thus, formally $\alpha = \Delta^{-1}\delta\omega$ and $\epsilon = \Delta^{-1}\delta d\omega$. Therefore, the operator $H_\Delta = \Delta^{-1}\delta$ (formally) satisfies $\omega = d(H_\Delta\omega) + H_\Delta d\omega$ so that it is a homotopy operator. Note that one has to solve a boundary value problem to obtain the zero-form α such that $d\alpha$ best approximates ω in L^2. This should be contrasted with the radial homotopy operator, which requires only simple integration with respect to a parameter. For instance, if the system (1) has polynomial nonlinearities in the original coordinates, the characteristic one-form ω_0 given by (9) will also have polynomial coefficients and thus the integration in (31) can be easily performed, yielding polynomial expressions for α and ϵ. The situation is usually much more difficult after applying an optimal approximate integrating factor β_0. The optimal characteristic one-form $\beta_0\omega$ will rarely be polynomial, and the result of integration in (31) might not be expressed in terms of elementary functions. This is one reason why we might not always be able to apply the optimal approximate integrating factor in practice, even if we find one.

5. Change of coordinates. In this section we prove that a change of coordinates based on the exact part of any characteristic one-form ω obtained with a homotopy operator H having its center at the origin defines a local diffeomorphism that takes the system (1) to a normal form that looks like a linearizable part perturbed by some nonlinear terms that depend linearly on the error one-form $\epsilon := Hd\omega$. This approach can be applied to both linearizable and nonlinearizable systems.

For exactly linearizable systems (1), we proceed as follows. First, we construct a characteristic one-form ω_0. Then we choose an exact integrating factor β and obtain a new characteristic form $\omega := \beta\omega_0$ such that $d\omega = 0$. We apply a homotopy operator H to get the zero form $\alpha := H\omega$. Then we use change of variables

$$\begin{aligned} z_1 &= \alpha, \\ z_2 &= L_f\alpha, \\ &\vdots \\ z_n &= L_f^{n-1}\alpha. \end{aligned} \tag{32}$$

The system (8) in new coordinates is

$$\begin{aligned} \dot{z}_1 &= z_2, \\ \dot{z}_2 &= z_3, \\ &\vdots \\ \dot{z}_{n-1} &= z_n, \\ \dot{z}_n &= p + ru, \end{aligned} \tag{33}$$

where

$$r = L_g L_f^{n-1}\alpha, \qquad p = L_f^n \alpha,$$

and the feedback $u = r^{-1}(u_{\text{new}} - p)$ makes it linear.

For a nonlinearizable system we proceed as follows. First, we construct a characteristic one-form ω_0. Then we choose an integrating factor β, either optimal or not, and obtain a new characteristic one-form $\omega := \beta\omega_0$. We apply a homotopy operator H to get the zero form $\alpha := H\omega$ and the corresponding error one-form $\epsilon := Hd\omega$. Then we use change of variables (32) (as for exactly linearizable systems) to get a normal form

$$\begin{aligned} \dot{z}_1 &= z_2 + e_1 u, \\ \dot{z}_2 &= z_3 + e_2 u, \\ &\vdots \\ \dot{z}_{n-1} &= z_n + e_{n-1} u, \\ \dot{z}_n &= p + ru + e_n u, \end{aligned} \tag{34}$$

where

$$\begin{aligned} e_1 &= L_g \alpha, \\ e_2 &= L_g L_f \alpha, \\ &\vdots \\ e_n &= L_g L_f^{n-1}\alpha - (-1)^{n-1}\omega(ad_f^{n-1} g), \\ r &= (-1)^{n-1}\omega(ad_f^{n-1} g), \\ p &= L_f^n \alpha. \end{aligned} \tag{35}$$

In the subsequent discussion we will need the following result.

LEMMA 5.1. *Let ω be any characteristic one-form for the system* (1). *Let i and j be nonnegative integers. Then*

1. $(L_f^i \omega)(ad_f^j g) = 0$ *for* $i + j < n - 1$.
2. $(L_f^i \omega)(ad_f^{n-1-i} g) = (-1)^i \omega(ad_f^{n-1} g)$ *for* $i = 0, 1, \ldots, n-1$.

Proof. 1. One can prove by induction the formula

$$(L_X^i \eta)(Y) = \sum_{l=0}^{l=i} (-1)^l \binom{i}{l} L_X^{i-l}(\eta(ad_X^l Y)). \tag{36}$$

In particular,

$$\begin{aligned} (L_f^i \omega)(ad_f^j g) &= \sum_{l=0}^{l=i} (-1)^l \binom{i}{l} L_X^{i-l}(\omega(ad_f^l(ad_f^j g))) \\ &= \sum_{l=0}^{l=i} (-1)^l \binom{i}{l} L_X^{i-l}(\omega(ad_f^{l+j} g)). \end{aligned} \tag{37}$$

Note that $\omega(ad_f^{l+j} g) = 0$ for $l = 0, \ldots, i$ if $i + j \leq n - 1$.

2. Apply the formula (37) for $j = n - 1 - i$, and note that all the terms $\omega(ad_f^{l+n-1-j}g)$ vanish except when $l = i$. □

To establish continuity relationships between noninvolutivity and nonlinearizability we express the nonlinear perturbation terms e_i in terms of the error one-form ϵ.

PROPOSITION 5.2. *Let ω be a characteristic form for* (1), *H be a homotopy operator on $\mathcal{M}$, $\alpha := H\omega$, and $\epsilon := Hd\omega$. Let $e_i, i = 1, \ldots, n$, and p be given by* (35). *Then*

$$
\begin{aligned}
e_1 &= -\epsilon(g), \\
e_2 &= -(L_f\epsilon)(g), \\
&\vdots \\
e_{n-1} &= -(L_f^{n-2}\epsilon)(g), \\
e_n &= -(L_f^{n-1}\epsilon)(g).
\end{aligned} \tag{38}
$$

Proof. It is a straightforward calculation using Lemma 5.1. □

Note that the above choice for e_n and r is not the only one possible. Actually, any choice that guarantees $r + e_n = L_g L_f^{n-1}\alpha$ with $e_n(0) = 0$ could be considered, for instance, $e_n = 0$ and $r = L_g L_f^{n-1}\alpha$. Our choice is dictated by the fact that it guarantees $r \neq 0$ on the whole $\mathcal{M}$ and $e_n = -(L_f^{n-1}\epsilon)(g)$.

One can also express the function $p := L_f^n\alpha$ using the error one-form ϵ as $p = (L_f^{n-1}\omega)(f) - (L_f^{n-1}\epsilon)(f)$.

A natural question to ask is whether the zero-form α together with its $n-1$ Lie derivatives along f is a well-defined change of coordinates. The main result of this section says that in a neighborhood of the origin, (32) indeed defines a local diffeomorphism. Before we prove it, we need some preliminary results.

LEMMA 5.3. *Let η be any smooth one-form and X and Y be any smooth vector fields on $\mathcal{M}$. Then*

1. $(L_X\eta)(Y) = L_X(\eta(Y)) - \eta([X, Y])$.
2. *If $\eta(0) = 0$ and $X(0) = 0$, then $(L_X^i\eta)(Y)(0) = 0$ for $i = 0, 1, 2, \ldots$.*

Proof. 1. See [12, §7.3].

2. For $i = 0$ the formula is true as $(L_X^0\eta)(Y)(0) = \eta(Y)(0) = 0$. Assume that the formula is true for $i = 0, \ldots, m$. Using statement 1 one easily shows that $(L_X^{m+1}\eta)(Y)(0) = d((L_X^m\eta)(Y))(X)(0) - (L_X^m\eta)([X, Y])(0)$. The first part of this expression is zero because $X(0) = 0$, and the second by assumption. By induction, the formula holds for all nonnegative i. □

PROPOSITION 5.4. *Assume that* $\dim \operatorname{span}\{g, ad_f g, \ldots, ad_f^{n-1}g\} = n \ \forall x$ *in a neighborhood of* 0 *in $\mathcal{M}$ (linear controllability). Let ω be any characteristic one-form for the system* (1). *Then the one-forms $\omega, L_f\omega, \ldots, L_f^{n-1}\omega$ are linearly independent in a neighborhood of the origin.*

Proof. It is sufficient to show that $(L_f^{n-1}\omega \wedge L_f^{n-2}\omega \wedge \cdots \wedge \omega)(0) \neq 0$. Since this form is smooth, it is enough to check that $(L_f^{n-1}\omega \wedge L_f^{n-2}\omega \wedge \cdots \wedge \omega)(g, ad_f g, \ldots, ad_f^{n-1}g)(0) \neq 0$. For this, note that $(L_f^{n-1}\omega \wedge L_f^{n-2}\omega \wedge \cdots \wedge \omega)(g, ad_f g, \ldots, ad_f^{n-1}) = \det S$, where S is an $n \times n$ matrix whose (i, j) entry is $(L_f^{i-1}\omega)(ad_f^{j-1}g)$. Now, by Lemma 5.1, S is an upper triangular matrix whose ith diagonal element is $(-1)^{n-i}\omega(ad_f^{n-1}g)$. Therefore, $\det S = (-1)^n(\omega(ad_f^{n-1}g))^n \neq 0$. □

Now we are ready to prove the main result of this section.

THEOREM 5.5. *Assume that* $\dim \operatorname{span}\{g, ad_f g, \ldots, ad_f^{n-1}g\} = n \ \forall x$ *in a neighborhood of 0 in $\mathcal{M}$ (linear controllability). Let H be any homotopy operator on $\mathcal{M}$ with the center at the origin such that $\phi(\lambda, 0) \equiv 0$ and ω be any characteristic one-form for the system* (1). *Set $\alpha := H\omega$. Then* (32) *defines a local diffeomorphism in a neighborhood of the origin.*

Proof. We will show that the differentials of the zero-forms $\alpha, L_f\alpha, \ldots, L_f^{n-1}\alpha$ are linearly independent at the origin (and thus, in a neighborhood of the origin). Let $\epsilon := Hd\omega$. Since $\omega = d\alpha + \epsilon$ and the Lie and exterior derivatives commute, we have $dL_f^i\alpha = L_f^i d\alpha = L_f^i(\omega - \epsilon) = L_f^i\omega - L_f^i\epsilon$. Hence, $dL_f^i\alpha(0) = L_f^i\omega(0) - L_f^i\epsilon(0)$. Note that $\epsilon(0) := (Hd\omega)(0) = 0$, since H is a homotopy operator with the center at 0 such that $\phi(\lambda, 0) \equiv 0$. Now, it follows from Lemma 5.3, statement 2, that $\epsilon, L_f\epsilon, \ldots, L_f^{n-1}\epsilon$ all vanish at the origin and hence $dL_f^i\alpha(0) = L_f^i\omega(0)$. Now the result follows from Proposition 5.4. □

We usually cannot guarantee a priori that the change of coordinates (32) will be valid in the whole $\mathcal{M}$. Some conditions for a map to be a global diffeomorphism are quoted in [14] and [20]. Below, we show an example of a system that admits a global transformation in R^3 to the normal form (6).

Example 5.1. Consider the system

$$\begin{aligned} \dot{x}_1 &= x_2 + h_1(x_3), \\ \dot{x}_2 &= x_3, \\ \dot{x}_3 &= u, \end{aligned} \tag{39}$$

where $h_1(\cdot)$ is any smooth function with $h_1'(0) = 0$. We have

$$\begin{aligned} \omega_0 &= dx_1 - h_1'(x_3)dx_2, \\ \alpha &= H\beta\omega_0 = x_1 - \frac{x_2 h_1(x_3)}{x_3}, \\ \epsilon &= \left(\frac{h_1(x_3) - x_3 h_1'(x_3)}{x_3^2}\right)(x_3 dx_2 - x_2 dx_3), \\ L_f\epsilon &= L_f^2\epsilon = 0. \end{aligned}$$

The system can be transformed by a global diffeomorphism

$$\begin{aligned} z_1 &= \alpha &&= x_1 - \frac{x_2 h_1(x_3)}{x_3}, \\ z_2 &= L_f\alpha &&= x_2, \\ z_3 &= L_f^2\alpha &&= x_3 \end{aligned} \tag{40}$$

to the form

$$\begin{aligned} \dot{z}_1 &= z_2 + \frac{z_2(h_1(z_3) - z_3 h_1'(z_3))}{z_3^2}u, \\ \dot{z}_2 &= z_3, \\ \dot{z}_3 &= u. \end{aligned} \tag{41}$$

The inverse transformation given by

$$\begin{aligned} x_1 &= z_1 + \frac{z_2 h_1(z_3)}{z_3}, \\ x_2 &= z_2, \\ x_3 &= z_3. \end{aligned} \tag{42}$$

□

In the case when the change of coordinates is not valid on the whole region $\mathcal{M}$, we have to restrict to a region on which the change of coordinates is valid. In the following discussion we assume that this has been done, and the restricted region is also called $\mathcal{M}$.

6. Estimates of the nonlinear part. In this section we estimate the nonlinear perturbation terms $e_1, \ldots, e_n$ using the error one-form ϵ. First, let us rewrite the equations (34) (in the usual matrix-vector notation) as

$$\dot{z} = Az + Bru + Bp + Eu, \tag{43}$$

where A and B are in the Brunovsky form, that is,

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & 0 \\ \vdots & & & 0 & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

and $E = (e_1, e_2, \ldots, e_n)^T$. ($e_i's$, r, and p are defined by (35).) We see that $r \neq 0$ on $\mathcal{M}$ (under the assumption that linear controllability holds on $\mathcal{M}$) and E depends *linearly* on ϵ and vanishes whenever ϵ does. We will choose $u = r^{-1}(u_{\text{new}} - p)$, where u_{new} is a new control variable. After this change of coordinates and control variable the system is of the form (6) with $Q := r^{-1}E$, $P := -r^{-1}pE$. In this section we obtain estimates on the uniform norm of Q and P (via estimates on r, p, and E) in terms of the error one form ϵ_β for any fixed β on any compact, contractible region $\mathcal{M}$.

Let h be a smooth vector field on $\mathcal{M}$ and l be a nonnegative integer. Let ζ be a k-form on $\mathcal{M}$. We define the C^0 norm $\|\zeta\|^0$ of ζ as $\|\zeta\|^0 := \sup|\zeta(x)|$ for $x \in \mathcal{M}$ (uniform norm on $\mathcal{M}$) and the C^l_h norm $\|\zeta\|^l_h$ as $\|\zeta\|^l_h := \sup(|\zeta(x)|^2 + |L_h\zeta(x)|^2 + \cdots + |L^l_h\zeta(x)|^2)^{1/2}$ for $x \in \mathcal{M}$ (uniform norm on $\mathcal{M}$, together with the first l Lie derivatives along h). It is immediately seen from Proposition 5.2 that whenever the one-forms $\epsilon, L_f\epsilon, \ldots, L^{n-1}_f\epsilon$ are small on $\mathcal{M}$, so is the term E on $\mathcal{M}$.

THEOREM 6.1. *Let ω be any characteristic one-form for the system* (1) *and ϵ be the error one-form corresponding to a given homotopy operator. Then the mapping $\epsilon \mapsto E$ is a continuous mapping from the space of smooth one forms equipped with the C^{n-1}_f norm on $\mathcal{M}$ into the space of smooth vector fields on $\mathcal{M}$ equipped with the C^0 norm (uniform norm on $\mathcal{M}$). In particular,*

$$\|E\|^0 \leq \|\epsilon\|^{n-1}_f \|g\|^0.$$

Proof. The proof is immediate in view of Proposition 5.2. □

Note that in the above result we could substitute for $\epsilon, L_f\epsilon, \ldots, L^{n-1}_f\epsilon$ their evaluations at (contractions by) the vector field g. Let h and v be smooth vector fields on $\mathcal{M}$ and l be a nonnegative integer. Let us define the $C^l_{h,v}$ seminorm $\|\zeta\|^l_{h,v}$ of a one-form ζ on $\mathcal{M}$ as $\|\zeta\|^l_{h,v} := \sup(|\zeta(v)(x)|^2 + |L_h\zeta(v)(x)|^2 + \cdots + |L^l_h\zeta(v)(x)|^2)^{1/2}$ over $x \in \mathcal{M}$. Note that $C^l_{h,v}$ is not quite a norm, since it may happen that $\|\zeta\|^l_{h,v} = 0$ even though $\zeta \neq 0$ (for example, $\|\omega_0\|^0_{f,g}$ for ω_0 being a characteristic one-form for (1)). However, it happens that for the one-forms on $\mathcal{M}$, the $C^{n-1}_{f,g}$ seminorm becomes a norm if the vector fields f and g satisfy the linear controllability condition of Theorem 1.1. This follows from the following result.

PROPOSITION 6.2. *Let ζ be a one-form on $\mathcal{M}$ and the vector fields f and g satisfy the linear controllability condition of Theorem* 1.1. *Then $\zeta = 0$ if and only if $\|\zeta\|^{n-1}_{f,g} = 0$.*

Proof. ($\Rightarrow$) The proof is obvious.

($\Leftarrow$) Note that $L^i_f\zeta(g) = L_f(L^{i-1}_f\zeta(g)) - L^{i-1}_f\zeta(ad_f g)$. We have $L^i_f\zeta(g) = 0$ for $i = 0, 1, \ldots, n-1$. In particular, $\zeta(g) = 0$. Thus, using Lemma 5.3, statement 1, we get

$0 = (L_f\zeta)(g) = L_f(\zeta(g)) - \zeta(ad_f g) = -\zeta(ad_f g)$. Continuing in the same fashion, we obtain $\zeta(ad_f^i g) = 0$ for $i = 0, 1, \ldots, n-1$. By the linear controllability assumption, the vector fields $ad_f^i g$ for $i = 0, 1, \ldots, n-1$ are linearly independent. The one-form ζ annihilates n linearly independent fields on an n-dimensional manifold. Thus $\zeta = 0$. □

The above result, when applied to the error one-form ϵ, yields an obvious fact that the $\epsilon = 0$ is equivalent with (32) being the linearizing change of coordinates for (1). The fact that we wanted to emphasize here is that, because of (38), the nonlinear perturbation terms e_i can be used to define a norm for the error one-form ϵ, thus making the relationship between a measure of noninvolutivity of the characteristic distribution D and a direct measure of nonlinearity of the system (1) in new coordinates explicit. Namely, we have the following proposition.

PROPOSITION 6.3. $\|E\|^0 = \|\epsilon\|_{f,g}^{n-1}$.

We conclude this section with establishing some upper bounds on the uniform norms $\|P\|^0$ and $\|Q\|^0$ of the nonlinear terms $Q := r^{-1}E$, $P := -r^{-1}pE$ in the system (6) after change of coordinates and preliminary feedback.

PROPOSITION 6.4. *Let ω be any characteristic form for the system* (1), $\alpha := H\omega$, $\epsilon := Hd\omega$. *Let* $\rho := \inf |\omega(ad_f^{n-1}g)(x)|$ *over* $x \in \mathcal{M}$ *and* $\varrho := \sup |L_f^n\alpha(x)|$ *over* $x \in \mathcal{M}$. *Then*

1.

$$\|P\|^0 \leq \frac{\varrho\|\epsilon\|_{f,g}^{n-1}}{\rho}. \tag{44}$$

2.

$$\|Q\|^0 \leq \frac{\|\epsilon\|_{f,g}^{n-1}}{\rho}. \tag{45}$$

Proof. The proof is immediate, in view of Proposition 6.3 and (35). □

7. Application to stabilization. In this section we will use the results of the previous section to study various locally stabilizing feedback laws for the system (1). The laws that we have in mind will be linear in new coordinates (32), with the gains chosen so that the linear part of the system (6) is asymptotically stable. We will then study robustness of such control laws when applied to the system (6). We will accomplish that studying Lyapunov functions that are quadratic in new coordinates. We shall examine how the nonlinear part of (6) affects the time derivative of the Lyapunov function. The continuity result of Theorem 6.1 will allow us to formulate some robustness criteria for stabilization.

The idea behind transforming a linearizable system (1) to an equivalent form (2) is to design control schemes for (2), which is much easier to analyze and control, and apply them to (1). For example, if Φ happens to be a global diffeomorphism from R^n into R^n, one can globally asymptotically stabilize the system (1). For this, one can choose new control variable $u_{\text{new}} = Kz$ (linear feedback in new variables) so that the closed-loop system

$$\dot{z} = (A + BK)z \tag{46}$$

is globally asymptotically stable. (Controllability of (2) is equivalent to possibility of arbitrary assignment of the eigenvalues of $(A + BK)$ by an appropriate choice of the feedback gain K.) Then $u = k(x) + l(x)u_{\text{new}} = k(x) + l(x)K\Phi(x)$ makes the closed-loop system

$$\dot{x} = f(x) + g(x)(k(x) + l(x)K\Phi(x)) \tag{47}$$

globally asymptotically stable, since Φ^{-1} is a diffeomorphism preserving the equilibrium point at the origin.

For nonlinearizable systems the best we can hope for using our approach is to transform (1) to (6) with P and Q small. Then we will try to use the new form (6) to design a locally stabilizing feedback—in this case we expect to improve the basin of attraction of the origin of the closed-loop system. We will choose $u_{\text{new}} = Kz$ (a feedback law linear in new variables) so that the mapping $A + BK$ is stable (has all eigenvalues with negative real parts) and analyze its robustness as a stabilizing law for (6); bounds on uniform norms for Q and P should help us to do so. Let us stress that we will actually use new coordinates z and new control u_{new} only as intermediate tools, and the control law $u = r^{-1}(u_{\text{new}} - p) = r^{-1}(Kz - p)$ will be expressed in the old coordinates x as $u = k_1(x)$ (where $k_1(x) := r(\Phi(x))^{-1}(K\Phi(x) - p(\Phi(x)))$) and applied to (1). Since Φ^{-1} is a diffeomorphism preserving the equilibrium point at the origin, it maps the basin of attraction of the equilibrium for

$$\dot{z} = (A + BK)z + P(z) + Q(z)Kz \tag{48}$$

to the basin of attraction of the equilibrium for

$$\dot{x} = f(x) + g(x)k_1(x). \tag{49}$$

Observe that, to express the feedback laws computed in new coordinates z in the original coordinates x, we don't even need to find the form (6) explicitly. It would be actually very difficult, if not impossible, to do so in general, since we would have to know the inverse transformation $x = \Phi^{-1}(z)$ in order to obtain the form (6).

Of course, we might not always be able to find the best integrating factor β_0 for ω_0 annihilating $D := \text{span}\{g, ad_f g, \ldots, ad_f^{n-2} g\}$ to begin with. Still, for any scaling factor β we can choose the corresponding zero-form α_β and its Lie derivatives along f as new coordinates. We can also find the corresponding error one-form ϵ_β and verify the bounds on the corresponding terms Q and P in (6) and decide if they are sufficiently small for our purpose.

THEOREM 7.1. *Assume that $z = \Phi(x)$ is a (global) diffeomorphism of $\mathcal{M}$ onto its image given by* (32). *Let $u_{\text{new}} = Kz$ be any linear feedback in new variables so that the linear part*

$$\dot{z} = (A + BK)z \tag{50}$$

of the system (6) *obtained from* (1) *after change of coordinates and preliminary feedback is asymptotically stable. Let N be a positive definite $n \times n$ matrix and M be the unique positive semidefinite solution of the Lyapunov equation*

$$(A + BK)^T M + M(A + BK) + N = 0. \tag{51}$$

Let

$$E(z) := P(z) + Q(z)Kz \tag{52}$$

and $\Omega_r = \{0\} \cup \{z \in \Phi(\mathcal{M}) : \langle z, Mz\rangle < r$ and $\langle z, Nz\rangle - 2\langle z, ME(z)\rangle > 0\}$. Define $r_{\max} := \sup\{r \geq 0 : \Omega_r \subseteq \Phi(\mathcal{M})\}$. Then $\Phi^{-1}(\Omega_{r_{\max}})$ is an invariant set contained in the basin of attraction of the origin of the system

$$\dot{x} = f(x) + g(x)k_1(x), \tag{53}$$

where $k_1(x) := r(\Phi(x))^{-1}(K\Phi(x) - p(\Phi(x)))$ (p and q are defined by (35)).

Proof. The linearizable part of the system in new coordinates z can be made asymptotically stable by feedback $u_{\text{new}} = Kz$ linear in new coordinates. One can define a quadratic Lyapunov

function $V(z) := \langle z, Mz\rangle$ with a negative time derivative $\dot{V}(z) = -\langle z, Nz\rangle$ solving the Lyapunov equation (51). The sets Ω_r are invariant sets for the closed-loop linear part (50). Now, the time derivative of Lyapunov function for the true system in new coordinates is $\dot{V}(z) = -(\langle z, Nz\rangle - 2\langle z, ME(z)\rangle)$. If this is negative, the sets $\Phi^{-1}(\Omega_{r_{\max}})$ are invariant sets for the closed-loop system (53). □

The above result simply states a sufficient condition for a region of $\mathcal{M}$ to be an invariant set contained in the basin of attraction of the origin of the system $\dot{x} = f(x) + g(x)k_1(x)$ and is well known. What is nice about the above result is that we can actually estimate the set $\Phi^{-1}(\Omega_{r_{\max}})$ in our approach. Namely, since we have estimates on the uniform norms $\|P(z)\|$ and $\|Q(z)\|$ of the nonlinear terms in the system (6), we obtain an upper bound on the uniform norm $E(z) = P(z) + Q(z)Kz$. Thus, we can check if the time derivative $-(\langle z, Nz\rangle + 2\langle z, ME(z)\rangle)$ of the Lyapunov function is negative on the region of interest. Moreover, since we expect $P(z)$ and $Q(z)$ to be small, so will $E(z)$. Since the first term in $-(\langle z, Nz\rangle + 2\langle z, ME(z)\rangle)$ is negative and the second term is small, the whole expression is negative in some neighborhood of the origin.

Let us define yet another measure of nonlinearity in new coordinates that is particularly suited for studying stabilization:

$$\eta_{afl}(z) := \frac{2\langle z, ME(z)\rangle}{\langle z, Nz\rangle} \text{ for } z \neq 0, \ \eta_{afl}(0) := 0. \tag{54}$$

Now we can replace the condition

$$2\langle z, ME(z)\rangle < \langle z, Nz\rangle \text{ for } z \neq 0 \tag{55}$$

with

$$\eta_{afl}(z) < 1. \tag{56}$$

Note that the quantity η_{afl} actually depends on the choice of characteristic one-form, the particular homotopy operator, the stabilizing feedback gain matrix K, and the matrix N. Observe that $|\eta_{afl}(z)| < 1$ means that the linear term dominates the nonlinear one in the time derivative $\dot{V}(z) = -(\langle z, Nz\rangle + 2\langle z, ME(z)\rangle)$ of the Lyapunov function $V(z) := \langle z, Mz\rangle$ at the particular point z, guaranteeing its negative sign. On the other hand $0 < \eta_{afl}(z)$ means that the nonlinearities contribute to making $\dot{V}(z)$ more positive and thus have a destibilizing effect, while $\eta_{afl}(z) < 0$ means that the nonlinearities try to make $\dot{V}(z)$ more negative and hence help to stabilize the system. Therefore, the following terminology is justified: we will say that the nonlinearities are *weak* (respectively, *strong*) at z if $|\eta_{afl}(z)| < 1$ (respectively, $|\eta_{afl}(z)| > 1$) and *friendly* (respectively, *unfriendly*) if $\eta_{afl}(z) < 0$ (respectively, $0 < \eta_{afl}(z)$).

Let us express this condition in terms of system (6) and (8). We have $E(z) = P(z) + Q(z)Kz = (Kz - p(z))(r^{-1}(z)E(z))$. Thus $2\langle z, ME(z)\rangle = 2\langle z, M(Kz - p(z))(r^{-1}(z)E(z))\rangle$, and (56) is equivalent to

$$\frac{2\langle z, M(Kz - p(z))(r^{-1}(z)E(z))\rangle}{\langle z, Nz\rangle} < 1 \text{ for } z \neq 0. \tag{57}$$

Using bounds on $\|P(z)\|$ and $\|Q(z)\|$ obtained in the previous section, one can formulate the following inequality, which implies the previous ones:

$$2\frac{(\varrho + |K||z|)\|\epsilon\|_{f,g}^{n-1}}{\rho} < \frac{\inf \sigma(N)}{\sup \sigma(M)}|z| \text{ for } z \neq 0, \tag{58}$$

where $\sigma(\cdot)$ denotes a spectrum of a matrix.

It is possible to combine the problems of designing a stabilizing feedback for the linear part of of the system (6) obtained from (1) after change of coordinates and preliminary feedback and construction of a Lyapunov function in a linear quadratic optimal control design: find u_{new} minimizing

$$\int_0^\infty (\langle z(t), Nz(t)\rangle + \langle u_{\text{new}}(t), Ru_{\text{new}}(t)\rangle)dt$$

for strictly positive definite R and a positive definite N. (To make life easier, we will assume that N is also strictly positive definite.) It is well known that the optimal control u_{new} has the form of linear feedback $u_{\text{new}} = Kz$ for $K = -R^{-1}B^TM$, where M is the unique positive definite solution of the Riccati equation

$$A^TM + MA - MBR^{-1}B^TM + N = 0. \tag{59}$$

Then $V(z) := \langle z(t), Mz(t)\rangle$ is the Lyapunov function for the closed-loop system (50) and $\dot V(z) = -\langle z(t), (N + MBR^{-1}B^TM)z(t)\rangle$.

Example 7.1. Consider the system

$$\begin{aligned} \dot x_1 &= x_2 + ax_3^3 + bx_1^3, \\ \dot x_2 &= x_3 + cx_1^2x_2, \\ \dot x_3 &= u. \end{aligned} \tag{60}$$

□

Note this is a particular case of the system considered in Example 3.1. We have $\omega_0 = dx_1 - 3ax_3^2dx_2$ and $d\omega_0 = 6ax_3dx_2 \wedge dx_3$. For scaling factor $\beta = 1$ we get $\alpha := H\omega = x_1 - ax_2x_3^2, \epsilon := Hd\omega = (2ax_3)(x_2dx_3 - x_3dx_2)$. New coordinates $z = \Phi(x)$ are given by $z_1 := \alpha = x_1 - ax_2x_3^2, z_2 := L_f\alpha = x_2 + bx_1^3 - acx_1^2x_2x_3^2, z_3 := L_f^2\alpha = x_3 + 3b^2x_1{}^5 + 3bx_1{}^2x_2 + cx_1{}^2x_2 - 2abcx_1{}^4x_2x_3{}^2 - ac^2x_1{}^4x_2x_3{}^2 - 2acx_1x_2{}^2x_3{}^2 + 3abx_1{}^2x_3{}^3 - acx_1{}^2x_3{}^3 - 2a^2cx_1x_2x_3{}^5$. Note that $\Phi(x)$ is only a local diffeomorphism around the origin, and it is impossible to find an inverse transformation. Thus, in the following discussion we express the nonlinear terms $E(z)$, $r(z)$, and $p(z)$ in old coordinates: $E(\Phi(x)) = [-2ax_2x_3, -2acx_1{}^2x_2x_3, -2acx_1x_2x_3\,(2bx_1{}^3 + cx_1{}^3 + 2x_2 - 4ax_3{}^3)]^T$,

$$\begin{aligned} r(\Phi(x)) &:= 1 + 9abx_1{}^2x_3{}^2 - 3acx_1{}^2x_3{}^2 - 18a^2cx_1x_2x_3{}^4, \\ p(\Phi(x)) &:= 15b^3x_1{}^7 + 21b^2x_1{}^4x_2 + 5bcx_1{}^4x_2 + c^2x_1{}^4x_2 + 6bx_1x_2{}^2 \\ &\quad + 2cx_1x_2{}^2 + 3bx_1{}^2x_3 + cx_1{}^2x_3 - 8ab^2cx_1{}^6x_2x_3{}^2 - 6abc^2x_1{}^6x_2x_3{}^2 \\ &\quad - ac^3x_1{}^6x_2x_3{}^2 - 10abcx_1{}^3x_2{}^2x_3{}^2 - 8ac^2x_1{}^3x_2{}^2x_3{}^2 - 2acx_2{}^3x_3{}^2 \\ &\quad + 21ab^2x_1{}^4x_3{}^3 - 4abcx_1{}^4x_3{}^3 - ac^2x_1{}^4x_3{}^3 + 12abx_1x_2x_3{}^3 \\ &\quad - 4acx_1x_2x_3{}^3 - 10a^2bcx_1{}^3x_2x_3{}^5 - 6a^2c^2x_1{}^3x_2x_3{}^5 - 4a^2cx_2{}^2x_3{}^5 \\ &\quad + 6a^2bx_1x_3{}^6 - 4a^2cx_1x_3{}^6 - 2a^3cx_2x_3{}^8. \end{aligned}$$

(All computations were done using Mathematica.) To design a locally stabilizing feedback, we have solved the linear quadratic regulator problem for the linear part of the system as mentioned above for N being the 3×3 identity matrix and $R = 1$. The optimal feedback gain matrix was $K = [-1, -2.41421, -2.41421]$, and eigenvalues of $A + BK$ were $-1, -0.707107 + i0.707107, -0.707107 - i0.707107$. The feedback law applied to the original system (1) was $u_{afl} := r(\Phi(x))^{-1}(K\Phi(x) - p(\Phi(x)))$. We choose the values of parameters $a = 0.01, b = 1, c = 5$, and $\mathcal{M} := \{|x_i| < 0.36, i = 1, 2, 3\}$. We checked that the condition (57) was satisfied on $\mathcal{M}$, with $\sup \eta_{afl}(\Phi(x)) \approx 0.45$. Thus, by Theorem 7.1, the corresponding set $\Phi^{-1}(\Omega_{r_{\max}})$ (defined in the formulation of Theorem 7.1) is in the basin of attraction of the

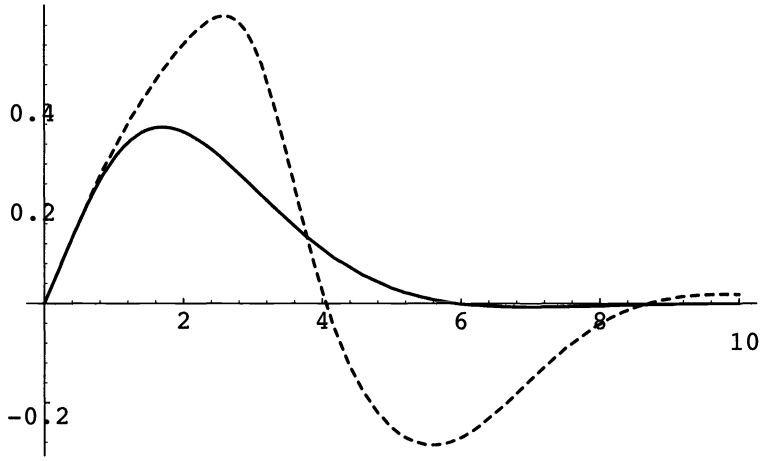


FIG. 1. $x_1(t)$ *for* $x_1(0) = 0, x_2(0) = 0.3, x_3(0) = 0.3$.

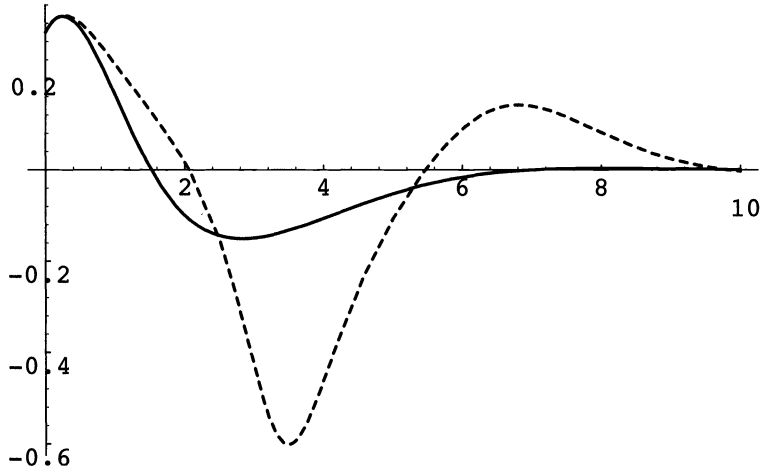


FIG. 2. $x_2(t)$ *for* $x_1(0) = 0, x_2(0) = 0.3, x_3(0) = 0.3$.

origin. As we have checked, the whole $\mathcal{M}$ was in the basin of attraction of the origin. The basin of attraction was actually much larger than $\mathcal{M}$, even though the condition (57) was not satisfied. (Note that Theorem 7.1 gives only an underestimate of the actual stability region.) For comparison, we considered the control based on Jacobian linearization $u_{\text{jac}} := Kx$ for the same gain matrix K. Note that the x and z coordinates agree up to first order and that both control schemes u_{afl} and u_{jac} yield the same linear part of the closed-loop system with eigenvalues $-1, -0.707107 + i0.707107, -0.707107 - i0.707107$. We checked that for u_{jac} condition (57) failed to hold on $\mathcal{M}$, with $\sup \eta_{\text{jac}}(x) \approx 5.6$ (11 times more than for u_{afl}), where $\eta_{\text{jac}}(x) := \frac{2\langle x, M(Kx)(E_{\text{jac}}(x))\rangle}{\langle x, Nx\rangle}$, $E_{\text{jac}} := [ax_3^3 + bx_1^3, cx_1^2x_2, 0]^T$. Not whole $\mathcal{M}$ was in the region of stability for u_{jac}, and the region of stability for u_{jac} was strictly contained in the the region of stability for u_{afl}. We present (in Figures 1–3) typical plots of the state variables as functions of time. (The continuous lines represent the time responses for u_{afl}, the dashed lines for u_{jac}.) Comparing those responses of our system for both control schemes, we see that u_{afl} offered faster convergence to the origin and less oscillatory responses than u_{jac}. We also plot (in Figures 4 and 5) the terms $\eta_{afl}(\Phi(x))$ and $\eta_{\text{jac}}(x)$ along trajectories, because they in some sense measure nonlinearity of the corresponding closed-loop systems. Observe that the strong and unfriendly nonlinearities prevail in the closed-loop system with u_{jac} control when compared to weak nonlinearities in the closed-loop system with u_{afl} control.

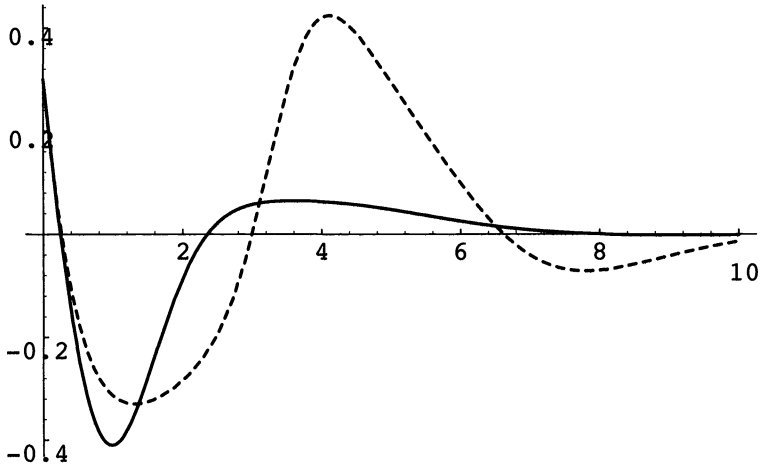


FIG. 3. $x_3(t)$ *for* $x_1(0) = 0, x_2(0) = 0.3, x_3(0) = 0.3$.

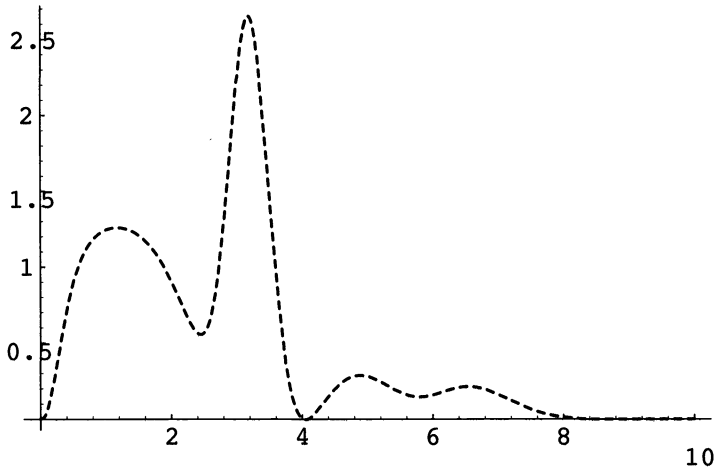


FIG. 4. η_{jac} *for* $x_1(0) = 0, x_2(0) = 0.3, x_3(0) = 0.3$.

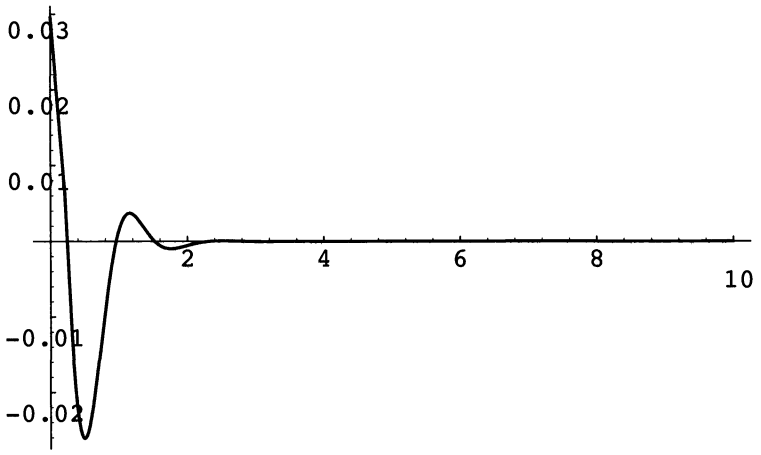


FIG. 5. η_{afl} *for* $x_1(0) = 0, x_2(0) = 0.3, x_3(0) = 0.3$.

Note that from §3 we actually know that the optimal integrating factor for $\omega_0 = dx_1 - ax_3^2 dx_2$ is $\beta = |\omega_0|^{-1} = (1 + a^2 x_3^4)^{-1/2}$. Observe that, with $a = 0.01$ and $|x_3| < 0.36$, we have $\beta \approx 1$ up to six decimal places. We have found the coresponding change of coordinates and performed simulations, but the results were indistinguishable from the case $\beta = 1$.

8. Conclusion. In this paper, we presented an approach for finding feedback linearizable systems that approximate a given single-input nonlinear system on a given compact region of the state space. We have shown that if the system is close to being involutive, then it is also close to being linearizable. We have applied this approach for design of locally stabilizing feedback laws for nonlinear systems that are close to being linearizable. The main idea was to study the characteristic one-forms rather than deal with the characteristic distribution directly. In this approach two issues have occurred: first, how to scale characteristic forms; second, how to approximate them by exact forms. We have presented some ideas on that subject and indicated some open problems.

Acknowledgment. We acknowledge the use of the "Differential Forms" Mathematica package created by Frank Zizza of Willamette University.

REFERENCES

[1] R. ABRAHAM, J. MARSDEN, AND T. RATIU, *Manifolds, Tensor Analysis, and Applications*, Addison-Wesley, Reading, MA, 1983.

[2] A. BANASZUK AND J. HAUSER, *Approximate feedback linearization: Approximate integrating factors*, in Proc. 1994 American Control Conf., Baltimore, MD, June 1994, pp. 1690–1694.

[3] A. BANASZUK, J. HAUSER, AND D. TAYLOR, *L^2-approximate feedback linearization using the Hodge decomposition theorem*, in Proc. International Symp. on Implicit and Nonlinear Systems, Fort Worth, TX, 1992, pp. 329–333.

[4] A. BANASZUK, A. ŚWIĘCH, AND J. HAUSER, *Approximate feedback linearization: least squares approximate integrating factors*, in Proc. 33rd IEEE Conf. on Decision and Control, Lake Buena Vista, FL, December 1994, pp. 1621–1626.

[5] ———, *Approximate feedback linearization: Higher order approximate integrating factors*, in Proc. 1995 NOLCOS, Lake Tahoe, CA, June 1995.

[6] ———, *Least squares integration of one-dimensional codistributions with application to approximate feedback linearization*, Math. Control Signals Systems, (1996), submitted.

[7] R. BRYANT, S. CHERN, R. GARDNER, H. GOLDSCHMIDT, AND P. GRIFFITHS, *Exterior Differential Systems*, Springer-Verlag, New York, 1991.

[8] D. EDELEN, *Applied Exterior Calculus*, Wiley, New York, 1985.

[9] H. FLANDERS, *Differential Forms with Applications to the Physical Sciences*, Academic Press, New York, 1963.

[10] J. HAUSER, *Nonlinear control via uniform system approximation*, Systems Control Lett., 17 (1991), pp. 145–154.

[11] J. HAUSER, S. SASTRY, AND P. KOKOTOVIC, *Nonlinear control via approximate input-output linearization: The ball and beam example*, IEEE Trans. Automat. Control, AC–37 (1992), pp. 392–398.

[12] N. HICKS, *Notes on Differential Geometry*, Van Nostrand, New York, 1965.

[13] L. HUNT AND J. TURI, *A new algorithm for constructing approximate transformations for nonlinear systems*, IEEE Trans. Automat. Control, AC–38 (1993), pp. 1553–1556.

[14] L. R. HUNT, R. SU, AND G. MEYER, *Global transformations of nonlinear systems*, IEEE Trans. Automat. Control, AC–28 (1983), pp. 24–31.

[15] B. JAKUBCZYK AND W. RESPONDEK, *On linearization of control systems*, Bull. Polish Acad. Sci. Math., XXVIII (1980), pp. 517–522.

[16] A. KRENER, *Approximate linearization by state feedback and coordinate change*, Systems Control Lett., 5 (1984), pp. 181–185.

[17] A. KRENER AND B. MAAG, *Controller and observer design for cubic systems*, in Modeling, Estimation and Control of Systems with Uncertainty, G. DiMasi, A. Gombani, and A. Kurzhansky, eds., Birkhäuser-Boston, Cambridge, MA, 1991, pp. 224–239.

[18] Z. XU AND J. HAUSER, *Higher order approximate feedback linearization about a manifold*, J. Math. Systems Estim. Control, 4 (1994), pp. 451–465.

[19] ———, *Higher order approximate feedback linearization about a manifold for multi-input systems*, IEEE Trans. Automat. Control, AC-40 (1995), pp. 833–840.

[20] G. ZAMPIERI, *Finding domains of invertibility for smooth functions by means of attraction basins*, J. Differential Equations, 1 (1993), pp. 11–19.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1555–1567, September 1996

005

PERTURBED OPTIMIZATION IN BANACH SPACES III: SEMI-INFINITE OPTIMIZATION*

J. FRÉDÉRIC BONNANS† AND ROBERTO COMINETTI‡

Abstract. This paper is devoted to the study of perturbed semi-infinite optimization problems, i.e., minimization over $\mathbb{R}^n$ with an infinite number of inequality constraints. We obtain the second-order expansion of the optimal value function and the first-order expansion of approximate optimal solutions in two cases: (i) when the number of binding constraints is finite and (ii) when the inequality constraints are parametrized by a real scalar.

These results are partly obtained by specializing the sensitivity theory for perturbed optimization developed in part I (cf. [*SIAM J. Control Optim.*, 34 (1996), pp. 1151–1171]) and deriving specific sharp lower estimates for the optimal value function which take into account the curvature of the positive cone in the space $C(\Omega)$ of continuous real-valued functions.

Key words. sensitivity analysis, marginal function, approximate solutions, directional constraint qualification, semi-infinite programming, epilimits

AMS subject classifications. 46N10, 47H19, 49K27, 49K40, 58C15, 90C31

1. Introduction.

This paper is the last of a trilogy devoted to the analysis of parametric optimization problems of the form

$$\min_x \{f(x,u) : G(x,u) \in K\},$$

with X and Y Banach spaces, K a closed convex subset of Y, and $f(x,u)$ and $G(x,u)$ mappings of class $\mathcal{C}^2$ from $X \times \mathbb{R}$ into $\mathbb{R}$ and Y, respectively. This third part is devoted to the study of the parametric semi-infinite optimization problem

$$(\mathrm{P}_u) \qquad \min_x \{f(x,u) : G(x,u)_\omega \geq 0, \ \forall \omega \in \Omega\},$$

where Ω is a compact metric space; $G(x,u) := \{G(x,u)_\omega\}_{\omega\in\Omega}$ belongs to $C(\Omega)$, the space of continuous functions on Ω endowed with the max norm; and the mapping $(x,u) \to (f(x,u), G(x,u))$ is of class $\mathcal{C}^2$ from $\mathbb{R}^n \times \mathbb{R}^+$ into $\mathbb{R} \times C(\Omega)$. Since

$$C_+(\Omega) := \{y \in C(\Omega) \ : \ y \geq 0\}$$

is a closed convex cone in the Banach space $C(\Omega)$, it follows that (P_u) is a particular case of the above abstract optimization problem.

Semi-infinite optimization problems occur in robust control theory, the design of filters, the design of devices having to respect some specifications in a certain range of pressure and temperature, and optimal control problems when the control has a finite-dimensional parametrization; see [16]. However, the wealth of applications is not the only motivation for studying semi-infinite optimization. In the past few years, a rather complete perturbation theory has been developed for optimization problems with a finite number of constraints, the so-called perturbed nonlinear programming problem; see [2], [6], [8], [19]. The theory of perturbed semi-infinite optimization problems, although it seems much easier than the general perturbation problem in Banach space, includes an essential difficulty related to the curvature of $C_+(\Omega)$. As a consequence, the standard second-order upper and lower estimates for the cost

*Received by the editors November 11, 1994; accepted for publication (in revised form) May 6, 1995. This research was supported by the French–Chilean ECOS Program and European Community contract 931091CL.

†INRIA-Rocquencourt, B.P. 105, 78153 Rocquencourt, France.

‡Universidad de Chile, Casilla 170/3 Correo 3, Santiago, Chile. The research of this author was prtially supported by Fondecyt grant 1940564.

do not, in general, coincide. Our main contribution here is to exhibit a sharp lower estimate that, in some cases, is equal to the parabolic upper estimate.

There is already a large body of literature on semi-infinite optimization; see the early references [3] and [11]. The recent review [9] describes in particular the so-called *reduction theory* that reduces (P_u) to an optimization problem with a finite number of constraints (see also [13]). This reduction is possible when the contact set includes a finite number of points and each of them can be expressed locally as a function of the data, typically a local solution of an optimization problem with finitely many constraints. Then the perturbation theory for nonlinear programming can be applied for deriving optimality conditions as well as for conducting a perturbation analysis; see the early reference [18].

In this paper we do not use any reduction device. In this way we may handle some cases where there is a continuum of binding constraints, especially when Ω is a one-dimensional interval. We are also able to treat the case of a finite number of binding constraints in cases where the reduction theory does not apply.

The paper is organized as follows. In §2 we discuss the directional qualification condition introduced in part I. We characterize it and show how to deduce a first-order upper estimate. Section 3 is devoted to the parabolic (second-order) upper estimates. There we combine the technique of parabolic estimates with the directional qualification condition and a characterization of second-order tangent sets to $C_+(\Omega)$, recently obtained in [7]. This upper estimate combined with the strong quadratic growth condition implies the upper Lipschitz property for the set of solutions. In §4 we discuss some sharp lower estimates. We use there specific properties of semi-infinite optimization, among them the fact that an extremal multiplier has a finite support. Then in §5 we recapitulate and state our main result.

2. Directional qualification. We start with some notations. The feasible set, value function, and set of solutions of (P_u) are denoted

$$F(u) := \{x \in \mathbb{R}^n \ : \ G(x,u) \geq 0\},$$
$$v(u) := \inf\{f(x,u) \ : \ x \in F(u)\},$$
$$S(u) := \{x \in F(u) \ : \ f(x,u) = v(u)\}.$$

Similarly, given any optimization problem (P), we define F(P), v(P), and S(P) as the feasible set, value function, and set of solutions of (P).

We recall that the dual space of $C(\Omega)$ is the set $M(\Omega)$ of bounded measures; see, e.g., [22]. If $(\lambda, y) \in M(\Omega) \times C(\Omega)$, then $\langle \lambda, y \rangle = \int_\Omega y(\omega) d\lambda(\omega)$. The support of $\lambda \in M(\Omega)$, denoted supp(λ), is defined as the complement of the greatest open subset of Ω over which $|\lambda|$ is null. The negative cone of $M(\Omega)$ is denoted $M_-(\Omega)$.

The Lagrangian function associated with (P_u) is

$$\mathcal{L}(x,\lambda,u) = f(x,u) + \int_\Omega G(x,u)_\omega d\lambda(\omega).$$

With $x \in F(u)$ we associate the set of Lagrange multipliers

$$\Lambda_u(x) := \{\lambda \in M_-(\Omega) \ : \ \text{supp}(\lambda) \subset Z(G(x,u)) \text{ and } \mathcal{L}'_x(x,\lambda,u) = 0\},$$

where, for $y \in C(\Omega)$, the set $Z(y)$ is the contact set defined as

$$Z(y) := \{\omega \in \Omega \ : \ y(\omega) = 0\}.$$

Let us fix a particular solution $x_0 \in S(0)$ and denote $\Lambda_0 := \Lambda_0(x_0)$ and $Z_0 := Z(G(x_0,0))$. The problem with linearized data (L) and its dual (D) are

$$\text{(L)} \qquad \min_d \{f'(x_0,0)(d,1) \ : \ G'(x_0,0)(d,1) \geq 0 \text{ on } Z_0\},$$

$$\text{(D)} \qquad \max_{\lambda \in \Lambda_0} \mathcal{L}'_u(x_0,\lambda,0).$$

We consider the directional qualification hypothesis

$$\text{(DCQ)} \qquad \exists\, \hat{d} \in \mathbb{R}^n \ : \ G'(x_0, 0)(\hat{d}, 1) > 0 \text{ on } Z_0,$$

which is to be compared to the standard qualification hypothesis (see [12])

$$\text{(CQ)} \qquad \exists\, \tilde{d} \in \mathbb{R}^n \ : \ G'_x(x_0, 0)\tilde{d} > 0 \text{ on } Z_0.$$

The following is essentially known.

LEMMA 2.1. *Condition* (CQ) *is equivalent to each of the following two conditions:*

(i) *There exists* $\tilde{d} \in \mathbb{R}^n$ *such that*

$$G(x_0, 0) + G'_x(x_0, 0)\tilde{d} > 0 \text{ on } \Omega.$$

(ii) *The set* Λ_0 *is nonempty and bounded.*

Proof. The equivalence between (CQ) and (i) is proven in [20]. That (CQ) implies (ii) follows from [23]. Let us prove that (ii) implies (CQ). If (CQ) does not hold, then the linear semi-infinite optimization problem

$$\min_{d,z}\{z : G'(x_0, 0)_\omega(d, 0) + z \geq 0, \ \forall \omega \in Z_0\}$$

has value 0. It follows that $(d, z) = (0, 0)$ is a solution of this problem at which the qualification condition is satisfied by the direction $(0, 1)$. By (i) $\Rightarrow$ (ii), there exists at least one multiplier $\hat{\lambda}$. Expressing the optimality conditions, we find that $\hat{\lambda} \in M_-(\Omega) \setminus \{0\}$, $\text{supp}(\hat{\lambda}) \subset Z_0$, and $\hat{\lambda} \circ G'_x(x_0, 0) = 0$. It follows that whenever $\lambda \in \Lambda_0$ and $t \in \mathbb{R}_+$, then $\lambda + t\hat{\lambda} \in \Lambda_0$, in contradiction to (ii). □

A similar result holds for condition (DCQ).

LEMMA 2.2. *Condition* (DCQ) *is equivalent to*

(i) *there exist* $\varepsilon > 0$ *and* $\tilde{d} \in \mathbb{R}^n$ *such that*

$$G(x_0, 0) + \varepsilon G'(x_0, 0)(\tilde{d}, 1) > 0 \text{ on } \Omega.$$

If in addition $\Lambda_0 \neq \emptyset$, *then* (DCQ) *is equivalent to*

(ii) *the set* $S(\mathrm{D})$ *is nonempty and bounded.*

Proof. Noting that (DCQ) is nothing but the standard qualification condition for the set of constraints $\{G(x, u) \geq 0; \ u \geq 0\}$, and applying Lemma 2.1, we obtain the equivalence of (DCQ) and (i).

Now assume that $\Lambda_0 \neq \emptyset$. That (DCQ) implies (ii) follows from [4, Prop. 3.1]. Conversely, if (DCQ) does not hold, we have

$$\alpha := \min_{d,z}\{z : G'(x_0, 0)_\omega(d, 1) + z \geq 0, \ \forall \omega \in Z_0\} \geq 0.$$

Considering the perturbation function

$$\varphi((d, z), h) = \begin{cases} z & \text{if } h(\omega) + G'(x_0, 0)_\omega(d, 1) + z \geq 0, \ \forall \omega \in Z_0 \\ +\infty & \text{otherwise,} \end{cases}$$

we obtain the dual problem

$$\min_{\lambda}\left\{-\int_\Omega G'_u(x_0, 0)d\lambda : \lambda \in M_-(\Omega), \text{supp}(\lambda) \subset Z_0, \int_\Omega d\lambda = -1, \lambda \circ G'_x(x_0, 0) = 0\right\}.$$

Applying [4, Thm. A2] we get the existence of an optimal solution $\hat{\lambda}$ for this problem, and we have

$$\int_\Omega G'_u(x_0, 0)d\hat{\lambda} = \alpha \geq 0.$$

It follows that for each $\lambda \in S(\mathrm{D})$ and every $t > 0$ we have $\lambda + t\hat{\lambda} \in S(\mathrm{D})$, contradicting the boundedness of $S(\mathrm{D})$ stated in (ii). □

From the above lemma and [5, Prop. 5.2], it follows that (DCQ) is a particular case of the abstract directional constraint qualification of part I.

PROPOSITION 2.3. *If* (DCQ) *holds, then*

$$\limsup_{u\downarrow 0} \frac{v(u) - v(0)}{u} \leq v(\mathrm{D}) = v(\mathrm{L}).$$

Proof. This follows from Propositions 2.1 and 3.1 in [4] and Lemma 2.2 above. □

Remark. The above statements hold when f and G are merely of class C^1.

3. Second-order upper estimates. Define a *path* as a mapping $u \to x_u$ from $\mathbb{R}_+$ to X, with $x_u \to x_0$ when $u \downarrow 0$. The path is said to be feasible if $G(x_u, u) \in K$ for u small enough. In the study of second-order upper estimates, we analyze feasible paths of the form

$$x_u := x_0 + ud + \frac{u^2}{2}z + o(u^2).$$

Feasibility of x_u implies some relations between the expansion of $G(x_u, u)$ and the geometry of $C_+(\Omega)$. Given a convex subset K of a Banach space Y, we define the first-order tangent set at $y \in K$ as

$$T_K(y) := \{h \in Y : \text{ there exists } o(t) \text{ such that } y + th + o(t) \in K\}.$$

Similarly, the second-order tangent set at $y \in K$ in the direction $h \in T(y)$ is

$$T_K^2(y, h) := \left\{z \in Y : \text{ there exists } o(t^2) \text{ such that } y + th + \frac{t^2}{2}z + o(t^2) \in K\right\}.$$

For the sake of simplicity we write $T := T_{C_+(\Omega)}$ and $T^2 := T^2_{C_+(\Omega)}$ and denote the terms of the second-order expansion of $f(x_u, u)$ and $G(x_u, u)$ as

$$\Psi_f(z, d) := f'_x(x_0, 0)z + f''(x_0, 0)(d, 1)(d, 1),$$

$$\Psi_G(z, d) := G'_x(x_0, 0)z + G''(x_0, 0)(d, 1)(d, 1).$$

Expanding $G(x_u, u)$ we obtain that if x_u is a feasible path, then

$$G'(x_0, 0)(d, 1) \in T(G(x_0, 0)), \tag{1}$$

$$\Psi_G(z, d) \in T^2(G(x_0, 0), G'(x_0, 0)(d, 1)), \tag{2}$$

and, when $d \in S(\mathrm{L})$, we get

$$v(u) \leq v(0) + uv(L) + \frac{u^2}{2}\Psi_f(z, d) + o(u^2). \tag{3}$$

In [4], it was shown that an upper estimate of the second-order variation of the cost is obtained by minimizing $\Psi_f(z, d)$ over those z satisfying (2). The purpose of this section is

to make explicit this bound in the case of semi-infinite programming. In the statement of our result, we use some expressions for the tangent sets of $K = C_+(\Omega)$. The first-order tangent cone is well known (see, e.g., [20]):

$$T(y) = \{h \in C(\Omega) : h \geq 0 \quad \text{on} \quad Z(y)\}.$$

In particular, the tangent cone at $G(x_0, 0)$ is

$$T(G(x_0, 0)) = \{h \in C(\Omega) : h \geq 0 \quad \text{on} \quad Z_0\}.$$

A formula for the second-order tangent set has been recently obtained in [7]. This formula uses the concept of lower epilimit that we now recall, referring to [1] for a detailed exposition. Let $(A_t)_{t>0}$ be a family of subsets of a Banach space Y. The upper limit of $(A_t)_{t>0}$ at $t = 0$ in the sense of Painlevé–Kuratowski is defined as

$$\limsup_{t\downarrow 0} A_t := \left\{ y \in Y : \liminf_{t\downarrow 0} d(y, A_t) = 0 \right\}.$$

The lower epilimit of a family $(f_t)_{t>0}$ of extended real-valued functions on the topological space K is defined as the function whose epigraph is $\limsup_{t\downarrow 0} \operatorname{epi} f_t$, where

$$\operatorname{epi} f_t := \{(x, r) \in K \times \mathbb{R} : f_t(x) \leq r\}$$

is the epigraph of f_t. An alternative characterization is given by

$$\text{e-}\liminf_{t\downarrow 0} f_t(x) = \sup_{V \in \mathcal{N}(x)} \liminf_{t\downarrow 0} \inf_{y \in V} f_t(y) = \liminf_{(t,y)\to(0^+,x)} f_t(y),$$

where $\mathcal{N}(x)$ is the set of neighborhoods of x.

PROPOSITION 3.1 (cf. [7]). *Let $y \in C_+(\Omega)$ and $h \in T(y)$. Then*

$$T^2(y, h) = \{h \in C(\Omega) : \ h + \tau(y, h) \geq 0\},$$

where $\tau(f, v)$ is the lower semicontinuous extended real-valued function defined by

$$\tau(f, v) := \text{e-}\liminf_{t\downarrow 0} \left[\frac{f + tv}{t^2/2} \right].$$

Equivalently, $\tau(f, v)$ is given by the formula

$$\tau(f, v)(\omega) = \begin{cases} 0 & \text{if } \omega \in \operatorname{int} Z(f) \text{ and } v(\omega) = 0, \\ -\theta(\omega) & \text{if } \omega \in \operatorname{bd} Z(f) \text{ and } v(\omega) = 0, \\ +\infty & \text{otherwise,} \end{cases} \tag{4}$$

where

$$\theta(\omega) := \limsup_{\substack{y\to\omega \\ f(y)>0}} \frac{[-v(y)]_+^2}{2f(y)}. \tag{5}$$

In view of this result, defining

$$\Xi(d, u)_\omega := \frac{G(x_0, 0)_\omega + uG'(x_0, 0)_\omega(d, 1)}{u^2/2},$$

$$\tau_d(\omega) := \text{e-}\liminf_{u\downarrow 0} \Xi(d, u)_\omega,$$

we get the characterization

$$T^2(G(x_0,0);G'(x_0,0)(d,1)) = \{h \in C(\Omega) :\ h + \tau_d \geq 0\}.$$

Writing $T^2(d)$ for the above set for brevity, we see that $T^2(d) \neq \phi$ if and only if $\tau_d > -\infty$. In such a case the support function of $T^2(d)$ can be characterized as

$$\sigma(\lambda, T^2(d)) := \sup\left\{\int_\Omega h(\omega)d\lambda(\omega) : h \in T^2(d)\right\} = -\int_\Omega \tau_d(\omega)d\lambda(\omega)$$

for all $\lambda \in S(D)$. (Since $\lambda \leq 0$ and τ_d is lower semicontinuous and nonpositive on $\mathrm{supp}(\lambda)$, the integral on the right-hand side above is well defined.) To make this equality always valid we define $\int_\Omega \tau_d(\omega)d\lambda(\omega) := +\infty$ whenever τ_d takes the value $-\infty$.

The function $-\tau_d$ may be interpreted as an *upper curvature function.*

Given $d \in S(\mathrm{L})$, the relations (1)–(3) suggest consideration of the subproblem

$$(\mathrm{L}_d) \qquad \min_z\{\Psi_f(z,d) : \Psi_G(z,d) + \tau_d \geq 0\},$$

with which we associate a dual formulation

$$(\mathrm{D}_d) \qquad \max_{\lambda \in S(\mathrm{D})} \mathcal{L}''(x_0,\lambda,0)(d,1)(d,1) + \int_\Omega \tau_d(\omega)d\lambda(\omega).$$

We also introduce the problem

$$(\mathrm{Q}) \qquad \min\{v(\mathrm{L}_d) : d \in S(\mathrm{L})\}.$$

Whenever $v'(0)$ exists, we define the upper and lower second-order Dini derivatives

$$(6) \qquad v''_+(0) := \limsup_{u\downarrow 0} 2[v(u) - v(0) - uv'(0)]/u^2,$$

$$(7) \qquad v''_-(0) := \liminf_{u\downarrow 0} 2[v(u) - v(0) - uv'(0)]/u^2.$$

PROPOSITION 3.2. *If* (DCQ) *holds, then* $v(\mathrm{L}_d) = v(\mathrm{D}_d)$ *for all* $d \in S(\mathrm{L})$, *and we have* $v(\mathrm{Q}) < +\infty$ *iff there exists* $d \in S(\mathrm{L})$ *such that* $\tau_d(\omega) > -\infty$ *for all* $\omega \in \Omega$. *Moreover, if* $S(\mathrm{L})$ *is nonempty and* $v(\mathrm{Q}) > -\infty$, *we then have*

$$(8) \qquad v(u) \leq v(0) + uv(\mathrm{L}) + \frac{u^2}{2}v(Q) + o(u^2).$$

In particular, if there exists $v'(0) = v(\mathrm{L})$, *we get*

$$(9) \qquad v''_+(0) \leq \inf_{d\in S(\mathrm{L})} \max_{\lambda\in S(\mathrm{D})} \left\{\mathcal{L}''(x_0,\lambda,0)(d,1)(d,1) + \int_\Omega \tau_d(\omega)d\lambda(\omega)\right\}.$$

Proof. This is a consequence of Propositions 2.1, 2.2, and 4.2 in [4]. □

By the above Proposition, $v(\mathrm{Q}) < +\infty$ iff $\tau_d > -\infty$ for some $d \in S(\mathrm{L})$. We show that a sufficient condition for this is a quadratic growth condition, recently introduced in [20].

LEMMA 3.3. *Suppose that* $G'(x_0,0)$ *is Lipschitz with respect to* ω *and assume that* $G(x_0,0)$ *satisfies the quadratic growth condition*

$$(\mathrm{QGC}) \qquad \exists\, c > 0 \text{ such that } G(x_0,0)_\omega \geq c\,\mathrm{dist}(\omega, Z_0)^2.$$

Then $\tau_d > -\infty$ *for all* $d \in S(\mathrm{L})$.

Proof. Using (4) it suffices to show that for all $\omega \in \text{bd } Z_0$ with $G'(x_0, 0)(d, 1)_\omega = 0$ we have $\theta(\omega) < +\infty$. To this end let L be a Lipschitz constant for $\omega \to G'(x_0, 0)(d, 1)_\omega$. For each $\omega \notin Z_0$ let ω_0 be a projection of ω onto Z_0. Then ω_0 lies on the boundary of Z_0 so that $G(x_0, 0)_{\omega_0} = 0$, and since $G'(x_0, 0)(d, 1) \in T(G(x_0, 0))$ we deduce $G'(x_0, 0)(d, 1)_{\omega_0} \geq 0$. We obtain

$$-G'(x_0, 0)(d, 1)_\omega \leq -G'(x_0, 0)(d, 1)_{\omega_0} + L\, d(\omega, w_0) \leq L\ \text{dist}(\omega, Z_0).$$

From this and (QGC) we deduce

$$[-G'(x_0, 0)(d, 1)_\omega]_+^2 \leq L^2\ \text{dist}(\omega, Z_0)^2 \leq \frac{L^2}{c} G(x_0, 0)_\omega,$$

which implies $\theta(\omega) \leq L^2/(2c) < +\infty$. □

Remark. Assuming (DCQ), we know by Lemma 2.2 that $S(\mathrm{D})$ is bounded. Let $d \in S(\mathrm{L})$. If (QGC) holds, as a consequence of the bound established for $\theta(\omega)$, the amount $\int_\Omega \tau_d(\omega) d\lambda(\omega)$ is bounded uniformly for $\lambda \in S(\mathrm{D})$, so $v(\mathrm{D}_d)$ is finite.

4. Stability of solutions. We state next a sufficient condition for ensuring a Lipschitz behavior of the (sub)optimal paths x_u of the perturbed problems. The result, which is a rather straightforward application of the preceding discussion and [4, Prop. 6.3], is based on the strong second-order sufficient condition

$$\text{(SOC)} \qquad \max_{\lambda \in S(\mathrm{D})} \mathcal{L}''_x(x_0, \lambda, 0)dd > 0 \ \text{ for all } d \text{ in } C \backslash \{0\},$$

where C denotes the critical cone

$$C := \{d \in \mathbb{R}^n \ :\ f'_x(x_0, 0)d \leq 0 \ \text{ and } \ G'_x(x_0, 0)d \geq 0 \text{ on } Z_0\}.$$

PROPOSITION 4.1. *Suppose that*

(i) (DCQ) *holds and* $\Lambda_0 \neq \emptyset$.

(ii) *There exists* $d \in S(\mathrm{L})$ *such that* $\tau_d(\omega) > -\infty\ \forall \omega \in \Omega$.

(iii) (SOC) *holds.*

Then every $O(u^2)$*-optimal path* x_u *satisfies* $x_u = x_0 + O(u)$.

Remark. Combining Lemma 3.3 and Proposition 4.1, we obtain a sufficient condition for Lipschitz behavior of solutions, similar to the result of [20].

Remark. Following [10], we may introduce the strong quadratic growth condition

$$\text{(SQG)} \qquad \exists \alpha > 0, c > 0,\ F(x, u) \geq v(0) + uv(L) + \alpha\, \text{dist}(x, S_0)^2 - cu^2,$$

where

$$F(x, u) := \begin{cases} f(x, u) & \text{if } G(x, u) \geq 0, \\ +\infty & \text{if not.} \end{cases}$$

Then we can show that (SOC) $\Rightarrow$ (SQG), and Proposition 4.1 is still valid if we replace assumption (iii) by (SQG).

5. Lower estimates. We recall for reference the following standard lower estimate, which is in fact a particular case of the general lower estimate of part I.

LEMMA 5.1. *Let us assume* (DCQ) *and suppose that there exists a path of* $o(u^2)$*-optimal solutions* x_u *satisfying* $x_u = x_0 + O(u)$. *Then* $\Lambda_0 \neq \emptyset$, $v'(0)$ *exists with* $v'(0) = v(\mathrm{L}) = v(\mathrm{D})$, $S(\mathrm{L}) \neq \emptyset$, *and we have*

$$v''_-(0) \geq \inf_{d \in S(\mathrm{L})} \max_{\lambda \in S(\mathrm{D})} \mathcal{L}''(x_0, \lambda, 0)(d, 1)(d, 1).$$

Proof. The existence of $v'(0)$, as well as the equality $v'(0) = v(\mathrm{L}) = v(\mathrm{D})$, follows from [4, Prop. 3.2]. The finiteness of $v'(0) = v(\mathrm{D})$ implies that $\Lambda_0 \neq \emptyset$, and [4, Prop. 3.3] gives $S(\mathrm{L}) \neq \emptyset$. The lower estimate on $v''_-(0)$ is then obtained by applying [4, Prop. 4.3(b)]. □

Comparing the previous lower estimate to the upper estimate (9) in §3, we observe a gap due to the curvature of $C_+(\Omega)$ at $G(x_0, 0)$. More precisely, if

$$\tau_d = 0 \text{ on } \operatorname{supp}(\lambda) \quad \forall \lambda \in S(\mathrm{D}),$$

then the two estimates coincide, but as one can see from (4) and (5), this may occur only in some very special situations.

We are then led to search for sharper lower estimates. We will obtain a lower estimate on $v''_-(0)$ involving the upper epilimit of $\Xi(d, u)$.

We recall the concept of upper epilimit, for which we refer again to [1]. Let $(A_t)_{t>0}$ be a family of subsets of a Banach space Y. The lower limit of $(A_t)_{t>0}$ at $t = 0$ in the sense of Painlevé–Kuratowski is defined as

$$\liminf_{t\downarrow 0} A_t := \left\{ y \in Y : \limsup_{t\downarrow 0} d(y, A_t) = 0 \right\}.$$

The upper epilimit of a family $(f_t)_{t>0}$ of extended real-valued functions on the topological space K is defined as the function whose epigraph is $\liminf_{t\downarrow 0} \operatorname{epi} f_t$. A useful formula is

$$\text{e-}\limsup_{t\downarrow 0} f_t(x) = \sup_{V\in\mathcal{N}(x)} \limsup_{t\downarrow 0} \inf_{y\in V} f_t(y). \tag{10}$$

When the upper and lower epilimits coincide at a given point, we shall say that the family of functions *epiconverges* at that point, and we shall denote

$$\text{e-}\lim_{t\downarrow 0} f_t(x) = \text{e-}\liminf_{t\downarrow 0} f_t(x) = \text{e-}\limsup_{t\downarrow 0} f_t(x).$$

We shall say that the family of functions epiconverges on a subset $K_0 \subset K$ if it epiconverges at each point of K_0.

The next proposition makes use of the set of extreme points of $S(\mathrm{D})$, which will be denoted $S^*(\mathrm{D})$. The result will be derived under a technical assumption ($\mathrm{H}_{\hat\omega}$), which will be further clarified afterwards. $B(\hat\omega, r)$ denotes the ball of center $\hat\omega$ and radius r.

PROPOSITION 5.2. *Assume* (DCQ) *and suppose that there exists an $o(u^2)$-optimal path such that $x_u = x_0 + O(u)$. Let $G(x_0, 0)$, $G'(x_0, 0)$, and $G''(x_0, 0)$ be Lipschitz with respect to ω and assume also that for every $d \in S(\mathrm{L})$, each $\lambda \in S^*(\mathrm{D})$, and every $\hat\omega \in \operatorname{supp}(\lambda)$, one has*

$$(\mathrm{H}_{\hat\omega}) \qquad \exists\, V \in \mathcal{N}(\hat\omega), \exists\, r > 0 \text{ s.t. } \inf_{\omega\in V} \Xi(d, u)_\omega = \inf_{\omega\in B(\hat\omega, ru)} \Xi(d, u)_\omega + o(1),$$

with $o(1)$ converging to 0 uniformly as $u \downarrow 0$. Then

$$v''_-(0) \geq \inf_{d\in S(L)} \max_{\lambda\in S(\mathrm{D})} \left\{ \mathcal{L}''(x_0, \lambda, 0)(d, 1)(d, 1) + \int_\Omega \bar\tau_d(\omega) d\lambda(\omega) \right\}, \tag{11}$$

where

$$\bar\tau_d(\omega) := \text{e-}\limsup_{u\downarrow 0} \Xi(d, u)_\omega.$$

For $\int_\Omega \bar\tau_d(\omega) d\lambda(\omega)$ we adopt the same convention as in §3: its value is the usual integral when $\bar\tau_d > -\infty$, and $+\infty$ when $\bar\tau_d$ takes the value $-\infty$. With this convention

$$\int_\Omega \bar{\tau}_d(\omega)d\lambda(\omega) = -\sigma(\lambda, \{h \in C(\Omega); h + \bar{\tau}_d \geq 0\}),$$

which is an upper semicontinuous function of $\lambda \in S(\mathrm{D})$.

The function $-\bar{\tau}_d$ may be interpreted as a *lower-curvature function.*

We discuss some consequences of this proposition, postponing the proof until the end of the section.

Comparing the bounds obtained for $v''_+(0)$ and $v''_-(0)$ in Propositions 3.2 and 5.2, we see that the only difference is between the terms $\text{e-lim inf}_{u\downarrow 0}\, \Xi(d, u)$ and $\text{e-lim sup}_{t\downarrow 0}\, \Xi(d, u)$. The statement below follows.

COROLLARY 5.3. *In addition to the assumptions of Proposition* 5.2, *let us suppose that* $\Xi(d, u)$ *epiconverges on* $\mathrm{supp}(\lambda)$ *for each* $\lambda \in S^*(\mathrm{D})$ *and* $d \in S(\mathrm{L})$. *Then there exists*

$$v''(0) = \inf_{d\in S(L)} \max_{\lambda\in S(\mathrm{D})} \left\{ \mathcal{L}''(x_0, \lambda, 0) + \int_\Omega \text{e-lim}_{u\downarrow 0}\, \Xi(d, u)_\omega d\lambda(\omega) \right\}.$$

It remains to find sufficient conditions to ensure the technical assumption $(\mathrm{H}_{\hat{\omega}})$ in Proposition 5.2. Lemma 5.5 below gives a result in this direction. We first need a technical lemma which describes the structure of extreme points of $S(\mathrm{D})$.

LEMMA 5.4. *Suppose* (DCQ) *and* $\Lambda_0 \neq \emptyset$. *Then* $S(\mathrm{D})$ *is the closed convex hull of* $S^*(\mathrm{D})$, *and any* $\lambda \in S^*(\mathrm{D})$ *is of the form*

$$\lambda = \sum_{i=1}^{p} \lambda_{\omega_i} \delta_{\omega_i} \tag{12}$$

with $p \leq n$, $\lambda_{\omega_i} < 0$, *and* δ_{ω_i} *being the Dirac mass at* ω_i. (*Recall that* n *is the dimension of the space to which* x *belongs.*)

Proof. By Lemma 2.2, the set $S(\mathrm{D})$ is nonempty and bounded. Being closed, it is weak* compact. The Krein–Milman theorem implies that $S(\mathrm{D})$ is the closed convex hull of its extreme points (see, e.g., [22]). Now, $S(\mathrm{D})$ is a face of Λ_0 so that the points in $S^*(\mathrm{D})$ are also extreme points of Λ_0, and the latter are known to be the sum of at most n Dirac masses (see [20]). □

LEMMA 5.5. *Suppose that* $G(x_0, 0)$, $G'(x_0, 0)$, *and* $G''(x_0, 0)$ *are Lipschitz with respect to* ω *and* (QGC) *holds. Under each of the following conditions, property* $(\mathrm{H}_{\hat{\omega}})$ *is satisfied* $\forall\, \hat{\omega} \in \mathrm{supp}(\lambda)$, $\forall\, \lambda \in S^*(\mathrm{D})$, $\forall\, d \in S(\mathrm{L})$:

(i) Z_0 *is a finite set.*

(ii) Ω *is an interval, and* Z_0 *is the union of finitely many intervals.*

Proof. By the previous lemma, $\lambda \in S^*(D)$ has a finite support included in Z_0. Then, using $d \in S(\mathrm{L})$, we obtain

$$G(x_0, 0)_{\hat{\omega}} = G'(x_0, 0)(d, 1)_{\hat{\omega}} = 0 \quad \forall\, \hat{\omega} \in \mathrm{supp}(\lambda).$$

From this we deduce that for all $V \in \mathcal{N}(\hat{\omega})$ we have for all $r > 0$, $u > 0$ small enough

$$\inf_{\omega\in V} \Xi(d, u)_\omega \leq \inf_{\omega\in B(\hat{\omega}, ru)} \Xi(d, u)_\omega \leq \Xi(d, u)_{\hat{\omega}} = 0. \tag{13}$$

Let V be a closed neighborhood of $\hat{\omega}$ such that $\mathrm{dist}(\omega, Z_0) = d(\omega, \hat{\omega})$ whenever $\omega \in V \setminus Z_0$: this is possible by (i) or (ii). Let us consider a Lipschitz constant L for

$$g(\omega) := G'(x, 0)(d, 1)_\omega,$$

and let ω_u minimize $\Xi(d, u)_\omega$ over $\omega \in V$. If $\omega_u \in Z_0$, then we get $\Xi(d, u)_{\omega_u} \geq 0$, which combined with (13) gives $(\mathrm{H}_{\hat{\omega}})$ with $o(1) \equiv 0$ and $r > 0$ arbitrary. Let us then assume that

$\omega_u \notin Z_0$. Then we may use (QGC) and (13) to obtain

$$c\ \mathrm{dist}(\omega_u, Z_0)^2 + ug(\omega_u) \le G(x_0, 0)_{\omega_u} + ug(\omega_u) = \frac{u^2}{2}\Xi(d, u)_{\omega_u} \le \frac{u^2}{2}\Xi(d, u)_{\hat{\omega}} = ug(\hat{\omega}).$$

Since $g(\omega)$ is Lipschitzian, we get

$$c\, d(\omega_u, \hat{\omega})^2 \le u(g(\hat{\omega}) - g(\omega_u)) \le uL\, d(\omega_u, \hat{\omega}),$$

and then $d(\omega_u, \hat{\omega}) \le uL/c$. This proves that minimizing $\Xi(d, u)_\omega$ over $\omega \in V$ is equivalent to minimizing it over $B(\hat{\omega}, uL/c)$, proving $(H_{\hat{\omega}})$ with $r = L/c$ and $o(1) \equiv 0$. □

We now return to the proof of Proposition 5.2, starting with the following lemma, which does not use the specific properties of semi-infinite optimization.

LEMMA 5.6. *Assume* (DCQ) *and suppose that there exists an* $o(u^2)$*-optimal path such that* $x_u = x_0 + O(u)$. *Then there exists* $u_k \downarrow 0$ *and* $d \in S(\mathrm{L})$ *such that for any* $\lambda \in S(\mathrm{D})$

$$v''_-(0) = \mathcal{L}''(x_0, \lambda, 0)(d, 1)(d, 1) - \lim_{k\to\infty} \int_\Omega \Psi_G(z^k, d)_\omega d\lambda(\omega), \tag{14}$$

where

$$z^k := \frac{x_{u_k} - x_0 - u_k d}{u_k^2/2}. \tag{15}$$

Proof. Let us take a sequence $u_k \downarrow 0$ such that

$$2(v(u_k) - v(0) - u_k v(L))/u_k^2 \to v''_-(0). \tag{16}$$

Denoting $x^k = x_{u_k}$ and passing to a subsequence we may also assume that

$$(x^k - x_0)/u_k \to d \tag{17}$$

for some $d \in \mathbb{R}^n$, which, by [4, Prop. 3.3], satisfies $d \in S(\mathrm{L})$.

Since $x_u = x_0 + O(u)$, a second-order expansion of G gives

$$G(x^k, u_k) = G(x_0, 0) + u_k G'(x_0, 0)(d, 1) + \frac{u_k^2}{2}\Psi_G(z^k, d) + o(u_k^2).$$

Since $d \in S(\mathrm{L})$ and $\lambda \in S(\mathrm{D})$ we have

$$\langle \lambda, G(x_0, 0)\rangle = \langle \lambda, G'(x_0, 0)(d, 1)\rangle = 0,$$

from which we get

$$\langle \lambda, G(x^k, u_k)\rangle = \frac{u_k^2}{2}\langle \lambda, \Psi_G(z^k, d)\rangle + o(u_k^2).$$

Taking into account that $\mathcal{L}(x_0, \lambda, 0) = v(0)$, $\mathcal{L}'_x(x_0, \lambda, 0) = 0$, and $\mathcal{L}'_u(x_0, \lambda, 0) = v'(0)$, we deduce

$$v(u_k) = f(x^k, u_k) + o(u_k^2) = \mathcal{L}(x^k, \lambda, u_k) - \langle \lambda, G(x^k, u_k)\rangle + o(u_k^2)$$

$$= v(0) + u_k v'(0) + \frac{u_k^2}{2}[\mathcal{L}''(x_0, \lambda, 0)(d, 1)(d, 1) - \langle \lambda, \Psi_G(z^k, d)\rangle] + o(u_k^2),$$

from which the conclusion follows. □

In order to get the best lower estimate, and recalling that the multiplier is nonpositive, we must minorize $\Psi_G(z^k, d)$ efficiently. Note that by expanding $G(x^k, u_k)$, we get the following relation, which we shall use later:

$$\Psi_G(z^k, d) + \Xi(d, u^k) \geq o(1), \tag{18}$$

where the inequality is to be understood in $C(\Omega)$, i.e., $o(1) \to 0$ uniformly when $u \downarrow 0$.

We now proceed with the proof of Proposition 5.2.

Proof. Let u_k, d, z^k be chosen as in Lemma 5.6. Consider the problem

$$\max_{\lambda \in S(\mathrm{D})} \left\{ \mathcal{L}''(x_0, \lambda, 0)(d, 1)(d, 1) + \int_\Omega \bar{\tau}_d(\omega) d\lambda(\omega) \right\}.$$

As the cost function is affine and upper semicontinuous and $S(\mathrm{D})$ is weak* compact, the maximum is attained at an extreme point $\lambda^* \in S^*(\mathrm{D})$. By Lemma 5.4, $\lambda^* = \sum_{i=1}^p \lambda_{\omega_i} \delta_{\omega_i}$.

Let us take $\hat{\omega} = w_i$. From (18) we have

$$\Psi_G(z^k, d)_{\hat{\omega}} + \Xi(d, u_k)_\omega \geq \Psi_G(z^k, d)_{\hat{\omega}} - \Psi_G(z^k, d)_\omega + o(1),$$

with $o(1)$ uniform with respect to ω. Minimize the right-hand side first and then the left-hand side for $\omega \in B(\hat{\omega}, ru_k)$ to obtain

$$\Psi_G(z^k, d)_{\hat{\omega}} + \inf_{\omega \in B(\hat{\omega}, ru_k)} \Xi(d, u_k)_\omega \geq \Psi_G(z^k, d)_{\hat{\omega}} - \sup_{\omega \in B(\hat{\omega}, ru_k)} \Psi_G(z^k, d)_\omega + o(1).$$

Now if we fix $r > 0$, because $G'(x_0, 0)$ and $G''(x_0, 0)$ are Lipschitzian, using the fact that $u_k z^k \to 0$, we get

$$\Psi_G(z^k, d)_{\hat{\omega}} - \sup_{\omega \in B(\hat{\omega}, ru_k)} \Psi_G(z^k, d)_\omega \geq O(u_k)\|z^k\| + o(1) = o(1),$$

which combined with the previous estimate gives

$$\Psi_G(z^k, d)_{\hat{\omega}} + \inf_{\omega \in B(\hat{\omega}, ru_k)} \Xi(d, u_k)_\omega \geq o(1).$$

Invoking assumption $(\mathrm{H}_{\hat{\omega}})$, we may select $V \in \mathcal{N}(\hat{\omega}), r > 0$ such that

$$\liminf_{k \to +\infty} \Psi_G(z^k, d)_{\hat{\omega}} + \limsup_{u \downarrow 0} \inf_{\omega \in V} \Xi(d, u)_\omega \geq 0,$$

and therefore, by (10)

$$\liminf_{k \to +\infty} \Psi_G(z^k, d)_{\hat{\omega}} + \bar{\tau}_d(\hat{\omega}) \geq 0.$$

Combining this estimate with (14) and noting that $\lambda_{\omega_i} < 0$, we obtain (11) as required. □

6. Conclusion. Let us assume that the original problem (P_0) has a unique solution

$$S(0) = \{x_0\} \tag{H1}$$

and that we have the uniform boundedness of solutions:

$$\exists\, r > 0,\ u_0 > 0 \text{ such that for all } u \leq u_0,\ S(u) \neq \emptyset \text{ and } S(u) \subset B(0, r). \tag{H2}$$

THEOREM 6.1 (main result). *Let us assume* (H1), (H2), (DCQ), (QGC), (SOC), *and* $S(\mathrm{L}) \neq \emptyset$. *Suppose also that* $G(x_0, 0)$, $G'(x_0, 0)$, *and* $G''(x_0, 0)$ *are Lipschitz with respect to*

ω. Suppose finally that for each $d \in S(\mathrm{L})$ *there exists* $\text{e-lim}_{u \downarrow 0} \Xi(d, u)$ *and also that one of the following conditions holds:*

(i) Z_0 *is finite.*

(ii) Ω *is an interval, and* Z_0 *is the union of finitely many intervals.*

Then we have

(a) *The value function has first- and second-order (right) derivatives given by* $v'(0) = v(L)$, $v''(0) = v(Q)$. *Moreover, if* $v(Q) > -\infty$, *we have the expansion*

$$v(u) = v(0) + uv(L) + \frac{u^2}{2} v(Q) + o(u^2).$$

(b) *The set of all limit points of* $(x_u - x_0)/u$, *where* x_u *ranges over all paths of* $o(u^2)$*-optimal solutions, is included in* $S(\mathrm{Q})$. *In particular, if* $S(\mathrm{Q})$ *is a singleton, i.e.,* $S(\mathrm{Q}) = \{d\}$, *and* x_u *is as above, then* $x_u = x_0 + ud + o(u)$.

(c) *Let* $d \in S(Q)$. *If there exists* $z \in S(\mathrm{L}_d)$ *(this is always the case when* (i) *holds), then there exists an* $o(u^2)$*-optimal path* $x_u = x_0 + ud + o(u)$.

(d) *Let* λ_u *be a multiplier associated with a solution* x_u *of* (P_u). *Then all weak* limit points of* λ_u *belong to* $S(\mathrm{D})$.

Proof. From (H1) and (H2) there exists $x_u \in S(u)$ which satisfies $x_u \to x_0$ when $u \downarrow 0$. Then, from Proposition 4.1 we get $x_u = x_0 + O(u)$, and part (a) follows by combining Lemmas 5.1 and 5.5 with Corollary 5.3.

If x_u is a path of $o(u^2)$-optimal solutions, expanding $f(x_u, u)$ as in the proof of Lemma 5.6, we obtain the first statement in (b). The second statement is an immediate consequence of the first.

We now prove (c). Let $d \in S(\mathrm{Q})$ and $z \in S(\mathrm{L}_d)$. Then there exists a feasible path $x_u = x_0 + ud + \frac{u^2}{2} z + o(u^2)$. Expanding $f(x_u, u)$ and $G(x_u, u)$, we obtain $v(\mathrm{Q}) = \Psi_f(z, d)$ as well as (2). The conclusion follows.

Assertion (d) is a consequence of [4, Prop. 3.3]. □

Concluding remarks. Our final result is an extension to semi-infinite optimization of the results of the sequence of papers [8], [19], [2], and [6] in the following sense: if Ω is a finite set, then we exactly recover the above-mentioned results up to the presence of equality constraints. However, there is no difficulty in adding a finite number of equality constraints to our formulation. We avoided it for the sake of clarity of exposition and in order to concentrate on the real difficulty, which is to handle an infinite number of constraints.

Some of our hypotheses, however, may seem unduly strong. First of all, we assume that $S(\mathrm{L})$ is nonempty. While this hypothesis is automatically satisfied when the contact set is finite (due to the standard theory of linear programming), we are not aware of general criteria allowing to check nonemptyness of $S(\mathrm{L})$ for semi-infinite programming. Performing an analysis of the variation of the solutions when $S(\mathrm{L})$ is empty is an open problem. Some of the results of part II might be useful for dealing with this case.

The other hypothesis that seems excessively strong is the alternative (i) or (ii). We need it in order to satisfy the geometrical hypothesis $(\mathrm{H}_{\hat{\omega}})$. Still, the most important contribution of this paper is to present a new way of obtaining sharp lower estimates of the cost, and we hope that the technique presented here can be improved in order to deal with more general contact sets.

REFERENCES

[1] H. ATTOUCH, *Variational Convergences for Functions and Operators*, Pitman Advanced Publishing Program, Boston, 1984.

[2] A. AUSLENDER AND R. COMINETTI, *First and second order sensitivity analysis of nonlinear programs under directional constraint qualification conditions*, Optimization, 21 (1990), pp. 351–363.

[3] A. BEN-TAL, M. TEBOULLE, AND J. ZOWE, *Second order necessary optimality conditions for semi-infinite programming problems*, in Semi-Infinite Programming, Proceedings, Lecture Notes in Control and Inform. Sci. 15, R. Hettich, ed., Springer-Verlag, Berlin, 1979, pp. 17–30.

[4] J. F. BONNANS AND R. COMINETTI, *Perturbed optimization in Banach spaces* I*: A general theory based on a weak directional constraint qualification*, SIAM J. Control Optim., 34 (1996), pp. 1151–1171.

[5] ———, *Perturbed optimization in Banach spaces* II*: A theory based on a strong directional qualification*, SIAM J. Control Optim., 34 (1996), pp. 1172–1189.

[6] J. F. BONNANS, A. D. IOFFE, AND A. SHAPIRO, *Expansion of exact and approximate solutions in nonlinear programming*, in Proc. French–German Conference in Optimization, D. Pallaschke and W. Oettli, eds., Lecture Notes in Econom. and Math. Systems, Springer-Verlag, New York, 1992, pp. 103–117.

[7] R. COMINETTI AND J-P. PENOT, *Tangent sets to unilateral convex sets*, C. R. Acad. Sci. Paris Sér. I Math., (1995), pp. 1631–1636.

[8] J. GAUVIN AND R. JANIN, *Directional behavior of optimal solutions in nonlinear mathematical programming*, Math. Oper. Res., 13 (1988), pp. 629–649.

[9] R. HETTICH AND K. O. KORTANEK, *Semi-infinite programming: Theory, methods, and applications*, SIAM Rev., 35 (1994), pp. 380–429.

[10] A. D. IOFFE, *On sensitivity analysis of nonlinear programs in Banach spaces: The approach via composite unconstrained optimization*, SIAM J. Optim., 4 (1994), pp. 1–43.

[11] ———, *Second order conditions in nonlinear nonsmooth problems of semi-infinite programming*, in Semi-Infinite Programming and Applications, Lecture Notes in Control and Inform. Sci. 215, A. V. Fiacco and K. O. Kortanek, eds., Springer-Verlag, Berlin, 1983, pp. 262–280.

[12] H. T. JONGEN, F. TWILT, AND G. W. WEBER, *Semi-infinite optimization: Structure and stability of the feasible set*, J. Optim. Theory Appl., 72 (1992), pp. 529–552.

[13] D. KLATTE, *Stability of stationary solutions in semi-infinite optimization via the reduction approach*, in Advances in Optimization: Proceedings of the 6th French–German Colloquium on Optimization, held at Lambrecht, FRG, June 2–8, 1991, Lecture Notes in Econom. and Math. Systems 382, W. Oettli and D. Pallaschke, eds., Springer-Verlag, Berlin, 1992, pp. 155–170.

[14] H. MAURER AND J. ZOWE, *First and second order necessary and sufficient optimality conditions for infinite dimensional programming problems*, Math. Prog., 16 (1979), pp. 98–110.

[15] J. P. PENOT, *Optimality Conditions for Minimax Problems, Semi-infinite Programming Problems and Their Relatives*, Report 92/16, UPRA, Laboratoire de Math. Appl., France.

[16] E. POLAK, *On the mathematical foundations of nondifferentiable optimization in engineering design*, SIAM Rev., 29 (1987), pp. 21–89.

[17] D. TORRALBA, *An Epigraphical Calculus for Quasi-monotone Sequences of Functions*, Dept of Mathematics, University of Montpellier II, Montpellier, France, 1993, submitted paper.

[18] A. SHAPIRO, *Second-order derivatives of extremal-value functions and optimality conditions for semi-infinite programming*, Math. Oper. Res., 10 (1985), pp. 207–219.

[19] ———, *Sensitivity analysis of nonlinear programs and differentiability properties of metric projections*, SIAM J. Control Optim., 26 (1988), pp. 628–645.

[20] ———, *On Lipschitzian stability of optimal solutions of parametrized semi-infinite programs*. Math. Oper. Res., 10 (1994), pp. 743–752.

[21] ———, *Directional differentiability of the optimal value function in convex semi-infinite programming*, Math. Prog., 70 (1995), pp. 149–157.

[22] K. YOSIDA, *Functional Analysis*, Springer-Verlag, Berlin, 1980.

[23] J. ZOWE AND S. KURCZYUSZ, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49–62.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1568–1591, September 1996

006

VISCOSITY SOLUTIONS AND VISCOSITY SUBDERIVATIVES IN SMOOTH BANACH SPACES WITH APPLICATIONS TO METRIC REGULARITY*

JONATHAN M. BORWEIN† AND QIJI J. ZHU‡

Abstract. In Gâteaux or bornologically differentiable spaces there are two natural generalizations of the concept of a Fréchet subderivative. In this paper we study the viscosity subderivative (which is the more robust of the two) and establish refined fuzzy sum rules for it in a smooth Banach space. These rules are applied to obtain comparison results for viscosity solutions of Hamilton–Jacobi equations in smooth spaces. A unified treatment of metric regularity in smooth spaces completes the paper. This illustrates the flexibility of viscosity subderivatives as a tool for analysis.

Key words. viscosity subderivative, fuzzy sum rule, viscosity solutions, Hamilton–Jacobi equations, smooth spaces, metric regularity

AMS subject classifications. 49J52, 49L25, 49J40, 49J50, 58C20

1. Introduction.

It is well known that the proximal limit formula for generalized derivatives plays a crucial role in Hilbert space nonsmooth and variational analysis. The reason is twofold: many properties of a nonsmooth function are determined by the (densely existing) proximal subderivatives, and proximal subderivatives are easier to handle than various other generalized derivatives. However, the proximal derivative concept depends crucially on the analysis of nearest points and, therefore, relies heavily on the inner product structure of the underlying space. Recent research [8, 10, 19] shows that, in fact, what is essential in this context is a "smooth" support function (rather than a nearest point) that corresponds to the "viscosity" (as opposed to the limit) subderivative concept. This makes it possible to do nonsmooth and variational analysis in smooth Banach spaces by using bornological subderivatives (compatible to the smoothness of the underlying space). Such a new technology is crucially important in studying problems on non–Fréchet-smooth spaces (see, for example, [10]).

One of the most important properties of the proximal subderivatives (and that of other generalized derivatives) is the sum rule. There is extensive research on this topic. We refer to Aubin and Frankowska [1], Clarke [11, 12], Deville and Haddad [20], Fabian [21], Ioffe [25, 26, 27, 28, 30], Jourani and Thibault [31], Kruger and Mordukhovich [32], Loewen [34], Mordukhovich [37], Mordukhovich and Shao [38], Rockafellar [41], Thibault [42], Ward and Borwein [43], and Warga [44, 45] and the references therein for sum rules for various generalized derivatives. The main purpose of this paper is to establish refined versions of the "fuzzy" sum rule given in [8] for viscosity bornological subderivatives and discuss its applications to viscosity solutions of Hamilton–Jacobi (HJ) equations and to the metric regularity problem. Roughly speaking, the major difference between our sum rules and that of [8] is the following observation: we can have certain control on the "size" of the bornological subderivatives in the sum. This observation is new even for sum rules in finite-dimensional spaces and is important for applications (e.g., in viscosity solution theory). A crucial tool in proving our "fuzzy" sum rules is the smooth variational principle proven in [9] that requires the underlying space to have a smooth equivalent norm. In most of the following results this condition can be weakened by using the smooth variational principle proven by Deville, Godefroy, and Zizler [18]: it suffices that the space has a smooth bump function.

*Received by the editors December 16, 1994; accepted for publication (in revised form) May 6, 1995.

†Department of Mathematics and Statistics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada (jborwein@cecm.sfu.ca). The research of this author was supported by the NSERC and the Shrum Endowment at Simon Fraser University.

‡Department of Mathematics and Statistics, Western Michigan University, Kalamazoo, MI 49008 (zhu@mathstat.wmich.edu).

Viscosity solutions were introduced by Crandall and Lions [15] to handle partial differential equations (in particular, HJ equations) that do not have any classical solution. Naturally the uniqueness of viscosity solution is one of the most important issues in the viscosity solution theory. The basic uniqueness result for finite-dimension problems was established in the pioneering paper of Crandall and Lions [15] and then developed in [13]. It is extended to Banach spaces with the Radon–Nikodým property in [16, part I]. In [19] and [20] uniqueness results are derived for HJ equations in smooth spaces with elegant short proofs using the smooth variational principle and fuzzy sum rules for viscosity subderivatives. However, the results in [19] and [20] require restrictive uniform continuity conditions to be imposed on the Hamiltonian. For HJ equations corresponding to optimal control problems involving general control equations, these uniform continuity conditions are not satisfied. By using our refined fuzzy sum rule we can prove a uniqueness theorem for an HJ equation under less restrictive conditions. It is applicable to infinite-horizon optimal control problems with a control equation that satisfies the usual Lipschitz condition in the state variable. We then apply this result to show that the value function of a class of infinite-horizon optimal control problems in a β-smooth space is the unique β-viscosity solution of the corresponding HJ equation. These results in particular applies to such a problem in L^1 which has a weak Hadamard smooth equivalent norm [6].

We also apply these fuzzy sum rules to give a unified treatment of metric regularity in smooth spaces. We prove a dual sufficient condition for metric regularity that is parallel to the dual conditions given in Ioffe [27] and Ginsburg and Ioffe [23] and deduce a primal condition that improves a similar condition in Borwein and Strojwas [5]. Then we show that various primal conditions discussed in [1, 2, 3, 5] can be deduced from our primary conditions. This illustrate the flexibility of viscosity subderivatives as a tool for analysis.

We introduce terminology and prove our refined fuzzy sum rules (Theorems 2.9–2.12) in §2. In §3 we discuss viscosity solutions to the HJ equations in smooth spaces, and in §4 we discuss metric regularity.

Finally let us remark that in bornologically differentiable spaces there are two natural generalizations of a Fréchet subderivative. It seems to us that in such spaces the viscosity subderivative (Definition 2.1) is usually the right generalization to use, rather than the limit definition.

2. Viscosity subderivatives and fuzzy sum formulae. Let X be a real Banach space with closed unit ball B and dual X^*. For a set S in X, we denote its diameter by $\operatorname{diam}(S) := \sup\{||x - y|| : x, y \in S\}$. A *bornology* β of X is a family of closed bounded and centrally symmetric subsets of X whose union is X, which is closed under multiplication by scalars and is directed upward (that is, the union of any two members of β is contained in some member of β). We will denote by X^*_β the dual space of X endowed with the topology of uniform convergence on β-sets. The most important bornologies are those formed by all (symmetric) bounded sets (the Fréchet bornology, denoted by F), weak compact sets (the weak Hadamard bornology, denoted by WH), compact sets (the Hadamard bornology, denoted by H), and finite sets (the Gâteaux bornology, denoted by G).

We will define a *convex bornology* as one that also contains all convex closures of the sets in the corresponding bornology. (In particular, any finite-dimensional subspace is included in the subspace spanned by some element of the convex bornology.) Note that the convex Gâteaux bornology lies strictly between the Gâteaux and Hadamard bornology, while for the Fréchet, weak Hadamard, and Hadamard bornologies the convex and nonconvex definitions are the same.

By a *function* we always mean an *extended-real-valued* function, usually lower (upper) semicontinuous and *proper* (that is to say, not everywhere equal to $+\infty$ ($-\infty$) and nowhere

equal to $-\infty$ ($+\infty$)). Given a function f on X, we say that f is β-differentiable at x and has a β-derivative $\nabla^\beta f(x)$ if $f(x)$ is finite and

$$t^{-1}(f(x+tu) - f(x) - t\langle \nabla^\beta f(x), u\rangle) \to 0$$

as $t \to 0$ uniformly in $u \in V$ for every $V \in \beta$. We say that a function f is β-smooth at x if $\nabla^\beta f : X \to X^*_\beta$ is continuous in a neighbourhood of x. It is not hard to check that a convex function f is β-smooth at x if and only if f is β-differentiable on a convex neighbourhood of x. Now we can define β-viscosity subderivatives and superderivatives.

DEFINITION 2.1. *Let f be a lower semicontinuous function and $f(x) < +\infty$. We say f is* β-viscosity subdifferentiable *and x^* is a* β-viscosity subderivative *of f at x if there exists a locally Lipschitz function g such that g is β-smooth at x, $\nabla^\beta g(x) = x^*$, and $f - g$ attains a local minimum at x. We denote the set of all β-viscosity subderivatives of f at x by $D_\beta f(x)$.*

Let f be a upper semicontinuous function and $f(x) > -\infty$. We say f is β-viscosity superdifferentiable *and x^* is a* β-viscosity superderivative *of f at x if there exists a locally Lipschitz function g such that g is β-smooth at x, $\nabla^\beta g(x) = x^*$, and $f - g$ attains a local maximum at x. We denote the set of all β-viscosity superderivatives of f at x by $D^\beta f(x)$.*

Remark 2.2. The concepts of viscosity subderivatives and superderivatives are introduced in [19] in slightly different forms where the β-smooth function g is required only to be upper semicontinuous and lower semicontinuous, respectively. We require g to be locally Lipschitz because it seems more convenient for applications.

Remark 2.3. By adding a constant we may always assume that the β-smooth function g in the above definition satisfies $g(x) = f(x)$.

Remark 2.4. The limit definition of the β-subdifferential $\partial_\beta f(x)$ of f at x is as follows: $x^* \in \partial_\beta f(x)$ if, for any $\varepsilon > 0$ and any $V \in \beta$, there exists a $\eta > 0$ such that

$$t^{-1}(f(x+th) - f(x)) - \langle x^*, h\rangle > -\varepsilon \quad \forall t \in (0, \eta),\ h \in V.$$

One can check that $D_\beta f(x) \subset \partial_\beta f(x)$. It is proven in [18] that $D_F f(x) = \partial_F f(x)$, provided that there exists a Lipschitz Fréchet-smooth bump function on X. However, the two concepts are different in general as shown by the following example.

Example 2.5. Let $f : R^n \to R$ ($n \geq 2$) be continuous, and suppose that f is Gâteaux but not Fréchet (weak Hadamard) differentiable at 0 (e.g., $f : R^2 \to R$ defined by $f(x, y) = xy^3/(x^2 + y^4)$ when $(x, y) \neq (0, 0)$ and $f(0, 0) = 0$ at point $(0, 0)$). Let $g(h) := -|f(h) - f(0) - \nabla^G f(0)h|$. Then g is locally uniformly continuous and

(1) $\partial_G g(0) = \{0\}$;

(2) $D_G g(0) = \emptyset$.

Proof. In fact, we can directly check that $\nabla^G g(0) = 0$; hence $\partial_G g(0) = \{0\}$. Since we always have $D_G g(0) \subset \partial_G g(0)$, either (2) is true or else $D_G g(0) = \{0\}$. In the latter case, there exists a locally Lipschitz and Gâteaux-differentiable (and therefore Fréchet-differentiable) function k such that $k(0) = g(0) = 0$, $\nabla^G k(0) = \nabla^G g(0) = 0$, and $k \leq g$ in a neighbourhood of 0. Thus,

$$\frac{|f(0+h) - f(0) - \nabla^G f(0)h|}{\|h\|} \leq -\frac{k(h) - k(0)}{\|h\|} \leq \frac{|k(h) - k(0)|}{\|h\|},$$

which implies that f is Fréchet differentiable at 0, a contradiction. □

DEFINITION 2.6 (see [8]). *Let $f_1, \ldots, f_N$ be lower semicontinuous functions and E be a closed subset of X. We say that $(f_1, \ldots, f_N)$ is* uniformly lower semicontinuous *on E if*

$$\inf_{x \in E} \sum_{n=1}^N f_n(x) \leq \liminf_{\varepsilon \to 0} \left\{ \sum_{n=1}^N f_n(x_n) : \|x_n - x_m\| \leq \varepsilon, x_n, x_m \in E, n, m = 1, \ldots, N \right\}.$$

We say that $(f_1, \dots, f_N)$ *is* locally uniformly lower semicontinuous *if, for any* $x \in \cap_{n=1}^N \operatorname{dom}(f_n)$, $(f_1, \dots, f_N)$ *is uniformly lower semicontinuous on a closed ball centered at* x.

The next useful proposition follows directly from the definition.

PROPOSITION 2.7. *Let* $(f_1, \dots, f_N)$ *be lower semicontinuous functions on* X *and* E *be a closed subset of* X. *Then the following conditions are sufficient for* $(f_1, \dots, f_N)$ *to be uniformly lower semicontinuous on* E:

1. *all but one of the functions are uniformly continuous on* E;
2. *one of the functions has compact level sets when restricted to* E;
3. X *is finite dimensional and* E *is bounded.*

The following lemma, first used in [8] in a different form, is an infinite-dimensional extension of the smooth penalization result [14, Lem. 3.1].

LEMMA 2.8. *Let* $(f_1, \dots, f_N)$ *be lower semicontinuous functions and* E *be a closed subset of* X. *Suppose that* $(f_1, \dots, f_N)$ *is uniformly lower semicontinuous on* E. *Define, for* $t > 0$,

$$M_t = \inf_{(x_1,\dots,x_N)\in E^N} \left[\sum_{n=1}^N f_n(x_n) + t \sum_{n,m=1}^N \|x_n - x_m\|^2 \right].$$

Assume that $(x_1^t, \dots, x_N^t)$ *satisfies*

$$\lim_{t\to\infty} \left(M_t - \left[\sum_{n=1}^N f_n(x_n^t) + t \sum_{n,m=1}^N \|x_n^t - x_m^t\|^2 \right] \right) = 0.$$

Then

(i) $\lim_{t\to\infty} t[\operatorname{diam}(\{x_1^t, \dots, x_N^t\})]^2 = 0$;

(ii) $\lim_{t\to\infty} M_t = \inf_{x\in E} \sum_{n=1}^N f_n(x)$.

Proof. Set

$$d_t := M_t - \left[\sum_{n=1}^N f_n(x_n^t) + t \sum_{n,m=1}^N \|x_n^t - x_m^t\|^2 \right].$$

Then $\lim_{t\to\infty} d_t = 0$. Evidently M_t increases with t and $M_t \le \inf_{x\in E} \sum_{n=1}^N f_n(x)$. Therefore, $M := \lim_{t\to\infty} M_t$ exists and $M \le \inf_{x\in E} \sum_{n=1}^N f_n(x)$. Moreover,

$$\begin{aligned} M_{t/2} &\le \sum_{n=1}^N f_n(x_n^t) + \frac{t}{2} \sum_{n,m=1}^N \|x_n^t - x_m^t\|^2 \\ &= \sum_{n=1}^N f_n(x_n^t) + t \sum_{n,m=1}^N \|x_n^t - x_m^t\|^2 - \frac{t}{2} \sum_{n,m=1}^N \|x_n^t - x_m^t\|^2 \\ &= M_t - d_t - \frac{t}{2} \sum_{n,m=1}^N \|x_n^t - x_m^t\|^2 \end{aligned}$$

or

$$2(M_t - M_{t/2} - d_t) \ge t \sum_{n,m=1}^N \|x_n^t - x_m^t\|^2 \ge t[\operatorname{diam}(\{x_1^t, \dots, x_N^t\})]^2.$$

Thus,

$$\lim_{t\to\infty} t[\operatorname{diam}(\{x_1^t, \dots, x_N^t\})]^2 = 0.$$

It remains to show that $M \geq \inf_{x\in E} \sum_{n=1}^N f_n(x)$. To do so, observe that $\|x_n^t - x_m^t\| \to 0$ when $t \to \infty$ and $(f_1, \ldots, f_N)$ is uniformly lower semicontinuous on E. Therefore,

$$M = \lim_{t\to\infty} \left(\sum_{n=1}^N f_n(x_n^t) + t \sum_{n,m=1}^N \|x_n^t - x_m^t\|^2 \right)$$
$$\geq \liminf_{t\to\infty} \sum_{n=1}^N f_n(x_n^t) \geq \inf_{x\in E} \sum_{n=1}^N f_n(x). \qquad \square$$

The next theorem is a refined fuzzy sum rule for the sum of several functions at its minimum. It refines [8, Prop. 4] in that it gives a bound for the β-subderivatives in the fuzzy sum. This bound is new even for fuzzy sum rules in finite-dimensional spaces and is important in application to the uniqueness result for HJ equations.

THEOREM 2.9. *Let X be a Banach space with an equivalent β-smooth norm and $f_1, \ldots, f_N$ be lower semicontinuous functions on X. Suppose that $(f_1, \ldots, f_N)$ is locally uniformly lower semicontinuous and $\sum_{n=1}^N f_n$ attains a local minimum at x. Then, for any $\varepsilon > 0$, there exist $x_n \in x + \varepsilon B$ and $x_n^* \in D_\beta f_n(x_n), n = 1, \ldots, N$, such that $|f_n(x_n) - f_n(x)| < \varepsilon$, $\|x_n^*\|\operatorname{diam}(\{x_1, \ldots, x_N\}) < \varepsilon, n = 1, 2, \ldots, N$, and*

$$\left\| \sum_{n=1}^N x_n^* \right\| < \varepsilon.$$

Proof. Let $\varepsilon > 0$ be given. Taking a smaller ε if necessary, we may assume that $(f_1, \ldots, f_N)$ is uniformly lower semicontinuous on $x+\varepsilon B$ and that $\sum_{n=1}^N f_n$ attains a minimum at x over $x + \varepsilon B$. Define

$$w_t(y_1, \ldots, y_N) := \sum_{n=1}^N f_n(y_n) + t \sum_{n,m=1}^N \|y_n - y_m\|^2.$$

Then, by Lemma 2.8,

$$\lim_{t\to\infty} \inf\{w_t(y_1, \ldots, y_N) : y_n \in x + \varepsilon B, n = 1, \ldots, N\}$$
$$= \inf_{y\in x+\varepsilon B} \sum_{n=1}^N f_n(y) = \sum_{n=1}^N f_n(x) = w_t(x, \ldots, x). \tag{1}$$

Choose $t_i \to \infty$ such that

$$w_{t_i}(x, \ldots, x) < \inf\{w_{t_i}(y_1, \ldots, y_N) : y_n \in x + \varepsilon B, n = 1, \ldots, N\} + \frac{\varepsilon^2}{Ni}.$$

Then by the Borwein–Preiss smooth variational principle [9], for each i there exist a β-smooth function ϕ_i and $x_n^i, n = 1, \ldots, N$, such that $w_{t_i} + \phi_i$ attains a local minimum at $(x_1^i, \ldots, x_N^i)$, $\|\nabla^\beta \phi_i\| < \varepsilon/N$, $\|x_n^i - x\| < \varepsilon/i$, and

$$w_{t_i}(x_1^i, \ldots, x_N^i) < \inf\{w_{t_i}(y_1, \ldots, y_N) : y_n \in x + \varepsilon B, n = 1, \ldots, N\} + \frac{\varepsilon^2}{Ni}. \tag{2}$$

For each n, the function

$$y \to w_{t_i}(x_1^i, \ldots, x_{n-1}^i, y, x_{n+1}^i, \ldots, x_N^i) + \phi_i(x_1^i, \ldots, x_{n-1}^i, y, x_{n+1}^i, \ldots, x_N^i)$$

attains a local minimum at $y = x_n^i$. Thus

$$x_{n_i}^* := -\nabla_{x_n}^{\beta} \phi_i(x_1^i, \ldots, x_N^i) - 2t_i \sum_{m=1}^{N} \nabla^{\beta} \|\cdot\|^2 (x_n^i - x_m^i) \in D_{\beta} f_n(x_n^i). \tag{3}$$

Summing $x_{n_i}^*$, $n = 1, \ldots, N$, yields

$$\sum_{n=1}^{N} x_{n_i}^* = -\sum_{n=1}^{N} \nabla_{x_n}^{\beta} \phi_i(x_1^i, \ldots, x_N^i) - 2t_i \sum_{n=1}^{N} \sum_{m=1}^{N} \nabla^{\beta} \|\cdot\|^2 (x_n^i - x_m^i).$$

Observing that $\| - \sum_{n=1}^{N} \nabla_{x_n}^{\beta} \phi_i(x_1^i, \ldots, x_N^i)\| \le \varepsilon$ and

$$\nabla^{\beta} \|\cdot\|^2 (x_n^i - x_m^i) + \nabla^{\beta} \|\cdot\|^2 (x_m^i - x_n^i) = 0$$

so that the double sum in the previous inclusion vanishes, we obtain

$$\left\| \sum_{n=1}^{N} x_{n_i}^* \right\| \le \varepsilon.$$

By (1) and (2)

$$\sum_{n=1}^{N} f_n(x) = \lim_{i \to \infty} \sum_{n=1}^{N} f_n(x_n^i).$$

Since f_n, $n = 1, \ldots, N$, are lower semicontinuous

$$\lim_{i \to \infty} f_n(x_n^i) = f_n(x), \quad n = 1, \ldots, N.$$

Moreover, Lemma 2.8 and (2) imply

$$\lim_{i \to \infty} t_i [\operatorname{diam}(\{x_1^i, \ldots, x_N^i\})]^2 = 0. \tag{4}$$

Since $\nabla^{\beta} \|\cdot\|^2(x)$ is bounded by $2\|x\|$, combining (4) and (3) yields

$$\lim_{i \to \infty} \|x_{n_i}^*\| \operatorname{diam}(\{x_1^i, \ldots, x_N^i\}) = 0 \quad \text{for } n = 1, \ldots, N.$$

Thus, when i is sufficiently large,

$$\|x_n^i - x\| < \varepsilon, \quad |f_n(x_n^i) - f_n(x)| < \varepsilon, \quad \text{and} \quad \|x_{n_i}^*\| \operatorname{diam}(\{x_1^i, \ldots, x_N^i\}) < \varepsilon$$

for $n = 1, 2, \ldots, N$. It remains to take $x_n = x_n^i$ and $x_n^* = x_{n_i}^*$, $n = 1, \ldots, N$. □

The next three theorems provide general forms of sum rules for β-viscosity subderivatives.

THEOREM 2.10. *Let X be a Banach space with an equivalent β-smooth norm. Let $f_1, \ldots, f_N$ be lower semicontinuous functions on X, with $x \in \cap_{n=1}^{N} \operatorname{dom}(f_n)$. Then, for any $x^* \in D_{\beta}(\sum_{n=1}^{N} f_n)(x)$, $\varepsilon > 0$, and any weak-star neighbourhood V of 0 in X^*, there exist $x_n \in x + \varepsilon B$, $x_n^* \in D_{\beta} f_n(x_n)$, $n = 1, \ldots, N$, such that $|f_n(x_n) - f_n(x)| < \varepsilon$, $\|x_n^*\| \operatorname{diam}(\{x_1, \ldots, x_N\}) < \varepsilon$, $n = 1, 2, \ldots, N$, and*

$$x^* \in \sum_{n=1}^{N} x_n^* + V.$$

Proof. Let $\varepsilon > 0$ and V be a weak-star neighbourhood of 0 in X^*. Fix $r > 0$, L a finite-dimensional subspace of X containing x, and V_β a β-neighbourhood of 0 in X^*_β such that $V_\beta + L^\perp + rB_{X^*} \subset V$. Since $x^* \in D_\beta(\sum_{n=1}^N f_n)(x)$, there exists a locally Lipschitz function g such that g is β smooth at x with $\nabla^\beta g(x) = x^*$ and $\sum_{n=1}^N f_n - g$ attains a local minimum at x. Choose $0 < \eta < \min(\varepsilon, r)$ such that $\|y - x\| < \eta < \varepsilon$ implies that $\nabla^\beta g(x) - \nabla^\beta g(y) \in V_\beta$, and let δ_L be the indicator function of L. Then $\sum_{n=1}^N f_n - g + \delta_L$ attains a local minimum at x. By Proposition 2.7 $(f_1, \ldots, f_N, -g, \delta_L)$ is locally uniformly lower semicontinuous. Applying Theorem 2.9 yields the existence of x_n, $n = 1, \ldots, N+2$, such that $\|x_n - x\| < \eta < \varepsilon, n = 1, \ldots, N+2, x_n^* \in D_\beta f(x_n), n = 1, \ldots, N, x_{N+1}^* = -\nabla^\beta g(x_{N+1})$, and $x_{N+2}^* \in D_\beta \delta_L(x_{N+2})$ satisfying the conclusion of Theorem 2.9, i.e., $|f_n(x_n) - f_n(x)| < \eta < \varepsilon$, $\|x_n^*\| \operatorname{diam}(\{x_1, \ldots, x_N\}) \le \|x_n^*\| \operatorname{diam}(\{x_1, \ldots, x_{N+2}\}) < \eta < \varepsilon$ for $n = 1, \ldots, N$, $|\delta_L(x_{N+2}) - \delta_L(x)| < \eta$, i.e., $x_{N+2} \in L$, and

$$\sum_{n=1}^N x_n^* - \nabla^\beta g(x_{N+1}) + x_{N+2}^* \in rB_{X^*}.$$

Note that $D_\beta \delta_L(x_{N+2}) = L^\perp$ and $x^* - \nabla^\beta g(x_{N+1}) \in V_\beta$. □

Using the method in [8], we can prove the following stronger result.

THEOREM 2.11. *Let β be a convex bornology and X be a Banach space with an equivalent β-smooth norm. Let $f_1, \ldots, f_N$ be lower semicontinuous functions and $x \in \cap_{n=1}^N \operatorname{dom}(f_n)$. Then, for any $x^* \in \partial_\beta(\sum_{n=1}^N f_n)(x)$, $\varepsilon > 0$, and any weak-star neighbourhood V of 0 in X^*, there exist $x_n \in x + \varepsilon B, x_n^* \in D_\beta f_n(x_n), n = 1, \ldots, N$, such that $|f_n(x_n) - f_n(x)| < \varepsilon$, $\|x_n^*\| \operatorname{diam}(\{x_1, \ldots, x_N\}) < \varepsilon, n = 1, 2, \ldots, N$, and*

$$x^* \in \sum_{n=1}^N x_n^* + V.$$

Proof. Let $\varepsilon > 0$ and V be a weak-star neighbourhood of 0 in X^*. Fix $r > 0$ and L a finite-dimensional subspace of X containing x such that $L^\perp + 2rB_{X^*} \subset V$. Let $x^* \in \partial_\beta(\sum_{n=1}^N f_n)(x)$. Then, for any $K \in \beta$,

$$\liminf_{t \to 0^+} \inf_{h \in tK} t^{-1} \left[\sum_{n=1}^N (f_n(x+h) - f_n(x)) - \langle x^*, h \rangle \right] \ge 0. \tag{5}$$

Choose a $K \in \beta$ containing the intersection of L with a ball around zero. Then (5) in particular implies that

$$\liminf_{t \to 0^+} \inf_{h \in tB \cap L} t^{-1} \left[\sum_{n=1}^N (f_n(x+h) - f_n(x)) - \langle x^*, h \rangle \right] \ge 0.$$

Since L is a finite-dimensional space, this is equivalent to

$$\liminf_{\|y-x\| \to 0} \inf_{y - x \in L} \|y - x\|^{-1} \left[\sum_{n=1}^N (f_n(y) - f_n(x)) - \langle x^*, y - x \rangle \right] \ge 0.$$

Thus, there exists $\eta < r$ such that the function

$$y \to \sum_{n=1}^N f_n(y) - \langle x^*, y \rangle + r\|y - x\| + \delta_L(y)$$

attains a minimum over $y \in x + \eta B$ at $y = x$.

By Proposition 2.7, the mapping $y \to (f_1(y), \dots, f_N(y), -\langle x^*, y\rangle, r\|y - x\|, \delta_L(y))$ is locally uniformly lower semicontinuous. Applying Theorem 2.9 yields that there exist $x_n, n = 1, \dots, N+3$, with $\|x_n - x\| < \eta < \varepsilon, n = 1, \dots, N+3$; $x_n^* \in D_\beta f(x_n), n = 1, \dots, N$; $x_{N+1}^* = -x^*$, $x_{N+2}^* \in rD_\beta\|x_{N+2} - x\|$, and $x_{N+3}^* \in D_\beta\delta_L(x_{N+3})$ such that $|f_n(x_n) - f_n(x)| < \eta < \varepsilon$, $\|x_n^*\|\text{diam}(\{x_1, \dots, x_N\}) \le \|x_n^*\|\text{diam}(\{x_1, \dots, x_{N+3}\}) < \eta < \varepsilon$ for $n = 1, \dots, N$, $|\delta_L(x_{N+3}) - \delta_L(x)| < \eta$, i.e., $x_{N+3} \in L$ and

$$\sum_{n=1}^{N} x_n^* - x^* + x_{N+2}^* + x_{N+3}^* \in rB_{X^*}.$$

Observing that $D_\beta\delta_L(x_{N+3}) = L^\perp$ and $rD_\beta\|x_{N+2} - x\| \subset rB_{X^*}$, we obtain

$$x^* \in \sum_{n=1}^{N} x_n^* + L^\perp + 2rB_{X^*} \subset \sum_{n=1}^{N} x_n^* + V. \qquad \square$$

THEOREM 2.12. *Let X be a Banach space with an equivalent β-smooth norm. Let $f_1, \dots, f_N$ be lower semicontinuous functions, with all but one of $f_n, n = 1, \dots, N$, locally uniformly continuous and $x \in \cap_{n=1}^N \text{dom}(f_n)$. Then, for any $x^* \in D_\beta(\sum_{n=1}^N f_n)(x)$, $\varepsilon > 0$, and any β-neighbourhood V of 0 in X_β^*, there exist $x_n \in x + \varepsilon B$, $x_n^* \in D_\beta f_n(x_n), n = 1, \dots, N$, such that $|f_n(x_n) - f_n(x)| < \varepsilon$, $\|x_n^*\|\text{diam}(\{x_1, \dots, x_N\}) < \varepsilon, n = 1, 2, \dots, N$, and*

$$x^* \in \sum_{n=1}^{N} x_n^* + V.$$

Proof. Let $\varepsilon > 0$ and V be a neighbourhood of 0 in X_β^*. Let $r > 0$, U be a neighbourhood of $0 \in X_\beta^*$ such that $U + rB_{X^*} \subset V$. Let $x^* \in D_\beta(\sum_{n=1}^N f_n)(x)$. Then there exists a β smooth function g such that $\nabla^\beta g(x) = x^*$ and $\sum_{n=1}^N f_n - g$ attains a local minimum at x. Choose $0 < \eta < \varepsilon$ such that $\|y - x\| < \eta < \varepsilon$ implies that $\nabla^\beta g(x) - \nabla^\beta g(y) \in U$. By Proposition 2.7 $(f_1, \dots, f_N, -g)$ is locally uniformly lower semicontinuous. Applying Theorem 2.9 yields that there exist $x_n, n = 1, \dots, N+1$, with $\|x_n - x\| < \eta < \varepsilon, n = 1, \dots, N+1$, $x_n^* \in D_\beta f_n(x_n), n = 1, \dots, N$, and $x_{N+1}^* = -\nabla^\beta g(x_{N+1})$ such that $|f_n(x_n) - f_n(x)| < \eta < \varepsilon$, $\|x_n^*\|\text{diam}(\{x_1, \dots, x_N\}) < \varepsilon, n = 1, \dots, N$, and

$$\left\| \sum_{n=1}^{N} x_n^* - \nabla^\beta g(x_{N+1}) \right\| < r.$$

Thus,

$$\begin{aligned} x^* &\in \sum_{n=1}^{N} x_n^* + \nabla^\beta g(x) - \nabla^\beta g(x_{N+1}) + rB_{X^*} \\ &\subset \sum_{n=1}^{N} x_n^* + U + rB_{X^*} \subset \sum_{n=1}^{N} x_n^* + V. \qquad \square \end{aligned}$$

Remark 2.13. As in [9] we call a lower semicontinuous function $f : X \to [-\infty, \infty]$ *s-Hölder subdifferentiable* at x ($s \in (0, 1]$) with subderivative $x^* \in X^*$ if f is finite at x and there exists a positive constant C_x such that

$$f(y) - f(x) - \langle x^*, y - x\rangle \ge -C_x\|y - x\|^{1+s}$$

for all y in a neighbourhood of x. We denote the set of s-Hölder subderivatives of f at x by $\partial^{H(s)} f(x)$. When $s = 1$, such subderivatives are called *Lipschitz smooth*, and in Hilbert space they coincide with Rockafellar's *proximal subderivatives*, written $\partial^{\pi} f(x)$ in [40]. Then all the above statements and proofs will still hold true in a Banach space with a power modulus of smoothness t^p [33] if we replace β with $H(p-1)$.

Remark 2.14. By the same argument as in [18, §VIII, Lem. 1.3] we can prove the following result.

LEMMA 2.15. *Let X be a Banach space that admits a bump function which is Lipschitzian and β-smooth. Then there exist a function $d : X \to R^+$ and a scalar $K > 1$ such that*

i) *d is bounded, Lipschitzian on X, and β-smooth on $X\backslash\{0\}$.*

ii) *$\|x\| \le d(x) \le K\|x\|$ if $\|x\| \le 1$ and $d(x) = 2$ if $\|x\| \ge 1$.*

Then, observing that we can replace the $\|\cdot\|^2$ term in Lemma 2.8 and Theorem 2.9 and their proofs with $d(\cdot)^2$ and using the Deville–Godefroy–Zizler smooth variational principle (cf. [18, 39]) in place of the Borwein–Preiss smooth variational principle, the condition that X has a β-smooth norm can be replaced by the weaker condition that X has a β-smooth Lipschitzian bump function.

Remark 2.16. Ioffe [25, 26] named Banach spaces that have the fuzzy sum rules for lower semicontinuous functions for the ε-Fréchet-subderivative and the ε-Dini-subderivative trustworthy (T) and weak trustworthy (WT) spaces, respectively. He proved that if a Banach space has a Fréchet- or Gâteaux-smooth norm, then it has property (T) or (WT). It is also known that (T) or (WT) is equivalent to the dense ε-Fréchet-subdifferentiability (S) or dense ε-Dini-subdifferentiability (WS) of lower semicontinuous functions [21]. In this spirit we will call a Banach space β-trustworthy (T_β) if it has the property stated in the conclusion of Theorem 2.10 (without requiring the bound on the β-subderivatives in the sum) and say that a Banach space is a (strong) dense β-subdifferentiability space if, for any lower semicontinuous function f, the set $((x, f(x)) \in \text{graph}(f))\, x \in \text{dom}(f)$ for which $D_\beta f(x) \neq \emptyset$ is dense in the (graph) domain of f. We will use notations S_β^+ and S_β for strong β-subdifferentiability and β-subdifferentiability spaces, respectively. We will denote the properties "has a β-smooth norm" and "has a β-smooth Lipschitz bump function" by H_β and H_β^-, respectively. Then Theorem 2.10 and Remark 2.14 tell us that $H_\beta \Longrightarrow H_\beta^- \Longrightarrow T_\beta$. It is obvious that $S_\beta^+ \Longrightarrow S_\beta$. Let f be a lower semicontinuous function in a T_β space. Then, for any $x \in \text{dom}(f)$, applying the T_β property to $f + \delta_{\{x\}}$ yields S_β^+. Therefore, we have

$$H_\beta \Longrightarrow H_\beta^- \Longrightarrow T_\beta \Longrightarrow S_\beta^+ \Longrightarrow S_\beta.$$

In [21] it is shown that $T \Longleftrightarrow S$ and $WT \Longleftrightarrow WS$. Using similar arguments we can show that $T_\beta \Longleftrightarrow S_\beta^+$. However, it is not clear if $S_\beta \Longleftrightarrow S_\beta^+$. Since conditions H_β and H_β^- are easier to check than S_β and S_β^+ and can yield a bound on the β-subderivatives in the fuzzy sum, we will not pursue this topic further here.

Adapting the arguments in [7, 8] and using Theorem 2.9 we can obtain corresponding viscosity versions of the sequential limit formulae for the g-subdifferential of a function and the g-normal cone of a closed set. In the following theorems, ∂_g, ∂_c, N_g, and N_c signify the g-subdifferential, Clarke generalized gradient, g-normal cone, and Clarke normal cone, respectively, and we refer the reader to [8] and the references therein for their definitions.

THEOREM 2.17. *Let X be a Banach space with an equivalent β-smooth norm. Let f be a lower semicontinuous proper function on X. Then for any $x \in X$*

$$\partial_g f(x) = \text{cl}^* \bigcup_{k=1}^{\infty} \{\text{w}^* - \lim_{n\to\infty} x_n^* : x_n^* \in D_\beta^k f(x_n),\ x_n \to_f x\}$$

and

$$\partial_c f(x) = \mathrm{cl}^* co \bigcup_{k=1}^{\infty} \{w^* - \lim_{n\to\infty} x_n^* : x_n^* \in D_\beta^k f(x_n),\ x_n \to_f x\} + \partial_c^\infty f(x).$$

where $D_\beta^k f(x)$ is the subset of $D_\beta f(x)$ for which the support function in the definition has a Lipschitz constant no greater than k.

THEOREM 2.18. *Let X be a Banach space with an equivalent β-smooth norm. Let S be a closed subset of X. Then for any $x \in S$*

$$N_g(S, x) = \mathrm{cl}^* \bigcup_{k=1}^{\infty} \{w^* - \lim_{n\to\infty} x_n^* : x_n^* \in kD_\beta d(S, x_n),\ x_n \to_S x\}$$

and

$$N_c(S, x) = \mathrm{cl}^* co \bigcup_{k=1}^{\infty} \{w^* - \lim_{n\to\infty} x_n^* : x_n^* \in kD_\beta d(S, x_n),\ x_n \to_S x\}.$$

In the following discussion we will use the formulae given in the above theorems as equivalent definitions for the g-normal cone in a β-smooth space. These formulae also have corresponding s-Hölder versions. For details see [10].

3. Viscosity solutions to HJ equations. Consider the partial differential equation

$$F(x, u, Du) = 0. \tag{6}$$

This equation encompasses the usual HJ equation associated with the optimal value function of certain optimal control problems. In general, (6) does not have a classical solution. Viscosity solutions were introduced by Crandall and Lions [15] to replace classical solutions. The original definition of viscosity solutions (cf. [15, 16]) is based on the Fréchet-subderivative. In [19], β-viscosity solutions are defined for problems on non–Fréchet-smooth spaces. We recall this definition below.

DEFINITION 3.1 (see [19]). *Let X be a Banach space with an equivalent β-smooth norm. A function $u : X \to R$ is a* β-viscosity subsolution *of* (6) *if u is upper semicontinuous and, for every $x \in X$ and every $x^* \in D^\beta u(x)$,*

$$F(x, u(x), x^*) \le 0.$$

A function $u : X \to R$ is a β-viscosity supersolution *of* (6) *if u is lower semicontinuous and, for every $x \in X$ and every $x^* \in D_\beta u(x)$,*

$$F(x, u(x), x^*) \ge 0.$$

A continuous function u is called a β-viscosity solution *if u is both a β-viscosity subsolution and a β-viscosity supersolution.*

Now we prove the main result of this section.

THEOREM 3.2. *Let X be a Banach space with an equivalent β-smooth norm. Suppose that $\gamma > 0$, $F(x, u, x^*) = \gamma u + H(x, x^*)$, and $H : X \times X_\beta^* \to R$ satisfy the following assumption:*

(A) *for any $x_1, x_2 \in X$ and $x_1^*, x_2^* \in X_\beta^*$,*

$$|H(x_1, x_1^*) - H(x_2, x_2^*)| \le \omega(x_1 - x_2, x_1^* - x_2^*) + M \max(||x_1^*||, ||x_2^*||)||x_1 - x_2||,$$

where $M > 0$ is a constant and $\omega : X \times X_\beta^* \to R$ is a continuous function with $\omega(0, 0) = 0$.

Let u and v be two uniformly continuous functions such that v is bounded below and u is bounded above. If u is a β-viscosity subsolution of (6) *and v is a β-viscosity supersolution of* (6), *then $u \le v$.*

Proof. Let $\varepsilon > 0$ be an arbitrary positive number. By assumption (A) there exist $\eta \in (0, \varepsilon)$ and a neighbourhood V of 0 in X_β^* such that $\|x_1 - x_2\| < 2\eta$ and $x_1^* - x_2^* \in V$ implies that

$$|H(x_1, x_1^*) - H(x_2, x_2^*)| < \varepsilon + M \max(||x_1^*||, ||x_2^*||)||x_1 - x_2||.$$

The function $v - u$ is uniformly continuous and bounded below. By the smooth variational principle there exist $x \in X$ and $x^* \in D_\beta(v-u)(x)$ such that $x^* + \frac{1}{2}V \subset V$ and $(v-u)(x) < \inf_X(v-u) + \varepsilon$. By Theorem 2.12 with $f_1 = v$ and $f_2 = -u$, there exist $x_1, x_2 \in X$, $x_1^* \in D_\beta v(x_1)$, and $x_2^* \in D^\beta u(x_2)$ satisfying

(i) $\|x_1 - x\| < \eta$ and $\|x_2 - x\| < \eta$;
(ii) $|v(x_1) - v(x)| < \varepsilon$ and $|u(x_2) - u(x)| < \varepsilon$;
(iii) $||x_1^*||||x_1 - x_2|| < \varepsilon$ and $||x_2^*||||x_1 - x_2|| < \varepsilon$;
(iv) $x_1^* - x_2^* - x^* \in \frac{1}{2}V$.

Since the function u is a viscosity subsolution of (6) we have

$$F(x_1, u(x_1), x_1^*) = \gamma u(x_1) + H(x_1, x_1^*) \le 0.$$

Similarly

$$F(x_2, v(x_2), x_2^*) = \gamma v(x_2) + H(x_2, x_2^*) \ge 0.$$

Therefore, observing that $||x_1 - x_2|| < 2\eta$ and $x_1^* - x_2^* \in V$,

$$\begin{aligned}\inf_X(v-u) &> (v-u)(x) - \varepsilon > v(x_2) - u(x_1) - 3\varepsilon \\ &\ge \gamma^{-1}[H(x_1, x_1^*) - H(x_2, x_2^*)] - 3\varepsilon \\ &\ge -\gamma^{-1}[\varepsilon + M \max(||x_1^*||, ||x_2^*||)||x_1 - x_2||] - 3\varepsilon \\ &\ge -[\gamma^{-1}(1+M) + 3]\varepsilon.\end{aligned}$$

As ε is arbitrary, $\inf_X(v-u) \ge 0$. □

COROLLARY 3.3. *Under the assumptions of Theorem* 3.2 *any uniformly continuous bounded β-viscosity solution to* (6) *is unique.*

Remark 3.4. Condition (A) is significantly weaker than the uniform continuity conditions imposed in [19, 20]. This allows us to apply Corollary 3.3 to the HJ equation corresponding to a general optimal control problem $P(x)$ in the sequel. In [16, part I], a uniqueness result is established for HJ equations in Banach spaces with the Radon–Nikodým property under somewhat weaker (but rather technical) conditions. How much we can weaken our requirements for the Hamiltonian in a β-smooth space is an interesting problem that we will not pursue further in this paper.

Remark 3.5. Definition 3.1 is slightly different from the original definition in [19] in that we require X to be β-smooth. We make this modification to avoid the following pathological examples.

Example 3.6. Let X be a β'-differentiable Banach space with a nowhere β-differentiable norm $\|\cdot\|$. Consider the case of (6) where F is the constant function 1. Then obviously this equation does not have any β'-viscosity solution, but $u(x) = (\|x\| + 2)/(\|x\| + 1)$ is a uniformly continuous bounded β-viscosity solution in the sense of [19, Def. III.2]. To see this, observe that u is trivially a supersolution. Since $u(x)$ is convex and nowhere β-differentiable

(otherwise $\|x\| = (2 - u(x))/(u(x) - 1)$ would be β-differentiable), $D^\beta u(x) = \emptyset$ for all $x \in X$. Thus, u is also a β-viscosity subsolution and, therefore, a β-viscosity solution in the sense of [19, Def. III.2].

Example 3.7. In Banach space X consider

$$V(x) + \|DV(x)\| = 0. \tag{7}$$

Assume that the norm $\|\cdot\|$ of X is nowhere β-differentiable and X has an equivalent β'-smooth norm. Then, in the sense of [19, Def. III.2],

1. $V = 0$ is the unique uniformly continuous bounded β'-viscosity solution of (7);
2. $V(x) = (\|x\| + 2)/(\|x\| + 1)$ is a uniformly continuous bounded β-viscosity solution of (7).

Proof. Step 1. If V is a uniformly continuous bounded β'-viscosity solution of (7), then u is nonnegative. In fact, if $-\Delta = \inf_X V < 0$, then since X has an equivalent β'-smooth norm, by the Borwein–Preiss smooth variational principle, there exists an $x \in X$ and a $p \in X^*$ such that

$$V(x) < \inf_X V + \frac{\Delta}{2}$$

and

$$p \in D_{\beta'} V(x), \quad \|p\| < \frac{\Delta}{2}.$$

This implies

$$V(x) + \|p\| < 0$$

so that V is not a β'-viscosity supersolution of (7), which is absurd.

Step 2. If V is a nonnegative β'-viscosity solution of (7), then V is a nonnegative β'-viscosity subsolution of (7). Thus, $D^{\beta'} V(x) = \{0\}$ whenever it is nonempty. By [10, Thm. 6.3] u is a constant. It is then obvious that V must be the zero function. This completes the proof of 1.

Step 3. We prove 2. It is trivial to observe that $V(x) = (\|x\|+2)/(\|x\|+1)$ is a β-viscosity supersolution of (7). Since $V(x)$ is convex and nowhere β-differentiable, $D^\beta V(x) = \emptyset$ for all $x \in X$. Thus, V is also a β-viscosity subsolution and, therefore, a β-viscosity solution of (7) according to [19, Def. III.2]. □

Note that (7) is the HJ equation corresponding to the optimal control problem

$$\text{minimize } J(x, u) := \int_0^\infty e^{-s} f(x(s), u(s)) ds$$

$$\text{subject to } \dot{x}(s) = u(s),$$

$$x(0) = x, \ u(s) \in B_X,$$

with $f = 0$. Since the optimal value function for this problem is identically 0, it is reasonable to expect 0 to be the unique viscosity solution of (7). Thus, it is reasonable to require X to be β-smooth in the definition of β-viscosity solutions.

As concrete examples of the spaces in Examples 3.6 and 3.7,

a. the sup norm of $C[0, 1]$ is Gâteaux smooth but nowhere weak Hadamard differentiable [6].

b. $L_1[0, 1]$ or $l_1(N)$ has an equivalent weak Hadamard smooth norm [6] but is not Asplund and, therefore, has a nowhere Fréchet-differentiable norm. Moreover, $L_1[0, 1]$ fails to have the Radon–Nikodým property.

Remark 3.8. Let X be a Banach space with a β-smooth equivalent norm. Then this norm is also β'-smooth for any $\beta' \subset \beta$. Thus, in Theorem 3.2 and Corollary 3.3 we can always reduce the requirement on the smoothness of the space by using a smaller bornology. However, the price to pay is that we have to impose stronger continuity conditions on F.

Let X be a Banach space with a β-smooth norm and U be a metric space. We now consider the following optimal control problem in X:

$$P(x): \quad \text{minimize } J(x, u) := \int_0^\infty e^{-\gamma s} f(x(s), u(s))ds$$
$$\text{subject to } \dot{x}(s) = g(x(s), u(s)),$$
$$x(0) = x, \ u \in \mathcal{U},$$

where $g : X \times U \to R$ is continuous and Lipschitz in x uniformly in U and there exists $K \in \beta$ such that $g(x, U) \subset K$ for all $x \in X$, $f : X \times U \to R$ is continuous, bounded, continuous in x uniformly in U, and

$$\mathcal{U} := \{u : u \text{ is measurable and } u(t) \in U \text{ for } t \in [0, \infty) \text{ a. e.}\}.$$

Under our assumptions, for given $x \in X$ and $u \in \mathcal{U}$, $\dot{x}(s) = g(x(s), u(s))$, $x(0) = x$ has a unique solution defined on $[0, \infty)$, denoted by $x(s, x, u)$.

Since f is bounded, the problem is well defined. We denote the value function of $P(x)$ by $V(x)$. Then we have the following.

THEOREM 3.9 (dynamic programming principle). *For any $t > 0$,*

$$V(x) = \inf_{u \in \mathcal{U}} \left\{ \int_0^t e^{-\gamma s} f(x(s, x, u), u(s))ds + e^{-\gamma t} V(x(t, x, u)) \right\}.$$

This theorem is standard; for a proof see, for example, [22].

Define $H : X \times X^* \to R$ by

$$H(x, p) = \sup_{u \in U} \{-\langle p, g(x, u)\rangle - f(x, u)\}.$$

We prove the following theorem.

THEOREM 3.10. *V is the unique β-viscosity solution of the HJ equation*

$$\gamma V(x) + H(x, DV(x)) = 0. \tag{8}$$

Proof. Since f is bounded, so is V. Using Gronwall's inequality, since f is continuous in x uniformly in U, we can show that V is uniformly continuous. One can directly check that $H(x, p)$ satisfies assumption (A). Therefore, uniqueness is a direct consequence of Corollary 3.3. We need only to show that V is a β-viscosity solution of (8).

a. Subsolution. Let y be an element of X such that $p \in D^\beta V(y)$. Then there exists a β-smooth locally Lipschitz function w such that $\nabla^\beta w(y) = p$, y is a (local) maximum of $V - w$, and $0 = (V - w)(y)$. Note that w depends only on y and p and is fixed in the subsequent discussion. By the dynamic programming principle, for any $u \in \mathcal{U}$ and $t > 0$,

$$w(y) = V(y) \le \int_0^t e^{-\gamma s} f(x(s, y, u), u(s))ds + e^{-\gamma t} V(x(t, y, u)). \tag{9}$$

Since w is β-smooth and locally Lipschitz at x, there exists $\eta > 0$ such that w is Lipschitz and $\nabla^\beta w$ exist in $y + \eta B$. Therefore, w is at least Hadamard differentiable and $\nabla^\beta w = \nabla^H w$ in $y + \eta B$. Observing that $\{x(s, y, u) : s \in [0, 1]\}$ is compact, when $t > 0$ is sufficiently small, we have

$$\frac{d}{dt}[e^{-\gamma t} w(x(t, y, u))] = -e^{-\gamma t}\gamma w(x(t, y, u)) + e^{-\gamma t}\langle \nabla^\beta w(x(t, y, u)), g(x(t, y, u), u(t))\rangle.$$

Thus, we can write (9) as

$$t^{-1}\int_0^t e^{-\gamma s}[\gamma w(x(s, y, u)) - \langle \nabla^\beta w(x(s, y, u)), g(x(s, y, u), u(s))\rangle - f(x(s, y, u), u(s))]ds \le 0.$$

Fixing an arbitrary $v \in U$ and setting $u(s) = v$ for all $s \in [0, t]$ yield

$$t^{-1}\int_0^t e^{-\gamma s}[\gamma w(x(s, y, u)) - \langle \nabla^\beta w(x(s, y, u)), g(x(s, y, u), v)\rangle - f(x(s, y, u), v)]ds \le 0.$$

Taking limits when $t \to 0$, observing that the integrand is continuous in s and $x(0, y, u) = y$, we obtain

$$\gamma V(y) - \langle p, g(y, v)\rangle - f(y, v) = \gamma w(y) - \langle \nabla^\beta w(y), g(y, v)\rangle - f(y, v) \le 0.$$

Therefore,

$$\gamma V(y) + H(y, p) \le 0,$$

that is to say, V is a β-viscosity subsolution of (8).

b. Supersolution. Now let y be an element of X such that $p \in D_\beta V(y)$. Then there exists a β-smooth function w such that $\nabla^\beta w(y) = p$, y is a (local) minimum of $V - w$, and $0 = (V - w)(y)$. By the dynamic programming principle, for each integer i, there exists $u^i \in \mathcal{U}$ such that

$$w(y) + \frac{1}{i^2} = V(y) + \frac{1}{i^2} \ge \int_0^{1/i} e^{-\gamma s} f(x(s, y, u^i), u^i(s))ds + e^{-\gamma/i} V\left(x\left(\frac{1}{i}, y, u^i\right)\right). \tag{10}$$

Arguments similar to those in the previous paragraph yield

$$\frac{1}{i} + i\int_0^{1/i} e^{-\gamma s}[\gamma w(x(s, y, u^i)) - \langle \nabla^\beta w(x(s, y, u^i)), g(x(s, y, u^i), u^i(s))\rangle - f(x(s, y, u^i), u^i(s))]ds \ge 0.$$

We rewrite this inequality as

$$\gamma w(y) + i\int_0^{1/i} [-\langle \nabla^\beta w(y), g(y, u^i(s))\rangle - f(y, u^i(s))]ds \ge h(i),$$

where

$$h(i) = -\frac{1}{i} + h_1(i) + h_2(i) + h_3(i)$$

and the h_j's are defined as follows:

$$h_1(i) := \gamma w(y) - i\int_0^{1/i} e^{-\gamma s}\gamma w(x(s, y, u^i))ds,$$

$$h_2(i) := i \int_0^{1/i} [e^{-\gamma s} \langle \nabla^\beta w(x(s, y, u^i)), g(x(s, y, u^i), u^i(s)) \rangle - \langle \nabla^\beta w(y), g(y, u^i(s)) \rangle] ds,$$

and

$$h_3(i) := i \int_0^{1/i} [e^{-\gamma s} f(x(s, y, u^i), u^i(s)) - f(y, u^i(s))] ds.$$

It is obvious that

$$-\langle \nabla^\beta w(y), g(y, u^i(s)) \rangle - f(y, u^i(s)) \leq H(y, p)$$

and, therefore,

$$\gamma V(y) + H(y, p) \geq h(i). \tag{11}$$

Since

$$\sup \left\{ \|x(s, y, u^i) - y\| : s \in \left[0, \frac{1}{i}\right] \right\} \to 0$$

when $i \to \infty$, $\lim_{i\to\infty} h_1(i) = \lim_{i\to\infty} h_3(i) = 0$. Observing that g is Lipschitz in x uniformly in U and $g(y, u^i(s))$ is in the β-set K, the β-smoothness of w at y yields $\lim_{i\to\infty} h_2(i) = 0$. Therefore, $\lim_{i\to\infty} h(i) = 0$. Sending i to infinity in (11), we obtain that V is a β-viscosity supersolution of (8) and, hence, a β-viscosity solution of (8). □

4. Metric regularity. Metric regularity is closely related to the open mapping property and plays an important role in studying exact penalization and necessary conditions for constrained minimization problems. In this section we apply the fuzzy sum rules derived in §2 to derive a necessary condition for metric regularity to fail; contraposition produces a sufficient condition for a function to be metrically regular at a given point. We deduce a dual condition parallel to the dual conditions given in Ioffe [27] and Ginsburg and Ioffe [23] and a β-smooth space version of the primal condition discussed in Borwein and Strojwas [5]. Many authors have discussed primal sufficient conditions. The condition that we give below (Theorem 4.9) appears to be the weakest (in β-smooth spaces) up to now. Various primal conditions discussed in [1, 2, 3, 5] are deduced as corollaries. To avoid complications in the notations we consider metric regularity for a single-valued function with respect to a closed set. This is in fact an entirely general formulation (see [1] and [29]). There are many discussions about regularity of multifunctions (see, for example, [31, 35] and the references therein). We will indicate at the end of this section how to handle multifunctions with our results. We recall the following definition of metric regularity.

DEFINITION 4.1. *Let X and Y be Banach spaces. One says that $h : X \to Y$ is* regular with respect to S at x_0 *if there exist positive constants ε and K such that, for all $x \in S \cap (x_0 + \varepsilon B)$ and $\xi \in h(x_0) + \varepsilon B$,*

$$d(x, S \cap h^{-1}(\xi)) \leq K \|h(x) - \xi\|.$$

DEFINITION 4.2. *Let X and Y be Banach spaces with β-smooth norms, S be a closed subset of X, and $h : X \to Y$ be differentiable at x_0. We say that x_0 is an* extremal with respect to (h, S) *if, for any $\eta > 0$, there exist a unit vector y^*, $y \in (x_0 + \eta B) \cap S$, and a constant L such that*

$$0 \in (\nabla h(x_0))^* y^* + L D_\beta d(y, S) + \eta B_{X^*}.$$

The next theorem shows that a function is not metrically regular at a point corresponding to the fuzzy extremal condition described above at the given point. The essential part of the proof is an application of the Ekeland variational principle that was first introduced in [24] for handling metric regularity problems and now becomes quite standard. Before stating the theorem, let us recall that a mapping $f : X \to Y$ is called strictly differentiable at x, provided that f is (Fréchet) differentiable at x, for each v,

$$\lim_{x' \to x, t \to 0^+} \left\| \frac{f(x' + tv) - f(x')}{t} - \langle \nabla f(x), v \rangle \right\| = 0,$$

and the convergence is uniform for $v \in B_X$. Note that f is strictly differentiable at x implies that f is locally Lipschitz at x.

THEOREM 4.3. *Let X and Y be Banach spaces with β-smooth norms, S be a closed subset of X, and $x_0 \in S$. Assume that $h : X \to Y$ is strictly differentiable at x_0. Then h is not regular with respect to S at x_0 implies that x_0 is an extremal with respect to (h, S).*

Proof. Let η be an arbitrary positive constant in $(0, 1)$. Since h is strictly differentiable at x_0, for any $d \in B_X$, there exists $\eta_1 < \eta/4$ such that $\|x - x_0\| < \eta_1$ and $t \in (0, \eta_1)$ implies that

$$\frac{h(x + td) - h(x)}{t} - \nabla h(x_0)d \in \frac{\eta}{2} B_Y \tag{12}$$

and

$$\|h(x) - h(x_0)\| < \frac{\eta^2}{128}.$$

Suppose h is not regular with respect to S at x_0. Then there exist an $s \in S \cap (x_0 + \frac{\eta_1}{4} B)$ and a $\xi \in h(x_0) + \frac{\eta^2}{128} B$ such that

$$d(s, S \cap h^{-1}(\xi)) > \frac{1}{\eta} ||h(s) - \xi||. \tag{13}$$

By inequality (13) $d(s, S \cap h^{-1}(\xi)) > 0$ and, therefore, $\|h(s) - \xi\| > 0$. Apply Ekeland's principle with

$$f(x) := ||h(x) - \xi||, \qquad \varepsilon := ||h(s) - \xi|| \leq \frac{\eta^2}{64},$$

$$\lambda := \frac{\eta}{8}, \qquad \bar{x} := s.$$

Then there exist $v \in S$ such that

1. $||s - v|| \leq \frac{\eta}{8}$.
2. v minimizes

$$x \to ||h(x) - \xi|| + \frac{\varepsilon}{\lambda} ||x - v||$$

on S.

Since h is strictly differentiable at x_0, it is locally Lipschitz at x_0. Let L' be a local Lipschitz constant for h at x_0. Then $L := L' + 1$ is a local Lipschitz constant for $x \to ||h(x)|| + \frac{\varepsilon}{\lambda} ||x - v||$. Thus, using the penalization result of Clarke [11] we have

2′. v locally minimizes

$$x \to ||h(x) - \xi|| + \frac{\varepsilon}{\lambda}||x - v|| + L \cdot d(x, S).$$

By Theorem 2.9 there exist y, z, and w such that

$$\begin{aligned} 0 &\in D_\beta||h(\cdot) - \xi||(y) + \frac{\varepsilon}{\lambda}D_\beta||w - v|| + LD_\beta d(z, S) + \frac{\eta}{8}B_{X^*} \\ &\subset D_\beta||h(\cdot) - \xi||(y) + LD_\beta d(z, S) + \frac{\eta}{4}B_{X^*}, \end{aligned} \tag{14}$$

where $\|y - v\| < \frac{\eta_1}{4}$ and $\|z - v\| < \frac{\eta_1}{4}$ and, therefore, $\|y - x_0\| < \eta_1$ and $\|z - x_0\| < \eta_1$. Without loss of generality we may assume that y were chosen close enough to s such that $\|h(y) - \xi\| > 0$. Then, for any $d \in B$, when t is small enough, $y_t^*(d) := \nabla^\beta||.||(h(y+td)-\xi)$ exists. Since $\|y_t^*(d)\| = 1$, by (12), for any $d \in B_X$, when t is sufficiently small,

$$\begin{aligned} \frac{\eta}{2} &\geq \left\langle y_t^*(d), \frac{h(y + td) - h(y)}{t} - \nabla h(x_0)d \right\rangle \\ &= \left\langle y_t^*(d), \frac{(h(y + td) - \xi) - (h(y) - \xi)}{t} \right\rangle - \langle y_t^*(d), \nabla h(x_0)d \rangle \\ &\geq \frac{||h(y + td) - \xi|| - ||h(y) - \xi||}{t} - \langle y_t^*(d), \nabla h(x_0)d \rangle. \end{aligned} \tag{15}$$

By inclusion (14) there exists $z^* \in LD_\beta d(z, S)$ such that $z^* = -u^* + v^*$, where $u^* \in D_\beta||h(.) - \xi||(y)$ and $v^* \in \frac{\eta}{4}B_{X^*}$. Observe that, for a fixed d, there exists a t_d such that when $t \in (0, t_d)$,

$$\frac{||h(y + td) - \xi|| - ||h(y) - \xi||}{t} > \langle u^*, d \rangle - \frac{\eta}{4} \geq \langle z^*, d \rangle - \frac{\eta}{2}.$$

Combining this with (15) we obtain that when $t \in (0, t_d)$

$$\eta \geq -\langle y_t^*(d), \nabla h(x_0)d \rangle - \langle z^*, d \rangle.$$

Denote $y^* = \nabla^\beta||.||(h(y) - \xi)$. Then $y_t^*(d) = \nabla^\beta||.||(h(y + td) - \xi) \rightharpoonup_{w^*} y^*$ as $t \to 0^+$ and $||y^*|| = 1$. (In fact, $y_t^*(d)$ converges to y^* in Y_β^*.) Taking limits in the above inequality as $t \to 0^+$ yields

$$\eta \geq -\langle y^*, \nabla h(x_0)d \rangle - \langle z^*, d \rangle.$$

Since this is true for all $d \in B$, we obtain

$$(\nabla h(x_0))^* y^* + z^* \in \eta B_{X^*}$$

or

$$0 \in (\nabla h(x_0))^* y^* + LD_\beta d(y, S) + \eta B_{X^*},$$

as was to be shown. □

We now turn to some sufficient conditions for h to be regular with respect to S at x_0. We consider a dual condition first. To do so we need the following definitions.

DEFINITION 4.4. *Let X be a Banach space with a β-smooth norm and S be a closed subset of X. We define the* β-normal cone (*denoted by $N_\beta(x, S)$*) *and the* β-tangent cone (*denoted by $T_\beta(x, S)$*) *of S at $x \in S$ by*

$$N_\beta(x, S) := \bigcup_{L>0} L \cdot D_\beta d\,(x, S)$$

and

$$T_\beta(x, S) := (N_\beta(x, S))^0.$$

Here we adopt the convention that $(\emptyset)^0 = X$.

Remark 4.5. Recall that the contingent cone $T_b(x, S)$ and the Clarke tangent cone $T_c(x, S)$ of S at $x \in S$ are defined by

$$T_b(x, S) := \left\{u : \liminf_{t\to 0^+} \frac{d(x + tu, S)}{t} = 0\right\}$$

and

$$T_c(x, S) := \left\{u : \limsup_{t\to 0^+, y\to x} \frac{d(y + tu, S)}{t} = 0\right\},$$

respectively. It is easy to see that $T_c(x, S) \subset T_b(x, S) \subset T_\beta(x, S)$ for any bornology β. Note also that $N_\beta(x, S)$ need not be closed. Also, it is known [8] that $T_c(x, S) = (N_g(x, S))^0$.

The next definition is a β-normal version of the finite-codimension condition defined in [23]. In a β-smooth space this condition is less restrictive than the definition in [23].

DEFINITION 4.6. *Let X be a Banach space with an equivalent β-smooth norm and Y be a Banach space. We say that $f : X \to Y$ has the* β-finite codimension property *with respect to S at x_0 if there is a weak-star closed subspace $V^* \subset Y^*$ of finite codimension and constants $\varepsilon, c > 0$ such that if*

(1) $x \in S$ *and* $\|x - x_0\| < \varepsilon$,

(2) $x^* \in N_\beta(x, S)$,

(3) $\|y^*\| = 1, d(y^*, V^*) < \varepsilon$,

then

$$\|(\nabla f(x_0))^* y^* + x^*\| \geq c.$$

The next theorem can be deduced from [23, Thm. 2.7]. We give a self-contained proof by using Theorem 4.3.

THEOREM 4.7. *Let X and Y be Banach spaces with β-smooth norms and S be a closed subset of X. Assume that h is strictly differentiable at x_0 and has the β-finite-codimension property with respect to S at x_0. Then*

$$cl\nabla h(x_0) T_c(x_0, S) = Y \tag{16}$$

implies that h is regular with respect to S.

Proof. In light of Theorem 4.3 we need only to show that x_0 is not an extremal with respect to (h, S). Assume on the contrary that x_0 is an extremal with respect to (h, S). Then, for each n, there exist a unit vector y_n^* and $y_n \in (x_0 + \frac{1}{n}B) \cap S$ such that

$$0 \in (\nabla h(x_0))^* y_n^* + L D_\beta d(y_n, S) + \frac{1}{n} B_{X^*}. \tag{17}$$

We may assume that $y_n^* \to_{w^*} y^*$. Then by the definition of the g-normal cone

$$-(\nabla h(x_0))^* y^* \in N_g(x_0, S).$$

In other words, y^* is normal to $\nabla h(x_0) T_c(x_0, S)$ (because $T_c(x_0, S) = N_g(x_0, S)^0$), which, together with (16), yields $y^* = 0$. Since h has the β-finite-codimension property with respect to S at x_0, there exists a weak-star closed subspace V^* of Y^* such that h, S, and V^* satisfy the conditions in Definition 4.4. Write $Y^* = V^* + W^*$, where W^* is a finite-dimensional space with $V^* \cap W^* = \{0\}$ and $y_n^* = v_n^* + w_n^*$ with $v_n^* \in V^*$ and $w_n^* \in W^*$. Since W^* is finite dimensional and $\|w_n^*\| \le \|y_n^*\| = 1$, without loss of generality we may assume that $w_n^* \to w^*$. Since $y^* = 0$, we have $w^* = 0$. Thus, $d(V^*, y_n^*) \to 0$. Observe that (17) implies that, for each n, there exists an $e_n^* \in \frac{1}{n} B_{X^*}$ such that

$$u_n^* := -(\nabla h(x_0))^* y_n^* + e_n^* \in L D_\beta d(y_n, S) \subset N_\beta(y_n, S).$$

When n is sufficiently large, the β-finite-codimension property of h with respect to S at x_0 implies that

$$0 < c \le \|(\nabla h(x_0))^* y_n^* + u_n^*\| = \|e_n^*\| \le \frac{1}{n} \quad \forall n,$$

which is a contradiction. □

Remark 4.8. Since $N_\beta(x, S) \subset N_g(x, S)$, a function f which satisfies the codimension condition in [23] also satisfies the β-finite-codimension property. Thus, the β-finite-codimension property is less restrictive than the codimension condition in [23]. However, we should note that the result in [23] was proven without the smoothness assumption on the underlying space.

Now we turn to a β-smooth space version of the primal condition discussed in [5].

THEOREM 4.9. *Let X and Y be Banach spaces with β-smooth norms and S be a closed subset of X. Assume that $h : X \to Y$ is strictly differentiable at x_0. Suppose that*

(i) $\operatorname{cl} \nabla h(x_0) T_c(x_0, S) = Y$,

(ii) *there exists a nonempty compact set $K \subset Y$ and positive numbers α, γ, and η such that $\alpha < \gamma$ and, for all $\|x - x_0\| \le \eta$,*

$$\gamma B_Y \subset \operatorname{cl} \{\nabla h(x_0)(T_\beta(x, S) \cap B_X) + \alpha B_Y\} + K.$$

Then h is regular with respect to S at x_0.

Proof. We need only to show that the condition of this theorem implies that of Theorem 4.7. Suppose that conditions (i) and (ii) are satisfied. By [4, Lem. 2.1] we may assume that K is finite dimensional. Let $V := \operatorname{span}(K)$ and

$$V^* := \{y^* \in Y^* : y^* \perp V\}.$$

Then V^* is of finite codimension. Let

$$\varepsilon := \min\left\{\eta, \frac{\gamma - \alpha}{2 \sup\{\|k\| : k \in K\}}\right\}$$

and $c := \gamma - \varepsilon \sup\{\|k\| : k \in K\} - \alpha$. Consider x, x^*, and y^* satisfying (1), (2), and (3) in Definition 4.12. Let ξ be an arbitrary element of γB_Y. Then by condition (ii), for any $\alpha' > \alpha$, there exist $u \in T_\beta(x, S) \cap B_X$, $b \in B_Y$, and $k \in K$ such that

$$\xi = \nabla h(x_0) u + \alpha' b + k.$$

Then

$$\begin{aligned}\langle y^*, -\xi\rangle &= \langle(\nabla h(x_0))^* y^*, -u\rangle + \alpha'\langle y^*, -b\rangle + \langle y^*, -k\rangle \\ &= \langle(\nabla h(x_0))^* y^* + x^*, -u\rangle + \langle x^*, u\rangle + \alpha'\langle y^*, -b\rangle + \langle y^*, -k\rangle \\ &\le \|(\nabla h(x_0))^* y^* + x^*\| + \alpha' + \varepsilon \sup\{\|k\| : k \in K\}.\end{aligned}$$

Taking the supremum on the left-hand side and letting $\alpha' \to \alpha$ lead to

$$\gamma \le \|(\nabla h(x_0))^* y^* + x^*\| + \alpha + \varepsilon \sup\{\|k\| : k \in K\}$$

or

$$c \le \|(\nabla h(x_0))^* y^* + x^*\|. \qquad \square$$

Remark 4.10. Since $T_c(x, S) \subset T_b(x, S) \subset T_\beta(x, S)$, the condition of this theorem is weaker than that of [5, Thm. 4.1] in β-smooth spaces. We should note that the results of [5, Thm. 4.1] are proven without the β-smoothness assumption. The same remark also applies to the corollaries below.

As shown in Borwein and Strojwas [5], in a β-smooth space, conditions (i) and (ii) in Theorem 4.9 are weaker than the sufficient conditions for regularity given in Borwein [3] for the cases where (i) Y is a finite-dimensional space, (ii) S is convex, and (iii) S is epi-Lipschitz-like (see [5] for the definition). Therefore we have the following corollary.

COROLLARY 4.11 (see [3, 5]). *Let X and Y be Banach spaces with β-smooth norms and S be a closed subset of X. Assume that $h : X \to Y$ is strictly differentiable at $x_0 \in S$. Then h is regular with respect to S at x_0, provided that one of the following conditions is satisfied:*

a. *S is convex and $0 \in \operatorname{core}\nabla h(x_0)(S - x_0)$.*

b. *S is epi-Lipschitz-like and $\nabla h(x_0)T_c(x_0, S) = Y$.*

c. *Y is finite dimensional and $\nabla h(x_0)T_c(x_0, S) = Y$.*

Next we show that the sufficient condition for regularity in terms of uniform sleekness of the set S given by Aubin and Frankowska [1] is also a direct consequence of Theorem 4.9. First we recall the definitions of sleekness and uniform sleekness.

DEFINITION 4.12. *We say S is* β-sleek *at x if*

$$\liminf_{y \to x} T_\beta(y, S) \supset T_b(x, S).$$

We say S is uniformly β-sleek *at x if, for $u \in T_b(x, S) \cap B$,*

$$\lim_{y \to x} d(u, T_\beta(y, S)) = 0$$

uniformly.

Remark 4.13. (a) By Remark 4.5 β-sleekness is weaker than sleekness defined in [1].

(b) Borwein and Ioffe [8] have shown that

$$T_c(x, S) \supset \liminf_{y \to x} T_\beta(y, S).$$

It is well known that $T_c(x, S) \subset T_b(x, S)$. Thus, if S is β-sleek at x, then $T_c(x, S) = T_b(x, S)$ (i.e., S is b-tangentially regular at x).

COROLLARY 4.14. *Let X and Y be Banach spaces with β-smooth norms and S be a closed subset of X. Assume that S is uniformly β-sleek at x_0 and $h : X \to Y$ is strictly differentiable at $x_0 \in S$. Then*

$$\nabla h(x_0)T_c(x_0, S) = Y \tag{18}$$

implies that h is regular with respect to S at x_0.

Proof. We need only to show that h and S satisfy condition (ii) in Theorem 4.9. By the Robinson–Ursescu theorem (see [1]) there exists a $\gamma > 0$ such that

$$\gamma B_Y \subset \nabla h(x_0)\left(T_c(x_0, S) \cap \frac{1}{2}B_X\right).$$

Take an $\varepsilon < \frac{1}{2}$ such that $\alpha := \varepsilon\|\nabla h(x_0)\| < \gamma$. The uniform sleekness of S at x_0 implies that there exists an η such that $\|x - x_0\| \leq \eta$ implies that

$$T_c(x_0, S) \cap \frac{1}{2}B_X \subset T_\beta(x, S) \cap B_X + \varepsilon B_X.$$

Therefore

$$\gamma B_Y \subset \nabla h(x_0)\left(T_c(x_0, S) \cap \frac{1}{2}B_X\right) \subset \operatorname{cl}\{\nabla h(x_0)(T_\beta(x, S) \cap B_X) + \alpha B_Y\},$$

as was to be shown. □

Recently Azé and Chou [2] derived a primal sufficient condition in terms of equi-circatangent cones. It turns out this condition can also be deduced directly from Theorem 4.9.

DEFINITION 4.15 (see [2]). *A cone $K \subset X$ is said to be* equi-circatangent *to S at $x \in S$ if*

$$\lim_{t \to 0^+, y \to_S x} e(K \cap B_X, t^{-1}(S - y)) = 0,$$

where $e(A, S) := \inf\{\lambda > 0 : A \subset S + \lambda B_X\}$.

The primal condition discussed in [2] (as specialized for single-valued functions) can be stated as the following corollary.

COROLLARY 4.16. *Let X and Y be Banach spaces with β-smooth norms and S be a closed convex subset of X. Assume that $h : X \to Y$ is strictly differentiable at x_0. Let K be an equi-circatangent cone of S at x_0. Suppose that*

$$\nabla h(x_0)K = Y.$$

Then h is regular with respect to S at x_0.

Proof. We show that the conditions of this corollary imply those of Theorem 4.9. Condition (i) is obvious. To show condition (ii), invoke the Robinson–Ursescu theorem: there exists a positive constant γ such that

$$\gamma B_X \subset \nabla h(x_0)\left(K \cap \frac{1}{2}B_X\right).$$

By the definition of K there exists $\eta > 0$ such that, for all $t \in (0, \eta)$ and $x \in (x_0 + \eta B_X) \cap S$,

$$K \cap B_X \subset \frac{S - x}{t} + \varepsilon B_X,$$

where $\varepsilon := \min(\frac{\gamma}{2\|\nabla h(x_0)\|}, \frac{1}{2})$. Then

$$K \cap \frac{1}{2}B_X \subset T_\beta(x, S) \cap B_X + \varepsilon B_X.$$

Therefore

$$\begin{aligned}\gamma B_Y &\subset \nabla h(x_0)\left(K \cap \frac{1}{2}B_X\right)\\ &\subset \nabla h(x_0)(T_\beta(x, S) \cap B_X) + \varepsilon \nabla h(x_0)B_X\\ &\subset \nabla h(x_0)(T_\beta(x, S) \cap B_X) + \frac{\gamma}{2}B_X. \qquad \square\end{aligned}$$

Finally let us discuss regularity for multifunctions.

DEFINITION 4.17. *Let X and Y be Banach spaces and F be a multifunction from X to Y. We say that F is* regular *at $(x_0, y_0) \in Gr(F)$ if there exist positive constants ε and γ such that, for all $x \in x_0 + \varepsilon B$, $y \in y_0 + \varepsilon B$, and $d(y_0, F(x)) < \varepsilon$,*

$$d(x, F^{-1}(y)) \leq \gamma d(y, F(x)).$$

The following well-known lemma (see, e.g., [1, 29]) will enable us to connect the regularity of a multifunction to that of a single-valued function and derive sufficient conditions for a multifunction to be regular through various of the results discussed before. We include a proof here for completeness.

LEMMA 4.18. *Let $p : X \times Y \to Y$ be defined by $p(x, y) = y$. For a multifunction F from X to Y to be regular at $(x_0, y_0) \in Gr(F)$, it suffices that p be regular with respect to $Gr(F)$ at (x_0, y_0).*

Proof. Since p is regular with respect to $Gr(F)$ at (x_0, y_0), there exist positive constants ε and γ such that, for all $(x, z) \in ((x_0, y_0) + \varepsilon B) \cap Gr(F)$ and $y \in y_0 + \varepsilon B$,

$$d((x, z), Gr(F) \cap p^{-1}(y)) \leq \gamma \|z - y\|.$$

We may assume that $\gamma > 1$. Let $\varepsilon_1 = \varepsilon/2$. If $\|x - x_0\| < \varepsilon_1$, $\|y - y_0\| < \varepsilon_1$, and $d(y_0, F(x)) < \varepsilon_1$, then taking the infimum with respect to $z \in F(x)$ yields

$$\begin{aligned}
\gamma d(y, F(x)) &= \gamma \inf\{\|z - y\| : z \in F(x)\} \\
&= \gamma \inf\{\|z - y\| : z \in F(x), \|z - y_0\| < \varepsilon\} \\
&\geq \inf_{z \in F(x)} d((x, z), Gr(F) \cap p^{-1}(y)) \\
&= \inf_{z \in F(x)} \inf\{\|x - x'\| + \|z - y'\| : x' \in F^{-1}(y'), y' = y\} \\
&= \inf_{z \in F(x)} \{d(x, F^{-1}(y)) + \|z - y\|\} \\
&= d(x, F^{-1}(y)) + d(y, F(x)).
\end{aligned}$$

Thus,

$$d(x, F^{-1}(y)) \leq (\gamma - 1) d(y, F(x)). \qquad \square$$

Using this lemma and the various sufficient conditions that we discussed before will lead to sufficient conditions for a multifunction to be regular. For example, combining Lemma 4.18 with Theorem 4.9 we get the following theorem.

THEOREM 4.19. *Let X and Y be Banach spaces with β-smooth norms, F be a multifunction from X to Y, and p be the projection defined in Lemma* 4.18. *Suppose that*

(i) $\operatorname{cl}[p(T_c(x_0, y_0; Gr(F)))] = Y$,

(ii) *there exists a nonempty compact set $K \subset Y$ and positive numbers α, γ, and η such that $\alpha < \gamma$ and*

$$\gamma B_Y \subset \operatorname{cl}\{p(T_c(x_0, y_0; Gr(F))) \cap B_X + \alpha B_Y\} + K$$

whenever $\|x - x_0\| \leq \eta$ and $\|y - y_0\| \leq \eta$.

Then F is regular at (x_0, y_0).

Acknowledgments. We thank P. D. Loewen, A. Ioffe, L. Thibault, and the referees for their helpful comments and criticism regarding an earlier version of this paper.

REFERENCES

[1] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.
[2] D. AZÉ AND C. C. CHOU, *On a Newton type iterative method for solving inclusions*, in Math. Oper. Res., to appear.
[3] J. M. BORWEIN, *Stability and regular points of inequality systems*, J. Optim. Theory Appl., 48 (1986), pp. 9–52.
[4] ———, *Epi-Lipschitz-like sets in Banach space: Theorems and examples*, Nonlinear Anal., 11 (1987), pp. 1207–1217.
[5] J. M. BORWEIN AND H. M. STROJWAS, *The hypertangent cone*, Nonlinear Anal., 13 (1989), pp. 125–144.
[6] J. M. BORWEIN AND S. FITZPATRICK, *A weak Hadamard smooth renorming of* $L_1(\Omega, \mu)$, Canad. Math. Bull., 36 (1993), pp. 407–413.
[7] ———, *Weak-star sequential compactness and bornological limit derivatives*, Convex Analysis: Special issue in celebration of R. T. Rockafellar's 60th birthday, part I, 2 (1995), pp. 59–68.
[8] J. M. BORWEIN AND A. IOFFE, *Proximal analysis in smooth spaces*, CECM research report 93-04 (1993), to appear in Set-Valued Anal.
[9] J. M. BORWEIN AND D. PREISS, *A smooth variational principle with applications to subdifferentiability and to differentiability of convex functions*, Trans. Amer. Math. Soc., 303 (1987), pp. 517–527.
[10] J. M. BORWEIN AND Q. J. ZHU, *Variational analysis in non-reflexive spaces and applications to control problems with* L^1 *perturbations*, Nonlinear Anal., 26 (1996); to appear.
[11] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.
[12] ———, *Methods of Dynamic and Nonsmooth Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math., Society for Industrial and Applied Mathematics, Philadelphia, 1989.
[13] M. G. CRANDALL, L. C. EVANS, AND P.-L. LIONS, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.
[14] M. G. CRANDALL, H. ISHII, AND P.-L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
[15] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
[16] ———, *Hamilton-Jacobi equations in infinite dimensions, Part* I. *Uniqueness of viscosity solutions*, J. Funct. Anal., 62 (1985), pp. 379–396; *Part* II. *Existence of viscosity solutions*, 65 (1986), pp. 368–405; *Part* III, 68 (1986), pp. 214–247; *Part* IV. *Unbounded linear terms*, 90 (1990), pp. 237–283; *Part* V, *B-continuous solutions*, 97 (1991), pp. 417–465.
[17] R. DEVILLE, *Stability of Subdifferential of Nonconvex Functions in Banach Spaces*, preprint.
[18] R. DEVILLE, G. GODEFROY, AND V. ZIZLER, *Smoothness and Renormings in Banach Spaces*, Pitman Monographs Surveys Pure Appl. Math. 64, John Wiley Sons, New York, 1993.
[19] ———, *A smooth variational principle with applications to Hamilton-Jacobi equations in infinite dimensions*, J. Funct. Anal., 111 (1993), pp. 197–212.
[20] R. DEVILLE AND E. M. E. HADDAD, *The Subdifferential of the Sum of Two Functions in Banach Spaces*, I. *First Order Case*, preprint.
[21] M. FABIAN, *On class of subdifferentiability spaces of Ioffe*, Nonlinear Anal., 12 (1988), pp. 63–74.
[22] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, 1975.
[23] B. GINSBURG AND A. D. IOFFE, *The maximum principle in optimal control of system governed by semilinear equations*, in Proceedings of the IMA Workshop on Nonsmooth Analysis and Geometric Methods in Deterministic Optim. Control, B. S. Mordukhovich and H. T. Sussmann, eds., IMA Vol. Math. Appl., Springer-Verlag, New York, 1995.
[24] A. D. IOFFE, *Regular points of Lipschitz mappings*, Trans. Amer. Math. Soc., 251 (1979), pp. 61–69.
[25] ———, *On subdifferentiability spaces*, Ann. New York Acad. Sci., 410 (1983), pp. 107–119.
[26] ———, *Subdifferentiability spaces and nonsmooth analysis*, Bull. Amer. Math. Soc., 10 (1984), pp. 87–90.
[27] ———, *Necessary conditions for nonsmooth optimization*, Math. Oper. Res., 9 (1984), pp. 159–189.
[28] ———, *Calculus of Dini subdifferentials of functions and contingent derivatives of set-valued maps*, Nonlinear Anal., 8 (1984), pp. 517–539.
[29] ———, *On the local surjection property*, Nonlinear Anal., 11 (1987), pp. 565–590.
[30] ———, *Proximal analysis and approximate subdifferentials*, J. London Math. Soc., 41 (1990), pp. 175–192.
[31] A. JOURANI AND L. THIBAULT, *Extensions of Subdifferential Calculus Rules in Banach Spaces and Applications*, preprint.
[32] A. Y. KRUGER AND B. S. MORDUKHOVICH, *Extremal points and Euler equations in nonsmooth optimization*, Dokl. Akad. Nauk. BSSR, 24 (1980), pp. 684–687. (In Russian.)

[33] J. LINDENSTRAUSS AND L. TZAFRIRI, *Classical Banach Spaces* II: *Function Spaces*, Springer-Verlag, Berlin, 1979.

[34] P. D. LOEWEN, *Optimal Control via Nonsmooth Analysis*, CRM Lecture Notes Series, Amer. Math. Soc., Summer School on Control, CRM, Université de Montréal, 1992, Amer. Math. Soc., Providence, 1993.

[35] B. MORDUKHOVICH, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.

[36] ———, *Maximum principle in problems of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.

[37] ———, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.

[38] B. MORDUKHOVICH AND Y. SHAO, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., to appear.

[39] R. R. PHELPS, *Convex Functions, Monotone Operators and Differentiability*, Lecture Notes in Math. 1364, Springer-Verlag, New York, Berlin, Tokyo, 1988, 2nd ed., 1993.

[40] R. T. ROCKAFELLAR, *Proximal subgradients, marginal values and augmented Lagrangians in nonconvex optimization*, Math. Oper. Res., 6 (1981), pp. 424–436.

[41] ———, *Extensions of subgradients and its applications to optimization*, Nonlinear Anal., 9 (1985), pp. 665–698.

[42] L. THIBAULT, *Subdifferentials of compactly Lipschitzian vector valued functions*, Ann. Math. Pura Appl. (4), 125 (1980), pp. 157–192.

[43] D. E. WARD AND J. M. BORWEIN, *Nonsmooth calculus in finite dimensions*, SIAM J. Control Optim., 25 (1987), pp. 1312–1340.

[44] J. WARGA, *Derivate containers, inverse functions, and controllability*, in Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976.

[45] ———, *Fat homeomorphisms and unbounded derivate containers*, J. Math. Anal. Appl., 81 (1981), pp. 545–560.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1592–1615, September 1996

007

OPTIMAL RELAXED CONTROLS FOR INFINITE-DIMENSIONAL STOCHASTIC SYSTEMS OF ZAKAI TYPE*

N. U. AHMED†

Abstract. In this paper, we present some new results on partially observed control problems for infinite-dimensional stochastic systems in Hilbert space using a fundamental result of Da Prato and Zabczyk on an infinite-dimensional Kolmogorov operator. We prove the existence of optimal relaxed controls for an infinite-dimensional Zakai equation following a semigroup approach and the theory of measurable selections. This result is also extended to differential inclusion. We also present some necessary conditions of optimality.

Key words. partially observed, infinite-dimensional stochastic systems, Zakai equations and inclusions, optimal relaxed controls, existence, necessary conditions

AMS subject classifications. 93E20, 93E03, 93E11, 49J20, 49J24, 49K20, 49K24

1. Introduction.

We consider the following controlled system governed by a pair of stochastic differential equations as described below:

$$\begin{aligned} dx &= Axdt + F(x)dt + B(x, u(t, y))dt + \sqrt{Q}dW, \qquad x(0) = x_0, \\ dy &= h(x, y)dt + \sigma_0(y)dw^0, \qquad y(0) = 0, \end{aligned} \tag{1.1}$$

where the first equation is defined on an infinite-dimensional Hilbert space H and the second is defined on a finite-dimensional Euclidian space R^d. The process x, which is generally not observable, is controlled through a controller u which exercises its control actions on the basis of available information about the process y which is physically measurable. The fundamental objective is to find from a suitable class of operators or maps, to be introduced shortly, a control law which minimizes the following cost functional:

$$J(u) \equiv E \int_I \ell(t, y(t), x(t), u(t, y))dt. \tag{1.2}$$

Generally A is an unbounded operator with domain and range in a separable Hilbert space H, F is a nonlinear continuous bounded operator in H, and B is also a continuous bounded map from $H \times \mathcal{Z}$ to H, where $\mathcal{Z}$ is a suitable set to be defined shortly. The map Q is a symmetric positive operator in H and W is a cylindrical Brownian motion with values in H. The operator h is a continuous bounded map from $H \times R^d$ to R^d and σ_0 is a continuous bounded map from R^d to the space of symmetric $d \times d$ matrices. Precise hypotheses will be introduced shortly.

Partially observed finite-dimensional control problems have been studied extensively by many authors over the last two decades [12]–[17], including the recent (1992) excellent book by Bensoussan [16] and the references therein. Fully observed infinite-dimensional control problems have been extensively studied by Barbu and Da Prato and Da Prato and his school through Hamilton–Jacobi–Bellman (HJB) equations [2, 3, 4]. Recently Zhu and Ahmed [5] considered HJB equations on Banach spaces related to fully observed stochastic control problems. However very little is known on infinite-dimensional partially observed control problems [7, 9]. In [9] the author considers partially observed stochastic differential inclusions using the theory of monotone operators and uses a stochastic approach. Recently Da Prato and Zabczyk discovered some very interesting results [1] regarding Kolmogorov operators and

*Received by the editors June 6, 1994; accepted for publication (in revised form) May 10, 1995.

†Department of Electrical Engineering, Department of Mathematics, University of Ottawa, Ottawa, ON, K1N 6N5, Canada.

the associated semigroups for infinite-dimensional stochastic differential equations on Hilbert spaces. This opens up the prospects of treating nonlinear filtering and control with partial information using an analytic approach in contrast to a stochastic approach [9]. Using their results Ahmed and Zabczyk [6] recently obtained some new results on nonlinear filtering of infinite-dimensional processes with finite-dimensional observation. In this paper we use the results of Da Prato and Zabczyk and those of Ahmed and Zabczyk as the starting point to study partially observed stochastic control problems in infinite-dimension, as stated above.

The rest of the paper is organized as follows: In §2 we discuss motivation and present some physical examples followed by basic notations. In §3, basic assumptions and some fundamental results due to Da Prato and Zabczyk and to Ahmed and Zabczyk are quoted for the convenience of readers. In §4, admissible controls are introduced and the partially observed control problem is transformed into a fully observed one. In §5, existence of optimal controls is proved. In §6, a similar result is proved for evolution inclusions. In §7, some necessary conditions of optimality are presented. We conclude the paper with §8 discussing the applicability and limitations of our results.

2. Motivation. The general problem of controlling the behavior of an unobservable random process based on the observation of a physically measurable process is known as a partially observed control problem. This is a general problem and arises in many physical situations. A classical example is an electrical communication problem where the received signal is corrupted by noise picked up from the source itself or (and) from the communication channel. The receiver is required to estimate the true message from the corrupted one. Often feedback communication is used to control the shape of the message signals in order to minimize error estimates at the receiver. This is a classical problem in finite-dimensional space that led to the Winer–Hoff equation and later Kalman filtering for linear dynamic systems and Kushner and Zakai equations for nonlinear systems. We present here two infinite-dimensional examples: one arising from ecological problems and another from an electromagnetic interference problem.

Ecological problem. Consider an aquatic system, like the Great Lakes, inhabited by various species of marine life, which is naturally affected by the presence of organic and inorganic agents. The concentration of organic and inorganic agents such as pollutants and nutrients in the water body can be described by a partial differential equation (PDE) as follows:

$$\begin{aligned} &\frac{\partial C}{\partial t} - D\Delta C + (\nabla C)v = b(C,u) + N, \qquad t \geq 0, \quad \xi \in \Omega, \\ &C|_{\partial\Omega} = 0, C(0,\xi) = C_0(\xi), \qquad \xi \in \Omega, \end{aligned} \tag{2.1}$$

where Ω is assumed to be an open, connected, bounded domain representing the aquatic body. C represents the concentration level of, say, m different organic and inorganic agents such as pollutants and nutrients. The function b represents the interactions between s different control agents u and the m different pollutants and nutrients C. The control may take the form of organized application of biological and biochemical agents reacting with the pollutants or simply physical removal of visible objects such as solid waste, algae, and other phytoplanktons. N is the distributed noise representing the additive effect of land run-offs from surrounding farmlands, summer cottages, acid rain, accidental oil spills, etc. The third term on the left of the equation represents transport of C due to water movement, where v is the given velocity vector as a function of space–time. For simplicity we shall assume that v is the steady-state velocity, that is, $v(t,x) = v(x)$ independent of time. The aquatic system is also inhabited by much important marine life such as microorganisms and fish. The stock of fish is subject to regulation by the Department of Fisheries. Assuming d different species of population and

denoting their biomass per unit volume by the d-vector y, a simplified model is given by

$$(d/dt)y = h(C, y) + N_0, \qquad t \geq 0, \tag{2.2}$$
$$y(0) = y_0,$$

where h represents the growth vector and N_0 the noise vector. For h we take the standard logistic growth function given by

$$h_i(C, y) \equiv \begin{cases} 0 & \text{for } y_i < 0; \\ r_i y_i(1 - (y_i/K_i)), & 0 \leq y_i \leq K_i; \\ \text{Exp}(-\delta_i K_i)(e^{-\delta_i(y_i - K_i)} - 1), & y_i > K_i, \end{cases} \tag{2.3}$$
$$K_i = K_i(C), \qquad i = 1, 2, \ldots, d.$$

The coefficient r_i is a positive constant representing the intrinsic growth rate of the ith species and the coefficient K_i, known as the environmental carrying capacity, is also positive and a nonincreasing (possibly decreasing) function of the concentration level of pollutants and nondecreasing (possibly increasing) function of the concentration level of nutrients in C. Above the carrying capacity the population decreases exponentially at the rate $\delta_i > 0$. The Department of Fisheries and Environment is interested in introducing a control program to promote marine life and water quality. For this purpose one may consider a simple cost integrand such as

$$\ell(y, C, u) \equiv \int_\Omega (Q_1(\xi)C(t, \xi), C(t, \xi))d\xi - \int_\Omega (Q_2(\xi)C(t, \xi), C(t, \xi))d\xi - (Q_3 y, y) + (Q_4 u, u),$$

where Q_1, Q_2 are symmetric, positive, semidefinite $m \times m$ matrix-valued functions bounded on Ω, Q_3 is a symmetric $d \times d$ positive, semidefinite matrix, and Q_4 is a positive definite $r \times r$ matrix. The r-dimensional control signifies r-different control actions including application of antipollutants, biological agents predating unwanted microorganisms, physical removal of solid waste, algae, etc. The cost functional may be taken as

$$J(u) \equiv E \int_I \ell(y, C, u)dt, \tag{2.4}$$

which is to be minimized. The first term promotes selective removal, the second promotes growth of possibly nutrient contents in C considered healthy for marine life, the third term promotes selective growth of marine life (like edible fish), and the fourth term represents the cost of administering controls.

The system (2.1)–(2.2) can be written as the abstract stochastic system (1.1) by choosing $H = L_2(\Omega, R^m)$, and A as the operator given by

$$D(A) \equiv \{\phi \in H : D\Delta\phi - (\nabla\phi)v \in H, \phi|_{\partial\Omega} = 0\}$$
$$= H^2(\Omega, R^m) \cap H_0^1(\Omega, R^m),$$

and setting

$$A\phi = D\Delta\phi - (\nabla\phi)v \text{ for } \phi \in D(A).$$

Define B as the Nemytskii operator corresponding to the vector function b by

$$(B(\phi, u), \psi) = \int_\Omega (b(\phi(\xi), u(\xi)), \psi(\xi))d\xi, \qquad \psi \in H,$$

and

$$E(N(t,\xi)N'(s,\eta)) = \delta(t-s)q(\xi,\eta), \qquad \xi,\eta \in \Omega,$$

where q is a positive symmetric kernel and δ is the Dirac measure. Let Q denote the integral operator

$$(Qz)(\xi) \equiv \int_\Omega q(\xi,\eta)z(\eta)d\eta, \qquad z \in H.$$

It is assumed that Q is a positive nuclear operator in H. For N_0, we take standard white noise in R^d so that

$$E(N_0(t)N_0'(s)) = \delta(t-s)I.$$

Hence the system (2.1)–(2.2) can be written as an abstract stochastic differential equation in $H \times R^d$ given by

$$\begin{aligned} dx &= Axdt + B(x,u)dt + \sqrt{Q}dW, \\ dy &= h(x,y)dt + dw^0, \end{aligned} \tag{2.5}$$

where W is a cylindrical Brownian motion in H and w^0 is a standard Brownian motion in R^d associated with the white noise N_0.

Electromagnetic interference. Power line harmonics interfere with nearby telephone lines. Similarly high-density multilayered printed circuit boards and multicored electrical cables experience (interline) interference. This kind of interference is known as "crosstalk" (see Khan and Costache [27, p. 9]). The mathematical model proposed for such systems can be described by a $2m$-dimensional first-order hyperbolic differential equation for currents and voltages associated with m transmission lines (of length l) subject to crosstalk as described below:

$$\begin{aligned} \partial i/\partial t &= -L^{-1}D_\xi v - L^{-1}Ri, \\ \partial v/\partial t &= -C^{-1}D_\xi i - C^{-1}Gv, \qquad t \geq 0, \xi \in (0,l) \equiv \Omega, \end{aligned} \tag{2.6}$$

where D_ξ denotes the first partial with respect to the spatial coordinate ξ; L, C, R, G are constant matrices representing the inductance, capacitance, resistance, and conductance parameters per unit length. For example, $L = L_0 + \tilde{L}$, $C = C_0 + \tilde{C}$, where L_0 is the self-inductance matrix (diagonal) and $\tilde{L}$ is the mutual inductance matrix. The boundary conditions are given by

$$\begin{aligned} v(t,0) + R_0 i(t,0) &= E(t), \\ C_1 \partial v(t,l)/\partial t + g(v(t,l)) &= i(t,l), \end{aligned} \tag{2.7}$$

where g represents the nonlinear $(i - v)$-characteristic of the n-vector terminal load and R_0 is the source-resistance matrix (diagonal) and E the source-voltage vector. Defining $I \equiv i$, $V \equiv v - E$, the first boundary condition in equation (2.7) can be reduced to a homogeneous boundary condition. Assuming distributed noise in the lines and boundary noise due to uncertain fluctuation of the demand and equipment noise, the system takes the form

$$\begin{aligned} \partial I/\partial t &= -L^{-1}D_\xi V - L^{-1}RI + N_1, \\ \partial V/\partial t &= -C^{-1}D_\xi I - C^{-1}GV + u_1(t) + N_2, \end{aligned} \tag{2.8}$$

with boundary conditions given by

(2.9)
$$V(t,0) + R_0 I(t,0) = 0,$$
$$\partial V(t,l)/\partial t + C_1^{-1} g(V(t,l) - G^{-1}C(u_1(t) + C_1^{-1}u_2(t)) + C_1^{-1}u_2(t) = C_1^{-1}I(t,l) + N_3,$$

where

$$\begin{aligned} u_1 &\equiv -C^{-1}GE - \dot{E}, \\ u_2 &\equiv C_1\dot{E}. \end{aligned} \tag{2.10}$$

Current probes placed at d-different locations along the transmission line can be used to measure the level of interference. This is described by a set of d-ordinary differential equations

$$\ell_k(d/dt)i_k = h_k(I, V) + \gamma_k, k = 1, 2, \ldots, d, \tag{2.11}$$

where h_k denotes the voltage induced in the induction coil of the current probe at the kth location and i_k is the current measured, where ℓ_k is the self-inductance of the coil and γ_k is the measurement noise.

The objective here is to choose the control signal $u \equiv \binom{u_1}{u_2}$ so that crosstalk is minimized while satisfying the desired electrical demand. The integrand of such an objective function can be taken as

$$\begin{aligned} \ell(y, I, V, \theta, , u) \equiv (1/2)\int_\Omega \{(\tilde{L}I, I) + (\tilde{C}V, V)\}d\xi + (1/2) \parallel \theta(t) - \theta_d(t) \parallel^2_{R^m} \\ + (1/2) \parallel u \parallel^2_{R^{2m}}, \end{aligned}$$

where $\tilde{L}, \tilde{C}$ are the coupling matrices of inductance and capacitance, $\theta(t) = V(t, l)$, θ_d is the desired load voltage, and u is the source signal. The first two terms give a measure of crosstalk (electromagnetic interference) which is measured in terms of electromagnetic energy and the remaining terms have the standard meaning. The cost functional is given by

$$J(u) = E\int_0^T \ell(y, I, V, \theta, u)dt.$$

Again we can write the system of equations (2.8)–(2.11) as an abstract stochastic differential equation of the form (1.1).

For the nonlinear $(i - v)$-characteristic of the terminal load we assume that g is Lipschitz continuous and there exists a symmetric positive definite matrix K and a number $\beta > 0$ such that

$$\sup\{\parallel g(\theta + z) - K\theta \parallel_{R^m}, \theta \in R^m\} \leq \beta(1 + \parallel z \parallel) \tag{2.12}$$

for all $z \in R^m$. For a mild nonlinear characteristic of the load, this is a reasonable approximation. Define $H \equiv L_2(\Omega, R^m) \times L_2(\Omega, R^m) \times R^m$ with the natural scalar product, and the operator A by

$$D(A) \equiv \{(\phi, \psi, \theta) \in H : \phi, \psi \in H^1(\Omega, R^m), \psi(0) + R_0\phi(0) = 0, \theta = \psi(l)\} \tag{2.13}$$

and

$$A(\phi, \psi, \theta) \equiv \begin{pmatrix} -L^{-1}R\phi - L^{-1}D_\xi\psi \\ -C^{-1}D_\xi\phi - C^{-1}G\psi \\ C_1^{-1}(\phi(l) - K\theta) \end{pmatrix}. \tag{2.14}$$

Define

$$
\begin{aligned}
b(\phi, \psi, \theta, u) &\equiv \begin{pmatrix} 0 \\ u_1 \\ -C_1^{-1} g(\theta - G^{-1} C(u_1 + C_1^{-1} u_2)) + C_1^{-1} K\theta - C_1^{-1} u_2 \end{pmatrix} \\
&\equiv \begin{pmatrix} 0 \\ u_1 \\ \tilde{g}(\theta, u) \end{pmatrix}, \\
N &\equiv \begin{pmatrix} N_1 \\ N_2 \\ N_3 \end{pmatrix}, \\
u &\equiv \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.
\end{aligned} \tag{2.15}
$$

Let $x(t) \equiv \operatorname{col}(I(t, \cdot), V(t, \cdot), \theta(t))$ denote the state vector at time t taking values from H, and $B(x, u)$ denote the Nemytskii operator associated with b. Assume that the transmission line noises (N_1, N_2) are independent of the electrical load (demand) noise N_3 and that the covariance operator Q corresponding to N is a positive nuclear operator in H. Let $y \equiv \operatorname{col}(i_1, i_2, \ldots, i_d)$ denote the measurement process governed by equation (2.11) and w^0 is the standard Brownian motion corresponding to the measurement noise $N_0 \equiv \operatorname{col}(\gamma_1, \gamma_2, \ldots, \gamma_d)$, $h \equiv \operatorname{col}((1/\ell_k)h_k, k = 1, 2, \ldots, d)$ and $\sigma_0 \equiv \operatorname{diag}((1/\ell_1), (1/\ell_2), \ldots, (1/\ell_d))$. Using the above notations we can rewrite the system of equations (2.8)–(2.11) as an abstract stochastic differential equation in $H \times R^d$:

$$
\begin{aligned}
dx &= Ax\,dt + B(x, u)dt + \sqrt{Q}\,dW, \\
dy &= h(x)dt + \sigma_0 dw^0.
\end{aligned} \tag{2.16}
$$

Further discussion of these examples is given in §8.

Basic Notations. For any pair of Banach spaces X, Y, $\mathcal{L}(X, Y)$ will denote the space of bounded linear operators from X to Y and $\mathcal{L}(X) \equiv \mathcal{L}(X, X)$. For any interval $I \equiv [0, T]$, $C(I, X)$ will denote the Banach space of continuous functions on I with values in X. Let H denote a separable Hilbert space, and $C_b(H)(B_b(H))$ the space of bounded continuous (measurable) functions on H. $\mathcal{M}_b(H)$ is the space of countably additive, bounded, signed measures on the measurable space $(H, \mathcal{B}(H))$ and $\mathcal{M}(H) \subset \mathcal{M}_b$ is the space of probability measures. For $\mu \in \mathcal{M}_b(H)$ and $\phi \in B_b(H)$ we use $\mu(\phi)$ to denote the functional $\int_H \phi(x)\mu(dx)$. For any topological space $\mathcal{Z}$, $2^{\mathcal{Z}} \setminus \emptyset$ will denote the space of nonempty subsets of $\mathcal{Z}$, and $c(\mathcal{Z})$, $(cc(\mathcal{Z}), cbc(\mathcal{Z}), kc(\mathcal{Z}))$ denotes the class of nonempty closed (closed convex, closed bounded convex, compact convex) subsets of $\mathcal{Z}$.

Let $(\Sigma, \mathcal{P})$ be an arbitrary measurable space and X a Polish space. A multifunction $F : \Sigma \longrightarrow 2^X \setminus \emptyset$ is said to be measurable (weakly measurable) if for every closed (open) set $C \subset X$ the set $F^-(C) \equiv \{\sigma \in \Sigma : F(\sigma) \cap C \neq \emptyset\} \in \mathcal{P}$.

Let X, Y be any two topological spaces and $F : X \mapsto c(Y)$ is a multifunction. F is said to be upper (lower) semicontinuous with respect to inclusion if for every $x_0 \in X$ and every open set $V \subset Y$ satisfying $F(x_0) \subset V$ $(V \cap F(x_0) \neq \emptyset)$, there exists an open set $U \subset X$ containing x_0 such that $F(x) \subset V$ $(F(x) \cap V \neq \emptyset)$ for all $x \in U$. If Y is a metric space with metric d, then one can introduce a metric d_H, called the Hausdorff metric, on $c(Y)$ as follows:

$$
d_H(C, D) \equiv \operatorname{Max}\{\operatorname{Sup}_{y \in D} d(C, y), \operatorname{Sup}_{z \in C} d(z, D)\}
$$

for $C, D \in c(Y)$. If (Y, d) is complete, then so is $(c(Y), d_H)$. $F : X \mapsto c(Y)$ is said to be continuous in the Hausdorff metric if, whenever $x_n \longrightarrow x$ in the topology of X,

$$\mathrm{Lim}_{n\to\infty} d_H(F(x_n), F(x)) = 0,$$

and it is said to be mild or quasi-upper semicontinuous if

$$\mathrm{Lim}_{n\to\infty} d^*(F(x_n), F(x)) \equiv \mathrm{Sup}\{d(z, F(x)), z \in F(x_n)\} \longrightarrow 0.$$

For other types of continuity see [20, 21].

Let $(\Omega, \mathcal{F}, \mathcal{F}_t \uparrow\subset \mathcal{F}, t \geq 0, P)$ denote a complete probability space furnished with an increasing family of right continuous complete sub σ-algebras $\mathcal{F}_t \subset \mathcal{F}$. All random processes considered in the paper will be assumed to be strongly $\mathcal{F}_t$-predictable processes unless stated otherwise.

3. Basic assumptions. For study of the control problem we shall make critical use of the Da Prato–Zabczyk semigroup [1] which is an extension of the Markov transition operator corresponding to the stochastic evolution equation

$$\begin{aligned} dx &= Axdt + F(x)dt + \sqrt{Q}dW, \\ x(0) &= x_0. \end{aligned} \tag{3.1}$$

For this we need the following assumptions:

(H1)

(a) A is the infinitesimal generator of a C_0-semigroup, $T(t), t \geq 0$ in H satisfying

$$\| T(t) \|_{\mathcal{L}(H)} \leq Me^{-\omega t}, \qquad t \geq 0, \omega > 0, M \geq 1.$$

(b) Q is a positive, symmetric, bounded operator in H so that the operator Q_t given by

$$Q_t x \equiv \int_0^t T(s)QT^*(s)xds, \qquad x \in H, t \geq 0$$

is nuclear for all $t \geq 0$ and $\mathrm{Sup}_{t\geq 0} Tr\, Q_t < \infty$.

(c) W is a cylindrical Wiener process with values in H with $\mathrm{Cov} W(1) = I$.

(H2) F is a bounded Lipschitz mapping from H to H.

(H3) For all $t \geq 0$, $\mathrm{Range}(T(t)) \subset \mathrm{Range}(Q_t^{1/2})$.

(H4) The operator-valued function $\Gamma(t) \equiv (Q_t^{-1/2}T(t)), t \geq 0$, is Laplace transformable.

Let $D\phi$ and $D^2\phi$ denote the first and the second Fréchet derivatives of the function $\phi : H \longrightarrow R^1$, whenever they exist as elements of H and $\mathcal{L}(H)$, respectively. Define the operators $\mathcal{A}_0$ and $\mathcal{A}$ by

$$\begin{aligned} \mathcal{A}_0\phi &\equiv (1/2)Tr(QD^2\phi) + (x, A^*D\phi), x \in H, \\ D(\mathcal{A}_0) &\equiv \{\phi \in C_b^2(H) : D^2\phi \in \mathcal{L}_1(H),\ \mathrm{Sup}_{x\in H} \| D^2\phi \|_{\mathcal{L}_1(H)} < \infty \\ &\qquad \text{and there exists } \psi \in C_b^2(H) : \phi(x) = \psi(A^{-1}x), x \in H\}, \\ \mathcal{B}\phi &\equiv \langle F(\cdot), D\phi(\cdot)\rangle, \phi \in W^{1,2}(H, \mu^0), \text{ and} \\ \mathcal{A} &\equiv (\bar{\mathcal{A}}_0 + \mathcal{B}), D(\mathcal{A}) = D(\bar{\mathcal{A}}_0), \end{aligned}$$

where $\mathcal{L}_1(H)$ is the space of nuclear operators in H and $C_b^k(H)$ is the space of bounded k-times Frechet differentiable functions on H. We consider the semigroup $S(t), t \geq 0$, corresponding

to the Kolmogorov operator associated with the nonlinear stochastic evolution equation (1.1). The following result is fundamental.

THEOREM 3.1. *Suppose the assumptions* (H1)–(H4) *hold. Then* (a) *the linear version of* (3.1), *with* $F = 0$, *has a unique invariant Gaussian measure* μ^0 *on* $\mathcal{B}(H)$; (b) *the operator* $\mathcal{A}$ *generates a* C_0*-semigroup of bounded linear operators,* $S(t), t \geq 0$, *in* $L_2(H, \mu^0)$ *and it is the extension of the original Markov transition operator (corresponding to system* (3.1)*) from* $B_b(H)$ *to* $L_2(H, \mu^0)$. *Further,* $D(\mathcal{A}) \subset W^{1,2}(H, \mu^0)$ *and for* $t > 0$, $S(t)$ *is a family of compact operators in* $L_2(H, \mu^0)$.

Proof. See Da Prato and Zabczyk [1, Thm. 2.10].

Consider the controlled system

$$\begin{aligned} dx &= Axdt + F(x)dt + B(x, u(t, y))dt + \sqrt{Q}dW, \qquad x(0) = x_0, \\ dy &= h(x, y)dt + \sigma_0(y)dw^0, \qquad y(0) = 0. \end{aligned} \tag{3.2}$$

To solve the control problem we need the solution of the associated filtering problem. Let $\mathcal{F}_t^y \equiv \sigma\{y(s), s \leq t\}$ denote the smallest σ-algebra generated by the observed process y up to time $t, t \geq 0$. Let $\phi : H \longrightarrow R$ be any continuous bounded function. The filtering problem is to find an $\mathcal{F}_t^y$-measurable process $\{\eta(t), t \geq 0\}$, such that

$$E\{(\eta(t) - \phi(x(t))^2 | \mathcal{F}_t^y\} = \min \quad \text{for all } t \in I \equiv [0, T].$$

It is well known that the best filter is given by

$$\begin{aligned} \eta^0(t) &= E\{\phi(x(t)) | \mathcal{F}_t^y\} \\ &= \int_H \phi(\xi) Q_t^y(d\xi) \equiv Q_t^y(\phi), \end{aligned} \tag{3.3}$$

where

$$Q_t^y(\chi_\Gamma) = P\{x(t) \in \Gamma | \mathcal{F}_t^y\}$$

for $\Gamma \in B(H)$ with $B(H)$ denoting the σ-algebra of Borel subsets of H. This solution suggests that we must find the conditional probability measure Q_t^y which is an $\mathcal{F}_t^y$-adapted (probability) measure-valued stochastic process. Recently it was proved [6] on the basis of the Da Prato–Zabczyk semigroup that Q_t^y satisfies the Kushner equation (in the weak sense) which is a nonlinear stochastic PDE in an infinite-dimensional space.

Let $I \equiv [0, T]$, $T < \infty$, and define $H_d \equiv H \times R^d$ with the obvious scalar product. Let $\mathcal{Z}$ denote a Polish space (a Hausdorff topological space for which there exists a metric, of countable type compatible with the topology, with respect to which it is a complete separable metric space). We introduce the following hypotheses:

(H5) B and $Q^{-1/2}B : H \times \mathcal{Z} \longrightarrow H$ are bounded Borel-measurable maps, Lipschitz in the first variable and continuous in the second.

(H6) $h : H_d \longrightarrow R^d$ is a bounded Lipschitz map and $\sigma_0 : R^d \longrightarrow \mathcal{L}(R^d)$ is a bounded Lipschitz map having bounded inverse.

Consider the canonical space $C(I, H_d) \equiv C$ and let $\mathcal{B}(C)$ denote the σ-algebra of Borel subsets of the topological space C. We shall need the following lemma.

LEMMA 3.2. *Suppose the hypotheses* (H1)–(H6) *hold and let* $u \in \mathcal{U}_{ad}$. *Let* ν^1 *denote the measure induced by the solution process of* (1.1) *on the measurable space* $(C, \mathcal{B}(C))$ *and* ν^0 *the measure induced by the same system with* $B = 0, h = 0$. *Then* ν^1 *is absolutely continuous with respect to* ν^0 *and* $d\nu^1 = q_T d\nu^0$, *where* q *denotes the Radon–Nikodým derivative and is given by*

$$q_t = \operatorname{Exp}\left\{\int_0^t \{(Q^{-1/2}B, dW) - (1/2) \parallel Q^{-1/2}B \parallel_H^2 dt\} + \int_0^t \{\langle \Gamma_0^{-1}h, dy\rangle - (1/2)\langle \Gamma_0^{-1}h, h\rangle dt\}\right\} \tag{3.4}$$

for all $t \in I \equiv [0, T]$, *where* $\Gamma_0 = \sigma_0\sigma_0^*$. *The process* $q_t, t \geq 0$, *is a continuous square integrable* $\mathcal{F}_t$*-martingale and* $E^0(q_t) = 1$ *for all* $t \geq 0$.

Proof. The proof is basically a consequence of the Girsanov theorem extended to Hilbert spaces (see [3]).

Conditioning with respect to the σ-algebra $\mathcal{F}_t^y$, we have

$$Q_t^y(\phi) = E^1(\phi(x(t))|\mathcal{F}_t^y) = \frac{E^0\{\phi(x(t))q_t|\mathcal{F}_t^y\}}{E^0\{q_t|\mathcal{F}_t^y\}} \equiv \frac{\mu_t^y(\phi)}{\mu_t^y(1)}. \tag{3.5}$$

The process μ_t^y is a measure-valued stochastic process possibly taking values from the Banach space of countably additive, bounded, signed measures, $\mathcal{M}_b(H)$. This is the unnormalized measure-valued process. It was shown in [6] that this process satisfies the so-called Zakai equation which is a linear stochastic PDE in an infinite-dimensional space.

THEOREM 3.3. *Under the assumptions* (H1)–(H6), *for any given* $\mathcal{F}_t^y$*-measurable control law* u *taking values from* $\mathcal{Z}$, *the measure-valued process* $\{\mu_t^y, t \geq 0\}$ *satisfies the following stochastic (partial) differential equation in the weak sense:*

$$\begin{aligned} d\mu_t(\phi) &= \mu_t(\mathcal{A}\phi)dt + \mu_t(L_u\phi)dt + \langle \mu_t(\phi h), \Gamma_0^{-1}(y(t))dy(t)\rangle \\ \mu_0(\phi) &= \Pi_0(\phi) \text{ for each test function } \phi \in D(\mathcal{A}), \end{aligned} \tag{3.6}$$

where $\Pi_0 \in \mathcal{M}(H)$ *is the probability law induced by the random variable* x_0 *and*

$$(L_u\phi)(x) \equiv (D\phi(x), B(x, u))_H. \tag{3.7}$$

Proof. Given a fixed $\mathcal{F}_t^y$-adapted control law, the proof is essentially the same as in [6].

Under suitable assumptions, equation (3.6) is equivalent to a stochastic evolution equation in the Hilbert space $L_2(H, \mu^0)$. This is given in the following result.

THEOREM 3.4. *Suppose the hypotheses* (H1)–(H6) *hold and the initial measure* Π_0 *is absolutely continuous with respect to the invariant measure* μ^0 *having a density* $\rho_0 \in L_2(H, \mu^0)$. *Then the Zakai equation* (3.6) *is equivalent to the stochastic evolution equation*

$$\begin{aligned} d\rho(t) &= \mathcal{A}^*\rho(t)dt + L_u^*(\rho(t))dt + G(\rho(t))dy(t), \qquad t \in I, \\ \rho(0) &= \rho_0 \end{aligned} \tag{3.8}$$

in the Hilbert space $L_2(H, \mu^0)$, *where* $G(\rho) \equiv \rho\Gamma_0^{-1}h$.

Proof. For proof see [6].

Later in the discussion we shall discuss the questions of existence and uniqueness of solutions of this equation. The major difficulty here, unlike in [6], where L_u^* does not arise, is that even though L_u is a nice operator mapping $W^{1,2}(H, \mu^0) \longrightarrow L_2(H, \mu^0)$, L_u^* is not.

4. Conversion to a fully observed control problem. In this section we prove that the partially observed control problem (1.1)–(1.2) is equivalent to a fully observed control problem on $L_2(H, \mu^0) \equiv \mathcal{H}$ involving the Zakai equation. Now we shall formally introduce the class of admissible controls.

Admissible Controls. Recall that a topological space $\mathcal{Z}$ is called a Polish space if it is metrizable of countable type and if there exists a metric, compatible with the topology of $\mathcal{Z}$, with respect to which $\mathcal{Z}$ is complete. In other words, embedded in a Polish space there may be more than one complete separable metric space. Let Γ be a Polish space, for example, a

closed subspace of $\mathcal{Z}$, and $\mathcal{M}(\Gamma)$ the space of Radon probability measures. This is also a Polish space. Let $Y \equiv C(I, R^d)$ and $\mathcal{B}(Y)$ denote the Borel σ-algebra on Y and for each $t \in I$, $\mathcal{B}_t(Y)$ denote the family of increasing subσ-algebras of the σ-algebra $\mathcal{B}(Y)$. Let $\mathcal{P} \equiv \mathcal{P}_\Sigma$ denote the σ-algebra of predictable (nonanticipating) subsets of the set $I \times Y \equiv \Sigma$. Let η be a probability measure on $\mathcal{B}(Y)$ and $\hat{\eta}$ the restriction of the product measure $dt \times \eta(dy)$ on the predictable σ-field $\mathcal{P}$. We assume that $\mathcal{P}$ has been completed with respect to the measure $\hat{\eta}$. Let $\mathcal{U}(\Sigma, \mathcal{M}(\Gamma))$ denote the class of functions (equivalence classes) from Σ to $\mathcal{M}(\Gamma)$ which are w^*-measurable with respect to the predictable σ-field $\mathcal{P}$. Let $cc(\mathcal{M}(\Gamma))$ denote the class of nonempty, closed, convex subsets of $\mathcal{M}(\Gamma)$ and $U : \Sigma \longrightarrow cc(\mathcal{M}(\Gamma))$ is a $\mathcal{P}$-measurable (nonanticipative) multifunction. Note that $U(t, y) = U(t, z)$ for all $y, z \in Y$ satisfying $y(s) = z(s)$ for $0 \leq s \leq t$. We take for the admissible controls the set

$$(4.1)\quad \mathcal{U}_{ad} \equiv \{u \in \mathcal{U}(\Sigma, \mathcal{M}(\Gamma)); u(t, y) \in U(t, y), \ \hat{\eta} \text{ almost everywhere (a.e.) on } \Sigma\}.$$

In other words, the admissible controls are given by the $\mathcal{P}$-measurable selections of the multifunction U.

Throughout the rest of the paper we assume that the operator B and the the cost integrand ℓ have the forms

$$(4.2)\quad \begin{aligned} B(x, u) &\equiv \int_\Gamma \tilde{B}(x, \gamma) u(d\gamma) \ \text{ for } x \in H \ \text{ and } \ u \in \mathcal{M}(\Gamma), \\ \ell(t, y, x, u) &\equiv \int_\Gamma \tilde{\ell}(t, y, x, \gamma)\, u(d\gamma) \ \text{ for } t \in I, y \in R^d, x \in H, u \in \mathcal{M}(\Gamma), \end{aligned}$$

where $\tilde{B} : H \times \Gamma \longrightarrow H$ and $\tilde{\ell} : I \times R^d \times H \times \Gamma \longrightarrow R \cup \{+\infty\}$ are generally Borel-measurable maps to be specified later.

Remark. Before we start with the control problem, some discussion on the choice of control space is warranted. The space $\mathcal{M}(\Gamma)$ has three possible choices. If Γ is an arbitrary set (even without any topology), one may choose $\mathcal{M}(\Gamma) \equiv M_{ba}(\Gamma)$, the space of bounded finitely additive (positive) measures, and in this case both $r \mapsto \tilde{B}(x, r)$ and $r \mapsto \tilde{\ell}(t, y, x, r)$ must be bounded H and scalar-valued functions, respectively. In case Γ is a normal topological space, one takes $\mathcal{M}(\Gamma) \equiv M_{rba}(\Gamma)$, the space of regular bounded finitely additive (positive) measures on the field generated by closed subsets of Γ. In this case the maps defined above must be continuous and bounded. In case Γ is a compact topological Hausdorff space, $\mathcal{M}(\Gamma) \equiv M_{rca}(\Gamma)$, regular countably additive (positive) measures on the σ-algebra of Borel subsets of Γ. In this case the maps defined above are merely continuous. For an excellent discussion on this topic see Fattorini [11]; Cutland and Lindstrom [14]. Since we consider Γ to be a Polish space which is clearly a normal topological space, either of the last two spaces are admissible in our case.

Now we are prepared to recast our original partially observed control problem as a fully observed control problem. Consider the stochastic differential equation

$$(4.3)\qquad dy = \sigma_0(y) dw^0, \ y(0) = 0.$$

Let η denote the measure induced by y on the path space Y or any other measure which is absolutely continuous with respect to η.

THEOREM 4.1. *Suppose that the hypotheses* (H1)–(H6) *hold and that the initial measure* Π_0 *is absolutely continuous with respect to the invariant measure* μ^0. *Then the partially observed control problem* (1.1)–(1.2) *is equivalent to the following fully observed control problem on the Hilbert space* $\mathcal{H} \equiv L_2(H, \mu^0)$: *find a control* $u \in \mathcal{U}_{ad}$ *such that*

$$(4.4)\qquad J(u) \equiv \int_Y \int_{I \times H} \{\ell(t, y(t), x, u(t, y)) \rho_t(x)\} \mu^0(dx) dt\, \eta(dy) \Longrightarrow \text{Inf},$$

where ρ *is the solution of the evolution equation*

$$d\rho(t) = \mathcal{A}^*\rho(t)dt + L_u^*\rho(t)dt + G(\rho(t))dy(t), t \geq 0, \rho(0) = \rho_0 \tag{4.5}$$

on the Hilbert space $L_2(H, \mu^0)$ *corresponding to the control law* u.

Proof. Let E^1 and E^0 denote the integrations with respect to the measures ν^1 and ν^0, respectively, on the canonical space $C \equiv C(I, H_d)$. Then in view of Lemma 3.2, we can write

$$\begin{aligned} J(u) &= E^1\left\{\int_I \ell(t, y(t), x(t), u(t, y))dt\right\} \\ &= E^0\left\{\left(\int_I \ell(t, y(t), x(t), u(t, y))dt\right) q_T(x, y, u)\right\} \\ &= E^0 \int_I \{E^0\{\ell(t, y(t), x(t), u(t, y))q_T(x, y, u)|\mathcal{F}_t\}\}dt \\ &= E^0 \int_I \{\ell(t, y(t), x(t), u(t, y))E^0\{q_T(x, y, u)|\mathcal{F}_t\}\}dt \\ &= E^0 \int_I \{\ell(t, y(t), x(t), u(t, y))q_t(x, y, u)\}dt. \end{aligned} \tag{4.6}$$

The first and the second equalities follow from Lemma 3.2 and the definition of the measures ν^1 and ν^0, respectively. The third and the fourth follow from the properties of conditional expectations and the $\mathcal{F}_t$-measurability of $\ell(t, y(t), x(t), u(t, y))$. The fifth follows from the fact that $q_t, t \geq 0$ is an $\mathcal{F}_t$-martingale (see Lemma 3.2). Now conditioning with respect to the σ-algebra $\mathcal{F}_t^y$ and recalling that

$$Q_t^y(\phi) = E^1(\phi(x(t))|\mathcal{F}_t^y) = \frac{E^0\{\phi(x(t))q_t|\mathcal{F}_t^y\}}{E^0\{q_t|\mathcal{F}_t^y\}} \equiv \frac{\mu_t^y(\phi)}{\mu_t^y(1)},$$

we can express $J(u)$ as

$$\begin{aligned} J(u) &= E^0 \int_I E^0\{\ell(t, y(t), x(t), u(t, y))q_t(x, y, u)|\mathcal{F}_t^y\}dt \\ &= E^0 \int_I \{\mu_t^y(\ell(t, y(t), x, u(t, y)))\}dt, \end{aligned} \tag{4.7}$$

where μ is any solution of the Zakai equation (3.6) of Theorem 3.3 corresponding to the control law u. Since by our assumption, Π_0 is absolutely continuous with respect to the invariant measure μ^0, it follows from Theorem 3.3 that $d\mu_t^y = \rho_t^y d\mu^0$, where ρ is the solution of equation (3.8) corresponding to the control u. Hence it follows from (4.7) that the cost-functional J can be written as

$$J(u) = E^0 \int_{I\times H} \{\ell(t, y(t), x, u(t, y))\rho_t^y(x)\}\mu^0(dx)dt. \tag{4.8}$$

Under the measure ν^0 the process x is independent of the process y and y is governed by the stochastic differential equation $dy = \sigma_0(y)dw^0$. Using the measure η as discussed earlier and Fubini's theorem, equation (4.8) can be rewritten as

$$\begin{aligned} J(u) &= \int_Y \left(\int_{I\times H} \{\ell(t, y, x, u(t, y))\rho_t^y(x)\}\mu^0(dx)dt\right)\eta(dy) \\ &= \int_\Sigma \left(\int_H \{\ell(t, y, x, u(t, y))\rho_t^y(x)\}\mu^0(dx)\right)\hat{\eta}(dtdy). \end{aligned} \tag{4.9}$$

The converse is easy and can be proved by reversing the above arguments. This completes the proof.

In the following section we shall prove that this control problem has a solution. Throughout the rest of the paper we assume that the measure $\hat{\eta}$ of (4.1) is constructed from the measure η induced by the process y as in the previous theorem or any other measure absolutely continuous with respect to η.

5. Existence of optimal controls. In this section we shall prove the existence of optimal relaxed feedback controls under the assumption that the operator A is merely the infinitesimal generator of a C_0-semigroup of negative type in H as given in (H1)(a). For this we must establish the existence of a solution of equation (4.5) under some general conditions. For convenience of notation we shall denote the Hilbert and Sobolev spaces introduced by Da Prato–Zabczyk [1] as follows:

Let $\mathcal{H} \equiv L_2(H, \mu^0)$, $V \equiv W^{1,2}(H, \mu^0)$, and let V^* denote the topological dual of V. Since the embedding $V \hookrightarrow \mathcal{H}$ is continuous and dense, identifying $\mathcal{H}$ with its own dual we obtain the Gelfand triple

$$V \hookrightarrow \mathcal{H} \hookrightarrow V^*,$$

where the embeddings are actually compact. Let $\langle \cdot, \cdot \rangle_{V^*,V}$ denote the duality pairing of elements of V^* with those of V, and $(\cdot, \cdot)_{\mathcal{H}}$ the scalar product in $\mathcal{H}$. Clearly, for $\xi, \zeta \in \mathcal{H}$,

$$(\xi, \zeta)_{\mathcal{H}} \equiv \int_H \xi(x)\bar{\zeta}(x)\mu^0(dx).$$

Note that for $\xi \in \mathcal{H}$ and $v \in V$, $\langle \xi, v \rangle_{V^*,V} = (\xi, v)_{\mathcal{H}}$. Let $L_2^e(I, \mathcal{H}) \equiv L_2^e(\mathcal{H})$ and $L_2^e(I, V^*) \equiv L_2^e(V^*)$ denote the Banach spaces of $\mathcal{F}_t^y$-predictable $\mathcal{H}$ and V^*-valued processes with respective norm topologies given by

$$\begin{aligned} \| \lambda \|_{L_2^e(\mathcal{H})} &\equiv \left(E \int_I \| \lambda(t) \|_{\mathcal{H}}^2 \, dt \right)^{1/2} \quad \text{for } \lambda \in L_2^e(\mathcal{H}), \\ \| \beta \|_{L_2^e(V^*)} &\equiv \left(E \int_I \| \beta(t) \|_{V^*}^2 \, dt \right)^{1/2} \quad \text{for } \beta \in L_2^e(V^*). \end{aligned} \tag{5.1}$$

Let $M_2(\mathcal{H})$ denote the class of $\mathcal{F}_t^y$-predictable $\mathcal{H}$-valued processes furnished with norm topology given by

$$\| \lambda \|_{M_2} \equiv \operatorname{Sup}\{(E \| \lambda(t) \|_{\mathcal{H}}^2)^{1/2}, t \in I\}.$$

It is easy to verify that M_2 is a Banach space.

Let us first consider the system

$$d\rho(t) = \mathcal{A}^*\rho(t)dt + \beta(t)dt + G(\rho(t))dy(t), \qquad t \geq 0, \rho(0) = \rho_0, \tag{5.2}$$

where $\beta \in L_2^e(V^*)$. We need the following lemma.

LEMMA 5.1. *Suppose the assumptions* (H1)–(H6) *hold and, in addition, with reference to assumption* (H4), *there exists a constant* $c > 0$ *and* $\alpha \in (0, 1/2)$ *so that*

$$\| \Gamma(t) \|_{\mathcal{L}(H)} \leq c/t^{\alpha} \quad \text{for } t > 0. \tag{5.3}$$

Consider the system (5.2) *with initial state* $\rho_0 \in \mathcal{H}$. *Then for every* $\beta \in L_2^e(V^*)$, *equation* (5.2) *has a unique mild solution* $\rho \in M_2(\mathcal{H})$. *Further, the solution map* $\beta \mapsto \rho$ *denoted by* $\rho = \Phi(\beta)$, *with values*

$$\rho(t) \equiv \Phi_t(\beta), \qquad t \in I$$

is weakly continuous from $L_2^e(V^*)$ *to* $M_2(\mathcal{H})$, *and* $\rho \in C(I, \mathcal{H})$ P *almost surely (a.s.).*

Proof. Using the dual semigroup $S^*(t), t \geq 0$ of the Da Prato–Zabczyk semigroup $S(t), t \geq 0$ corresponding to the infinitesimal generator $\mathcal{A}$ as defined here, we can write the evolution equation as an integral equation:

$$\rho(t) = S^*(t)\rho_0 + \int_0^t S^*(t-\tau)\beta(\tau)d\tau + \int_0^t S^*(t-\tau)G(\rho(\tau))dy(\tau). \tag{5.4}$$

We prove that this equation has a unique solution having the properties as stated in the theorem. Define

$$z(t) \equiv S^*(t)\rho_0 + \int_0^t S^*(t-\tau)\beta(\tau)d\tau, \qquad t \geq 0. \tag{5.5}$$

It is clear that the first term belongs to $M_2(\mathcal{H})$. For the second term, it follows from Theorem 2.10 of Da Prato–Zabczyk [1] that for $t > 0$, the semigroup $S(t) : \mathcal{H} \longrightarrow V$ is a bounded linear map. Hence for $t > 0$, the adjoint semigroup $S^*(t)$ is also a bounded linear map from V^* to $\mathcal{H}$. Further, by virtue of the assumption on the operator $\Gamma(t), t > 0$ (see (5.3)), it follows from some computation, using the perturbation series $S(t) \equiv \sum_{n\geq 0} S_n(t)$ as in [1, Prop. 2.8, Thm. 2.10], that

$$\int_I \| S^*(t) \|^2_{\mathcal{L}(V^*,\mathcal{H})}\, dt \leq c_2 < \infty,$$

where c_2 is a constant depending on c, α, and T. This justifies the following estimate:

$$\mathrm{Sup}\{E \| z(t) \|^2_{\mathcal{H}}, t \in I\} \leq 2M^2 \| \rho_0 \|^2_{\mathcal{H}} + 2c_2 \| \beta \|^2_{L_2^e(V^*)}, \tag{5.6}$$

where $M \equiv \mathrm{Sup}\{\| S(t) \|_{\mathcal{L}(\mathcal{H})}, t \in I\}$. Hence $z \in M_2(\mathcal{H})$. Using this fact and the uniform bound of $\Gamma_0^{-1}h$ over $H \times R^d$, one can establish that the right-hand expression of equation (5.4) defines an operator Ψ whose mth iterate, for m sufficiently large, is a contraction in $M_2(\mathcal{H})$ and hence it has a unique fixed point, thereby proving that equation (5.4) has a unique solution. Thus the solution map Φ is well defined on $L_2^e(V^*)$. For continuity, it is easy to verify that for any pair of data $\gamma, \beta \in L_2^e(V^*)$, the solutions $\Phi(\gamma), \Phi(\beta)$ satisfy the following inequality:

$$\mathrm{Sup}\{E(\| \Phi_t(\gamma) - \Phi_t(\beta) \|^2_{\mathcal{H}}), t \in I\} \leq K \| \gamma - \beta \|^2_{L_2^e(V^*)}, \tag{5.7}$$

where $K > 0$ is a suitable constant depending only on M, T, $\mathrm{Sup} \| \Gamma_0^{-1}h \|$. Thus Φ is a continuous map from $L_2^e(V^*)$ to $M_2(\mathcal{H})$. Since Φ is an affine map, this implies that it is also weakly continuous. In particular, Φ is also weakly continuous from $L_2^e(V^*)$ to $L_2^e(\mathcal{H})$. For almost sure continuity of the trajectories, we use the C_0-property of the semigroup S^* and the Lebesgue-dominated convergence theorem and the fact that $\int_0^t S^*(t-s)\beta(s)ds \in \mathcal{H}$, P a.s. for each $\beta \in L_2^e(V^*), t \in I$. This completes the proof of the Lemma.

Remark. If the assumption on α in Lemma 5.1 is relaxed by $\alpha \in (0, 1)$ then z, given by

$$z(t) \equiv \int_0^t S^*(t-s)\beta(s)ds, t \in I, \beta \in L_2^e(V^*),$$

only belongs to $L_2^e(\mathcal{H})$ and in this case we can only prove that equation (5.2) has a unique solution in $L_2^e(\mathcal{H})$ and the solution map $\beta \longrightarrow \Phi(\beta)$ is weakly continuous from $L_2^e(V^*)$ to $L_2^e(\mathcal{H})$. Assumption (5.3) is not required for weak solutions [28].

In the next theorem we present an existence result for the system (4.5) and prove that the family of solutions

$$\Xi \equiv \{\rho^u, u \in \mathcal{U}_{ad}\}$$

of the system (4.5) corresponding to the set of admissible controls $\mathcal{U}_{ad}$ is a bounded subset of $M_2(\mathcal{H})$.

THEOREM 5.2. *Under the assumptions of Lemma* 5.1, *for each initial state* $\rho_0 \in \mathcal{H}$ *and admissible control* $u \in \mathcal{U}_{ad}$, *the system* (4.5) *has a unique mild solution* $\rho \in M_2(\mathcal{H})$. *Further, the solution family* Ξ *corresponding to the set of admissible controls* $\mathcal{U}_{ad}$ *is a bounded subset of* $M_2(\mathcal{H})$.

Proof. Under the hypothesis (H5), for each $v \in M(\Gamma)$, the operator L_v given by

$$(L_v\phi)(x) \equiv (D\phi(x), B(x, v))_H \equiv \int_\Gamma ((D\phi)(x), \tilde{B}(x, \gamma))_H \, v(d\gamma), \qquad x \in H$$

is a bounded linear operator from V to $\mathcal{H}$. Indeed there exists a constant b independent of v such that

$$\| L_v\phi \|_{\mathcal{H}} \leq b \| D\phi \|_{\mathcal{H}} \leq b \| \phi \|_V \quad \text{for all } \phi \in V \text{ and } v \in M(\Gamma).$$

Hence the dual operator L_v^* is also a bounded linear operator from $\mathcal{H}$ to V^*. Indeed, for all $v \in M(\Gamma)$, we have

$$|\langle L_v^*\lambda, \phi\rangle_{V^*,V}| = |\langle \lambda, L_v\phi\rangle_{\mathcal{H}}| \leq b \| \lambda \|_{\mathcal{H}} \| \phi \|_V \quad \text{for all } \phi \in V \text{ and } \lambda \in \mathcal{H}.$$

Hence, for each $u \in \mathcal{U}_{ad}$, it follows from measurability of u and continuity of B that for each $\lambda \in L_2^e(\mathcal{H})$, $\beta \equiv L_u^*\lambda$ is weakly measurable, and hence by virtue of separability of the Gelfand triple, it is strongly measurable. Thus it follows from the above inequality that $\beta \equiv L_u^*\lambda \in L_2^e(V^*)$ and $\| L_u^*\lambda \|_{L_2^e(V^*)} \leq b \| \lambda \|_{L_2^e(\mathcal{H})}$. Thus the existence and uniqueness of a solution $\rho^u \in M_2$ corresponding to any given admissible control u follows from application of Lemma 5.1. Indeed using the map Φ as introduced in Lemma 5.1, one can easily verify that the question of existence and uniqueness of solution of equation (4.5) is equivalent to the question of existence of a unique fixed point of the composition operator $\Psi \equiv \Phi o L_u^*$. By repeated iteration one can verify that the nth iterate of Ψ, for n large enough, is a contraction in the Banach space M_2 and hence the existence and uniqueness follows. For boundedness of the solution family Ξ, one can verify using equation (5.4) and replacing β by $L_u^*(\rho)$ that

$$E\big(\| \rho^u(t) \|_{\mathcal{H}}^2\big) \leq 9 \left\{ M^2 \| \rho_0 \|_{\mathcal{H}}^2 + (b^2c_2 + c_3) \int_0^t E(\| \rho^u(s) \|_{\mathcal{H}}^2) ds \right\}, \quad t \in I, \tag{5.8}$$

where

$$b \equiv \operatorname{Sup}\{\| B(x, v) \|, x \in H, v \in \mathcal{M}(\Gamma)\},$$

$$c_2 \equiv \int_I \| S^*(t) \|_{\mathcal{L}(V^*,\mathcal{H})}^2 \, dt,$$

$$c_3 \equiv \operatorname{Sup}\{|\sigma_0^{-1}(z)h(x, z)|_{R^d}^2, x \in H, z \in R^d\}.$$

The inequality (5.8) holds uniformly with respect to $u \in \mathcal{U}_{ad}$. Hence by Gronwall's lemma, it follows from (5.8) that there exists a constant $c_4 > 0$ such that

$$\operatorname{Sup}\{E \| \rho^u(t) \|_{\mathcal{H}}^2, t \in I, u \in \mathcal{U}_{ad}\} \leq c_4 < \infty.$$

In other words Ξ is a bounded subset of M_2. This completes the proof. See also [28] for weak solutions.

Now we are prepared to consider the question of existence of optimal controls. Define for $t \in I$, $y \in Y$, $\lambda \in \mathcal{H}$, $v \in \mathcal{M}(\Gamma)$,

$$\begin{aligned} \hat{\ell}(t, y, \lambda, v) &\equiv \int_H \ell(t, y, x, v)\lambda(x)\mu^0(dx) \\ &\equiv \int_H \int_\Gamma \tilde{\ell}(t, y, x, \gamma)v(d\gamma)\lambda(x)\mu^0(dx). \end{aligned} \tag{5.9}$$

Thus the control problem (4.4)–(4.5), as stated in Theorem 4.1, is equivalent to the following: find a control $u \in \mathcal{U}_{ad}$ such that

$$J(u) \equiv \int_{\Sigma} \{\hat{\ell}(t, y, \rho_t, u(t, y))\}\hat{\eta}(dt, dy) \Longrightarrow \text{Inf}, \tag{5.10}$$

where ρ is the solution of the evolution equation

$$d\rho(t) = \mathcal{A}^*\rho(t)dt + L_u^*\rho(t)dt + G(\rho(t))dy(t), \qquad t \geq 0, \rho(0) = \rho_0. \tag{5.11}$$

For an arbitrary $v \in \mathcal{M}(\Gamma)$, recall the definition of the operator L_v and its dual L_v^*:

$$L_v\phi \equiv (D\phi(\cdot), B(\cdot, v))_H \equiv \int_{\Gamma} (D\phi(\cdot), \tilde{B}(\cdot, r))_H \, v(dr) \text{ for } \phi \in V,$$

$$L_v^*\lambda \equiv \beta \text{ for } \lambda \in \mathcal{H} \text{ such that } \langle \beta, \phi \rangle_{V^*,V} = (\lambda, L_v\phi)_{\mathcal{H}} \text{ for all } \phi \in V.$$

Define the set-valued map $\mathcal{Q} : \Sigma \times \mathcal{H} \longrightarrow 2^{R \times V^*} - \emptyset$ as follows:

(5.12)

$$\mathcal{Q}(t, y, \lambda) \equiv \{(r, \beta) \in R \times V^* : r \geq \hat{\ell}(t, y, \lambda, v) \text{ and } \beta = L_v^*\lambda \text{ for some } v \in U(t, y)\}.$$

For any $\lambda^0 \in \mathcal{H}$, let $N_\epsilon(\lambda^0)$ denote the ϵ-neighborhood of λ^0 in $\mathcal{H}$. The multifunction $\mathcal{Q}$ is said to satisfy the weak Cesari property on $\Sigma \times \mathcal{H}$ if, for each $\lambda^0 \in \mathcal{H}$,

$$\bigcap_{\epsilon > 0} C\ell Co\, \mathcal{Q}(t, y, N_\epsilon(\lambda^0)) \subset \mathcal{Q}(t, y, \lambda^0), \quad \text{for } (t, y) \in \Sigma.$$

If a multifunction satisfies the Cesari property then it must necessarily be closed convex-valued. On the other hand a closed, convex-valued, Hausdorff continuous multifunction always satisfies the Cesari property. In fact, the Cesari property holds for upper semicontinuous (even less, quasi-upper semicontinuous), closed, convex-valued multifunctions [20, 21]. Our first existence result is given in the following theorem. For its proof we adopt a similar procedure as in [8, 9].

THEOREM 5.3. *Consider the optimal control problem* (5.10)–(5.11) *and suppose the following assumptions hold in addition to the basic hypotheses* (H1)–(H6)*:*

(a1) $U : (\Sigma, \mathcal{P}) \longmapsto cc(\mathcal{M}(\Gamma))$ *is weakly measurable.*

(a2) *The integrand* $\hat{\ell}(t, y, \lambda, v)$ *is* $\mathcal{P}$*-measurable in the first two variables, and continuous in the last two arguments, and further there exists a real number* $\delta \geq 0$ *and an* $h \in L_1(\Sigma, \hat{\eta}; R)$ *such that*

$$\hat{\ell}(t, y, \lambda, v) + \delta \parallel \lambda \parallel_{\mathcal{H}}^2 \geq h(t, y), \quad \hat{\eta}\text{-a.e. } \textit{for all} \quad \lambda \in \mathcal{H} \quad \textit{and } v \in U(t, y).$$

(a3) *The set-valued map* $\mathcal{Q}$ *satisfies the weak Cesari property on* $\Sigma \times \mathcal{H}$*. Then there exists an optimal control for the problem.*

Proof. First note that for $\rho \in L_2^e(\mathcal{H})$ and $u \in \mathcal{U}_{ad}$, the functional

$$\tilde{J}(u, \rho) \equiv \int_{\Sigma} \hat{\ell}(t, y, \rho, u)\hat{\eta}(dt, dy)$$

is a well-defined, extended, real-valued function on $\mathcal{U}_{ad} \times L_2^e(\mathcal{H})$. Let Φ denote the solution map as defined in Lemma 5.1 and

$$\mathcal{D} \equiv \{(u, \rho) \in \mathcal{U}_{ad} \times \Xi : \rho = \Phi(\mathbf{L}_{\mathbf{u}}^*\rho)\}$$

denote the set of admissible control-state pairs. Clearly the cost functional J is the restriction of $\tilde{J}$ to $\mathcal{D}$. Thus it suffices to prove the existence of a pair $(u^0, \rho^0) \in \mathcal{D}$ such that

$$\tilde{J}(u^0, \rho^0) \leq \tilde{J}(u, \rho) \quad \text{for all} \quad (u, \rho) \in \mathcal{D}.$$

Since the solution set Ξ is a bounded subset of $L_2^e(\mathcal{H})$, it follows from assumptions (a1) and (a2) that

$$\text{Inf}\{\tilde{J}(u, \rho), (u, \rho) \in \mathcal{D}\} \equiv m_0 > -\infty.$$

Clearly if $m_0 = +\infty$ there is nothing to prove. So we assume that $m_0 < +\infty$. Let $(u^n, \rho^n) \in \mathcal{D}$ be a minimizing sequence for the functional $\tilde{J}$ restricted to $\mathcal{D}$. That is,

$$\lim_{n\to\infty} \tilde{J}(u^n, \rho^n) = m_0. \tag{5.13}$$

Define $\beta^n \equiv L^*_{u^n}\rho^n$ so that $\rho^n = \Phi(\beta^n)$. Since the operator L^*_u is a bounded operator from $L_2^e(\mathcal{H})$ to $L_2^e(V^*)$ uniformly with respect to $u \in \mathcal{U}_{ad}$, as observed in the proof of Theorem 5.2, $\{\beta^n\}$ is contained in a bounded subset of $L_2^e(V^*)$. Note that $L_2^e(V^*)$ is a reflexive Banach space, in fact, a Hilbert space. Thus there exists a subsequence of the sequence $\{\beta^n\}$, relabeled as $\{\beta^n\}$, and an element $\beta^0 \in L_2^e(V^*)$ such that

$$\beta^n \xrightarrow{w} \beta^0 \text{ in } L_2^e(V^*). \tag{5.14}$$

Clearly by virtue of weak continuity of the solution map Φ (see Lemma 5.1), we also have

$$\rho^n \equiv \Phi(\beta^n) \xrightarrow{w} \Phi(\beta^0) \equiv \rho^0 \text{ in } L_2^e(\mathcal{H}). \tag{5.15}$$

Thus by Mazur's theorem there exists a finite convex combination of $\{\beta^n\}$ that converges strongly to β^0 in $L_2^e(V^*)$. In particular, for each integer k, there exists an integer n_k, a set of integers $\{i = 1, 2, \ldots, m(k)\}$ and a set of nonnegative numbers $\{\alpha_{k,i}, 1 = 1, 2, 3, \ldots, m(k)\}$ satisfying

$$\sum_{i=1}^{m(k)} \alpha_{k,i} = 1 \quad \text{for all } k$$

such that

$$\psi_k \xrightarrow{s} \beta^0 \text{ in } L_2^e(V^*), \tag{5.16}$$

where

$$\psi_k(t, y) \equiv \sum_{i=1}^{m(k)} \alpha_{k,i}\, \beta^{n_k+i}(t, y), \quad (t, y) \in \Sigma.$$

Corresponding to the above sequence, define the sequence $\{C_k \equiv C_k(t, y), (t, y) \in \Sigma\}$ as follows:

$$\begin{aligned} C_{n_k+i}(t, y) &\equiv \hat{\ell}\big(t, y, \Phi(\beta^{n_k+i})(t, y), u^{n_k+i}(t, y)\big), \\ C_k(t, y) &\equiv \textstyle\sum_{i=1}^{m(k)} \alpha_{k,i}\, C_{n_k+i}(t, y). \end{aligned} \tag{5.17}$$

Define

$$C_0(t, y) \equiv \liminf_{k\to\infty} C_k(t, y), \qquad (t, y) \in \Sigma. \tag{5.18}$$

Using assumption (a2) and the boundedness of the set Ξ, it is easy to verify that $\liminf C_k(t, y)$ is well-defined $\hat{\eta}$-a.e. on Σ. Therefore by Fatou's lemma we have

$$\int_\Sigma C_0(t, y)\hat{\eta}(dt, dy) \leq \liminf_{k\to\infty} \int_\Sigma C_k(t, y)\hat{\eta}(dt, dy). \tag{5.19}$$

Clearly by virtue of (5.13)

$$\tilde{J}(u^{n_k+i}, \rho^{n_k+i}) \overset{k\to\infty}{\longrightarrow} m_0,$$

and hence it follows from (5.17) that

$$\lim_{k\to\infty} \int_\Sigma C_k(t, y)\hat{\eta}(dt, dy) = m_0. \tag{5.20}$$

Thus from (5.19) and (5.20) we obtain

$$\int_\Sigma C_0(t, y)\hat{\eta}(dt, dy) \leq m_0. \tag{5.21}$$

On the other hand it follows from our assumption (a2) and the boundedness of the set Ξ that there exists an $\tilde{h} \in L_1(\Sigma, \hat{\eta}; R)$ dependent on h such that $C_0(t, y) \geq \tilde{h}(t, y)$, $\hat{\eta}$-a.e. Using this fact along with (5.21) we have $C_0 \in L_1(\Sigma, \hat{\eta}; R)$. Now we show that

$$(C_0(t, y), \beta^0(t, y)) \in \mathcal{Q}(t, y, \Phi(\beta^0)) \equiv \mathcal{Q}(t, y, \rho^0(t, y)), \quad \hat{\eta}\text{-a.e.}$$

Define

$$N_1 \equiv \{(t, y) \in \Sigma : |C_0(t, y)| < \infty \text{ and } \lim_{k\to\infty} \| \psi_k(t, y) - \beta^0(t, y) \|_{V^*} = 0\},$$
$$M_k \equiv \{(t, y) \in \Sigma : u^k(t, y) \notin U(t, y)\} \text{ and } N_0 \equiv \bigcup_{k\geq 1} M_k.$$

Since $C_0 \in L_1(\Sigma, \hat{\eta}; R)$ and $\psi_k \overset{s}{\longrightarrow} \beta^0$ we have $\hat{\eta}(\Sigma \setminus N_1) = 0$. Set $N_2 \equiv (\Sigma \setminus N_0)$ and $N_3 \equiv N_1 \cap N_2$. By the definition of admissible controls, $\hat{\eta}(N_0) = 0$. Thus $\hat{\eta}(\Sigma \setminus N_3) = 0$. In other words $\hat{\eta}(N_3) = \hat{\eta}(\Sigma)$. Define $\Sigma_0 \equiv N_3 \bigcap \{(t, y) \in \Sigma : t \neq \{0, T\}\}$. Therefore, for $(t, y) \in \Sigma_0$, there exists a subsequence, possibly dependent on (t, y), of the sequence $\{C_k\}$, again denoted by $\{C_k\}$, such that

$$C_k(t, y) \longrightarrow C_0(t, y) \quad \text{for } (t, y) \in \Sigma_0. \tag{5.22}$$

Choosing the corresponding subsequence for the sequence $\{\psi_k\}$, we have

$$\psi_k(t, y) \overset{s}{\longrightarrow} \beta^0(t, y) \text{ in } V^* \quad \text{for } (t, y) \in \Sigma_0. \tag{5.23}$$

Thus for every $(t, y) \in \Sigma_0$ and $\epsilon > 0$, there exists an integer $\tilde{k} \equiv \tilde{k}(t, y)$ such that, for $k > \tilde{k}$,

$$\Phi(\beta^{n_k+i})(t, y) \in N_\epsilon(\Phi(\beta^0)(t, y)),$$

where $N_\epsilon(\lambda)$ denotes the ϵ-neighborhood of any point $\lambda \in \mathcal{H}$. Clearly

$$\mathcal{Q}(t, y, \Phi(\beta^{n_k+i})) \subset \mathcal{Q}(t, y, N_\epsilon(\Phi(\beta^0))) = \mathcal{Q}(t, y, N_\epsilon(\rho^0)) \quad \text{for } k > \tilde{k} \text{ and } (t, y) \in \Sigma_0.$$

It follows from the definition of $\mathcal{Q}$ that

$$(C_{n_k+i}(t, y), \beta^{n_k+i}(t, y)) \in \mathcal{Q}(t, y, \Phi(\beta^{n_k+1})(t, y)).$$

Thus for $k > \tilde{k}$, using (5.16)–(5.17), we obtain

$$(C_k(t, y), \psi_k(t, y)) \in Co\ \mathcal{Q}(t, y, N_\epsilon(\Phi(\beta^0)(t, y)) \quad \text{for } (t, y) \in \Sigma_0.$$

Using these facts, it follows from (5.22) and (5.23) that for $(t, y) \in \Sigma_0$,

$$(C_0(t, y), \beta^0(t, y)) \in C\ell Co\ \mathcal{Q}\ (t, y, N_\epsilon(\Phi(\beta^0)(t, y)))$$

for every $\epsilon > 0$, and hence

$$(C_0(t, y), \beta^0(t, y)) \in \bigcap_{\epsilon>0} C\ell Co\ \mathcal{Q}(t, y, N_\epsilon(\Phi(\beta^0)(t, y)).$$

Therefore, by the weak Cesari property (a3),

$$(C_0(t, y), \beta^0(t, y)) \in \mathcal{Q}(t, y, \Phi(\beta^0)(t, y)) = \mathcal{Q}(t, y, \rho^0(t, y)) \text{ for all } (t, y) \in \Sigma_0,$$

and hence $\hat{\eta}$-a.e. on Σ. This implies that for every $(t, y) \in \Sigma_0$ there exists $\tilde{u}(t, y) \in U(t, y)$ such that

$$C_0(t, y) \geq \hat{\ell}(t, y, \rho^0(t, y), \tilde{u}(t, y)) \text{ for } (t, y) \in \Sigma_0,$$
$$\beta^0(t, y) = (L^*_{\tilde{u}}\, \rho^0)(t, y) \text{ for } (t, y) \in \Sigma_0.$$

Since $\hat{\eta}(\Sigma_0) = \hat{\eta}(\Sigma)$, we have

$$C_0(t, y) \geq \hat{\ell}(t, y, \rho^0(t, y), \tilde{u}(t, y)), \ \hat{\eta}\text{-a.e. on } \Sigma,$$
$$\beta^0(t, y) = (L^*_{\tilde{u}}\, \rho^0)(t, y) \ \hat{\eta}\text{-a.e. on } \Sigma.$$

In view of (5.21), the question that remains to be settled is whether or not a $\mathcal{P}$-measurable substitute for $\tilde{u}$ can be found. We prove this using the theory of measurable selections. Define for $(t, y) \in \Sigma_0$, the set-valued map

$$\Lambda(t, y) \equiv \left\{v \in U(t, y) : C_0(t, y) \geq \hat{\ell}(t, y, \rho^0(t, y), v) \text{ and } \beta^0(t, y) = (L^*_v(\rho^0))(t, y)\right\}.$$

Clearly this set is nonempty. We prove that it has a $\mathcal{P}$-measurable selection. A most general result in this direction states that a weakly measurable set-valued function (for definition see §2) with closed values, from an arbitrary measurable space to a Polish space, has measurable selection [10, Thm. 4.1, pp. 867]. Since $\mathcal{M}(\Gamma)$ is a Polish space, it suffices to verify that Λ is closed valued and weakly measurable. To prove closedness, let $v_n \in \Lambda(t, y)$ and suppose $v_n \xrightarrow{w^*} v_0$ in $\mathcal{M}(\Gamma)$. Since $U(t, y) \in cc(\mathcal{M}(\Gamma))$ we have $v_0 \in U(t, y)$. Further, it follows from continuity of $\nu \longrightarrow \hat{\ell}(t, y, \lambda, \nu)$ that

$$\hat{\ell}(t, y, \rho^0(t, y), v_n) \longrightarrow \hat{\ell}(t, y, \rho^0(t, y), v_0), \ \hat{\eta}\text{-a.e.}$$

Similarly for any $\phi \in V$, we have

$$\begin{aligned}\langle \beta^0(t, y), \phi\rangle_{V^*,V} &= \langle L^*_{v_n} \rho^0(t, y), \phi\rangle_{V^*,V} \\ &= (\rho^0(t, y), L_{v_n}\phi)_{\mathcal{H}} \longrightarrow (\rho^0(t, y), L_{v_0}\phi)_{\mathcal{H}} \\ &= \langle L^*_{v_0}\rho^0, \phi\rangle_{V^*,V}\ \hat{\eta}\text{-a.e.}\end{aligned}$$

Since $\phi \in V$ is arbitrary, this implies that Λ has closed values. In fact $\Lambda : \Sigma \longrightarrow cc(\mathcal{M}(\Gamma))$. Now we prove measurability. For simplicity of notation, we denote $\sigma \equiv (t, y) \in \Sigma$. Define the multifunctions

$$\Lambda_0(\sigma) \equiv \{v \in \mathcal{M}(\Gamma) : \hat{\ell}(\sigma, \rho^0(\sigma), v) - C_0(\sigma) \leq 0\},\ \sigma \in \Sigma,$$
$$\Lambda_1(\sigma) \equiv \{v \in \mathcal{M}(\Gamma) : (L^*_v\rho^0)(\sigma) - \beta^0(\sigma) = 0\},\ \sigma \in \Sigma.$$

Then

$$\Lambda(\sigma) = \big(\Lambda_0(\sigma) \cap U(\sigma)\big) \bigcap \big(\Lambda_1(\sigma) \cap U(\sigma)\big). \tag{5.24}$$

We show that each component is $\mathcal{P}$-measurable. Let $M_0 \subset \mathcal{M}(\Gamma)$ be any closed subset. Since $\mathcal{M}(\Gamma)$ is a Polish space, there exists a countable dense subset M_{00} of M_0 so that we have

$$\begin{aligned} \Lambda_0^-(M_0) &\equiv \{\sigma \in \Sigma : \Lambda_0(\sigma) \cap M_0 \neq \emptyset\} \\ &= \bigcup_{v \in M_{00}} \{\sigma \in \Sigma : \hat{\ell}(\sigma, \rho^0(\sigma), v) - C_0(\sigma) \leq 0\}. \end{aligned} \tag{5.25}$$

Since ρ^0 and C_0 are $\mathcal{P}$-measurable it follows from our assumption (a2) that each of the components in (5.25) is $\mathcal{P}$-measurable and hence $\Lambda_0^-(M_0) \in \mathcal{P}$. Similarly, using the separability of V we can also write

$$\begin{aligned} \Lambda_1^-(M_0) &\equiv \{\sigma \in \Sigma : \Lambda_1(\sigma) \cap M_0 \neq \emptyset\} \\ &= \bigcup_{v \in M_{00}} \bigcap_{\phi \in V_0} \big\{\sigma \in \Sigma : \langle (L_v^* \rho^0)(\sigma), \phi \rangle = \langle \beta^0(\sigma), \phi \rangle\big\}, \end{aligned} \tag{5.26}$$

where V_0 is a countable dense subset of V. This shows that both Λ_0 and Λ_1 are $\mathcal{P}$-measurable. Since for set-valued maps measurability implies weak measurability and by our assumption U is weakly measurable, we conclude that Λ, given by (5.24), is weakly measurable. Thus there exists a $\mathcal{P}$-measurable selection u^* of Λ which is a substitute for $\tilde{u}$. This completes the proof.

Remark. Assumption (5.3) of Lemma 5.1 can be replaced by hypothesis 5 of Da Prato and Zabczyk [1], in which case the operator $-\mathcal{A}$ is coercive with respect to the triple $\{V, \mathcal{H}, V^*\}$. In this case weak solutions can be exploited instead of mild solutions.

6. Optimal control of Zakai inclusion. So far we have considered the map $B(x, u)$ to be single valued. If it is a multivalued map, equation (1.1) turns into a stochastic differential inclusion:

$$\begin{aligned} &dx \in Axdt + F(x)dt + B(x, u(t, y))dt + \sqrt{Q}dW, x(0) = x_0, \\ &dy = h(x, y)dt + \sigma_0(y)dw^0, y(0) = 0. \end{aligned} \tag{6.1}$$

We consider F to be single valued and B a multivalued map. This class of systems may arise from evolution inequalities or from parametric uncertainty of system coefficients [9, 18]. Here we shall use the notion of a solution for differential inclusions as given in Ahmed [8, 9]. Corresponding to an admissible control u, a process x is a mild solution of (6.1) if there exists a predictable process $z \in L_2^e(H)$ such that x is a mild solution of (1.1) with z substituted for B and that $z(t) \in B(x(t), u(t, y))$ for almost all $t \in I$ -P a.s.

We present here an existence result similar to Theorem 5.2. First we translate the above problem into a differential inclusion of Zakai type.

For each $v \in M(\Gamma)$, define

$$\mathcal{B}(v) \equiv \{\text{ all } \mu^0\text{-measurable Borel selections } b : b(x) \in B(x, v), \mu^0 \text{ a.e. on } H\}.$$

We assume that for each $v \in M(\Gamma)$, the set $\mathcal{B}(v)$ is nonempty. A sufficient condition for this is that there exists $\zeta \in L_1(H, \mu^0)$ such that

$$\| B(x, v) \|^0 \equiv \text{ Sup}\{\| \beta \|_H, \beta \in B(x, v)\} \leq \zeta(x), x \in H.$$

Then define the multivalued operator $L_{\mathcal{B}(v)}$ as follows: for each $\phi \in V$,

$$L_{\mathcal{B}(v)}\phi \equiv \{g \in \mathcal{H} : g = (D\phi(\cdot), b(\cdot))_H \equiv L_b\phi \text{ for some } b \in \mathcal{B}(v)\}.$$

The corresponding adjoint family $L^*_{\mathcal{B}(v)}$ is given by

$$L^*_{\mathcal{B}(v)}\zeta \equiv \{\beta \in V^* : \langle \beta, \phi\rangle_{V^*,V} = (\zeta, L_b\phi)_{\mathcal{H}}, \text{ for some } b \in \mathcal{B}(v) \text{ for all } \phi \in V\}, \qquad \zeta \in \mathcal{H}.$$

The equation (5.11) now turns into an inclusion. For any admissible control law u we have the differential inclusion in $\mathcal{H}$:

$$d\lambda \in \mathcal{A}^*\lambda dt + L^*_{\mathcal{B}(u(t,y))}\lambda dt + G(\lambda)dy, \lambda(0) = \rho_0, \qquad t \in I, \tag{6.2}$$

which is associated with the differential inclusion (6.1) in H. An element $\rho \in M_2(\mathcal{H}) \subset L^e_2(\mathcal{H})$ is a mild solution of the evolution inclusion (6.2) if there exists a $\beta \in L^e_2(V^*)$ such that

$$\begin{aligned} &d\rho = \mathcal{A}^*\rho dt + \beta dt + G(\rho)dy, \qquad \rho(0) = \rho_0 \\ &\text{and } \beta(t) \in (L^*_{\mathcal{B}(u)}\rho)(t) \ \hat{\eta}\text{-a.e.} \end{aligned} \tag{6.3}$$

Let $\Phi : \beta \mapsto \rho$ denote the solution map giving $\rho = \Phi(\beta)$ as the solution of the first equation of (6.3) corresponding to any $\beta \in L^e_2(V^*)$. For each $u \in \mathcal{U}_{ad}$, define the multivalued map

$$\hat{C}_u(\beta) \equiv \{\gamma \in L^e_2(V^*) : \gamma(t) \in (L^*_{\mathcal{B}(u(t,y))}\Phi(\beta))(t), \ \hat{\eta}\text{-a.e.}\}.$$

The question of existence of a solution of the evolution inclusion (6.2) is equivalent to the question of existence of a fixed point of the multivalued map $\hat{C}_u$. We present here the following existence result for optimal control of (6.2) and (5.10).

THEOREM 6.1. *Consider the system* (6.2) *along with the cost functional* (5.10). *In addition to the assumptions* (H1)–(H6), *suppose the following assumptions hold:*

(b1) $U : (\Sigma, \mathcal{P}) \mapsto cc(\mathcal{M}(\Gamma))$ *is measurable.*

(b2) $B : H \times \mathcal{M}(\Gamma) \mapsto cbc(H)$ *and there exists a constant* $b_0 < \infty$ *such that* $\| B \|^0 \equiv \text{Sup}\{\| b(x) \|_H, x \in H, b \in \mathcal{B}(v), v \in \mathcal{M}(\Gamma)\} \le b_0$.

(b3) *The cost integrand* $\hat{\ell}$ *satisfies* (a2) *of Theorem* 5.3.

(b4) *The set-valued map* $\mathcal{Q}$ *given by*

$$\mathcal{Q}(t, y, \lambda) \equiv \{(r, \beta) \in R \times V^* : r \ge \hat{\ell}(t, y, \lambda, v) \text{ and } \beta \in L^*_{B(v)}\lambda \text{ for some } v \in U(t, y)\} \tag{6.4}$$

satisfies the weak Cesari property. Then there exists an optimal control for the problem.

Proof. The major part of the proof is identical to that of Theorem 5.3. It is required only to show that the evolution inclusion (6.2) has solutions. This follows if, for every admissible control u, the fixed point set $Fix(\hat{C}_u)$ of the multivalued map $\hat{C}_u$ is nonempty. Under the assumptions on B and U, it can be shown that $\hat{C}_u : L^e_2(V^*) \mapsto cbc(L^e_2(V^*))$. Further $\beta \longrightarrow \hat{C}_u(\beta)$ is upper semicontinuous with respect to inclusion. By virtue of boundedness of the set-valued map B, (see assumption (b2)), there exists a constant $0 < r < \infty$ independent of u such that $\hat{C}_u : D_r \subset L^e_2(V^*) \mapsto D_r$, where D_r is a closed ball of radius r in $L^e_2(V^*)$ with its center at the origin. Hence $\hat{C}_u(\beta) \cap D_r \neq \emptyset$ for all $\beta \in D_r$ and $u \in \mathcal{U}_{ad}$. Since $L^e_2(V^*)$ is a reflexive Banach space endowed with the weak topology, it is a locally convex, topological vector space. Hence it follows from a generalized version of the Kakutani–Fan fixed-point theorem [19, Thm. 9.B, p. 452] that the fixed-point set, $Fix(\hat{C}_u)$ is nonempty for every $u \in \mathcal{U}_{ad}$. Hence, for each $u \in \mathcal{U}_{ad}$, the evolution inclusion (6.2) has solutions. Let

$$\Xi \equiv \{\rho \in L^e_2(\mathcal{H}) : \rho = \Phi(\beta), \text{ for } \beta \in Fix(\hat{C}_u), u \in \mathcal{U}_{ad}\}$$

denote the solution set. Then the set of admissible control-state pairs is given by

$$\mathcal{D} \equiv \{(u, \rho) \in \mathcal{U}_{ad} \times \Xi : \rho = \Phi(\beta), \beta \in \hat{\mathbf{C}}_{\mathbf{u}}(\beta)\}. \tag{6.5}$$

From here on, the rest of the proof is identical to that of Theorem 5.3. This completes the proof.

Remark. In case the evolution inclusion arises from parametric uncertainty, it is natural to consider the min-max problem (see [18]) rather than the minimum problem considered here in Theorem 6.1. One defines the solution set corresponding to a fixed admissible control law u as follows:

$$\mathcal{X}(u) \equiv \{\rho \in L_2^e(\mathcal{H}) : \rho = \Phi(\beta) \text{ for some } \beta \in Fix(\hat{C}_u)\}.$$

The cost functional is then given by

$$J_0(u) \equiv \text{Sup}\{J(u, \rho), \rho \in \mathcal{X}(u)\}.$$

The problem is to find a control law that minimizes this functional. In other words optimal control is the one that minimizes the maximum risk.

Remark. If h of equation (6.1) is multivalued, the operator G of the inclusion (6.2) is also multivalued. This situation may arise if the measurement dynamics also has parametric uncertainties. Since this appears in the diffusion term the problem becomes much more difficult.

7. Necessary condions of optimality. In this section we present a result on the necessary conditions of optimality. This is essentially a stochastic minimum principle. We shall only state the result without proof. In principle the proof is based on arguments similar to those of [22]–[25].

Set $\sigma \equiv (t, y) \in \Sigma$ and define the Hamiltonian

$$H(\sigma, \rho, \phi, v) \equiv \langle \rho, L_v\phi\rangle_{\mathcal{H}} + \hat{\ell}(\sigma, \rho, v) \equiv \int_\Gamma \{\langle \rho, (B(\cdot, \xi), D\phi)\rangle + \bar{\ell}(\sigma, \rho, \xi)\}v(d\xi), \tag{7.1}$$

where $\bar{\ell}(\sigma, \rho, \xi) \equiv \int_H \tilde{\ell}(\sigma, x, \xi)\rho(x)\mu^0(dx)$. Introduce the function

$$g(\sigma, x) \equiv \int_\Gamma \tilde{\ell}(\sigma, x, \xi)u^o(\sigma)(d\xi),$$

where u^0 is the optimal control law, and define the abstract vector-valued function

$$g^o(\sigma) \equiv g(\sigma, \cdot),$$

taking values from $L_1((\Sigma, \hat{\eta}), \mathcal{H})$. The Hamiltonian

$$H : \Sigma \times \mathcal{H} \times V \times M(\Gamma) \longrightarrow R \cup \{+\infty\}.$$

THEOREM 7.1. *Let $\{u^o, \rho^o\} \in \mathcal{U}_{ad} \times L_2^e(\mathcal{H})$. In order that $\{u^o, \rho^o\}$ be an optimal pair, it is necessary that there exists an element $\phi^o \in L_2^e(V)$ so that*

$$\int_\Sigma H(\sigma, \rho^o, \phi^o, u)\hat{\eta}(d\sigma) \geq \int_\Sigma H(\sigma, \rho^o, \phi^o, u^o)\hat{\eta}(d\sigma) \text{ for all } u \in \mathcal{U}_{ad}, \tag{7.2}$$

where the pair $\{\rho^o, \phi^o\}$ are the mild solutions of the following equations:

$$\begin{aligned} d\rho &= \mathcal{A}^*\rho dt + L^*_{u^o}\rho dt + G(\rho)dy, \qquad t \geq 0, \rho(0) = \rho_0, \\ d\phi &= -(\mathcal{A}\phi + L_{u^o}\phi)dt + g^o dt - G(\phi)dy, \qquad \phi(T) = 0. \end{aligned} \tag{7.3}$$

Remark. The last equation in (7.3) is a backward stochastic evolution equation which naturally arises whenever one attempts to develop a stochastic minimum principle as the necessary condition of optimality. The question of existence of solutions of such equations has been considered in several papers. For details see [22]–[25].

Pointwise necessary conditions of optimality can be derived from (7.2) provided certain conditions are met. Let $\mathcal{B}(R^d)$ denote the σ-algebra of Borel subsets of R^d, and set $\Theta(t)(G) \equiv P\{y(t) \in G\}$ for any $G \in \mathcal{B}(R^d)$, where y is the unique solution of the stochastic differential equation $dy = \sigma_0(y)dw^0$, $y(0) = 0$. Define the Young measure $\hat{\Theta}$ on $\mathcal{B}(I \times R^d)$ as follows: for any $K \in \mathcal{B}(I \times R^d)$,

$$\int_K \hat{\Theta}(dtd\zeta) = \int_K \Theta(t)(d\zeta)dt.$$

Let Γ be a compact Polish space and $U : I \times R^d \longrightarrow cc(M(\Gamma))$ be a measurable multifunction. For the admissible controls we take

$$\mathcal{U}_{ad} \equiv \{u : I \times R^d \longrightarrow M(\Gamma) \text{ so that } u \text{ is } w^*\text{-measurable and } u(t, y) \in U(t, y)\ \hat{\Theta}\text{-a.s.}\}.$$

Define the Hamiltonian as follows:

$$H(t, y, \rho, \phi, v) \equiv \langle \rho, L_v\phi \rangle + \hat{\ell}(t, y, \rho, v),\ v \in U(t, y).$$

Note that H maps $I \times R^d \times \mathcal{H} \times V \times M(\Gamma)$ to $R \cup \{+\infty\}$. Then, for the given admissible controls, one can use (7.2) to derive the pointwise necessary condition of optimality given by

$$(7.5)\ H(t, y, \rho^o(t), u, \phi^o(t)) \geq H(t, y, \rho^o(t), u^o(t, y), \phi^o(t)) \quad \text{for all } u \in U(t, y),\ \hat{\Theta}\text{-a.e.}$$

8. Comments on applications.

Ecological problem. Considering the ecological problem, it is not difficult to verify that if $\beta \equiv \sup\{\| v(\xi) \|, \xi \in \Omega\} < \infty$, and $D \equiv \operatorname{diag}(d_1, d_2, \ldots, d_m)$ with $d_k > 0, k = 1, 2, \ldots, m$, then the operator A, as defined in this example, is the infinitesimal generator of an analytic semigroup $T(t), t \geq 0$ in H. If β is sufficiently small then the semigroup is exponentially stable and condition (H1)(a) is satisfied. For this problem, considering the source of noise, the covariance Q is practically nuclear and hence by virtue of exponential stability of the semigroup T, Q_t satisfies (H1)(b). (H1)(c) is an assumption of the model. Hence the existence of a unique invariant measure μ^0 follows. For this example, $F = 0$ and (H2) is trivially satisfied. Assumptions (H3) and (H4) are rather technical and are required for the proof of the main result of Da Prato and Zabczyk [1, Thm. 3.1] on which our result is based. Since the environmental agencies monitoring the system will never permit growth of C beyond a certain predetermined level, for all practical purposes we can assume B to satisfy assumption (H5). For example, we can replace B by $B_r(x, u) \equiv B(P_r(x), u)$, where P_r is the retraction of the ball $S_r \equiv \{x \in H : \| x \| \leq r\}$ defined as

$$P_r(x) \equiv \begin{cases} x & \text{for } x \in S_r, \\ (r/\| x \|)x & \text{otherwise.} \end{cases}$$

One can choose the ceiling r as large as required. With this modification we can admit any function $B(x, u)$ which is locally Lipschitz in $x \in H$ and continuous and bounded in u on closed bounded sets $\Gamma \subset L_\infty(\Omega, R^s_+)$, where R^s_+ denotes the positive orthant of R^s and s is the number of distinct control agents as described in §2. Clearly the function h as defined by

equation (2.3) is bounded and Lipschitz. Since $L_1(\Omega, R^s)$ is separable, any closed bounded set $\Gamma \subset L_\infty(\Omega, R^s_+)$ with the relative w^* topology is a Polish space. Thus with relaxed controls, all the conditions (a1)–(a3) of Theorem 5.3 hold. Conditions (a1) and (a2) are trivial. For (a3), note that continuity of the maps B and the cost integrand ℓ with respect to the state and control, all convexified by relaxed controls, imply that the Cesari map $\mathcal{Q}$ (see equation (5.12)) is continuous in the Hausdorff metric and closed convex-valued and hence satisfies the weak Cesari property.

Electrical problem. Now we consider the electrical "crosstalk" problem. The standard norm for $H \equiv L_2(\Omega, R^m) \times L_2(\Omega, R^m) \times R^m$ is induced by the scalar product

$$(y, z) \equiv (y_1, z_1)_{L_2(\Omega, R^m)} + (y_2, z_2)_{L_2(\Omega, R^m)} + (y_3, z_3)_{R^m}.$$

For the electrical problem, however, it is convenient to use the energy norm induced by the scalar product

$$(y, z)_e \equiv (Ly_1, z_1)_{L_2(\Omega, R^m)} + (Cy_2, z_2)_{L_2(\Omega, R^m)} + (C_1 y_3, z_3)_{R^m}.$$

Since the matrices L, C, C_1 are symmetric and positive definite, the two norms are equivalent. We assume that H is furnished with the energy-related scalar product $(y, z) \equiv (y, z)_e$ and omit the subscript e. It is easy to verify that the operator A (see §2, electrical problem) is closed and densely defined with domain and range in H and that it is strictly m-dissipative. Hence by the Lumer–Phillips theorem (see [26]), A is the infinitesimal generator of a C_0-semigroup, $T(t), t \geq 0$, of contractions in H. Further, by use of the energy norm, it is not difficult to verify that

$$(Ay, y) = -(Ry_1, y_1)_{L_2(\Omega, R^m} - (Gy_2, y_2)_{L_2(\Omega, R^m)} - (R_0 y_1(0), y_1(0))_{R^m} - (Ky_3, y_3)_{R^m}.$$

Since $\{R, G, R_0, K\}$ are all symmetric and positive definite and the two norms are equivalent, there exists an $\alpha > 0$ such that

$$(Ay, y) \leq -\alpha \parallel y \parallel^2 \quad \text{for all } y \in D(A).$$

Hence A generates a C_0-semigroup $T(t), t \geq 0$ of contractions satisfying $\parallel T(t) \parallel \leq e^{-\alpha t}$, for all $t \geq 0$. For this problem, the controls are finite dimensional and we may assume that u takes values from a closed bounded set $\Gamma \subset R^{2m}$, not necessarily convex. Since $\tilde{g}$ is Lipschitz and bounded, the operator $B(x, u)$ is Lipschitz and bounded on $H \times \Gamma$. The function h in equation (2.16) may not be uniformly bounded on H. In fact the voltages induced in the induction coils of the probe may increase with the increase of currents and voltages on the transmission lines and the frequency. But in all practical instrumentation, the measurement devices saturate if overdriven and hence whenever line currents and voltages exceed a certain limit, the induced voltages h_k will saturate. Hence for all practical purposes, h can be replaced by $\tilde{h}_r(x) \equiv h(P_r x)$ for a suitable $0 < r < \infty$, where, again, P_r denotes the retraction map with respect to the ball $S_r \equiv \{x \in H : \parallel x \parallel \leq r\}$. With these modifications all the assumptions of Theorem 5.3 hold with relaxed controls replacing the ordinary controls.

Remark. Even though from practical considerations, the theoretical results developed here are applicable to many applied problems, from a theoretical standpoint the theory is certainly not completely satisfactory. The main limitation arises from the assumption of boundedness of the operator-valued functions B and h. But this limitation of the theory has not been overcome even for finite-dimensional problems. As an example, for unbounded h, no satisfactory existence theorem for the solution of the Zakai equation is known. Thus it remains an open problem to develop theoretical results that allow unbounded operators B and h.

REFERENCES

[1] G. Da Prato and J. Zabczyk, *Regular densities of invariant measures in Hilbert spaces*, J. Func. Anal., 130 (1995), pp. 427–449.

[2] V. Barbu and G. Da Prato, *Hamilton Jacobi Equations in Hilbert spaces*, Pitman Res. Notes Math. Ser. 86, Harlow, UK, 1983.

[3] G. Da Prato and J. Zabczyk, *Stochastic equations in infinite dimension*, Encyclopedia Math. Appl., 44, Cambridge University Press, Cambridge, UK, 1992.

[4] G. Da Prato, *Parabolic Equations in Infinitely Many Variables*, Scuola Normale Superiore, Pisa, 1992, preprint 140.

[5] Q. Zhu and N. U. Ahmed, *Some results on parabolic equations in Banach space*, Nonlinear Anal., 24 (1995), pp. 1305–1319.

[6] N. U. Ahmed and J. Zabczyk, *Nonlinear Filtering for Semilinear Stochastic Differential Equations on Hilbert Spaces*, Institute of Mathematics, Polish Academy of Sciences, Warszawa, Poland, 1994, preprint 522.

[7] N. U. Ahmed, *Relaxed Controls for Stochastic Boundary Value Problems in Infinite Dimension*, Lecture Notes in Control and Inform. Sci. 149, Springer-Verlag, New York, 1990, pp. 1–10.

[8] ———, *Existence of optimal relaxed controls for a class of systems governed by differential inclusions on a Banach space*. 50 (1986), pp. 213–237.

[9] ———, *Optimal relaxed controls for nonlinear infinite dimensional stochastic differential inclusions*, in International Symposium on Optimal Control of Infinite Dimensional Systems, Lecture Notes in Pure and Appl. Math. 180, Marcel Dekker, New York and Basel, 1994, pp. 1–19.

[10] D. H. Wagner, *Survey of measurable selection theorems*, SIAM J. Control Optim., 15 (1977), pp. 859–903.

[11] H. O. Fattorini, *Relaxed controls, differential inclusions, existence theorems, and the maximum principle in nonlinear infinite dimensional control theory*, in Evolution Equations, Control Theory, and Biomathematics (Han sur Lesse, 1991), Lecture Notes in Pure and Appl. Math. 155, Marcel Dekker, New York, 1994, pp. 185–204.

[12] W. H. Fleming and M. Nisio, *On stochastic relaxed control for partially observed diffusions*, Nagoya Math. J., 93 (1984), pp. 71–108.

[13] W. H. Fleming and E. Pardoux, *Optimal control for partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 261–285.

[14] N. J. Cutland and T. Lindstrom, *Random relaxed controls and partially observed stochastic systems*, Acta Appl. Math., 32 (1993), pp. 157–182.

[15] O. Hijab, *Partially observed control of Markov processes* IV, J. Funct. Anal., 109 (1992), pp. 215–256.

[16] A. Bensoussan, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992.

[17] V. S. Borkar, *Existence of optimal controls for partially observed diffusions*, Stochastics, 11 (1983), pp. 103–141.

[18] N. U. Ahmed and X. Xiang, *Admissible relaxation in optimal control problems for infinite dimensional uncertain systems*, J. Appl. Math. Stochastic Anal., 5 (1993), pp. 227–236.

[19] E. Zeidler, *Nonlinear Functional Analysis and its Applications, Vol 1: Fixed Point Theorems*, English ed., Springer-Verlag, New York, Berlin, Heidelberg, 1991.

[20] M. Kisielewicz, *Differential Inclusions and Optimal Control*, PWN-Polish Scientific Publishers, Warsaw, Kluwer Academic Publishers, Dordrecht, Boston, London, 1991.

[21] N. U. Ahmed and K. L. Teo, *Optimal Control of Distributed Parameter Systems*, North–Holland, New York, Oxford, 1981.

[22] N. U. Ahmed, *Stochastic control on Hilbert space for linear evolution equations with random operator-valued coefficients*, SIAM J. Control Optim., 19 (1981), pp. 401–430.

[23] X. Li, *Optimal Control for Infinite Dimensional Systems*, Lecture Notes in Control and Inform. Sci. 159, Springer-Verlag, New York, 1991, pp. 96–105.

[24] S. G. Peng, *A general stochastic maximum principle*, SIAM J. Control Optim., 28 (1990), pp. 966–979.

[25] X. Y. Zhou, *Maximum principle, dynamic programming and their connections in deterministic controls*, J. Optim. Theory Appl., 65 (1990), pp. 363–373.

[26] N. U. Ahmed, *Semigroup Theory with Applications to Systems and Control*, Pitman Res. Notes Math. Ser. 246, Longman, Harlow, UK, 1991.

[27] R. L. Khan and G. I. Costache, *Finite element method applied to modeling crosstalk problems on printed circuit boards*, IEEE Trans. on Electromagnetic Compatibility, 31 (1989), pp. 5–15.

[28] N. U. Ahmed and J. Zabczyk, *Partially Observed Optimal Controls for Nonlinear Infinite Dimensional Stochastic Systems*, 1996, manuscript.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1616–1649, September 1996

008

STABILIZATION BY CONSTRAINED CONTROLS*

GEORGI V. SMIRNOV†

Abstract. A stabilization problem for a general nonlinear control system is considered. In particular the control corresponding to the equilibrium position may belong to the boundary of the control set. A linear control system is considered as a first approximation for the original problem. The right-hand side of the linear system generates a set-valued map of a special type known as a convex process. This set-valued map has a number of properties similar to those of a linear operator. They allow one to establish necessary and sufficient conditions for solvability of the regulator design problem for the first approximation and to construct a Lyapunov function. Based on these results the nonlinear stabilization problem is investigated. Different statements of the regulator design problem are studied. Stabilization problems for some mechanical systems are considered to illustrate the regulator design techniques. The properties of transient characteristics (the "peak" effect) are discussed for a linear stabilization problem under controllability conditions.

Key words. stabilization, Lyapunov function, constrained control, discontinuous right-hand side

AMS subject classifications. 93D15, 93D20, 93D30, 34A60, 93C05, 93C10

Introduction.

Consider a control system

$$\dot{x} = f(x, u), \ u \in U, \tag{1}$$

where U is the control set. Assume that there exists $u_0 \in U$ such that $f(0, u_0) = 0$. In this case we say that $x = 0$ is an equilibrium position of system (1) and u_0 is a control corresponding to the equilibrium position. Our aim is to find a function $u = u(x)$ defined in a neighborhood of the origin and satisfying the following conditions:

1. $u(0) = u_0$, and
2. all trajectories of the differential equation

$$\dot{x} = f(x, u(x)) \tag{2}$$

starting at a neighborhood of the origin tend to the equilibrium position $x = 0$.

The function $u(x)$ is referred to as a regulator or a stabilizer, and the problem of its construction is called a regulator design or stabilization problem.

The regulator design problem has been largely studied in the case of control systems where controls range over a vector space. Such problems serve as mathematical models for a large number of applications, and rather developed and satisfactory theory is now available for them (see Bacciotti [3], for example). This theory is also applicable in the presence of the control constraint $u \in U$ if the control u_0 corresponding to the equilibrium position is an interior point of the set U. Indeed, since the stabilization problem is of local nature we can expect that the stabilizer $u(x)$ varies in a neighborhood of u_0, and the constraint $u \in U$ is of no importance. However, in many technological applications processes possessing some extreme properties are of great interest. The main characteristic feature of the regulator design problem for such processes is that the control corresponding to the equilibrium position belongs to the boundary of the set U. In this case the control constraint is inevitably involved, and the problem is considerably more difficult. This paper is devoted to problems of this type.

There are many motivations that lead researchers to consider such problems. First of all, the Pontryagin maximum principle tells us that stabilization of optimal processes is a problem of this type. Usually optimal processes are not stabilizable and it is very important to determine

*Received by the editors January 26, 1994; accepted for publication (in revised form) May 12, 1995.

†Departamento de Matematica, Universidade de Evora, Apartado 94, P-7001 Evora codex, Portugal (smirnov@uevora.pt).

the set of deviations starting at which the system can be stabilized to the optimal process. The stabilization problem studied here can be considered as the first step in this direction.

A similar problem arises if we deal with a mechanical system subjected to a unilateral force. Consider, for instance, an oscillator and suppose that we have to stabilize it applying a force only in one direction. The corresponding mathematical model is

$$\ddot{x} = -x + u,$$
$$0 \le u \le 1.$$

The control $u_0 = 0$ corresponding to the equilibrium position $x = 0$ belongs to the boundary of the control set.

Now consider a guided missile which moves in a plane. We can vary the thrust of its jet propulsion in value and direction. If we neglect the angular motion and size of the missile and suppose that the velocity of the missile is always directed along the longitudinal axis, then the motion of the missile's mass center is described by the following equations:

$$\ddot{x} = -\sigma\dot{x} + \frac{u^1\dot{x} - u^2\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}},$$
$$\ddot{y} = -\sigma\dot{y} + \frac{u^1\dot{y} + u^2\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}},$$
$$(u^1, u^2) \in U = \left\{(u^1, u^2) \mid \sqrt{(u^1)^2 + (u^2)^2} \le b,\ \frac{u^2}{u^1} \le \tan\eta,\ u^1 \ge 0\right\},$$

where $\sigma > 0$ stands for a coefficient of air resistance, b is the maximal thrust, and η is the maximal angle between the longitudinal axis of the missile and that of the jet propulsion. Our aim is to stabilize the motion of the object along the x-axis with the constant maximal speed $\dot{x} = b/\sigma$. Thus we again obtain a stabilization problem where control $(u_0^1, u_0^2) = (b, 0)$ corresponding to the equilibrium position belongs to the boundary of the control set.

In mathematical biology and economics we also face stabilization problems with constrained controls (see Gouzé [10] and Berman, Neumann, and Stern [5]).

Classic regulator design theory cannot be applied to the above problems because of control constraints. The optimal control theory involves control constraints, but the problem of construction of the feedback control $u(x)$ that guarantees optimality of trajectories in a certain sense is too complex to be solved analytically, applying the Pontryagin maximum principle or studying the Bellman equation. Computational methods of optimal control theory are more suitable to finding optimal trajectories than to solving synthesis problems. Thus it is not worth considering the regulator design problem in the frame of optimal control theory. Therefore a reasonable statement of the regulator design problem and a rational combination of analytical and computational methods are required in order to solve the problem.

In classic stabilization theory we meet the statement of the problem where neither optimality nor finiteness of the stabilization time are required. The only requirement the regulator has to satisfy is that the equilibrium position is asymptotically stable. It seems natural to develop mathematical theory for this "weak" statement of the regulator design problem, if we aim to consider control systems of general type (1).

The origin of this setting of the problem and the method of its solution is the classical Lyapunov [14] stability theory. For practical problems it is important to consider not only the fact of asymptotic stability of an equilibrium position but also an estimate of a region of attraction, that is, the set of initial points starting at which trajectories tend to zero. It is possible with the help of the Lyapunov functions method. Unfortunately, there are no methods

to construct a Lyapunov function for a control system of general type. Therefore we have to "simplify" the system under consideration, that is, to find a less complex control system which is close to the original one in a neighborhood of the origin. To this end we consider the "first approximation" of system (1), that is, the linear control system

$$\dot{x} = Cx + w, \quad w \in K, \tag{3}$$

where $C = \nabla_x f(0, u_0)$ is an $n \times n$ matrix and K is a convex cone spanned by the set $f(0, U)$. Control systems of this type were first studied by Korobov and his students [12, 13] and then in more general form by Aubin, Frankowska, and Olech [2]. For control system (3) we derive necessary and sufficient conditions guaranteeing the existence of a Lyapunov function and prove that the latter is equivalent to the solvability of the regulator design problem for system (3). The proof is based on an investigation of the set-valued map $x \to Cx + K$. A map of this type, referred to as a convex process (see Rockafellar [16]), is a multivalued analogy of a linear operator. A number of properties similar to the Jordan theorem are established for it. The analysis of the structure of the convex process $x \to Cx + K$ is the basis for the construction of a Lyapunov function $V(x)$. To solve the regulator design problem we can choose the control $u(x)$ at the point x from the following condition:

$$V(x + \tau f(x, u(x))) = \min_{u \in U} V(x + \tau f(x, u)), \tag{4}$$

where $\tau > 0$.

A multivalued version of the Jordan theorem for convex processes of general type appeared in Smirnov [18], where it was applied to solve the regulator design problem for differential inclusions. Here we specify the results from [18] for the case of control systems. This allows us to simplify the proofs and to establish many new properties of the developed regulator design techniques. Computational aspects of this approach are discussed in Bushenkov and Smirnov [7, 8], where many examples are considered.

The regulator design method derived from our analysis contains many parameters which we can dispose of as needed. Optimization methods or heuristics allow one to obtain a regulator with required transient characteristics. For example, in the case of the single-input linear control system

$$\dot{x} = Ax + bu, \quad u \in R, \tag{5}$$

under controllability condition, we obtain a regulator which is a linear feedback control $u(x) = \langle d, x \rangle$ such that the spectrum of the linear system

$$\dot{x} = (A + bd^T)x \tag{6}$$

is concentrated at any given point $\lambda \in R$. A linear stabilizer of this type satisfies the following extremal property. It ensures a minimal overshoot among linear feedback control laws with the spectra of (6) $\{\lambda_1, \ldots, \lambda_n\}$ satisfying the conditions: $\mathrm{Re}\lambda_i \leq \lambda$ and $|\lambda_i - \lambda| < \epsilon$, where $\epsilon > 0$ is sufficiently small.

The paper is organized as follows. In §1 we discuss different statements of the stabilization problem. Section 2 is devoted to the regulator design problem for the first approximation. The nonlinear stabilization problem is considered in §3. Examples are given in §4. In §5 we consider the stabilization problem for the first approximation under controllability conditions. The connection between stabilizability and weak asymptotic stability is discussed in §6.

Throughout this paper we denote by R the set of real numbers and by R^n the usual n-dimensional space of vectors $x = (x^1, \ldots, x^n)$, where $x^i \in R$, $i = \overline{1, n}$. The inner product

of two vectors x and y in R^n is expressed by $\langle x, y\rangle = x^1y^1 + \cdots + x^ny^n$. The norm of a vector $x \in R^n$ is defined by $|x| = \langle x, x\rangle^{1/2}$. If C is an $m \times n$ real matrix, then the transposed matrix is denoted by C^T. The unit linear operator from R^n to R^n will be denoted by E. We denote by B_n the unit ball in R^n: $B_n = \{x \in R^n \mid |x| \leq 1\}$. The open unit ball in R^n is denoted by $\overset{\circ}{B}_n = \{x \in R^n \mid |x| < 1\}$. Let $A \subset R^n$. The distance function $d(\cdot, A) : R^n \to R$ is denoted by $d(x, A) = \inf\{|x - a| \mid a \in A\}$. Let $\lambda \in R$. Then put $\lambda A = \{\lambda a \mid a \in A\}$. For two sets A and B in R^n, their sum is denoted by $A + B = \{a + b \mid a \in A,\ b \in B\}$. The closure and interior of A are denoted by $\mathrm{cl}A$ and $\mathrm{int}A$, respectively. The boundary of A is denoted by $\mathrm{bd}A = \mathrm{cl}A \setminus \mathrm{int}A$. The convex hull of A is denoted by $\mathrm{co}A$. The support function of a set A is denoted by

$$S(x^*, A) = \sup\{\langle x^*, a\rangle \mid a \in A\}.$$

Let $K \subset R^n$ be a convex cone. The conjugate cone of K is defined by

$$K^* = \{x^* \mid \langle x, x^*\rangle \geq 0,\ \forall x \in K\}.$$

Let $f : R^n \to R$ be a function. The directional derivative of f at x with respect to a vector v is denoted by

$$Df(x)(v) = \lim_{\lambda \downarrow 0} \frac{f(x + \lambda v) - f(x)}{\lambda}$$

if it exists.

Let $f : R \to R$ be a continuous function. The Lyapunov exponent (see Lyapunov [14]) of f is defined by

$$\chi[f(\cdot)] = -\limsup_{t \to \infty} \frac{1}{t} \ln |f(t)|.$$

The Lyapunov exponent has the following properties:

1. $\chi[(f + \varphi)(\cdot)] \geq \min\{\chi[f(\cdot)], \chi[\varphi(\cdot)]\}$,
2. $\chi[(f\varphi)(\cdot)] \geq \chi[f(\cdot)] + \chi[\varphi(\cdot)]$,
3. $\chi[(f\varphi)(\cdot)] = \chi[f(\cdot)]$, where $0 < a \leq \varphi(t) \leq b < \infty$.

If $f : R \to R^n$ is a vector function, then the Lyapunov exponent is defined as the minimal value of the Lyapunov exponents of the components

$$\chi[f(\cdot)] = \min\{\chi[f^1(\cdot)], \ldots, \chi[f^n(\cdot)]\}.$$

1. Statement of the problem. Consider a control system

$$\dot{x} = f(x, u), \quad u \in U \subset R^k. \tag{7}$$

Assume that $f : R^n \times U \to R^n$ is a continuous function differentiable with respect to x and that for any arbitrary compact set $X \in R^n$ there exists a constant $l > 0$ such that $|\nabla_x f(x, u)| \leq l$ for all $(x, u) \in X \times U$. Let $f(x, U) \subset R^n$ be a convex set for all $x \in R^n$, and let $u_0 \in U$ be a point such that $f(0, u_0) = 0$. Our aim is to find a map $u : R^n \to U$ defined in a neighborhood of the origin and satisfying the following conditions:

1. $u(0) = u_0$,
2. the equilibrium point $x = 0$ of the differential equation

$$\dot{x} = f(x, u(x)) \tag{8}$$

is asymptotically stable.

The regulator design problem described above is very general and intuitive and is not yet suitable for a mathematical consideration. We should specify the exact meaning of "asymptotic stability" and what we mean by "solution" of (8). As we will see, different specifications of these notions can lead to substantially different developments.

The first issue we must consider is how smooth the stabilizer $u = u(x)$ should be. In classic regulator design theory (see Sontag [19], for example) the control system

$$\dot{x} = g(x, u), \quad u \in R^k, \tag{9}$$

is studied. Suppose that the function g is smooth. Consider the following linear control system associated with (9):

$$\dot{x} = Cx + Bw, \quad w \in R^k, \tag{10}$$

where $C = \nabla_x g(0, u_0)$ and $B = \nabla_u g(0, u_0)$. System (10) is stabilizable if and only if there exists a matrix D such that the equilibrium position $x = 0$ of the linear system

$$\dot{x} = (C + BD)x$$

is asymptotically stable. If (10) is stabilizable, then any linear feedback law $w(x) = Dx$ which stabilizes (10) provides a local stabilizer for system (9): $u(x) = u_0 + Dx$. Thus, we can expect that under suitable assumptions a smooth and even an analytical regulator exists for control system (9).

The situation completely changes as soon as control constraints are involved. Let us consider the following example.

Example: Absence of smooth stabilizability. Suppose that the stabilization problem for the control system

$$\begin{aligned} \dot{x}^1 &= x^2 + u^1, \\ \dot{x}^2 &= x^1 + u^2, \end{aligned} \tag{11}$$

where $u = (u^1, u^2) \in U = \{(u^1, u^2) \mid u^1 \leq 0, \ u^2 \geq 0\}$ has a smooth solution, that is, there exists a smooth regulator $u = u(x) = (u^1(x^1, x^2), u^2(x^1, x^2))$ such that $u(0) = 0$, and the equilibrium position $x^1 = 0, \ x^2 = 0$ of the system

$$\begin{aligned} \dot{x}^1 &= x^2 + u^1(x^1, x^2), \\ \dot{x}^2 &= x^1 + u^2(x^1, x^2) \end{aligned} \tag{12}$$

is asymptotically stable. Show that $\nabla_x u(0) = 0$. Assume that there exists $x_0 \in R^2$ such that $\nabla_x u(0)x_0 = w \neq 0$. Then $\nabla_x u(0)(-x_0) = -w$. Assume for the sake of being definite that $w^1 \neq 0$, and consider the function $\varphi(\tau) = u^1(\tau x_0)$. Observe that $\varphi'(0) = w^1 \neq 0$. By the inverse function theorem there exists φ^{-1} defined in a neighborhood of zero. This implies that $u^1(\tau x_0)$ changes sign, if τ changes sign. Since the set U does not contain vectors (u^1, u^2) with $u^1 > 0$, we come to a contradiction. Hence $\nabla_x u(0) = 0$. Thus we conclude that the linearization of system (12) at zero is

$$\begin{aligned} \dot{x}^1 &= x^2, \\ \dot{x}^2 &= x^1. \end{aligned}$$

The matrix of this system has the eigenvalue 1, that is, system (12) is not stable. Consequently, there is no smooth stabilizer for system (11). As we shall see from Theorem 7 there exists a Lipschitz continuous stabilizer for this system.

Different statements of the problem. The example shows that we can expect the existence of at most a Lipschitz continuous stabilizer, if control constraints are involved. Therefore, the first reasonable statement of the regulator design problem is to find a *Lipschitzian control* $u(x)$ that guarantees asymptotic stability of zero solution to differential equation (8). Lipschitz continuity of $u(x)$ implies that equation (8) has a unique, classical solution for each initial condition.

Another, weaker statement of the stabilization problem can be obtained if the Lipschitz condition is replaced by the continuity of $u(x)$. The latter implies that equation (8) has a classical solution (not unique, in general) for each initial point. In this case we mean by asymptotic stability the following: the equilibrium point $x = 0$ of differential equation (8) is said to be asymptotically stable if for any $\epsilon > 0$ there exists $\delta > 0$ such that for all $x_0 \in \delta B_n$ *each* solution to differential equation (8) with $x(0) = x_0$ exists for $t \in [0, \infty)$ and satisfies the conditions $|x(t)| < \epsilon$ for all $t \in [0, \infty)$ and $\lim x(t) = 0$ as $t \to \infty$.

The above problems can be solved only under rather restrictive assumptions on the map $u \to f(x, u)$. Therefore, a quite natural goal is to find a control $u(x)$ such that the function $x \to f(x, u(x))$ is *continuous* and the equilibrium position $x = 0$ of differential equation (8) is asymptotically stable.

Practical experience shows that discontinuous control laws are of great importance. For example, optimal synthesis is usually discontinuous, relay stabilization systems are widely used in engineering, etc. For this reason we are interested in considering regulators such that the function $x \to f(x, u(x))$ is not continuous in general. To this end we invoke Filippov's notion of solution to a differential equation with discontinuous right-hand side (see Aubin and Cellina [1] and Filippov [9]).

The problem we shall mainly consider is to find a control $u(x)$ such that the equilibrium position $x = 0$ of a differential equation with *discontinuous* right-hand side (8) is asymptotically stable.

Recall the definition of the Filippov solution and the notion of asymptotic stability for discontinuous differential equations. Let $\varphi : R^n \to R^n$ be a bounded function satisfying $\varphi(0) = 0$, and let $\Phi : R^n \to R^n$ be the set-valued map defined by

$$\Phi(x) = \bigcap_{\eta > 0} \operatorname{cl} \operatorname{co} \varphi(x + \eta B_n).$$

An absolutely continuous function $x(\cdot)$ is called a Filippov solution to the differential equation

$$\dot{x} = \varphi(x) \tag{13}$$

if and only if it satisfies the differential inclusion

$$\dot{x} \in \Phi(x)$$

almost everywhere.

The equilibrium point $x = 0$ of the differential equation (13) is said to be asymptotically stable if, for any $\epsilon > 0$, there exists $\delta > 0$ such that for all $x_0 \in \delta B_n$, each Filippov solution to the differential equation (13) with $x(0) = x_0$ exists for $t \in [0, \infty)$ and satisfies the conditions $|x(t)| < \epsilon$ for all $t \in [0, \infty)$ and $\lim x(t) = 0$ as $t \to \infty$.

From now on, by the *stabilization* or *regulator design* problem we mean this formulation, if there is no other specification.

In technological applications a piecewise constant control law is usually used instead of the stabilizer $u(x)$ we dealt with before. The control is chosen in discrete moments of time $0, \sigma, 2\sigma, \ldots$, where $\sigma > 0$, and we have

$$u_\sigma(t) \equiv u(x(k\sigma)), \quad t \in [k\sigma, (k+1)\sigma).$$

This leads us to another problem where the goal is to find a control $u(x)$ such that all the trajectories of the nonautonomous differential equation

$$\dot{x} = f(x, u_\sigma(t))$$

tend to zero. This statement of the problem is also very important, since it substantiates practical realizations of stabilizers, and we discuss it in detail in §3.

Informal outline of the approach. First, we investigate the stabilization problem for the first approximation of system (7), that is, for the linear control system

$$\dot{x} = Cx + w, \quad w \in K, \tag{14}$$

where $C = \nabla_x f(0, u_0)$ is an $n \times n$ matrix and K is the closed convex cone spanned by the set $f(0, U)$. Consider the set-valued map $x \to Cx + K$ associated with control system (14). The properties of this map are similar to those of a linear operator. In particular, a multivalued version of the Jordan theorem can be established. More precisely, there exists a minimal invariant subspace I, that is, a minimal subspace such that for any $x \in I$ we have $Cx + K \subset I$. The map $x \to Cx + K$ considered as a map from the factor space R^n/I into R^n/I is a linear operator denoted by $\tilde{C}$. The meaning of I in terms of control is that only movements in linear manifolds parallel to I are affected by a control, while movements in the factor space R^n/I are completely determined by properties of $\tilde{C}$. Asymptotic stability of $\tilde{C}$ is necessary for the first approximation to be stabilizable. If this operator is not asymptotically stable, the control system (14) cannot be stabilized by any control $w(x)$.

Consider the structure of I. For the sake of simplicity suppose that $I = R^n$. Let λ_0 be the maximal real eigenvalue of C^T that corresponds to an eigenvector contained in the polar cone K^*. Assume that $\lambda > \lambda_0$. Then for any x there exist vectors $y_0, \ldots, y_k$ such that

$$\begin{aligned} y_0 &\in -K, \\ \lambda y_1 &\in Cy_1 + K, \\ y_1 + \lambda y_2 &\in Cy_2 + K, \\ &\vdots \\ y_{k-1} + \lambda y_k &\in Cy_k + K, \\ x &= y_k + y_0. \end{aligned} \tag{15}$$

We can interpret inclusions (15) as follows. Vectors y_1 and y_j, $j > 1$, can be considered as an eigenvector and joined (principal) vectors of the set-valued map $x \to Cx + K$, respectively. Inclusions (15) imply that the subspace I is a "cyclic" subspace of the set-valued map $x \to Cx + K$ corresponding to one "Jordan block."

Now we are in a position to explain how to choose the control $w(x)$ to stabilize the control system (14). We establish that the stabilizability of (14) is equivalent to the condition $\lambda_0 < 0$. If the first approximation is stabilizable, we fix $\lambda \in (\lambda_0, 0)$. For any $x \in R^n$ we find a finite set of vectors $y_0, \ldots, y_k$ satisfying (15).

Suppose, first, that $y_k = 0$. Then $x = y_0 \in -K$. Consequently, for any $\epsilon > 0$ we can choose $\tau > 0$ such that $\tau Cx \in \epsilon B_n$. Put $w_0(x) = -\frac{1}{\tau}x$. Then we have $w_0(x) \in K$ and

$$x + \tau(Cx + w_0(x)) \in \epsilon B_n.$$

It means that there exists a velocity of system (14) almost exactly directed from x to the origin.

Now assume that $y_0 = 0$, that is, $x = y_k$. Suppose that $k = 1$, which implies that x is an eigenvector of the set-valued map $x \to Cx + K$. Since $\lambda < 0$, we conclude that there exists

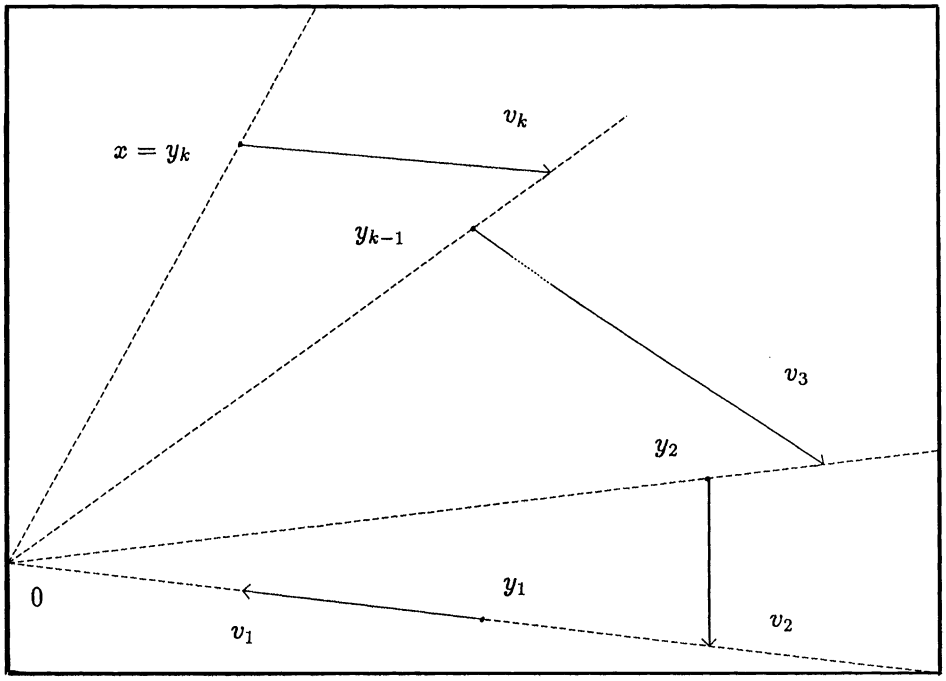


FIG. 1. *Choosing stabilizing controls.*

a vector v_1, a velocity of (14), directed from x to the origin. Indeed, let $w_1(x) = \lambda x - Cx$. Then we obtain $w_1(x) \in K$ and

$$x + \frac{1}{|\lambda|}(Cx + w_1(x)) = 0.$$

Consider the case $k > 1$. Since $\lambda < 0$, we have

$$\frac{1}{|\lambda|} y_{k-1} \in y_k + \frac{1}{|\lambda|}(Cy_k + K)$$

or, in other words,

$$\frac{1}{|\lambda|} y_{k-1} = y_k + \frac{1}{|\lambda|}(Cy_k + w_k(x)),$$

where $w_k(x) \in K$. This implies that there exists a vector v_k, a velocity of (14), directed from the point y_k to the ray spanned by y_{k-1}. The same is true for the vectors $y_{k-1}, \ldots, y_2$. Thus, starting at $x = y_k$ and choosing a suitable control we can reach the ray spanned by the eigenvector y_1 and then move along it (see Figure 1). In the general case when $x = y_k + y_0$ with nonzero y_k and y_0, we take $w(x)$ as a convex combination of $w_k(x)$ and $w_0(x)$. Since K is a convex cone, we have $w(x) \in K$. It is clear intuitively that trajectories of the system

$$\dot{x} = Cx + w(x)$$

tend to zero. The above informal considerations show the main ideas of the regulator design method we use in this paper.

We use similar reasoning to prove the existence of a Lyapunov function $V(x)$ in R^n and a number $\theta > 0$ satisfying the following condition: for all $x \in R^n$ there exists a vector $v \in Cx + K$ such that $DV(x)(v) + \theta V(x) \leq 0$. If $|x|$ is sufficiently small, then there exists a vector $v \in f(x, U)$ such that $DV(x)(v) + \frac{1}{2}\theta V(x) \leq 0$. The map $u(x)$ is defined to make $V(x)$ a Lyapunov function for differential equation (8). This implies the asymptotic stability of the equilibrium point $x = 0$. The proof of the existence of the function $V(x)$ is constructive and can serve as the basis for a numerical regulator design algorithm.

2. The first approximation. In this section we investigate the first approximation of system (7), that is, the linear control system

$$\dot{x} = Cx + w, \quad w \in K, \tag{16}$$

where $C = \nabla_x f(0, u_0)$ is an $n \times n$ matrix and K is the closed convex cone spanned by the set $f(0, U)$. For control system (16) we derive necessary and sufficient conditions guaranteeing the existence of a number $\theta > 0$ and a convex positive, positively homogeneous function $V(x)$ in R^n satisfying the following condition: for any $x \in R^n$ there exists a vector $v \in Cx + K$ such that $DV(x)(v) + \theta V(x) \leq 0$.

We now consider the linear differential equation

$$\dot{x} = -C^T x. \tag{17}$$

Let Λ^+ be the subspace such that a solution to (17) with the initial condition $x(0) = x \in \Lambda^+$ has a nonnegative Lyapunov exponent.

The following result contains necessary and sufficient conditions of solvability for the regulator design problem for linear system (16).

THEOREM 1. *The following conditions are equivalent:*

1. *The regulator design problem for control system* (16) *is solvable.*
2. *For any $x_0 \in R^n$ there exists a trajectory $x(\cdot)$ of system* (16) *such that $x(t) \to 0$ as $t \to \infty$.*
3. *The matrix C^T has neither eigenvectors which correspond to nonnegative real eigenvalues and are contained in the cone K^* nor proper invariant subspaces contained in the subspace $\Lambda^+ \cap K^* \cap -K^*$.*
4. *There exist numbers $\tau > 0$, $\delta \in (0, 1)$ and a convex positively homogeneous function $V(x)$ such that $V(0) = 0$, $V(x) > 0$, if $x \neq 0$, and for any $x \in R^n$ a vector $w \in K$ satisfying*

$$V(x + \tau(Cx + w)) \leq \delta V(x)$$

can be found.

First we prove auxiliary results.

LEMMA 1. *Assume that for any $x_0 \in R^n$ there exists a trajectory $x(\cdot)$ of system* (16) *with $x(0) = 0$ such that $x(t) \to 0$ when $t \to \infty$. Then there exist numbers $a > 0$ and $\gamma > 0$ such that for any point $x_0 \in R^n$ a trajectory of system* (16) *with $x(0) = x_0$ satisfying the condition*

$$|x(t)| \leq a|x_0|e^{-\gamma t}, \ t \geq 0, \tag{18}$$

can be found.

Proof. Let us consider a simplex $\Sigma = \text{co}\{z_0, \ldots, z_n\}$ containing a unit ball centered at zero. There are trajectories $x_k(\cdot)$ satisfying the conditions

$$x_k(0) = z_k, \quad \lim_{t \to \infty} x_k(t) = 0, \quad k = \overline{0, n}.$$

There exists a number $\tau \geq 0$ such that $|x_k(\tau)| \leq 1/e$ for all $k = \overline{0, n}$. Let $y \in \operatorname{bd} B_n$. Then there exist numbers $\lambda_k \geq 0$, $k = \overline{0, n}$, such that $\sum_{k=0}^{n} \lambda_k = 1$ and $y = \sum_{k=0}^{n} \lambda_k z_k$. Obviously, the function $x(\cdot, y) = \sum_{k=0}^{n} \lambda_k x_k(\cdot)$ is a trajectory of system (16) and satisfies the inequality $|x(\tau, y)| \leq 1/e$. Hence, for any $y \in R^n$ the function $x_y(\cdot) = |y| x(\cdot, y/|y|)$ is a trajectory of system (16) and satisfies the inequality $|x_y(\tau)| \leq |y|/e$. Set

$$\gamma = 1/\tau, \quad a = e \max\{|x_k(t)| \mid t \in [0, \tau], \ k = \overline{0, n}\}.$$

For every $x_0 \in R^n$ we define a trajectory $x(\cdot)$ of system (16) satisfying the initial condition $x(0) = x_0$ as follows:

$$x(t) = x_{x(m\tau)}(t - m\tau), \quad t \in [m\tau, (m+1)\tau], \ m = 0, 1, \ldots.$$

It is easy to check that the trajectory $x(\cdot)$ satisfies condition (18). □

Let us study some properties of the set-valued map $x \to Cx + K$. Denote by J the maximal invariant subspace of C^T contained in $K^* \cap -K^*$ and by I the orthogonal complement to J.

The subspace I turns out to be invariant by the set-valued map $x \to Cx + K$ in the following sense.

LEMMA 2. *For all $x \in I$ we have*

$$Cx + K \subset I.$$

Proof. Let $x \in I$, $x^* \in J$, $w \in K$. Since $C^T x^* \in J$, $x^* \in J \subset K^* \cap -K^*$, we have

$$\langle x^*, Cx + w \rangle = \langle C^T x^*, x \rangle + \langle x^*, w \rangle = 0. \qquad \square$$

Remark. A description of the minimal invariant subspace for convex processes of general form appeared in Aubin, Frankowska, and Olech [2].

It follows from Lemma 2 that $K \subset I$ and the subspace I is invariant by the linear operator C. We denote by C_I the restriction of C to the subspace I, i.e., $C_I = C \mid_I$. Let C_J be the transposed operator of the restriction of C^T to the subspace J, i.e., $C_J = (C^T \mid_J)^T$. The cone K considered as a subset of I is denoted by K_I. The unit linear operator from I to I is denoted by E_I. Since $R^n = I \times J$, every $x \in R^n$ can be represented as $x = (x_I, x_J)$, where $x_I \in I$ and $x_J \in J$.

LEMMA 3. *The linear operator C_I^T has no nontrivial invariant subspaces contained in the cone K_I^*.*

Proof. Suppose there is a nontrivial subspace $L \subset K_I^* \cap -K_I^*$ invariant by the linear operator C_I^T. Let us prove that $C^T(L, J) \subset (L, J)$. Since $(L, J) \subset (K_I^* \cap -K_I^*, J) = K^* \cap -K^*$, this will contradict the definition of J as the maximal invariant by C^T subspace contained in the subspace $K^* \cap -K^*$. Let $x_I^* \in L$, $x_I \in L^\perp$. Then

$$\langle C^T(x_I^*, 0), (x_I, 0) \rangle = \langle (x_I^*, 0), C(x_I, 0) \rangle = \langle (x_I^*, 0), (C_I x_I, 0) \rangle$$

$$= \langle x_I^*, C_I x_I \rangle = \langle C_I^T x_I^*, x_I \rangle = 0.$$

Since $C^T(0, J) \subset (0, J)$, we obtain $C^T(L, J) \subset (L, J)$. This achieves the proof. □

Denote by λ_0 the maximal real eigenvalue of C^T which corresponds to an eigenvector contained in the cone K^*. If there is no such eigenvector, we put $\lambda_0 = -\infty$.

LEMMA 4. *Suppose that μ is a real eigenvalue of C_I^T which corresponds to an eigenvector y_I^* contained in K_I^*. Then $\mu \leq \lambda_0$.*

Proof. If μ is an eigenvalue of C^T corresponding to an eigenvector from J, then we have nothing to prove. If this is not the case, then $(C^T - \mu E)J = J$.

Observe that for all $y_I \in I$

$$\langle C^T(y_I^*, 0) - \mu(y_I^*, 0), (y_I, 0)\rangle = \langle (y_I^*, 0), C(y_I, 0)\rangle - \mu\langle y_I^*, y_I\rangle$$

$$= \langle (y_I^*, 0), (C_I y_I, 0)\rangle - \mu\langle y_I^*, y_I\rangle = \langle C_I^T y_I^*, y_I\rangle - \mu\langle y_I^*, y_I\rangle = 0.$$

Thus, $\mu(y_I^*, 0) = C^T(y_I^*, 0) + (0, z_J^*)$ for some $z_J^* \in J$. In other words, $C^T(y_I^*, 0) - \mu(y_I^*, 0) \in J$.

There exists a vector p_J^* such that $(0, z_J^*) = (C^T - \mu E)(0, p_J^*)$. This implies that

$$\mu(y_I^*, p_J^*) = C^T(y_I^*, p_J^*).$$

Consequently, $\mu \leq \lambda_0$. □

LEMMA 5. *Let $y = Cx$, where $y = (y_I, y_J)$, $x = (0, x_J)$. Then $y_J = C_J x_J$.*

Proof. Let x_J, $x_J^* \in J$. Then

$$\langle C(0, x_J), (0, x_J^*)\rangle = \langle (0, x_J), C^T(0, x_J^*)\rangle$$

$$= \langle (0, x_J), (0, C^T|_J\, x_J^*)\rangle = \langle C_J x_J, x_J^*\rangle$$

$$= \langle (0, C_J x_J), (0, x_J^*)\rangle.$$

Since x_J, $x_J^* \in J$ are arbitrary, the proof ensues. □

For all real λ we define convex cones

$$L_k(\lambda) = -\mathrm{co}\bigcup_{i=0}^{k}(C_I - \lambda E_I)^{-i} K_I, \qquad k = \overline{0, \infty},$$

contained in the subspace I. We observe that $L_k(\lambda) \subset L_m(\lambda)$ if $k < m$.

We shall use the following generalization of the well-known Perron positive matrix theorem (see Berman and Plemmons [4], for example).

THEOREM 2. *Let $K \subset R^n$ be a nonzero convex closed cone which does not contain a line and let $C : R^n \to R^n$ be a linear operator. If $Cx \in K$ for all $x \in K$ then there exists an eigenvector of C contained in the cone K and corresponding to a nonnegative eigenvalue.*

THEOREM 3. *If $\lambda > \lambda_0$, then there exists a number k such that the equality*

$$L_k(\lambda) = I$$

holds.

Proof. Since a polyhedron can be chosen as a neighborhood of the origin in the subspace I, it is sufficient to prove that

$$L_\infty(\lambda) = I.$$

The latter is equivalent with the equality

$$L_\infty^*(\lambda) = \{0\}.$$

Suppose this equality is not true. Since $\lambda > \lambda_0$, we have

$$(C_I - \lambda E_I)I - K_I = I. \tag{19}$$

Indeed, if $(C_I - \lambda E_I)I - K_I \neq I$, then there exists a nonzero vector $x^* \in I$ such that $\langle (C_I - \lambda E_I)x, x^*\rangle \leq \langle z, x^*\rangle$ for all $x \in I$, $z \in K_I$. Taking $x = 0$, we obtain $x^* \in K_I^*$. Let $z = 0$. Then we have $\langle x, (C_I^T - \lambda E_I)x^*\rangle = 0$ whenever $x \in I$. Hence $x^* \in K_I^*$ is an eigenvector of C_I^T corresponding to the eigenvalue λ. By Lemma 4, $\lambda \leq \lambda_0$, a contradiction.

From (19) we obtain

$$(C_I - \lambda E_I)I - L_k(\lambda) = I.$$

By Corollary 16.3.2 in [16, p. 143] and Corollary 16.4.2 in [16, p. 146], we have

$$L^*_{k+1}(\lambda) \subset ((C_I - \lambda E_I)^{-1}\mathrm{cl}L_k(\lambda) - K)^* = (C_I^* - \lambda E_I)L_k^*(\lambda) \cap -K^*, \quad k = \overline{0, \infty}.$$

Obviously,

$$(C_I^T - \lambda E_I)^{-1}L^*_\infty(\lambda) \subset L^*_\infty(\lambda). \tag{20}$$

The cone $L^*_\infty(\lambda)$ does not contain a line. Indeed, if L is the maximal subspace contained in $L^*_\infty(\lambda)$, then $(C_I^T - \lambda E_I)^{-1}L \subset L$. This implies that L is invariant by the linear operator C_I^T. Since $L^*_\infty(\lambda) \subset K_I^*$, invoking Lemma 3, we derive a contradiction.

From inclusion (20) and Theorem 2 it follows that there exists a nonzero vector $x_I^* \in L^*_\infty(\lambda)$ and a number $\mu \geq 0$ such that

$$(C_I^T - \lambda E_I)^{-1}x_I^* = \mu x_I^*.$$

Since $\lambda > \lambda_0$, we conclude that $\mu \neq 0$. Consequently,

$$C_I^T x_I^* = \left(\lambda + \frac{1}{\mu}\right)x_I^*.$$

By Lemma 4, $\lambda + 1/\mu \leq \lambda_0$, a contradiction. □

Proof of Theorem 1. Condition 2 is a trivial consequence of the first condition. Assume that Condition 2 is fulfilled. Suppose that the third condition does not hold. This implies that the differential equation

$$\dot{x}^*(t) = -C^T x^*(t)$$

has a nontrivial solution satisfying the inclusion $x^*(t) \in K^*$, for all $t \geq 0$, and such that $\chi[x^*(\cdot)] \geq 0$. Let $x_0 \in R^n$. By Lemma 1 there exists a trajectory $x(\cdot)$ of (16) with $x(0) = x_0$ such that $\chi[x(\cdot)] > 0$. Let $w(\cdot)$ be a control corresponding to the trajectory $x(\cdot)$. We observe that

$$\langle x(t), x^*(t)\rangle = \langle x_0, x^*(0)\rangle + \int_0^t \frac{d}{ds}\langle x(s), x^*(s)\rangle ds$$

$$= \langle x_0, x^*(0)\rangle + \int_0^t (\langle Cx(s) + w(s), x^*(s)\rangle + \langle x(s), -C^T x^*(s)\rangle) ds$$

$$= \langle x_0, x^*(0)\rangle + \int_0^t \langle w(s), x^*(s)\rangle ds \geq \langle x_0, x^*(0)\rangle.$$

Since

$$\langle x(t), x^*(t)\rangle \leq |x(t)|\,|x^*(t)| \to 0 \text{ as } t \to \infty,$$

we obtain $0 \geq \langle x_0, x^*(0)\rangle$. Since x_0 is an arbitrary vector we conclude that $x^*(0) = 0$. This contradicts the nontriviality of $x^*(\cdot)$. Thus, the third condition is a consequence of the second one.

Now, suppose that the third condition holds. We shall derive Condition 4 from it. The third condition means that $\lambda_0 < 0$. Fix $\lambda \in (\lambda_0, 0)$. Let $\Sigma \subset I$ be a polyhedron containing

the ball $B_I = B_n \cap I$. Denote by x_i^Σ, $i = \overline{0, m}$, the vertices of the polyhedron. By Theorem 3, $x_i^\Sigma \in L_{k_i}(\lambda)$ for some k_i. Moreover, x_i^Σ does not belong to $L_k(\lambda)$, if $k < k_i$. From the definition of the cones $L_{k_i}(\lambda_i)$ we derive the existence of a finite set of nonzero vectors $\{y_{i,j}\}_{j=0}^{k_i}$ in I that satisfy the following inclusions:

$$
\begin{aligned}
y_{i,0} &\in -K_I, \\
\lambda y_{i,1} &\in C_I y_{i,1} + K_I, \\
y_{i,1} + \lambda y_{i,2} &\in C_I y_{i,2} + K_I, \\
&\vdots \\
y_{i,k_i-1} + \lambda y_{i,k_i} &\in C_I y_{i,k_i} + K_I, \\
x_i^\Sigma &= y_{i,k_i} + y_{i,0}.
\end{aligned} \tag{21}
$$

Fix a number $\alpha > 1$. Let M_I be the convex hull of the points

$$
z_{i,0} = y_{i,0}, \quad i = \overline{0, m};
$$

$$
z_{i,j} = (\alpha/|\lambda|)^{k_i - j} y_{i,j}, \quad j = \overline{1, k_i}, \quad i = \overline{0, m}.
$$

Let us consider the linear operator $C_J : J \to J$. The third condition implies that C_J is an asymptotically stable linear operator. There exists a positive definite quadratic form $W : J \to R$ which is a Lyapunov function for the differential equation

$$
\dot{x}_J = C_J x_J
$$

(see Lyapunov [14]). Denote by M_J the ellipsoid $\{x \in J \mid W(x) \leq 1\}$. Let $\omega > 0$. Consider the convex compact set

$$
\mathcal{M}_\omega = M_I \times \omega M_J \subset I \times J = R^n.
$$

We shall prove that the Minkowski function $V(x)$ of the set $\mathcal{M}_\omega$ is the function to be found whenever $\omega > 0$ is sufficiently small.

Since $B_I \subset \Sigma$, we conclude that $\frac{1}{2} B_I \subset M_I$. Obviously, $w_{i,0} = -4|C_I z_{i,0}| z_{i,0} \in K_I$ and

$$
z_{i,0} + \frac{1}{(4|C_I z_{i,0}|)} (C_I z_{i,0} + w_{i,0}) \in \frac{1}{4} B_I \subset \operatorname{int} M_I
$$

when $i = \overline{0, m}$. For each $i = \overline{0, m}$, the vector $w_{i,1} = \lambda z_{i,1} - C_I z_{i,1}$ belongs to the cone K_I and satisfies the inclusion

$$
z_{i,1} + |\lambda|^{-1} (C_I z_{i,1} + w_{i,1}) = 0 \in \operatorname{int} M_I.
$$

For all points $z_{i,j}$, $j > 1$, $i = \overline{0, m}$, we consider the vectors $w_{i,j} = (1/\alpha)|\lambda| z_{i,j-1} + \lambda z_{i,j} - C_I z_{i,j}$. Obviously, we have $w_{i,j} \in K_I$ and

$$
z_{i,j} + |\lambda|^{-1} (C_I z_{i,j} + w_{i,j}) = \frac{1}{\alpha} z_{i,j-1} \in \operatorname{int} M_I.
$$

From the above reasoning we conclude that there exist numbers $\tau_I > 0$ and $\delta_I \in (0, 1)$ such that for each $x_I \in \operatorname{bd} M_I$ there exists a vector $w \in K_I$ satisfying the inclusion

$$
x_I + \tau_I (C_I x_I + w) \in \delta_I M_I.
$$

Moreover, there exist numbers $\tau_J > 0$ and $\delta_J \in (0, 1)$ such that for all $x_J \in \operatorname{bd} M_J$ the inclusion

$$
x_J + \tau_J C_J x_J \in \delta_J M_J
$$

is fulfilled. Set $\tau = \min\{\tau_I, \tau_J\}$, $\delta = \max\{\delta_I, \delta_J\}$ and choose ω from the interval $(0, (1 - \delta)/(4\tau|C|b))$, where $b = \max\{|x_J| \mid x_J \in \mathrm{bd}M_J\}$.

Let $x = (x_I, x_J) \in \mathrm{bd}\mathcal{M}_\omega$. Then there exists a number $\eta \in [0, 1]$ such that $x_I \in \mathrm{bd}(\eta M_I)$. Observe that there exists a vector $w \in K$ satisfying the inclusion

$$(x_I, 0) + \tau(C(x_I, 0) + w) \in (\delta\eta M_I, 0) \subset (\delta M_I, 0). \tag{22}$$

On the other hand, by Lemma 5 we have

$$(0, x_J) + \tau C(0, x_J) = (0, x_J) + \tau(y_I, C_J x_J) \in \tau(y_I, 0) + (0, \delta\omega M_J). \tag{23}$$

Since

$$|y_I| \le |C(0, x_J)| \le |C|\omega b \le \frac{1 - \delta}{4\tau},$$

we conclude that $\tau y_I \in \frac{1}{2}(1 - \delta)M_I$. Summation of (22) and (23) yields

$$x + \tau(Cx + w) \in \left(\frac{1 + \delta}{2} M_I, \delta\omega M_J\right) \subset \frac{1 + \delta}{2}\mathcal{M}_\omega.$$

Thus, we obtain Condition 4 of the theorem.

Finally, let Condition 4 be fulfilled. We shall derive Condition 1. Consider the set-valued maps

$$G(x) = \{v \in R^n \mid x + \tau v \in \delta V(x)\mathcal{M}\} \quad \text{and} \quad H(x) = G(x) \cap (Cx + K),$$

where $\mathcal{M} = \{x \in R^n \mid V(x) \le 1\}$. Condition 4 implies $H(x) \ne \emptyset$ for all $x \in R^n$. Let $w : R^n \to K$ be an arbitrary function satisfying the inclusion $Cx + w(x) \in G(x)$. Obviously, $|Cx + w(x)| \le LV(x)$, where $L = \tau^{-1} \max\{V(x) \mid \delta z - x/V(x)| \mid z \in \mathcal{M},\ x \in \mathrm{bd}\mathcal{M}\}$. Consider the set-valued map

$$W(x) = \bigcap_{\eta > 0} \mathrm{cl\ co}\ w(x + \eta B_n).$$

Let $w \in W(x)$. Then we have

$$DV(x)(Cx + w) \le \frac{[V(x + \tau(Cx + w)) - V(x)]}{\tau} \le -\frac{1 - \delta}{\tau} V(x).$$

Hence any Filippov solution $x(\cdot)$ to the differential equation

$$\dot{x} = Cx + w(x) \tag{24}$$

satisfies

$$V(x(t)) \le V(x(0))e^{-\theta t}, \quad t \ge 0,$$

where $\theta = (1 - \delta)/\tau$. Consequently, the zero equilibrium position of differential equation (24) is asymptotically stable. This ends the proof. □

COROLLARY 1. *Assume that the matrix C^T has no proper invariant subspaces contained in the subspace $K^* \cap -K^*$ and one of the following conditions holds:*

1. *The matrix C^T has no eigenvectors corresponding to nonnegative real eigenvalues and contained in the cone K^*.*

2. *There exist $\lambda < 0$ and k such that $L_k(\lambda) = R^n$.*

Then the regulator design problem for control system (16) *is solvable.*

Proof. Since the matrix C^T has no proper invariant subspaces contained in the subspace $K^* \cap -K^*$, we see that $I = R^n$. The reasoning used to prove the implication $3 \Rightarrow 4$ achieves the proof. □

Remark 1. The proof of the implication $3 \Rightarrow 4$ can be modified as follows. Consider negative numbers $\lambda_1, \ldots, \lambda_m$ such that $\lambda_i \geq \lambda_0$, $i = \overline{1, m}$. Assume that vectors $y_{i,0}, \ldots, y_{i,k_i}$ satisfy (21) with $\lambda = \lambda_i$. Then we can continue the proof without any changes. Different collections $\lambda_1, \ldots, \lambda_m$ yield different Lyapunov functions and hence different stabilizers. The choice of these parameters provides additional possibilities for generating regulators with desired transient characteristics.

Remark 2. If $\lambda > \lambda_0$ we derive from Theorem 3 that the system of inclusions (21) is compatible for a finite k_i. To estimate the number of operations needed to construct a regulator we have to estimate $k = \max\{k_i \mid i = \overline{1, m}\}$. From now on the number $k = k(\lambda)$ is called the λ-*dimension* of the subspace I with respect to the set-valued map $x \to Cx + K$. In §5 we derive an estimate for $k(\lambda)$ in a general case. Here we only note that if the cone K is a subspace, then $\sum_{i=0}^{n-1} (C_I - \lambda E_I)^{-i} K_I$ spans the whole subspace I, thanks to the Cayley–Hamilton theorem. This implies that $k(\lambda) \leq n - 1$.

3. Stabilization of nonlinear systems. We now proceed to consider the regulator design problem for a nonlinear control system

$$\dot{x} = f(x, u), \quad u \in U. \tag{25}$$

The main stabilization problem. The following theorem contains sufficient conditions for stabilizability of system (25) at first approximation.

THEOREM 4. *Assume that the regulator design problem for the first approximation is solvable. Then there exist a neighborhood Ω of the origin and a map $u : \Omega \to U$ that satisfies the following conditions:*

1. $u(0) = u_0$.
2. *The equilibrium point $x = 0$ of the differential equation*

$$\dot{x} = f(x, u(x)) \tag{26}$$

is asymptotically stable; moreover, there exist constants $a > 0$ and $\theta > 0$ such that

$$|x(t)| \leq a|x(0)|e^{-\theta t}, \quad t \geq 0, \tag{27}$$

$$|\dot{x}(t)| \leq a|x(0)|e^{-\theta t}, \quad t \geq 0 \tag{28}$$

for any solution $x(\cdot)$ to (26) with sufficiently small $|x(0)|$.

Proof. By Theorem 1 there exist numbers $\tau > 0$, $\delta \in (0, 1)$, and a convex positively homogeneous function $V(x)$ such that $V(0) = 0$, $V(x) > 0$, if $x \neq 0$, and for any $x \in \mathrm{bd}\mathcal{M}$ ($\mathcal{M} = \{x \mid V(x) \leq 1\}$) there exists a vector $v \in Cx + K$ satisfying $x + \tau v \in \delta\mathcal{M}$.

Let x_0 and v_0 be vectors from the sets $\mathrm{bd}\mathcal{M}$ and $Cx_0 + K$, respectively, which satisfy $x_0 + \tau v_0 \in \delta\mathcal{M}$. If $\epsilon_1 > 0$ is sufficiently small, then the inclusions $x \in x_0 + \epsilon_1 B_n$, $v \in v_0 + \epsilon_1 B_n$ imply that $x + \tau v \in \frac{1}{2}(1 + \delta)\mathcal{M}$.

We need the following technical lemma.

LEMMA 6. *For any $v_0 \in Cx_0 + K$ the equality*

$$\lim_{\rho \downarrow 0,\ x \to x_0} \rho^{-1} d(\rho v_0,\ f(\rho x, U)) = 0$$

is fulfilled.

Proof. Let $\eta > 0$. There exist $u \in U$ and $\beta > 0$ such that $|v_0 - Cx_0 - \beta f(0,u)| < \eta$. If $\rho\beta < 1$, then taking into account convexity of the set $f(\rho x_0, U)$ and the inequality $|\nabla_x f(x,u)| \le l$, $x \in B_n$, we have

$$d(\rho v_0, f(\rho x, U)) \le d(\rho(Cx_0 + \beta f(0,u)), f(\rho x_0, U))$$
$$+\rho\eta + \rho l|x - x_0| \le |\rho\nabla_x f(0,u_0)x_0 + \rho\beta f(0,u)$$
$$-f(\rho x_0, u_0) - \rho\beta(f(\rho x_0, u) - f(\rho x_0, u_0))| + \rho\eta + \rho l|x - x_0|$$
$$\le |\rho\nabla_x f(0,u_0)x_0 - f(\rho x_0, u_0)| + \rho\beta(|f(0,u) - f(\rho x_0, u)|$$
$$+|f(\rho x_0, u_0)|) + \rho\eta + \rho l|x - x_0|.$$

Since η is an arbitrary positive number, dividing by ρ and taking the limit as $\rho \downarrow 0$ and $x \to x_0$, we obtain the result. □

End of the proof of Theorem 4. By the above lemma there exists $\epsilon_2 > 0$ such that the inequality

$$\rho^{-1} d(\rho v_0,\ f(\rho x, U)) \le \epsilon_1$$

holds for all $x \in x_0 + \epsilon_2 B_n$, $\rho \in (0, \epsilon_2)$. Set $\epsilon_0 = \min\{\epsilon_1, \epsilon_2\}$. Let $x \in \mathrm{bd}\mathcal{M} \cap (x_0 + \epsilon_0 B_n)$, $\rho \in (0, \epsilon_0)$. Then there exists a vector $v(\rho, x) \in f(\rho x, U)$ satisfying $|v_0 - \rho^{-1} v(\rho, x)| \le \epsilon_1$. We observe that

$$x + \tau\rho^{-1} v(\rho, x) \in \frac{1+\delta}{2}\mathcal{M}.$$

Now, we cover every point $x_0 \in \mathrm{bd}\mathcal{M}$ by such an ϵ_0-neighborhood and choose a finite subcovering. Let $\bar{\epsilon}$ be the minimal radius of a subcovering element. If $V(x) < \bar{\epsilon}$, then by the above considerations we conclude that there is a vector $v \in f(x, U)$ satisfying

$$x + \tau v \in \frac{1+\delta}{2} V(x)\mathcal{M}.$$

Let us consider the set-valued maps

$$G(x) = \left\{ v \in R^n \mid x + \tau v \in \frac{1+\delta}{2} V(x)\mathcal{M} \right\}, \quad H(x) = G(x) \cap f(x, U)$$

defined on the set $\Omega = \{x \mid V(x) < \bar{\epsilon}\}$. We take any single-valued map $\varphi(x) \in H(x)$. Since $\varphi(0) = 0$, there exists a map $u : \Omega \to U$ satisfying the conditions $u(0) = u_0$ and $\varphi(x) = f(x, u(x))$.

Consider the set-valued map

$$\Phi(x) = \bigcap_{\eta > 0} \mathrm{cl\ co}\ \varphi(x + \eta B_n).$$

Let $v \in \Phi(x)$. Then we have

$$DV(x)(v) \le \frac{[V(x + \tau v) - V(x)]}{\tau} \le -\theta V(x),$$

where $\theta = (1-\delta)/(2\tau)$. Hence any Filippov solution $x(\cdot)$ to differential equation (26) satisfies

$$V(x(t)) \le V(x(0))e^{-\theta t}, \quad t \ge 0, \tag{29}$$

whenever $x(0) \in \Omega$. By construction

$$\dot{x}(t) \in G(x) = \tau^{-1}\left(\frac{1+\delta}{2}V(x(t))\mathcal{M} - x(t)\right)$$

$$\subset \tau^{-1}e^{-\theta t}V(x(0))\left(\frac{1+\delta}{2}\mathcal{M} - \mathcal{M}\right). \tag{30}$$

From (29) and (30) we see that there exists a constant $a > 0$ such that (27) and (28) are satisfied for any solution $x(\cdot)$ to (26) with sufficiently small $|x(0)|$. The theorem is proved. □

Remark. For example, $\varphi(x) \in f(x, U)$ can be chosen from the condition

$$\min_{u \in U} V(x + \tau f(x, u)) = V(x + \tau\varphi(x))$$

if U is a compact. For many control systems this minimization problem can be easily solved using numerical methods (see [7, 8]).

An estimate for the region of attraction. Theorem 4 contains sufficient conditions for solvability of the regulator design problem. However it does not allow one to estimate the *region of attraction*, that is, the domain where the regulator is defined. Below we obtain such an estimate under additional assumptions on the control system. Assume that

1. the function $f(\cdot, u)$ is twice differentiable and $|\nabla_x f(x, u)| \le l$ and $|\nabla_{xx} f^i(x, u)| \le M$ for all $(x, u) \in R^n \times U$, $i = \overline{1, n}$;

2. the set U is compact and

$$0 < \sigma = \min\{|v| \mid v \in f(0, U),\ W(v) = 1\},$$

where W is the Minkowski function of the set $f(0, U)$.

Note that the latter assumption is fulfilled if $f(0, U)$ is a polyhedron, for example.

Under the assumptions of Theorem 4 there exist numbers $\tau > 0$, $\delta \in (0, 1)$, and a convex positively homogeneous function $V(x)$ such that $V(0) = 0$, $V(x) > 0$, if $x \neq 0$, and for any $x \in \mathrm{bd}\mathcal{M}$ ($\mathcal{M} = \{x \mid V(x) \le 1\}$) there is a vector $v \in Cx + K$ satisfying $x + \tau v \in \delta\mathcal{M}$.

Observe that there exist numbers $a > 0$ and $b > 0$ such that

$$aB_n \subset \mathcal{M} \subset bB_n.$$

For the Minkowski function $V(x)$ we have

$$\frac{|x|}{b} \le V(x) \le \frac{|x|}{a}.$$

Set

$$\bar{\rho} = \frac{a(1-\delta)\sigma}{b^2(\sqrt{n}M\tau\sigma + (4l + a/(\tau b))(\delta + |E + \tau C|))}.$$

THEOREM 5. *Assume that the conditions of Theorem* 4 *and Assumptions* 1 *and* 2 *are satisfied. Then for all* $x \in \rho\mathcal{M}$ *there exists* $u(x) \in U$ *such that*

$$x + \tau f(x, u(x)) \in \frac{1+\delta}{2}V(x)\mathcal{M}.$$

Proof. Let $\rho \in (0, \bar{\rho}]$ and $x \in \mathrm{bd}\rho\mathcal{M}$. There exist $x_0 \in \mathrm{bd}\mathcal{M}$ and $v_0 \in Cx_0 + K$ satisfying $x = \rho x_0$ and $x_0 + \tau v_0 \in \delta\mathcal{M}$, respectively. We claim that there exist $\beta \ge 0$ and

$u \in U$ such that $v_0 = Cx_0 + \beta f(0, u)$ and $W(f(0, u)) = 1$. Indeed, the vector $v_0 - Cx_0$ can be represented as $\lim \beta_i f(0, u_i)$, where $\beta_i \geq 0$, $u_i \in U$, and $W(f(0, u_i)) = 1$. We observe that $|f(0, u_i)| \geq \sigma > 0$. Consequently, the sequence β_i is bounded from above and, without loss of generality, converges to β. Since the set U is compact, taking the limit, we obtain the required statement.

Now we prove the inequality

$$d(v_0,\ \rho^{-1} f(\rho x_0, U)) \leq a\frac{1-\delta}{2\tau}. \tag{31}$$

Since the set $f(\rho x_0, U)$ is convex, we have

$$\begin{aligned} &d(Cx_0 + \beta f(0, u),\ \rho^{-1} f(\rho x_0, U)) \\ &\leq |Cx_0 + \beta f(0, u) - \rho^{-1}(\alpha f(\rho x_0, u) + (1-\alpha) f(\rho x_0, u_0))|, \end{aligned} \tag{32}$$

where $\alpha \in [0, 1]$.

The inclusion $x_0 + \tau(Cx_0 + \beta f(0, u)) \in \delta\mathcal{M}$ implies that $|x_0 + \tau(Cx_0 + \beta f(0, u))| \leq \delta b$. Hence

$$\tau\beta|f(0, u)| \leq |x_0 + \tau(Cx_0 + \beta f(0, u))| + |x_0 + \tau Cx_0| \leq \delta b + |E + \tau C|b.$$

Thus,

$$\beta \leq (\delta + |E + \tau C|)(\tau\sigma)^{-1} b. \tag{33}$$

Observe that $\beta\rho < 1$.

Set $\alpha = \beta\rho$. The right side of (32) is less than or equal to

$$\begin{aligned} &|Cx_0 - \rho^{-1} f(\rho x_0, u_0)| + \beta|f(0, u) - f(\rho x_0, u)| + \beta|f(\rho x_0, u_0)| \\ &\leq \rho\left(\frac{\sqrt{n}}{2}|x_0|^2 M + 2\beta|x_0|l\right) \leq \rho\left(\frac{\sqrt{n}}{2} b^2 M + 2\beta bl\right). \end{aligned}$$

Combining this inequality with (33) we derive (31).

Thus, there exists $v \in f(\rho x_0, U)$ such that $x_0 + (\tau/\rho)v \in \frac{1}{2}(1+\delta)\mathcal{M}$ or, in other words,

$$x + \tau f(x, u(x)) \in \frac{1+\delta}{2} V(x)\mathcal{M}.$$

The theorem is proved. □

Continuous stabilizers. Theorem 4 shows that there exists a stabilizer $u(x)$ such that the origin is the asymptotically stable equilibrium point of differential equation (26) with a discontinuous right-hand side. A natural question arises: is there a stabilizer which makes the right-hand side of (26) a continuous function? A positive answer is given by the following theorem.

THEOREM 6. *Assume that in addition to the conditions of Theorem* 4 *the set* $f(x, U)$ *is closed for any* x. *Then there exists a regulator* $u(x)$ *such that the map* $x \to f(x, u(x))$ *is continuous.*

Proof. As in the proof of Theorem 4 we observe that there exist numbers $\tau > 0, \delta \in (0, 1)$, and a convex positively homogeneous function $V(x)$ such that $V(0) = 0$, $V(x) > 0$, if $x \neq 0$, and for any x from the neighborhood of the origin Ω the set-valued map

$$H(x) = \tau^{-1}\left(\frac{1+\delta}{2} V(x)\mathcal{M} - x\right) \bigcap f(x, U),$$

where $\mathcal{M} = \{x \mid V(x) \leq 1\}$, is nonempty. Let us consider the set-valued map

$$\tilde{H}(x) = \tau^{-1}\left(\frac{3+\delta}{4}V(x)\mathcal{M} - x\right) \bigcap f(x, U).$$

Obviously, the map $\tilde{H}(x)$ is upper semicontinuous and has convex compact values. Let $v(x)$ be the projection of the origin onto $\tilde{H}(x)$. By Theorem 1.2.3 in [1, p. 49] $\tilde{H}(x)$ is lower semicontinuous. From Theorem 1.7.1 in [1, p. 70] we see that $v(x)$ is a continuous function. Hence there exists $u(x) \in U$ satisfying $v(x) = f(x, u(x))$. Asymptotic stability of the zero equilibrium position of the differential equation

$$\dot{x} = v(x) = f(x, u(x))$$

can be proved similarly to the last part of Theorem 4. □

Remark. From the computational point of view the construction of such a regulator $u(x)$ is much more complex than the procedure described in the remark after Theorem 4.

Under more restrictive assumptions on the map $u \to f(x, u)$ it is possible to establish continuity of the stabilizer $u(x)$. For example, applying the inverse function theorem we immediately obtain the following result.

COROLLARY 2. *Assume, in addition to the conditions of Theorem 6, that the map $u \to f(x, u)$ is differentiable. If there exists an inverse linear operator $(\nabla_u f(0, u_0))^{-1}$, then there exists a continuous stabilizer $u(x)$.*

Lipschitzian stabilizers. Now let us establish sufficient conditions for the existence of a Lipschitzian stabilizer.

We need the following auxiliary result.

Let $X \subset R^n$ be a compact set and let $U \subset R^k$ and $M \subset R^m$ be convex sets. Let $f : R^n \to R^m$ be a continuous function and let $\Lambda : R^k \to R^m$ be a linear operator.

LEMMA 7. *Assume that for any $x \in X$ there exists $u \in U$ satisfying*

$$f(x) + \Lambda u \in M.$$

Then for any $\epsilon > 0$ there exists a Lipschitzian map $u : X \to U$ such that

$$f(x) + \Lambda u(x) \in M + \epsilon B_m$$

whenever $x \in X$.

Proof. Let $\epsilon > 0$. Consider a point $x_0 \in X$. There exists $u_0 \in U$ satisfying

$$f(x_0) + \Lambda u_0 \in M.$$

Since f is continuous, there exists $\delta_0 > 0$ such that $|f(x) - f(x_0)| < \epsilon$ whenever $|x - x_0| < \delta_0$. We cover every point $x_0 \in X$ by such a δ_0-neighborhood and choose a finite subcovering $\{\Omega_i = x_i + \delta_i \overset{\circ}{B}_n\}_{i=1}^I$. Let $\{u_i\}_{i=1}^I$ be the corresponding vectors from U.

There exists a Lipschitz partition of unity $\{p_i(x)\}_{i=1}^I$ subordinated to this subcovering (see [1, Thm. 0.1.2]), that is, a family of functions $\{p_i(x)\}_{i=1}^I$ defined on X and satisfying the following conditions:

1. $p_i(\cdot)$ is Lipschitz for all $i = \overline{1, I}$;
2. $p_i(x) > 0$ for $x \in \Omega_i \cap X$ and $p_i(x) = 0$ for $x \in X \setminus \Omega_i$;
3. for each $x \in X$, $\sum_{i=1}^I p_i(x) = 1$.

Set

$$u(x) = \sum_{i=1}^{I} p_i(x) u_i.$$

Let $x \in X$. Then $p_i(x) > 0$ if and only if $|x - x_i| < \delta_i$. Hence

$$f(x) + \Lambda u(x) = \sum_{i=1}^{I} p_i(x)(f(x) + \Lambda u_i) \in \sum_{i=1}^{I} p_i(x)(f(x_i) + \Lambda u_i + \epsilon B_m)$$
$$\subset \sum_{i=1}^{I} p_i(x)(M + \epsilon B_m) = M + \epsilon B_m.$$

The lemma is proved. □

Assume that the right-hand side of (25) is affine, that is,

$$f(x, u) = f_0(x) + \sum_{q=1}^{N} u^q f_q(x) \tag{34}$$

with $u = (u^1, \ldots, u^N) \in U$, where $U \subset R^N$ is convex. Denote by $F(x)$ the matrix whose columns are the vectors $f_1(x), \ldots, f_N(x)$.

THEOREM 7. *Assume that the regulator design problem for the first approximation of system* (25) *with affine right-hand side* (34) *is solvable. Then there exists a Lipschitzian regulator* $u(x)$.

Proof. Consider the control system

$$\dot{x} = Cx + F(0)(u - u_0), \ u \in U.$$

By Theorem 1 there exist a convex positively homogeneous function $V(x)$, numbers $\tau > 0$, $\delta \in (0, 1)$, and a map $\tilde{u} : \mathcal{M} \to U$ $(\mathcal{M} = \{x \mid V(x) \leq 1\})$ such that $V(0) = 0$, $V(x) > 0$, if $x \neq 0$ and

$$x + \tau(Cx + F(0)(\tilde{u}(x) - u_0)) \in \delta\mathcal{M}$$

whenever $x \in \mathrm{bd}\mathcal{M}$. By Lemma 7 there exists a Lipschitzian map $\hat{u} : \mathrm{bd}\mathcal{M} \to U$ such that

$$x + \tau(Cx + F(0)(\hat{u}(x) - u_0)) \in \frac{1+\delta}{2}\mathcal{M}$$

for all $x \in \mathrm{bd}\mathcal{M}$. Let $V(x) < 1$. Then

$$x + \tau\left(Cx + V(x)F(0)\left(\hat{u}\left(\frac{x}{V(x)}\right) - u_0\right)\right) \in \frac{1+\delta}{2}V(x)\mathcal{M}.$$

Set

$$u(x) = (1 - V(x))u_0 + V(x)\hat{u}\left(\frac{x}{V(x)}\right).$$

If $V(x)$ is sufficiently small, then we have

$$f_0(x) + F(x)u_0 - Cx \in \frac{1-\delta}{8}V(x)\mathcal{M}$$

and

$$V(x)(F(x) - F(0))\left(\hat{u}\left(\frac{x}{V(x)}\right) - u_0\right) \in \frac{1-\delta}{8}V(x)\mathcal{M}.$$

Therefore

$$x + \tau(f_0(x) + F(x)u(x))$$

$$= x + \tau\left(f_0(x) + F(x)u_0 + V(x)F(x)\left(\hat{u}\left(\frac{x}{V(x)}\right) - u_0\right)\right)$$

$$\in x + \tau\left(Cx + V(x)F(0)\left(\hat{u}\left(\frac{x}{V(x)}\right) - u_0\right)\right) + \frac{1-\delta}{4}V(x)\mathcal{M} \subset \frac{3+\delta}{4}V(x)\mathcal{M}.$$

Consequently $V(x)$ is a Lyapunov function for the differential equation

$$\dot{x} = f_0(x) + F(x)u(x).$$

The theorem is proved. □

Note that the right-hand side of the first approximation is always affine. Therefore we obtain the following corollary.

COROLLARY 3. *Assume that the regulator design problem for the first approximation* (16) *is solvable. Then there exists a Lipschitzian stabilizer* $w(x)$ *for system* (16).

Now apply Theorem 7 to the general case. Suppose that the right-hand side of (25) is differentiable in u, and that U is a convex set. Consider the system

$$\dot{x} = f(x, u_0) + \nabla_u f(x, u_0)(u - u_0), \quad u \in U. \tag{35}$$

COROLLARY 4. *Assume that the regulator design problem for the first approximation of system* (35) *is solvable. Then there exists a Lipschitzian stabilizer* $u(x)$ *that solves the regulator design problem for system* (25).

Proof. By Theorems 1 and 7 there exist a convex positively homogeneous function $V(x)$, numbers $\tau > 0$, $\delta \in (0, 1)$, and a Lipschitzian map $u : \mathcal{M} \to U$ $(\mathcal{M} = \{x \mid V(x) \leq 1\})$ such that $V(0) = 0$, $V(x) > 0$, if $x \neq 0$ and

$$x + \tau(f(x, u_0) + \nabla_u f(x, u_0)(u(x) - u_0)) \in \delta V(x)\mathcal{M}$$

for all x sufficiently close to zero. If $V(x)$ is sufficiently small, then we have

$$x + \tau f(x, u(x)) = x + \tau(f(x, u_0) + f(x, u(x)) - f(x, u_0))$$

$$\in x + \tau(f(x, u_0) + \nabla_u f(x, u_0)(u(x) - u_0)) + \frac{1-\delta}{2}V(x)\mathcal{M} \subset \frac{1+\delta}{2}V(x)\mathcal{M}.$$

Consequently, $V(x)$ is a Lyapunov function for the differential equation

$$\dot{x} = f(x, u(x)). \qquad \square$$

Piecewise constant controls. In practical problems a piecewise constant control law is usually used instead of the stabilizer $u(x)$ considered before. The control is chosen in discrete moments of time $0, \sigma, 2\sigma, \ldots$, where $\sigma > 0$, and we have

$$u_\sigma(t) \equiv u(x(k\sigma)), \quad t \in [k\sigma, (k+1)\sigma),$$

where $u(x)$ is the regulator constructed in Theorem 4. The following theorem substantiates this control law.

THEOREM 8. *If $\sigma > 0$ is sufficiently small, then any trajectory $x(\cdot)$ of the differential equation*

$$\dot{x} = f(x, u_\sigma(t)) \tag{36}$$

tends to zero whenever $|x(0)|$ is sufficiently small.

Proof. Let $x(\cdot)$ be a solution to (36) and $t \in [k\sigma, (k+1)\sigma)$. Since $V(x) \le |x|/a$, the function $V(x)$ is Lipschitzian with the constant $1/a$. Indeed,

$$V(x_1) = V(x_2 + x_1 - x_2) \le V(x_2) + V(x_1 - x_2) \le V(x_2) + \frac{|x_1 - x_2|}{a}.$$

If $x(t)$ is sufficiently close to zero, then we have

$$\frac{d}{dt}V(x(t)) \le \frac{1}{\tau}[V(x(t) + \tau f(x(t), u_\sigma(t))) - V(x(t))]$$

$$\le \frac{1}{\tau}[V(x(k\sigma) + \tau f(x(k\sigma), u(x(k\sigma)))) - V(x(k\sigma))] + \frac{2+\tau l}{a\tau}|x(t) - x(k\sigma)|$$

$$\le -\frac{1-\delta}{2\tau}V(x(k\sigma)) + \frac{2+\tau l}{a\tau}|x(t) - x(k\sigma)|$$

(see the proof of Theorem 4). Since

$$|x(t) - x(k\sigma)| \le l\int_{k\sigma}^{t} |x(s) - x(k\sigma)|ds + \sigma|f(x(k\sigma), u(x(k\sigma))|,$$

we derive from the Gronwall inequality

$$|x(t) - x(k\sigma)| \le \sigma|f(x(k\sigma), u(x(k\sigma))|e^{l\sigma}.$$

Moreover, if N is a sufficiently large number, then $|f(x, u(x))| \le NV(x)$ for all $x \in \Omega$. Thus we obtain

$$\frac{d}{dt}V(x(t)) \le -\frac{1-\delta}{2\tau}V(x(k\sigma)) + \sigma\frac{2+\tau l}{a\tau}Ne^{l\sigma}V(x(k\sigma)).$$

Let $\sigma = \min\{1,\ (1-\delta)ae^{-l}/(4(2+\tau l)N)\}$. Then we have

$$\frac{d}{dt}V(x(t)) \le -\frac{1-\delta}{4\tau}V(x(k\sigma)), \quad \text{if } t \in [k\sigma, (k+1)\sigma).$$

This implies that $V(x(t)) < V(x(k\sigma))$, if $t \in (k\sigma, (k+1)\sigma)$. Consequently we obtain

$$\frac{d}{dt}V(x(t)) \le -\frac{1-\delta}{4\tau}V(x(t)).$$

Thus, any trajectory $x(\cdot)$ of differential equation (36) tends to zero whenever $|x(0)|$ is sufficiently small. □

4. Examples. In this section we study the regulator design problem for two simple control systems considered in the introduction.

Stabilization of an oscillator subjected to a unilateral force. The motion of an oscillator subjected to a unilateral force is described by the following equations:

$$\begin{aligned} &\dot{x}^1 = x^2, \\ &\dot{x}^2 = -x^1 + u, \\ &0 \le u \le 1. \end{aligned} \tag{37}$$

The control $u_0 = 0$ corresponds to the equilibrium position. Thus the first approximation of system (37) is given by

$$\begin{aligned} &\dot{x}^1 = x^2, \\ &\dot{x}^2 = -x^1 + w, \\ &w \ge 0. \end{aligned}$$

In this case

$$C = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

and

$$K = \{(w^1, w^2) \mid w^1 = 0,\ w^2 \ge 0\}.$$

The transposed matrix C^T has neither nontrivial subspaces nor eigenvectors contained in the conjugate cone $K^* = \{(w^{1*}, w^{2*}) \mid w^{2*} \ge 0\}$. Therefore, by Theorem 3 for all negative λ the cone $L_k(\lambda)$ coincides with the hole space for some $k = k(\lambda)$. By Corollary 1 and Theorem 4 the stabilization problem for system (37) is solvable.

Let us show that $k(\lambda) \to \infty$ as $\lambda \to \infty$. Indeed, by definition

$$L_k(\lambda) = -\operatorname{co} \bigcup_{i=0}^{k} (C - \lambda E)^{-i} K.$$

If $\lambda < 0$, then we have

$$(C - \lambda E)^{-i} = \left(\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} + |\lambda| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)^{-i}$$

$$= \left(\frac{1}{\lambda^2 + 1} \right)^i \left(\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} + |\lambda| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)^i = \left(\frac{1}{\lambda^2 + 1} \right)^i \sum_{j=0}^{i} C_i^j |\lambda|^{i-j} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^j$$

$$= \left(\frac{1}{\lambda^2 + 1} \right)^i \left[\left(\sum_{p=0} (-1)^p C_i^{2p} |\lambda|^{i-2p} \right) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right.$$

$$\left. + \left(\sum_{p=0} (-1)^p C_i^{2p+1} |\lambda|^{i-2p-1} \right) \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \right]. \tag{38}$$

Denote $e_1 = (1, 0)$, $e_2 = (0, 1)$, $l_i(\lambda) = -(C - \lambda E)^{-i} e_2$. Observe that the cone $L_k(\lambda)$ is spanned by the convex hull of the points $\{l_0(\lambda), \ldots, l_k(\lambda)\}$. From (38) we obtain

$$\langle e_1, l_i(\lambda) \rangle = \left(\frac{1}{\lambda^2 + 1} \right)^i \left(\sum_{p=0} (-1)^p C_i^{2p+1} |\lambda|^{i-2p-1} \right).$$

Obviously, for any k there exists a $\Lambda < 0$ such that $\langle e_1, l_i(\lambda) \rangle \ge 0$, $i = \overline{1, k}$, whenever $\lambda < \Lambda$. Hence $L_k(\lambda) \ne R^2$. Thus $k(\lambda) \to \infty$ as $\lambda \to -\infty$.

Stabilization of a missile uniform motion. Now consider the model of a missile (see the introduction). The motion of the missile's mass center is described by the following equations:

$$
\begin{aligned}
&\ddot{z}^1 = -\sigma \dot{z}^1 + \frac{u^1 \dot{z}^1 - u^2 \dot{z}^2}{\sqrt{(\dot{z}^1)^2 + (\dot{z}^2)^2}},\\
&\ddot{z}^2 = -\sigma \dot{z}^2 + \frac{u^1 \dot{z}^2 + u^2 \dot{z}^1}{\sqrt{(\dot{z}^2)^2 + (\dot{z}^2)^2}},\\
&(u^1, u^2) \in U = \{(u^1, u^2) \mid \sqrt{(u^1)^2 + (u^2)^2} \le b,\ \frac{u^2}{u^1} \le \tan \eta,\ u^1 \ge 0\},
\end{aligned}
\tag{39}
$$

where $\sigma > 0$ stands for a coefficient of air resistance, b is the maximal thrust, and η is the maximal angle between the longitudinal axis of the missile and that of the jet propulsion. Our aim is to stabilize the uniform motion of the object under consideration along the z^1-axis with the constant maximal speed $\dot{z}^1 = b/\sigma$. The control $(u_0^1, u_0^2) = (b, 0)$ corresponds to the uniform motion we are to stabilize.

Let us introduce new variables $x^1 = \dot{z}^1 - b/\sigma$, $x^2 = z^2$, $x^3 = \dot{z}^2$. System (39) now can be written in the form

$$
\begin{aligned}
&\dot{x}^1 = -\sigma \left(x^1 + \frac{b}{\sigma}\right) + \frac{u^1 (x^1 + \frac{b}{\sigma}) - u^2 x^3}{\sqrt{(x^1 + \frac{b}{\sigma})^2 + (x^3)^2}},\\
&\dot{x}^2 = x^3,\\
&\dot{x}^3 = -\sigma x^3 + \frac{u^1 x^3 + u^2 (x^1 + \frac{b}{\sigma})}{\sqrt{(x^1 + \frac{b}{\sigma})^2 + (x^3)^2}},\\
&(u^1, u^2) \in U,
\end{aligned}
\tag{40}
$$

and we obtain the regulator design problem in the usual form: to find a control $u(x)$ stabilizing system (40) to the zero equilibrium position.

Consider the first approximation of system (40)

$$
\begin{pmatrix} \dot{x}^1 \\ \dot{x}^2 \\ \dot{x}^3 \end{pmatrix} = \begin{pmatrix} -\sigma & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix} + \begin{pmatrix} w^1 \\ w^2 \\ w^3 \end{pmatrix}, \tag{41}
$$

where $(w^1, w^2, w^3) \in K = \{(w^1, w^2, w^3) \mid w^1 \le 0,\ w^2 = 0\}$. Obviously, the transposed matrix

$$
C^T = \begin{pmatrix} -\sigma & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}
$$

has the eigenvector $(-1,0,0)$ corresponding to the eigenvalue $-\sigma$, the eigenvector $(0,0,1)$ corresponding to the eigenvalue 0, and the invariant subspace spanned by the vectors $(0,0,1)$ and $(0,1,0)$ corresponding to one Jordan block of the matrix C^T with the eigenvalue 0. Since

$$
K^* = \{(w^{1*}, w^{2*}, w^{3*}) \mid w^{1*} \le 0,\ w^{3*} = 0\},
$$

we observe that the invariant space spanned by the vectors $(0, 0,1)$ and $(0,1,0)$ is not contained in the cone K^*. The vector $(-1,0,0)$ belongs to the cone K^*, but it corresponds to the negative eigenvalue $-\sigma$. By Corollary 1 and Theorem 4 the regulator design problem for system (40) is solvable.

In the previous example the λ-dimension depends on λ. In the example under consideration, for all $\lambda \in (-\sigma, 0)$ we have

$$L_1(\lambda) = \{(y^1, y^2, y^3) \mid \lambda y^2 = y^3,\ (\sigma + \lambda)y^1 \leq 0\} - K = R^3.$$

The results of computer simulation for these two examples can be found in [7, 8].

5. Regulator design under controllability conditions. In this section we consider the regulator design problem for the linear control system

$$\dot{x} = Cx + u,\ u \in K, \tag{42}$$

where K is a closed convex cone. We suppose that system (42) is controllable.

Recall the following result (Korobov [12] and Aubin, Frankowska, and Olech [2]).

THEOREM 9. *The following conditions are equivalent:*

1. *Linear control system* (42) *is controllable.*
2. *The matrix C^T has neither eigenvectors in K^* nor proper invariant subspaces contained in K^*.*

Remark. Another necessary and sufficient condition of controllability appeared in Brammer [6]. The case of a single-input system was considered by Saperstone and Yorke [17].

Theorem 9 combined with Corollary 1 implies that the regulator design problem for system (42) is solvable. Moreover, by Theorem 3 for any $\lambda < 0$ there exists a positive integer $k = k(\lambda)$ such that

$$L_k(\lambda) = -\operatorname{co}\bigcup_{i=0}^{k}(C - \lambda E)^{-i}K = R^n.$$

As was pointed out in Remark 2 at the end of §2, the λ-dimension characterizes the efficiency of the regulator design method. Our goal is to estimate the λ-dimension of R^n with respect to the set-valued map $x \to Cx + K$.

An estimate for the λ-dimension. To clarify the idea of the estimate note that for $\lambda < 0$ the cone

$$L_k(\lambda) = -\sum_{i=0}^{k}(C - \lambda E)^{-i}K = -\sum_{i=0}^{k}(E + |\lambda|^{-1}C)^{-i}\left(\bigcup_{\alpha \geq 0}\alpha W\right),$$

where $W = K \cap B_n$, can be considered as a cone spanned by the reachable set of the discrete control system

$$x_{i+1} = \left(E + \frac{1}{|\lambda|}C\right)^{-1}x_i - \frac{1}{|\lambda|}u,\ u \in W,\ i = \overline{0, k-1}. \tag{43}$$

The discrete system (43) approximates the control system

$$\dot{x} = -Cx - u,\ u \in W,\ t \in [0, T]. \tag{44}$$

The system

$$\dot{x} = Cx + u,\ u \in W, \tag{45}$$

is locally controllable (see [2, Lem. 5.7]). For systems (43) and (44) we denote the sets reachable from the origin by $\mathcal{A}_k(0)$ and $\mathcal{A}_T(0)$, respectively. Since system (45) is controllable,

there exist $T > 0$ and $\eta > 0$ such that $\eta B_n \subset \mathcal{A}_T(0)$. If k and $|\lambda|$ are sufficiently large, then $0 \in \operatorname{int}\mathcal{A}_k(0)$. The latter amounts to saying $L_k(\lambda) = R^n$. Thus, we can estimate the number $k = k(\lambda)$ via the time T.

Consider a linear control system

$$\dot{x} = Cx + u, \ u \in U, \tag{46}$$

where C is an $n \times n$ matrix and U is a convex compact set satisfying $U \subset B_n$. Consider the discrete control system

$$x_{i+1} = (E - \tau C)^{-1}x_i + \tau u_i, \ u_i \in U, \ i = \overline{0, k-1}. \tag{47}$$

The following lemma shows that any trajectory $x(\cdot)$ of system (46) can be approximated by a trajectory of discrete system (47).

LEMMA 8. *Let a number $\tau > 0$ and a positive integer k be such that $T/\tau < k \le (T/\tau)+1$ and $T/k < |C|^{-1}$. Then for any trajectory $x(\cdot)$ of system* (46) *with $x(0) = 0$ there exists a trajectory of discrete system* (47) $x_0, \dots, x_k$, $x_0 = 0$ *such that*

$$|x_k - x(T)| \le \frac{T}{k}(2 + |C|)^2 e^{2|C|T}.$$

Proof. Let $x(\cdot)$ be a trajectory of (46) with $x(0) = 0$. Applying the mean value theorem, from (46) we have

$$\begin{aligned} x(i\tau + \tau) - x(i\tau) &\in C\int_{i\tau}^{i\tau+\tau} x(t)dt + \tau U \\ &= C\int_{i\tau}^{i\tau+\tau}(x(i\tau+\tau) + \int_{i\tau+\tau}^{t}\dot{x}(s)ds)dt + \tau U \\ &= \tau C x(i\tau+\tau) + \tau U + C\int_{i\tau}^{i\tau+\tau}\int_{i\tau+\tau}^{t}\dot{x}(s)dsdt \end{aligned} \tag{48}$$

for all $i = \overline{0, k-1}$. Obviously,

$$|\dot{x}| \le |C|\int_0^t e^{(t-s)|C|}ds + 1 = e^{t|C|}. \tag{49}$$

From (48) and (49) we have

$$\begin{aligned} &x(i\tau+\tau) \\ &\in (E-\tau C)^{-1}x(i\tau) + \tau(E-\tau C)^{-1}U + (E-\tau C)^{-1}|C|\int_{i\tau}^{i\tau+\tau}\int_{t}^{i\tau+\tau} e^{s|C|}dsdtB_n \\ &\subset (E-\tau C)^{-1}x(i\tau) + \tau U + \tau(\tau|C| + \tau^2|C|^2 + \cdots)B_n \\ &+ |E-\tau C|^{-1}e^{T|C|}(\tau - |C|^{-1}(1 - e^{-\tau|C|}))B_n \subset (E-\tau C)^{-1}x(i\tau) + \tau U + \tau^2\gamma B_n, \end{aligned}$$

where $\gamma = |C|(1 + \exp(T|C|))/(1 - \tau|C|)$.

Define the trajectory $x_0, \dots, x_k$ of discrete system (47) approximating the trajectory $x(\cdot)$ by induction. Set $x_0 = 0$. If x_i is already determined, we choose $x_{i+1} \in (E - \tau C)^{-1}x_i + \tau U$ from the condition

$$|x_{i+1} - x(i\tau+\tau)| = d(x(i\tau+\tau), (E-\tau C)^{-1}x_i + \tau U).$$

Observe that

$$\begin{aligned} |x_{i+1} - x(i\tau+\tau)| &\le |(E-\tau C)^{-1}||x_i - x(i\tau)| + \tau^2\gamma \\ &\le (1 - \tau|C|)^{-1}|x_i - x(i\tau)| + \tau^2\gamma. \end{aligned}$$

By induction we have

$$|x_k - x(T)| \le ((1-\tau|C|)^{-(k-1)} + \cdots + (1-\tau|C|)^{-1} + 1)\tau^2\gamma$$
$$= \tau^2\gamma\frac{(1-\tau|C|)^{-k}-1}{(1-\tau|C|)^{-1}-1} \le \frac{T}{k}(2+|C|)^2 e^{2|C|T}.$$

The lemma is proved. □

THEOREM 10. *Assume that there exist* $T > 0$ *and* $\eta > 0$ *such that* $\eta B_n \subset \mathcal{A}_T(0)$. *If* $\lambda < -\max\{|C|, \eta^{-1}(2+|C|)^2\exp(2|C|T)\}$ *and* $k > T|\lambda|$, *then*

$$L_k(\lambda) = R^n.$$

Proof. For any point $b \in \eta B_n$ one can find a trajectory $x(\cdot)$ of control system (44) with $x(0) = 0$ and satisfying $x(T) = b$. Lemma 8 implies that the reachable set $\mathcal{A}_k(0)$ of discrete system (43) satisfies $\eta B_n \subset \mathcal{A}_k(0) + \beta B_n$, where $\beta < \eta$. Let $x^* \in \mathrm{bd}B_n$. Then we have

$$S(x^*, (\eta-\beta)B_n) = \eta - \beta = S(x^*, \eta B_n) - \beta \le S(x^*, \mathcal{A}_k(0) + \beta B_n) - \beta$$
$$= S(x^*, \mathcal{A}_k(0)) + S(x^*, \beta B_n) - \beta = S(x^*, \mathcal{A}_k(0)).$$

Hence, $(\eta-\beta)B_n \subset \mathcal{A}_k(0)$. Consequently, $L_k(\lambda) = \mathrm{cone}\mathcal{A}_k(0) = R^n$. □

Remark. We have reproved Theorem 3 for $\lambda_0 = -\infty$.

Single-input control systems. Now we proceed to study the regulator design problem for a single-input linear control system

$$\dot{x} = Cx + bu, \quad u \in R, \tag{50}$$

where $b \in R^n$ is a vector. We suppose that system (50) is controllable. By Kalman's criterion the controllability is equivalent with the linear independence of the vectors $b, Cb, \ldots, C^{n-1}b$. Let Σ be a polyhedron satisfying $0 \in \mathrm{int}\Sigma$. Following the proof of Theorem 1, for any $\lambda < 0$ we generate a polyhedron $\mathcal{M}$. Let $V(x)$ be the Minkowski function of $\mathcal{M}$. For system (50) we can analytically solve the minimization problem

$$\min_{u\in R} V(x + \tau(Cx + bu)) \tag{51}$$

considered in §3 (see the remark after Theorem 4) to determine the stabilizer $u(x)$. The result is contained in the following theorem.

THEOREM 11. *Let* $\lambda < 0$ *and* $\tau = 1/|\lambda|$. *Assume that the matrix* $C - \lambda E$ *is nonsingular. Then the regulator obtained as the solution to minimization problem* (51) *is a linear feedback* $u(x) = \langle c, x\rangle$ *such that the linear system*

$$\dot{x} = (C + bc^T)x \tag{52}$$

has the spectrum $\lambda_1 = \lambda_2 = \cdots = \lambda_n = \lambda$.

Proof. Since system (50) is controllable, the vectors $b_i = (C-\lambda E)^{-i}b$, $i = \overline{0, n-1}$, form a basis in R^n. Indeed, by Theorem 3 there exists a positive integer k such that

$$\mathrm{co}\bigcup_{i=0}^{k}(C-\lambda E)^{-k}(Rb) = R^n.$$

From the Cayley–Hamilton theorem $k = n-1$. Thus, for any $x \in R^n$ there exist β^i, $i = \overline{0, n-1}$ such that

$$x = \sum_{i=0}^{n-1}\beta^i b_i. \tag{53}$$

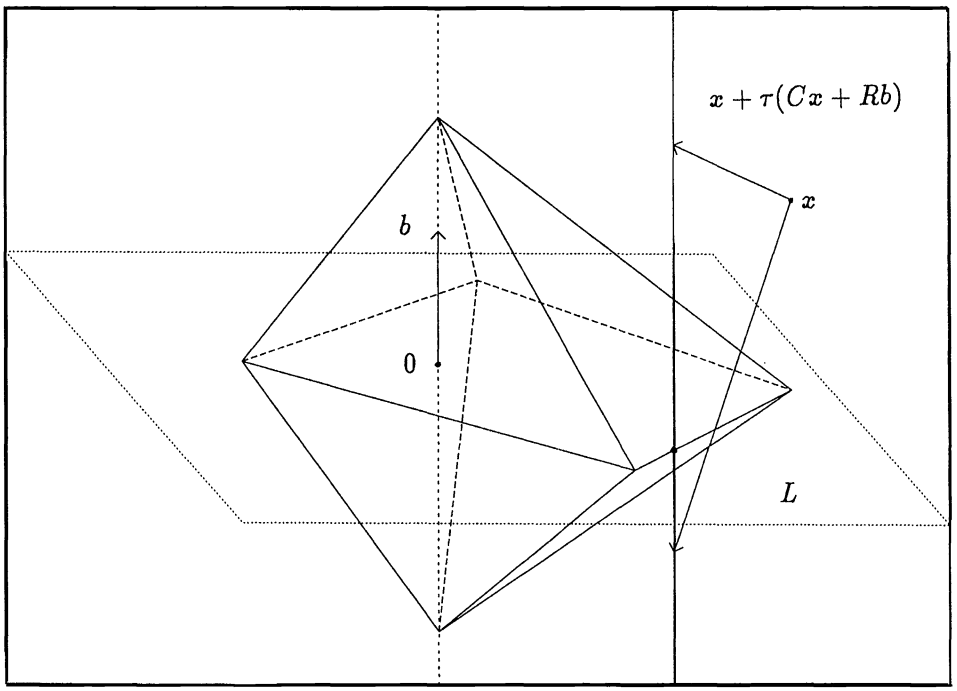


FIG. 2. *The stabilizer for a single-input system.*

Thus any vertex of the polyhedron Σ can be represented in the form (53). This implies that the polyhedron $\mathcal{M}$ is the convex hull of a polyhedron contained in the subspace L spanned by the vectors $b_1, \ldots, b_{n-1}$ and of two points contained in the one-dimensional subspace spanned by $b_0 = b$. Consequently, the Minkowski function V in (51) has the minimum at the point $u(x)$ such that the vector $x + \tau(Cx + bu(x))$ belongs to the subspace L (see Figure 2). To determine $u(x)$ observe that for all $i = \overline{0, n-1}$ we have

$$Cb_i = \sum_{k=0}^{n-1} \gamma_i^k b_k. \tag{54}$$

From (53) and (54) we have

$$x + \tau(Cx + bu)$$
$$= \sum_{i=1}^{n-1} \beta^i b_i + \tau \sum_{i=0}^{n-1} \beta^i \sum_{k=1}^{n-1} \gamma_i^k b_k + \left(\beta^0 + \tau \sum_{i=0}^{n-1} \beta^i \gamma_i^0 + \tau u \right) b_0.$$

Hence

$$u(x) = -\sum_{i=0}^{n-1} \beta^i \gamma_i^0 - \beta^0/\tau. \tag{55}$$

Thus $u(x)$ is a linear function of x, i.e., $u(x) = \langle c, x \rangle$ for some $c \in R^n$. Differentiating (53) and using (54) and (55), we obtain

$$\sum_{i=0}^{n-1} \dot{\beta}^i b_i = \sum_{i=0}^{n-1} \beta^i \sum_{k=1}^{n-1} \gamma_i^k b_k - \beta^0 b_0/\tau. \tag{56}$$

Since the vectors b_i, $i = \overline{0, k}$, are linear independent, from (56) we have

$$\dot{\beta}^0 = -\beta^0/\tau, \tag{57}$$

$$\dot{\beta}^i = \sum_{j=0}^{n-1} \beta^j \gamma_j^i, \qquad i = \overline{1, n-1}. \tag{58}$$

Subtracting λb_i from both sides of (54), we obtain

$$b_{i-1} = \sum_{k=0}^{n-1} \gamma_i^k b_k - \lambda b_i.$$

Consequently, $\gamma_i^i = \lambda$, $\gamma_i^{i-1} = 1$, $i = \overline{1, n-1}$, and $\gamma_i^k = 0$, $k = \overline{1, n-1}$, $i = \overline{1, n-1}$, $k \neq i$, $k \neq i-1$. Since $\tau = 1/|\lambda|$, (57) and (58) can be written in the form

$$\frac{d}{dt}\begin{pmatrix} \beta^0 \\ \beta^1 \\ \vdots \\ \vdots \\ \beta^{n-1} \end{pmatrix} = \begin{pmatrix} \lambda & 0 & 0 & \dots & 0 \\ \gamma_0^1 & \lambda & 1 & \dots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 1 \\ \gamma_0^{n-1} & 0 & 0 & \dots & \lambda \end{pmatrix} \begin{pmatrix} \beta^0 \\ \beta^1 \\ \vdots \\ \vdots \\ \beta^{n-1} \end{pmatrix}.$$

The spectrum of the system is, obviously, $\lambda_1 = \lambda_2 = \cdots = \lambda_n = \lambda$. □

There exists a linear feedback $u(x) = \langle c, x \rangle$ such that the closed-loop system

$$\dot{x} = (C + bc^T)x \tag{59}$$

has an arbitrary given spectrum (see [19, Thm. 4.1.7], for example). Consequently the solutions of (59) can damp to the zero equilibrium position with an arbitrary given speed determined by the real parts of the spectrum. It turns out that the trajectories of closed-loop systems with fast damping significantly deviate from the equilibrium position during a short initial time interval. In other words, trajectories have "peaks" before fast damping. This effect has been largely studied (see Polotski [15], Izmailov [11], and Sussmann and Kokotovic [20], for example). Let us show that the regulator obtained as the solution to (51) ensures a minimal overshoot among linear feedback control laws with the spectra $\{\lambda_1, \dots, \lambda_n\}$ satisfying the conditions $\mathrm{Re}\lambda_i \leq \lambda < 0$, $|\lambda_i - \lambda| < \epsilon$, where $\epsilon > 0$ is sufficiently small.

Recall the following result from Izmailov [11].

Denote by $x(t, x_0)$ the solution to (59) satisfying $x(0, x_0) = x_0$.

THEOREM 12. *There exists $\gamma = \gamma(C, b) > 0$ such that if the spectrum $\{\lambda_1, \dots, \lambda_n\}$ of system* (59) *satisfies* $\mathrm{Re}\lambda_i \leq \lambda < 0$, $i = \overline{1, n}$, *then we have*

$$\sup_{0 \leq t \leq 1/\lambda} \sup_{x_0 \in \mathrm{bd} B_n} |x(t, x_0)| \geq \gamma \frac{|\lambda_1 \dots \lambda_n|}{\max |\lambda_i|}.$$

As follows from Theorem 12, the "peak" effect always takes place in linear systems. An important practical problem is to choose a linear feedback with the minimal overshoot and the damping speed greater than or equal to a given value. This problem is very complex and its complete solution is unknown, but some results in this direction have been obtained by Polotski [15]. The following two theorems can be proved with the help of techniques similar to the ones developed in [15].

Consider linear control system (50). Let the linear feedback $u = \langle c, x \rangle$ be such that linear system (59) has the spectrum $\{\lambda_1, \dots, \lambda_n\}$, where $\lambda_1, \dots, \lambda_n$ are different complex numbers. Denote the matrix $A = C + bc^T$ by $A(\lambda_1, \dots, \lambda_n)$ to emphasize its dependence on $\lambda_1, \dots, \lambda_n$.

To characterize the "peak" effect introduce the function

$$p(\lambda_1, \ldots, \lambda_n) = \max_{t \geq 0} \max_{|x| \leq 1} |e^{A(\lambda_1, \ldots, \lambda_n)t} x|$$

defined for complex numbers satisfying $\mathrm{Re}\lambda_i < 0$, $i = \overline{1, n}$. The norm in the definition of p is not necessarily Euclidean.

THEOREM 13. *If* $\lambda_1 = \cdots = \lambda_n = \lambda < 0$, *then we have*

$$p(\lambda, \ldots, \lambda) \leq (\mathrm{const})\lambda^{n-1}.$$

Theorem 12 shows that

$$p(\lambda_1, \ldots, \lambda_n) \geq (\mathrm{const})\lambda^{n-1}$$

whenever the spectrum $\{\lambda_1, \ldots, \lambda_n\}$ satisfies $\mathrm{Re}\lambda_i \leq \lambda < 0$. Thus, from Theorem 13 we see that the spectrum $\lambda_1 = \cdots = \lambda_n = \lambda$ guarantees the minimal overshoot at least up to a factor. It is natural to expect that for any spectrum $\{\lambda_1, \ldots, \lambda_n\}$ satisfying $\mathrm{Re}\lambda_i \leq \lambda < 0$ we have

$$p(\lambda, \ldots, \lambda) \leq p(\lambda_1, \ldots, \lambda_n)$$

whenever $|\lambda|$ is sufficiently large. Unfortunately, we have no proof of this hypothesis, but its local variant can be proved.

Note that the function $p(\cdot)$ is directionally differentiable.

THEOREM 14. *For any complex vector* $(\lambda_1, \ldots, \lambda_n)$ *satisfying* $\mathrm{Re}\lambda_i \leq 0$ *and any real number* $\lambda < 0$ *there exists a number* $\nu < 0$ *such that*

$$Dp(\lambda, \ldots, \lambda)(\lambda_1, \ldots, \lambda_n) = \nu \frac{\mathrm{Re}\lambda_1 + \cdots + \mathrm{Re}\lambda_n}{n}$$

whenever $|\lambda|$ *is sufficiently large.*

From this theorem we derive a local variant of the optimality hypothesis.

COROLLARY 5. *Let* $\lambda_1, \ldots, \lambda_n$ *be a spectrum. If* $\mathrm{Re}\lambda_i < 0$ *and* $\lambda < 0$, *then*

$$Dp(\lambda, \ldots, \lambda)(\lambda_1, \ldots, \lambda_n) > 0$$

whenever $|\lambda|$ *is sufficiently large.*

We know from Theorem 11 that in the case under consideration the regulator design algorithm (51) generates a linear feedback corresponding to the spectrum $\{\lambda, \ldots, \lambda\}$ and, hence, generates a linear feedback with the *minimal overregulation*, at least in a local sense.

6. Weak asymptotic stability and stabilizability. In this section we investigate the connection between weak asymptotic stability and stabilizability and describe the class of control systems stabilizable at first approximation.

Recall the notion of weak asymptotic stability (see [9] for details). The equilibrium position $x = 0$ of a control system

$$\dot{x} = f(x, u), \quad u \in U \tag{60}$$

is called *weakly asymptotically stable* if, given $\epsilon > 0$, there exists $\delta > 0$ such that for any $x_0 \in \delta B_n$ *at least one* trajectory $x(\cdot)$ of (60) with $x(0) = x_0$ satisfies the conditions $|x(t)| < \epsilon$ for all $t \geq 0$ and $\lim x(t) = 0$ as $t \to \infty$.

Obviously, if the stabilization problem for system (60) is solvable, then the origin is its weakly asymptotically stable equilibrium position. As we know from Theorem 1, for the first

approximation these conditions are equivalent. In general stabilizability cannot be derived from weak asymptotic stability.

Example. Consider the following control system with the control set consisting of one point:

$$(\dot{x}^1, \dot{x}^2) = \begin{cases} (x^1, x^1(x^2)^{1/3}), & (x^1, x^2) \in \Omega_1, \\ -(x^1, x^2), & (x^1, x^2) \in \Omega_2, \end{cases} \tag{61}$$

where $\Omega_1 = \{(x^1, x^2) \mid x^1 \geq 0,\ |x^2| < (x^1)^2\}$, $\Omega_2 = R^2 \setminus \Omega_1$. Let $(x_0^1, x_0^2) \in \Omega_2$. Then the trajectory

$$(x^1(t), x^2(t)) = e^{-t}(x_0^1, x_0^2)$$

of system (61) obviously tends to zero when $t \to \infty$.

Let $(x_0^1, x_0^2) \in \Omega_1$ and $x_0^2 \geq 0$ (the case $x_0^2 \leq 0$ is similar to that one). Then the trajectory

$$(x^1(t), x^2(t)) = (x_0^1 e^t, \left(\frac{2}{3}\left(x_0^1(e^t - 1) + \frac{3}{2}(x_0^2)^{2/3}\right)^{3/2}\right)$$

satisfies $x^2(t) = (\frac{2}{3}(x^1(t) + \frac{3}{2}(x_0^2)^{2/3} - x_0^1))^{3/2}$. We see that there exists t^* such that $x^2(t^*) = (x^1(t^*))^2$ if (x_0^1, x_0^2) is sufficiently close to the origin. Hence $(x^1(t^*), x^2(t^*)) \in \Omega_2$. Thus, the equilibrium position $(x^1, x^2) = (0, 0)$ of system (61) is weakly asymptotically stable. Nevertheless, it is not asymptotically stable. Indeed, let $x_0^1 > 0$. Then the trajectory $(e^t x_0^1, 0)$ of system (61) does not tend to the origin. Thus the stabilization problem is not solvable.

In order to derive stabilizability from a stability concept we introduce the following definition. We say that the equilibrium position $x = 0$ of control system (60) is *weakly exponentially stable* if there exist positive constants a, θ, and δ such that for each $x_0 \in \delta B_n$ *at least one* trajectory $x(\cdot)$ of system (60) with $x(0) = x_0$ satisfies

$$|x(t)| \leq a|x_0|e^{-\theta t}, \quad t \geq 0,$$

$$|\dot{x}(t)| \leq a|x_0|e^{-\theta t}, \quad t \geq 0.$$

From Theorem 4 we see that if a system is stabilizable at first approximation, then its zero equilibrium position is weakly exponentially stable. Below we prove an inverse statement.

THEOREM 15. *Let the right-hand side of system* (60) *be twice differentiable in x, and let the derivatives $\nabla_x f(x, u)$ and $\nabla^2_{xx} f(x, u)$ be continuous in (x, u). Assume that U is a compact set and that there exists a unique $u_0 \in U$ satisfying $f(0, u_0) = 0$. If the equilibrium position $x = 0$ of system* (60) *is weakly exponentially stable, then the stabilization problem for the first approximation is solvable.*

Proof. Let $\bar{x} \in R^n$. Since the equilibrium position $x = 0$ is weakly exponentially stable, there exist trajectories $x_k(\cdot)$ of system (60) with $x_k(0) = \bar{x}/k$ satisfying

$$|x_k(t)| \leq \frac{a}{k}|\bar{x}|e^{-\theta t}, \quad t \geq 0, \tag{62}$$

$$|\dot{x}_k(t)| \leq \frac{a}{k}|\bar{x}|e^{-\theta t}, \quad t \geq 0, \tag{63}$$

where $a > 0$ and $\theta > 0$ and k is large enough. Introduce a new time $\tau = 1 - (t + 1)^{-1}$. Obviously τ varies in $[0, 1)$ when t varies in $[0, \infty)$. Let

$$y_k(\tau) = x_k(t(\tau)), \quad \tau \in [0, 1), \ k = 1, 2, \ldots.$$

Then we have

$$\frac{d}{d\tau}y_k(\tau) = (1-\tau)^{-2} f(y_k(\tau), u_k(t(\tau))), \tag{64}$$

where $u_k(\cdot)$, $k = 1, 2, \ldots$, are the controls corresponding to $x_k(\cdot)$. From (62) and (63) we obtain

$$|y_k(\tau)| \le \frac{a}{k}|\bar{x}|e^{-\theta((1-\tau)^{-1}-1)}, \quad \tau \in [0, 1), \tag{65}$$

$$|\dot{y}_k(\tau)| \le \frac{a}{k}|\bar{x}|(1-\tau)^{-2}e^{-\theta((1-\tau)^{-1}-1)}, \quad \tau \in [0, 1). \tag{66}$$

Consequently $y_k(\tau) \to 0$ and $\dot{y}_k(\tau) \to 0$ as $\tau \to 1$. Put by definition $y_k(1) = 0$, $k = 1, 2, \ldots$. From (65) and (66) we see that the sequence $\{ky_k(\cdot)\}_{k=1}^{\infty}$ is bounded in sup-norm and is equicontinuous. By the Ascoli–Arzelà theorem it contains a uniformly convergent subsequence. Without loss of generality $ky_k(\cdot)$ uniformly converge to a continuous function $y(\cdot)$.

From (64) and (66) we see that there exists $M > 0$ such that

$$k\dot{y}_k(\tau) \in k(1-\tau)^{-2} f(y_k(\tau), U) \cap MB_n$$

$$\subset k(1-\tau)^{-2}(f(k^{-1}y(\tau), U) + l|y_k(\tau) - k^{-1}y(\tau)|B_n) \cap MB_n. \tag{67}$$

Let $\tau \in (0, 1)$ be such that all functions $y(\cdot)$ and $y_k(\cdot)$, $k = 1, 2, \ldots$, are differentiable at τ. From (67) we see that there exist $v_k(\tau)$, $k = 1, 2, \ldots$, such that

$$v_k(\tau) \in k(1-\tau)^{-2} f(k^{-1}y(\tau), U), \quad k = 1, 2, \ldots,$$

and

$$k\dot{y}_k(\tau) \in v_k(\tau) + (1-\tau)^{-2} l|ky_k(\tau) - y(\tau)|B_n, \quad k = 1, 2, \ldots.$$

Let $w_k(\tau) \in U$, $k = 1, 2, \ldots$, be such that

$$v_k(\tau) = k(1-\tau)^{-2} f(k^{-1}y(\tau), w_k(\tau)), \quad k = 1, 2, \ldots.$$

Since U is compact, without loss of generality $w_k(\tau) \to w_\infty(\tau) \in U$. Observe that $v_k(\tau)$ are bounded. Hence we have

$$f(0, w_\infty(\tau)) = \lim_{k\to\infty} f(k^{-1}y(\tau), w_k(\tau)) = \lim_{k\to\infty} k^{-1}v_k(\tau)(1-\tau)^2 = 0.$$

Since $u_0 \in U$ is a unique solution to the equation $f(0, u_0) = 0$, we obtain $w_\infty(\tau) = u_0$.

From (67) and the equality

$$v_k(\tau) = k(1-\tau)^{-2}[f(k^{-1}y(\tau), u_0) + (f(0, w_k(\tau)) - f(0, u_0))$$

$$+ (f(k^{-1}y(\tau), w_k(\tau)) - f(0, w_k(\tau))) - (f(k^{-1}y(\tau), u_0) - f(0, u_0))]$$

$$= k(1-\tau)^{-2}[k^{-1}Cy(\tau) + (f(0, w_k(\tau)) - f(0, u_0))$$

$$+ k^{-1}(\nabla_x f(0, w_k(\tau)) - \nabla_x f(0, u_0))y(\tau) + r(k, \tau)],$$

where $\lim_{k\to\infty} \sup_\tau kr(k,\tau) = 0$, we have

$$k\dot{y}_k(\tau) \in (1-\tau)^{-2}(Cy(\tau)+K) \cap MB_n + \epsilon_k(\tau)B_n,$$

where $\epsilon_k(\tau) \to 0$ as $k \to \infty$. Applying Lemma 2.7.3 in [9, p. 82], we obtain

$$\dot{y}(\tau) \in (1-\tau)^{-2}(Cy(\tau)+K), \ \tau \in [0,1).$$

From (65) we have

$$|y(\tau)| \le a|\bar{x}|e^{-\theta((1-\tau)^{-1}-1)}, \quad \tau \in [0,1).$$

Thus, the function $x(t) = y(\tau(t))$ satisfies

$$\dot{x}(t) \in Cx(t) + K$$

and

$$|x(t)| \le a|\bar{x}|e^{-\theta t}, \quad t \ge 0.$$

Applying Theorem 1, we obtain the result. □

Remark. The uniqueness of $u_0 \in U$ satisfying $f(0,u_0) = 0$ is essential. Indeed, the zero equilibrium position of the control system $\dot{x} = ux$, $|u| \le 1$, $u_0 = 0$, is weakly asymptotically stable. The control $u \equiv -1$ is a constant stabilizer. Nevertheless the first approximation is $\dot{x} = 0$.

COROLLARY 6. *Under the conditions of Theorem* 15 *the stabilization problem for system* (60) *is solvable.*

REFERENCES

[1] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, New York, 1984.

[2] J.-P. AUBIN, H. FRANKOWSKA, AND C. OLECH, *Controllability of convex processes*, SIAM J. Control Optim., 24 (1986), pp. 1192–1211.

[3] A. BACCIOTTI, *Local Stabilizability of Nonlinear Control Systems*, World Scientific, River Edge, NJ, 1992.

[4] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, San Diego, CA, 1979.

[5] A. BERMAN, M. NEUMANN, AND R. J. STERN, *Nonnegative Matrices in Dynamic Systems*, Wiley-Interscience, New York, 1989.

[6] R. F. BRAMMER, *Controllability in linear autonomous systems with positive controllers*, SIAM J. Control, 10 (1972), pp. 339–353.

[7] V. A. BUSHENKOV AND G. V. SMIRNOV, *A new approach to the regulator design problem*, Optimization Methods and Software, 1 (1992), pp. 1–12.

[8] ———, *On a computational method for regulator design*, Izv. Ross. Akad. Nauk, Tekhn. Kibernet., 2 (1992), pp. 20–38. (In Russian.)

[9] A. F. FILIPPOV, *Differential Equations with Discontinuous Right-Hand Sides*, Kluwer Academic Publishers, Norwell, MA, 1988.

[10] J.-L. GOUZÉ, *Stabilisation globale de systèmes non-linéaires par un contrôle positif*, Lecture Notes in Control and Inform. Sci. 144, Springer-Verlag, New York, 1990, pp. 324–331.

[11] R. N. IZMAILOV, *The "peak" effect in stationary linear systems with scalar inputs and outputs*, Automat. Remote Control, 48 (1987), pp. 1018–1024.

[12] V. I. KOROBOV, A. P. MARINIC, AND E. N. PODOLSKII, *Controllability of linear autonomous systems in the presence of constraints on the control*, Differential Equations, 11 (1975), pp. 1465–1475.

[13] V. I. KOROBOV AND NGUEN KHOA SHON, *σ-controllability of linear autonomous systems in the presence of constraints on the control*, Differential Equations, 16 (1980), pp. 242–248.

[14] A. M. LYAPUNOV, *Problème Général de la Stabilité du Mouvement*, Ann. of Math. Stud., 17, Princeton Univ. Press, Princeton, NJ, Oxford Univ. Press, London, 1947.

[15] V. N. POLOTSKI, *On the maximal errors of an asymptotic state identifier*, Automat. Remote Control, 39 (1978), pp. 1116–1121.

[16] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[17] S. SAPERSTONE AND J. YORKE, *Controllability of linear oscillatory systems using positive controls*, SIAM J. Control Optim., 9 (1971), pp. 253–262.
[18] G. V. SMIRNOV, *Weak asymptotic stability for differential inclusions*, I, II, Automat. Remote Control, 51 (1990), pp. 901–908, pp. 1052–1058.
[19] E. D. SONTAG, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Springer-Verlag, Berlin, New York, 1990.
[20] H. J. SUSSMANN AND P. V. KOKOTOVIC, *The peaking phenomenon and the global stabilization of nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 424–440.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1650–1665, September 1996

009

MINIMAX RENDEZVOUS ON THE LINE*

WEI SHI LIM† AND STEVE ALPERN‡

Abstract. Suppose that n players are placed randomly on the real line at consecutive integers, and faced in random directions. Each player has maximum speed one, cannot see the others, and doesn't know his relative position. What is the minimum time M_n required to ensure that all the players can meet together at a single point, regardless of their initial placement? We prove that $M_2 = 3$, $M_3 = 4$, and M_n is asymptotic to $n/2$. We also consider a variant of the problem which requires players who meet to stick together, and find in this case that three players require 5 time units to ensure a meeting. This paper is thus a minimax version of the rendezvous search problem, which has hitherto been studied only in terms of minimizing the expected meeting time.

Key words. rendezvous, search game

AMS subject classifications. 90B40, 90D26

1. Introduction.

In this paper we ask how much time is needed to ensure that n players, placed randomly onto consecutive integers on the line, can all meet together at a single point. We assume that they cannot see each other and can move at unit speed. We also assume that they have no common notion of a positive direction on the line, or equivalently are each placed pointing in a random direction. We seek the minimum time M_n by which some n-tuple of strategies guarantees a group meeting regardless of the initial placement of the players. For example, it is clear that $M_2 \leq 3$ because if one player remains still while the other goes, say, one unit to the right followed by two to the left, then the meeting time is 1 or 3.

The problem considered here is a search game, and fits into the framework of [5] and [9], except that our problem is far from being zero sum. More specifically, our problem can be seen as a minimax version of the *rendezvous search problem* [1] which has previously been analyzed only in terms of expected-time minimization. However there are many applications in search theory where expected-time minimization is not the most appropriate solution concept, and minimax is one of several others that have been studied in various contexts. It seems that minimax is appropriate even in the original rendezvous problem of Schelling [10], where two parachutists have to meet after a simultaneous landing in a large field. If the overall plan involves their moving out together from this field at time t_1, then they must be dropped into the field no later than t_1 minus the minimax rendezvous time. For a simulation approach to Schelling's problem, see [12].

As the search region is the line, our work may be compared to several recent investigations of rendezvous search on the line in an expected-time context. The line as the search region was first studied in the original paper on rendezvous search [1] and improved on in [3], but only in a "symmetric" version in which the two players were required to use the same mixed strategy. Clearly the symmetric version of the rendezvous problem is not appropriate to the minimax context because it allows the possibility that both players might use the same pure strategy. (They would never meet in the case where they are initially pointed in the same direction.) The asymmetric version of rendezvous search on the line was first studied in [2], where it was shown that two players initally placed one unit apart could meet in least expected time 13/8. From our current perspective it is worth noting that the optimal strategies in that context had a maximum meeting time of 3. (It is interesting to note that two-person rendezvous on the line is in some sense dual to *high-low search* [4], since in the former the searcher knows the distance

*Received by the editors December 7, 1994; accepted for publication (in revised form) May 17, 1995.

†Faculty of Business Administration, National University of Singapore, 10 Kent Ridge Crescent, S 119260 Singapore (fbalimws@leonis.nus.sg).

‡Mathematics Department, London School of Economics, Houghton Street, London WC2A 2AE, England (alpern@lse.ac.uk).

but not the direction of the other player, while in the latter the direction but not the distance is learned after each guess.) More recently the problem of three-player rendezvous on the line was studied in [6]. No exact value was obtained for the least expected time required for all three to meet, but it was shown that two (out of three) could meet in least expected time 47/48.

The present paper finds the first exact value for a full three-player rendezvous problem, i.e., the time needed for all three to meet at a single point. We do this in a minimax context and find that 4 is the least time required to ensure that three players placed at unit distances apart can meet, that is $M_3 = 4$. In order for all three to meet this quickly, the two players who first meet must in some cases split up to find the remaining player, before regrouping in a threesome. The problem facing the two who meet first is thus similar to that studied in a different context [11], [7], [8]. We also consider an important variant of the problem, namely "sticky" rendezvous, where players who meet are required to remain together. We find that when players' strategies are thus restricted, they need 5 time units to ensure full three-player rendezvous. Our final result concerns the behavior of M_n for large values of n, the time required for large numbers of players to meet. We find that M_n is asymptotic to $n/2$. This asymptotic value is the same as for the least expected time in the symmetric problem, as proved in [6], although of course the problem and solution concept are entirely different.

The paper is organized as follows. In §2 we give a precise formulation of the minimax rendezvous problem, in its unrestricted and "sticky" forms. In §3 we analyze the two-player problem in an extended form that will arise for three players after two of them meet. In particular we establish that $M_2 = 3$. In §4 we solve the sticky version of the three-person problem, proving that sticky players require 5 time units to guarantee a three-way meeting. In §5 we use similar arguments to show that without this restriction three players can meet in 4 time units, $M_3 = 4$. The final section, §6, establishes the asymptotic result that $\lim_{n\to\infty} M_n/n = 1/2$.

2. The minimax rendezvous time. In this section we give the definitions of the n-player rendezvous problems Γ_n and the associated minimax rendezvous time M_n. We also define their "sticky" counterparts, in which players who meet must thenceforth remain together.

The problem (or game) Γ_n begins with a random placement of the n players onto the first n integers (or, equivalently, any n consecutive integers). The initial position of player i is denoted by c_i, where $\{c_1, c_2, \ldots, c_n\} = \{1, 2, \ldots, n\}$. Since the players do not have a common notion of a positive direction on the real line, we assume they are initially faced in independent random directions. We use c_{n+i} to denote the direction that player i is initially pointing, from an observer's fixed global view, where $c_{n+i} \in \{+1, -1\}$. Let $C = C_n$ denote the set of all $n!2^n$ *initial configurations* of the form $c = \{c_1, c_2, \ldots, c_{2n}\}$. Note that the first n coordinates denote position, while the next n denote direction.

A strategy for player i in the game Γ_n is a rule that gives his motion (relative to his starting point and starting direction) as a function of the information he receives from players he may meet. A strategy profile is simply an n-tuple of strategies, one for each player. A strategy profile together with an initial configuration determines completely the motions of all the players. If $\bar{S}$ is a strategy profile and c is an initial configuation, we define $T_c(\bar{S})$ to be the first time (if any) that all n players following the profile $\bar{S}$ meet together at a single point, when the initial configuration is c.

In particular, a strategy for player i must say how that player should move before he meets anyone. This part of the strategy is called the Stage 1 strategy. A Stage 1 strategy s_i is a (speed-one) path belonging to the set

$$P = \left\{p : \Re^+ \to \Re,\ p(0) = 0,\ |p(t_1) - p(t_2)| \le |t_1 - t_2|\right\}.$$

The Stage 1 path of a player following strategy s_i when the initial configuration is c is given by $c_i + c_{n+i}s_i(t)$. In a recent paper [6] it is shown that when initial distances between adjacent

players are integers, any strategy profile can be modified to one in a subset of P called P^* where players have piecewise linear paths, with slopes ± 1, and which turn only at times $k/2$, where k is an integer. This modification does not postpone any meeting between players and hence does not postpone the final meeting of all the players. For this reason we will further assume that throughout the game all players are restricted to paths in P^*. Observe that this assumption implies that $T_c(\bar{S})$ is always half of an integer. A useful notational device for describing Stage 1 strategies in P^* is simply to list the slopes in successive half-units of time. Thus $s = [+1, -1, +1, -1, \ldots]$ describes a path that oscillates between its starting point and a point a half-unit above the start. More generally, if $s = [x_1, x_2, \ldots]$ then we have (with [ ] denoting integer part)

$$s(t) = \frac{1}{2}\sum_{m=1}^{[2t]} x_m + x_{[2t]+1}\left(\frac{2t - [2t]}{2}\right).$$

One final remark regarding player strategies in Γ_n is that we may assume without loss of generality that they all begin with $+1$, that is, they go in the forward direction for the first half-unit of time. This assumption is valid because it does not restrict the actual player motion, since the player may be initially pointed either way.

The maximum rendezvous time for a strategy profile $\bar{S}$, denoted simply $T(\bar{S})$, is defined as

$$T(\bar{S}) = \max_{c \in C} T_c(\bar{S}).$$

Then we may define the minimax rendezvous time M_n as

$$M_n = \min_{\bar{S}} T(\bar{S}),$$

where the minimum is taken over all strategy profiles $\bar{S}$. We note that the index n does not appear on the right side of the above equation, but of course the player-number parameter n is implicit in all the definitions in this section. The existence of the minimum follows from our assumption that all player paths must belong to P^*.

In §4 we will consider a further restriction on player paths, namely, that when players meet they stick together. All the above definitions remain valid, with this assumption. The sticky version of the n-player game is denoted $\tilde{\Gamma}_n$ and the sticky minimax time is denoted $\tilde{M}_n$.

3. Two-player minimax rendevous. In this section we consider the minimax rendezvous problem on the line for the case of two players. In passing, we will prove that the minimax rendezvous time for the standard two-player problem Γ_2 is 3, but we will mainly be concerned with a more general two-person problem denoted $\Gamma(\alpha, \beta)$. We are forced to consider this more general problem because this is what the three-player "sticky" problem (discussed in the next section) collapses to after two players meet.

The problem $\Gamma(\alpha, \beta)$ is an asymmetric information rendezvous game defined as follows. Player I is placed at some point of the line (which we take as the origin 0) and pointed facing up (the line is taken to be vertical). Player II is then faced in a random direction either a distance α above player I or a distance β below player I (i.e., at α or at $-\beta$). Thus player II knows only that his partner is a distance α or β away, while player I knows that his partner is either α above (forward) or β below. If $\alpha = \beta = 1$ then the information is symmetric and indeed $\Gamma(1, 1)$ is the same as the problem we called Γ_2. If player I chooses a strategy $s_1 = f \in P$ and player II chooses a strategy $s_2 = g \in P$ (we cannot restrict strategies to P^* unless α and

β are integers) then the maximum rendezvous time $T^{\alpha,\beta}(f, g)$ is the first time t when the path $f(t)$ has intersected the four paths given below:

$$L_1(t) = \alpha - g(t),$$
$$L_2(t) = \alpha + g(t),$$
$$L_3(t) = -\beta - g(t),$$
$$L_4(t) = -\beta + g(t).$$

That is, $T^{\alpha,\beta}(f, g)$ is the maximum of $t_j = t_j(f, g) = \min\{t : f(t) = L_j(t)\}$ $\forall j = 1, \ldots, 4$. The minimax rendezvous time for $\Gamma(\alpha, \beta)$ is the minimum of $T^{\alpha,\beta}$ $\forall (f, g) \in P \times P$.

Note that any strategy pair (f, g) determines an ordering of the meeting times t_j, i.e., there is a permutation $\sigma = \sigma(f, g)$ of $\{1, 2, 3, 4\}$ such that

$$t_{\sigma(1)} \leq t_{\sigma(2)} \leq t_{\sigma(3)} \leq t_{\sigma(4)} = T^{\alpha,\beta}(f, g).$$

In such a case we will say that (f, g) has permutation type σ. If α and β are integers and $(f, g) \in P^* \times P^*$ then the permutation type is unique and $t_{\sigma(j+1)} \geq t_{\sigma(j)} + .5$.

There is a complementary notion (introduced in [2]) by which each permutation σ determines a canonical strategy pair (F_σ, G_σ), such that in the interval $t_{\sigma(j-1)} < t < t_{\sigma(j)}$, $j = 1, \ldots, 4$ (with $\sigma(0)$ defined as 0), the path F_σ and the path $L_{\sigma(j)}$ (which depends on G_σ) are moving towards each other at maximum speed. It is shown in [2, Thm. 3 and its proof] that *the canonical strategy pair (F_σ, G_σ) minimizes all the meeting times t_j, within the class of strategy pairs of permutation type σ*. It follows that the minimax rendezvous time, as well as the least expected rendezvous time, will be attained for some canonical strategy. However, there may also be noncanonical strategy pairs which achieve the minimax rendezvous time.

To see how this definition defines a unique strategy pair (F_σ, G_σ), we illustrate the construction for the identity permutation $\bar{\sigma}$, using the simpler notation $\bar{f} = F_{\bar{\sigma}}$ and $\bar{g} = G_{\bar{\sigma}}$. The path $\bar{f}$ is pictured in Figure 1 for the parameters $\alpha = 1$ and $\beta = 2$, where it is called f_1 (the meeting times there are .5, 1, 2.5, and 4). The strategies begin with the player I path $\bar{f}$ and the player II possible path $L_1(t) = \alpha - \bar{g}(t)$ moving towards each other at maximum speed. Since L_1 is above $\bar{f}$ at time zero, this means that $\bar{f}' = 1$ and $L_1' = -1$, or $\bar{g}' = +1$, from time zero until the first meeting time $t_1 = \alpha/2$. At this time L_2 is above $\bar{f}$ and so $\bar{f}' = +1$ and $L_2' = -1$, or $\bar{g}' = -1$, from time $t_1 = \alpha/2$ until $\bar{f}$ meets L_2 at time $t_2 = \alpha$. Note that $\bar{f}(\alpha) = \alpha$ and $\bar{g}(\alpha) = 0$. At time $t_2 = \alpha$, $L_3(\alpha) = -\beta - \bar{g}(\alpha) = -\beta$ is below $\bar{f}(\alpha) = \alpha$ and so for the next time interval,

$$t_2 < t < t_3 = t_2 + (\alpha - (-\beta))/2 = (3\alpha + \beta)/2,$$

$\bar{f}' = -1$, and $L_3' = +1$, or $\bar{g}' = -1$. At time t_3, $\bar{f}$ is at $(\alpha - \beta)/2$, while L_4 is lower, at $-(\alpha + 3\beta)/2$. Hence $\bar{f}$ goes down and L_4 goes up (or $\bar{g}$ goes up) throughout the interval

$$t_3 < t < t_4 = t_3 + \left(\frac{\alpha - \beta}{2} - \left(-\frac{\alpha + 3\beta}{2}\right)\right)/2 = 2\alpha + \beta.$$

The data just derived are presented in the first data row of Table 1 which gives the four meeting times $t_{\sigma(j)}$ for the canonical strategies (F_σ, G_σ) corresponding to various permutations. While this table contains just six of the $4! = 24$ possible permutations, it will in fact be sufficient to calculate the minimax rendezvous time. The data in this table will be useful in the following theorem.

THEOREM 1. *The minimax rendezvous time for the problem $\Gamma(\alpha, \beta)$, $\alpha \leq \beta$, is*

$$\min_{f,g \in P \times P} T^{\alpha,\beta}(f, g) = 2\alpha + \beta.$$

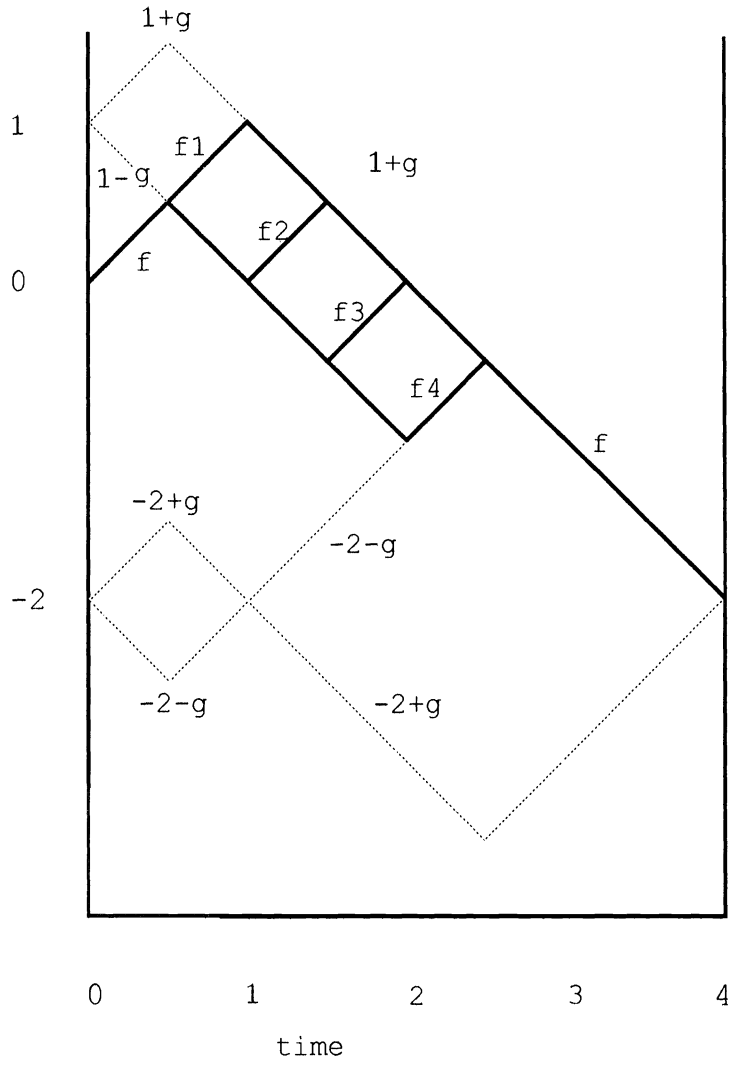


FIG. 1. *Optimal strategies for the problem* $\Gamma(1, 2)$.

TABLE 1
Meeting times for canonical strategies (F_σ, G_σ).

σ	$t_{\sigma(1)}$	$t_{\sigma(2)}$	$t_{\sigma(3)}$	$t_{\sigma(4)} = T^{\alpha,\beta}(F_\sigma, G_\sigma)$
(1, 2, 3, 4)	$\alpha/2$	α	$(3\alpha+\beta)/2$	$2\alpha+\beta$
(1, 2, 4, 3)	$\alpha/2$	α	$(3\alpha+\beta)/2$	$2\alpha+\beta$
(1, 3, 2, 4)	$\alpha/2$	$(2\alpha+\beta)/2$	$(3\alpha+\beta)/2$	$2\alpha+\beta$
(1, 3, 4, 2)	$\alpha/2$	$(2\alpha+\beta)/2$	$\alpha+\beta$	$(3\alpha+3\beta)/2$
(1, 4, 2, 3)	$\alpha/2$	$(\alpha+\beta)/2$	$\alpha+\beta$	$(3\alpha+3\beta)/2$
(1, 4, 3, 2)	$\alpha/2$	$(\alpha+\beta)/2$	$\alpha+\beta$	$(3\alpha+3\beta)/2$

Furthermore, if $T^{\alpha,\beta}(f, g) = 2\alpha + \beta$ *and* $\alpha < \beta$, *then the permutation type of* (f, g) *is* (1, 2, 3, 4), (1, 2, 4, 3), *or* (1, 3, 2, 4).

Proof. According to the result of [2] quoted above, the minimax rendezvous time is the minimum of $T^{\alpha,\beta}(F_\sigma, G_\sigma)$ as σ varies over the permutations of $\{1, 2, 3, 4\}$. We claim that

it is sufficient to consider only the six permutations listed in Table 1 (those where the player II path $\alpha - g(t)$ is intersected first), where the minimum (given that $\alpha \leq \beta$) is $2\alpha + \beta$. The cases which intersect the player II path $\alpha + g(t)$ first will give the same results, that is, the same meeting times $t_{\sigma(j)}$, the same F, and a sign reversal for G. The cases which intersect either of the paths $-\beta \pm g(t)$ first will give a similar table with α and β interchanged. So the only new maximum rendezvous time appearing in such a table would be $2\beta + \alpha$, which is not less than $2\alpha + \beta$. When $\alpha < \beta$, only the first three data rows of the table give the minimum $2\alpha + \beta$. □

As a special case of the above result when $\alpha = \beta = 1$, we have the following solution to the standard two-person minimax rendezvous problem.

COROLLARY 1. $M_2 = 3$.

The three-person "sticky" rendezvous problem $\tilde{\Gamma}_3$ to be analyzed in the next section may reduce to the problem $\Gamma(1, 2)$ in certain cases. For this reason we explicitly give the four optimal player I strategies in P^* for $\Gamma(1, 2)$ in the following result.

COROLLARY 2. *The minimax rendezvous value for the problem* $\Gamma(1, 2)$ *is* 4. *Furthermore if* $T^{1,2}(f, g) = 4$, *for* $(f, g) \in P^* \times P^*$ *then* f *is one of the four strategies in* P^* *which satisfy*

$$f(1/2) = 1/2, \quad f(2.5) = -.5, \quad \textit{and} \quad f(4) = -2, \quad \textit{i.e.},$$

$$\begin{aligned} f_1 &= [+1, +1, -1, -1, -1, -1, -1, -1], \\ f_2 &= [+1, -1, +1, -1, -1, -1, -1, -1], \\ f_3 &= [+1, -1, -1, +1, -1, -1, -1, -1], \\ f_4 &= [+1, -1, -1, -1, +1, -1, -1, -1]. \end{aligned}$$

Each of these has maximum rendezvous time of 4 *when paired with the player* II *strategy* $g = [+1, -1, -1, -1, -1, +1, +1, +1]$. *(See Figure* 1.*)*

Proof. The reader should first verify the obvious fact that the f_k are indeed the only four strategies in P^* which satisfy the three conditions (including of course $f(0) = 0$, which is part of the definition of P^*). According to Theorem 1 we can only find optimal strategies for the permutation types (1, 2, 3, 4), (1, 2, 4, 3), and (1, 3, 2, 4).

Case 1, σ *is* (1, 2, 3, 4) *or* (1, 2, 4, 3). Suppose $T^{1,2}(f, g) = 4$ for $(f, g) \in P^* \times P^*$, and

$$t_1 < t_2 < t_3, t_4.$$

We may assume that $t_1 = .5$ and $f(.5) = g(.5) = .5$ because in the alternative case that $f(.5) = -.5$ the resulting problem at time .5 is either $\Gamma(1, 2)$ (if I is told that II moved down) or $\Gamma(2, 1)$ (if I is told that II moved up), and therefore (by Theorem 1) requires an *additional* 4 units of time to ensure meeting. Since $g(.5) = .5$ it follows that $g(t) \geq .5 - (t - .5) = 1 - t$, for $t \geq .5$. Hence after f has intercepted $1 + g(t)$ at time t_2, we must have that

$$f(t) \geq 2 - t \text{ for } t \geq t_2. \tag{1}$$

Since f intersects $1 + g$ at time t_2 and $-2 + g$ at time $t_4 \leq 4$, and speeds are bounded by 1, we must have

$$t_4 - t_2 \geq |(1 + g(t_2)) - (-2) + g(t_2)|/2 = 1.5 \text{ or } t_2 \leq t_4 - 1.5 \leq 2.5. \tag{2}$$

Suppose that strict inequality holds in (1), that $f(s) = 2 - s + p$ for some $p > 0$ and some time $s < 4$. Then a player I starting at position $2 + p$ at time zero, following path $2 - t + p$ until time $t = s$, and then following path f, could ensure meeting paths $-2 \pm g$ by time 4. But this

would mean that the minimax rendezvous time for the problem $\Gamma(0, 4+p) = \Gamma(4+p, 0)$ is not more than 4, whereas Theorem 1 says it is equal to $4+p$. Hence our assumption that $f(t)$ could be larger than $2-t$ was false, and (1) must hold with equality, that is,

$$f(t) = 2 - t \text{ for } t \geq 2.5. \tag{3}$$

We showed earlier that $f \in P^*$ must go through (.5,.5) and (3) shows further that it must go through (2.5, −.5) and (4, −2). It follows that it must be one of the four strategies f_k. (Actually it cannot be f_4 in this case, but we do not need to prove this fact.)

Case 2, σ *is* (1, 3, 2, 4). Suppose $T^{1,2}(f, g) = 4$ for $(f, g) \in P^* \times P^*$, and

$$t_1 < t_3 < t_2 < t_4 = 4.$$

Since speeds are bounded by 1, it follows that

$$\begin{aligned} 4 - t_2 = t_4 - t_2 \geq |L_4(t_2) - L_2(t_2)|/2 = |(-2 + g(t_2) - (1 + g(t_2)))| = 1.5, \text{ or} \\ t_2 \leq 2.5, \text{ and hence } t_3 \leq t_2 - .5 \leq 2. \end{aligned} \tag{4}$$

By the same reasoning as (4), we have

$$t_3 - t_1 \geq 1.5, \quad \text{so } t_3 \geq t_1 + 1.5 \geq 2. \tag{5}$$

The only solution to (4) and (5) is $t_1 = .5$, $t_3 = 2$, $t_2 = 2.5$, $t_4 = 4$, which are the times for the canonical strategy pair with this permutation. Hence f must be $F_\sigma = f_4$. □

The optimal strategies for the problem $\Gamma(1, 2)$ are drawn in Figure 1. They are also drawn in a stacked form in Figure 3, which will be discussed later.

We conclude this section with an analysis of some two-person rendezvous problems where one of the players (taken to be I) knows the direction of the other. Since this person will clearly move at speed one in this direction, these are really one-person problems. The only strategic variable is the path of player II. These results are called lemmas because they will be used in §5 in the following way: When two players meet and do not know the direction of the other, they will each be assigned a direction and will assume the remaining player lies in that direction. A lower bound on the maximum time taken to find the remaining player, assuming he is in this direction, is the minimax value of the game in which the direction is known. It is these minimax values that we now calculate. These are all very simple results.

LEMMA 1. *Suppose that player* I *is placed facing up at* 0 *and player* II *is either placed facing down at* 1 *or facing up at* 2. *Then the minimax rendezvous time is* 3/2. *Furthermore this maximum meeting time occurs if and only if* I *moves up at speed one and* II *uses a strategy* $h \in P$ *satisfying* $h(3/2) = -1/2$. *There are three strategies* $h \in P^*$, *defined up to time* 3/2, *satisfying this condition. In the notation giving the slopes of the paths in successive time intervals of length* 1/2 *these paths* (*as shown in Figure* 2) *are as follows:*

$$\begin{aligned} h_1 &= [+1, -1, -1], \\ h_2 &= [-1, -1, +1], \\ h_3 &= [-1, +1, -1]. \end{aligned}$$

Proof. If player II uses strategy h, the maximum meeting time $T(h)$ is the time required for the path t (of player I) to meet both possible paths of II, that is, $1 - h(t)$ and $2 + h(t)$. This is the same as the time required for $h \in P$ to meet both $1 - t$ and $t - 2$. Clearly if $h(3/2) = -1/2$, then it meets both these paths at time 3/2. If it meets either of these paths before this time, the earliest it can meet the other is 3/2, since these two paths are approaching each other at combined speed 2. Furthermore if $h(3/2) > -1/2$, then it cannot yet have

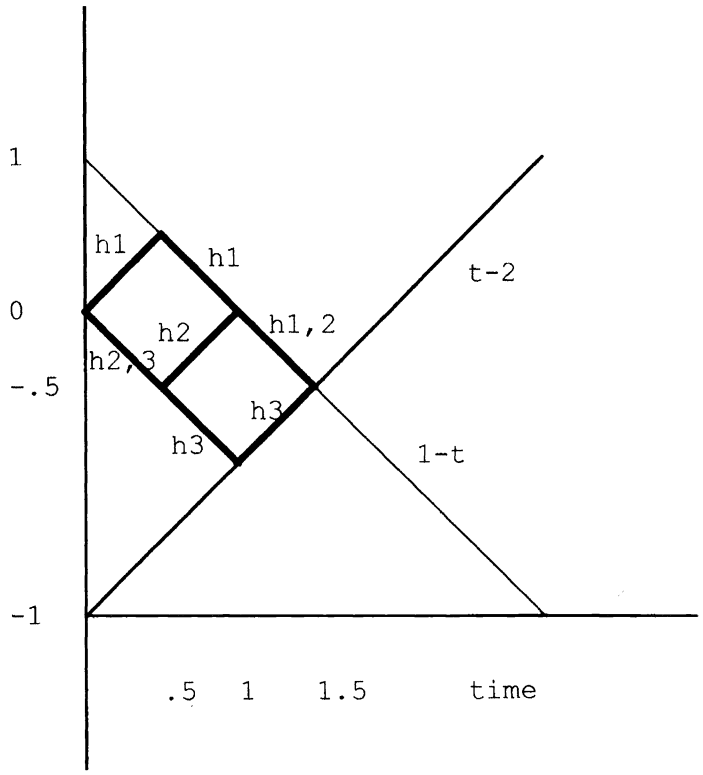


FIG. 2. *Optimal paths derived in Lemma* 1.

intersected $t-2$; if $h(3/2) < -1/2$, then it cannot yet have intersected $1-t$. Thus only paths with $h(3/2) = -1/2$ can achieve a maximum meeting time of $3/2$. The three paths stated in the lemma are the only ones in P^* satisfying this condition. These paths are drawn in bold in Figure 2. □

The following two lemmas are even easier, as they give minimax times when player I knows not only the direction but also the initial distance to player II. They may appear too obvious to bother stating, but we do so because they will be used repeatedly in §5, without specific mention. (The first is actually a corollary of Theorem 1, with $\alpha = 0$.)

LEMMA 2. *If player* I *is placed (at time* 0*) facing up at position* 0 *and player* II *is placed in a random direction at position* $\beta > 0$*, at any time prior to time* β*, then the minimax rendezvous time is* β*. Call this problem* $\Gamma'(\beta)$.

LEMMA 3. *If player* I *is placed (at time* 0*) facing up at position* 0 *and player* II *is placed in a known direction (say up) at position* $\beta > 0$*, at time* δ*,* $-\beta \leq \delta \leq \beta$*, then the minimax rendezvous time is* $(\beta + \delta)/2$*. Call this problem* $\Gamma''(\beta, \delta)$.

4. Sticky three-person rendezvous. We are now in a position to attack the problem $\tilde{\Gamma}_3$. Recall that in this problem three players are randomly placed onto the integers 1, 2, and 3, and faced in random directions. Once two players meet, they must stick together while trying to locate the third. The players' strategy paths are assumed to belong to P^*. The main result of this section is the determination of the minimax rendezvous time $\tilde{M}_3$ for this problem.

THEOREM 2. *The minimax rendezvous time* $\tilde{M}_3$ *for the sticky three-person problem* $\tilde{\Gamma}_3$ *is* 5.

Proof. We first show that $\tilde{M}_3 \leq 5$ by exhibiting a simple strategy triple which guarantees three-player rendezvous by time 5. The simplest version is that two of the players remain still

(until they are met by the moving player) while the third moves forward, taking along any player he meets, until he reaches an integer location with no player on it. He then reverses direction, similarly taking along any player he meets, until he has accumulated both of the other players. The case with maximum rendezvous time is when the moving player starts in the middle, and in this case the rendezvous time is 5. Since the strategy of staying still in Stage 1 does not belong to P^*, it has to be modified. The modification is simply to oscillate between the starting point and a point 1/2 unit forward. The analysis for the modified strategy is essentially the same and it also has a maximum rendezvous time of 5.

To demonstrate that $\tilde{M}_3 \geq 5$, we assume that there is a strategy triple S^* with $T(S^*) = 4.5$, and then show that this assumption leads to a contradiction. Since for strategies involving paths in P^* intersections can occur only at integer multiples of $1/2$, this will establish that $\tilde{M}_3 \geq 5$.

Let $S = (s_1, s_2, s_3)$ be the Stage 1 strategies for S^*. We may assume that each is simply the identity function t for $t \leq 1/2$. Observe that for any of the three players $j = 1, 2, 3$, there is an initial configuration $c = c(j)$ for which the two players other than j meet at time .5. (For example $c(2) = (1, 3, 2, +1, +1, -1)$.) Let g_j denote the strategy (path) followed by the two who meet in case $c(j)$ from time 1/2 onwards. We normalize this so that the position of these two players at time $t + 1/2$ is $g_j(t)$ plus their position when they meet at time .5. Thus g_j belongs to P^* (takes value 0 at 0). Similarly let $\tilde{s}_j$ denote the remainder of player j's path from time $1/2$ onwards, $\tilde{s}_j(t) = s_j(t + 1/2) - s_j(1/2)$. Thus $\tilde{s}_j$ also belongs to P^* (takes value 0 at 0). Note that the situations of player j and of the remaining two players are the same as that of players I and II in the game $\Gamma(1, 2)$; the paired players are either 1 unit above player j (that is in the direction he was initially pointed) or 2 units below him. Hence it follows that

$$T_{c(j)}(S^*) = 1/2 + T^{1,2}(\tilde{s}_j, g_j),$$

where $T^{1,2}$ is the minimax rendezvous time defined in the previous section for the game $\Gamma(1, 2)$. Our assumption that $T(S^*) = 4.5$ implies that $T_{c(j)}(S^*) \leq 4.5$ and by the above that $T^{1,2}(\tilde{s}_j, g_j) \leq 4$. It follows from Corollary 2 that $T^{1,2}(\tilde{s}_j, g_j) = 4$ and that $\tilde{s}_j$ belongs to the set of optimal strategies for player I in $\Gamma(1, 2)$, that is, to the set $\{f_1, f_2, f_3, f_4\}$.

Since the above argument holds for each player $j = 1, 2, 3$, we have shown that the Stage 1 paths of S^* must be optimal for the problem $\Gamma(1, 2)$ from time $1/2$ onwards, that is,

$$\tilde{s}_j \in \{f_1, f_2, f_3, f_4\}, \quad \text{for } j = 1, 2, 3.$$

It now follows that there is a case $\bar{c}$ for which none of the three players meet (not even two of them) by time 4.5, that is $T_{\bar{c}}(S^*) > 4.5$, which contradicts our assumption. To see that such a case (initial configuration) $\bar{c}$ exists, first look at Figure 3. This shows a drawing of the four paths f_1, f_2, f_3, f_4 with each preceeded by a slope 1 diagonal for time $1/2$. The lower-indexed functions are started at higher positions on the line, and there are no intersections by time 4.5. The general algorithm for choosing $\bar{c}$ as a function of (s_1, s_2, s_3) is very simple: point all the players up, and place the players using lower-indexed $\tilde{f}_k$'s higher. If two players are using the same f_k, then of course it doesn't matter which of these is placed higher. For example, if $(\tilde{s}_1, \tilde{s}_2, \tilde{s}_3) = (f_1, f_4, f_3)$, then $\bar{c} = (3, 1, 2, +1, +1, +1)$. □

5. Three-player rendezvous (unrestricted). In this section we show that three players placed on adjacent integers can ensure a three-way meeting by time 4, that is $M_3 = 4$. This is a savings of 1 time unit over the 5 needed in the sticky case considered in the previous section. The novel feature considered here is that players who meet can separate to find the third (although the game does not end until all three are together).

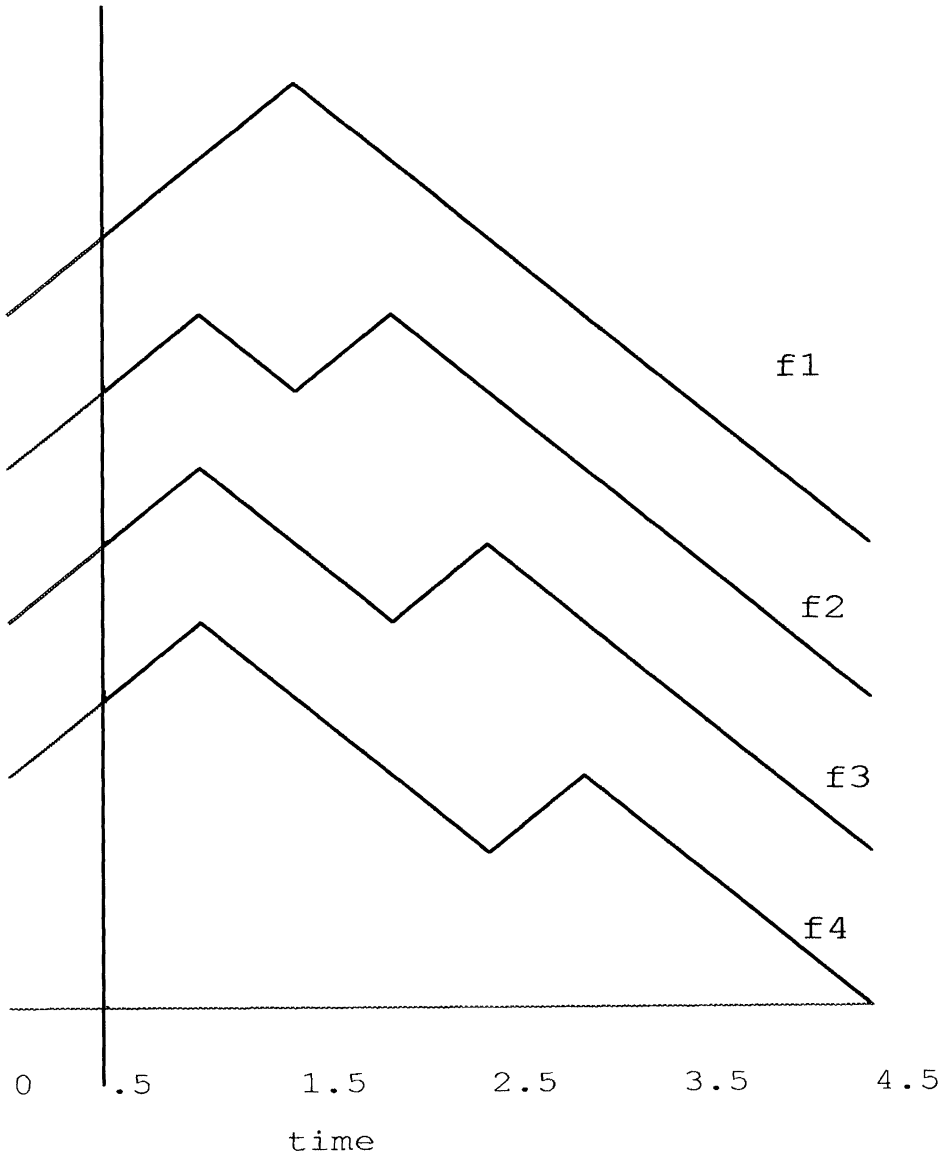


FIG. 3. *A nonintersecting stacking of the strategies f_k, starting at time .5.*

LEMMA 4. $M_3 \leq 4$.

Proof. We exhibit a strategy profile with a maximum rendezvous time of 4. The Stage 1 strategies are the same as for the optimal sticky rendezvous: Player I follows the path (given in slope form for a half-unit time interval) $[+1, +1, -1, -1, -1, -1]$ up to time 3. The other two players follow a path which oscillates between their start and a point half a unit away, such as $[+1, -1, +1, -1, +1, -1, +1, -1]$. If player I has not met another player by time 1, then he can conclude the other players were both behind him, so when he reverses direction at time 1 he continues forever in this direction, bringing with him the first player he meets (at time 3), and meeting the second at time 4. If he first meets another player at time 1, the two who meet know that the remaining player is either 1 above or 2 below, and will be there at every integer time. So one of them (say player I) goes 1 above and then reverses, while the

other goes 2 below and then reverses. If either finds the remaining player he asks that player to stick with him. Thus the two who originally met will meet again in 3 more time units, at player I's starting point. Furthermore one of the two is sure to have brought the remaining player along with him. Finally, if player I first meets a player at time 1/2, he can ignore this and bring that player back to that player's start. This puts the two who met in the situation analyzed above. Thus in any case the rendezvous time is not more than 4. □

LEMMA 5. *$M_3 \geq 3.5$. Furthermore any strategy profile for the game Γ_3 which has a maximum rendezvous time of 3.5 must have all its Stage 1 paths, up to time 2, belonging to the set $\{\tilde{h}_1 \tilde{h}_2, \tilde{h}_3\}$ defined as follows. The path $\tilde{h}_k$ is the path h_k of Lemma 1, preceded by a forward speed-one motion for $t \leq 1/2$. That is, $\tilde{h}_k(t) = t$, $t \leq 1/2$, and $\tilde{h}_k(t) = 1/2 + h_k(t - 1/2)$, $t \geq 1/2$. In the notation giving the slope in each half-unit of time, these paths are*

$$\begin{aligned} \tilde{h}_1 &= [+1, +1, -1, -1], \\ \tilde{h}_2 &= [+1, -1, -1, +1], \\ \tilde{h}_3 &= [+1, -1, +1, -1]. \end{aligned}$$

Proof. As in the proof of Theorem 2, we begin by assuming an initial configuration such that the two players other than player j meet each other at time $1/2$. One of these players (call each of these player I) must go up to find the remaining player j (call him player II). Renormalize the line so that the origin (0) is where the two players have met. Assuming II is above this, he is either at 1 facing down (if he started facing down) or at 2 facing up (if he started facing up). Hence by Lemma 1 the earliest that the player I who goes up can guarantee finding II, assuming he is up, is (additional) time 1.5. It follows that the earliest the two agents of Player I (the one going up and the one going down) can meet together, bringing along player II, is $T = 1/2 + 2(1.5) = 3.5$. Furthermore, it follows from the second part of Lemma 1 that in order for this time to be achieved, player II must be following one of the paths h_k from time $1/2$. Since we are assuming that strategies for Γ_3 begin by going up for time $1/2$, it follows that the player we are calling II and j must use a strategy $\tilde{h}_k$ up to time 2. But since this argument applied to any player $j = 1, 2, 3$, we are done. □

LEMMA 6. *Any strategy for the game Γ_3, whose Stage 1 paths (up to time 2) belong to the set $\{\tilde{h}_1, \tilde{h}_2, \tilde{h}_3\}$, has maximum rendezvous time at least $M = 4$.*

Proof. Since order does not matter, there are ten strategy triples in $\{\tilde{h}_1, \tilde{h}_2, \tilde{h}_3\}^3$. We divide these into four types. For each type we stop the action at some time T_0 and assume a certain set of initial configurations. We then give a lower bound on the maximum remaining time, which when added to T_0 is at least 4.

Type 1: All three use the same strategy. This type covers the three strategy profiles $(\tilde{h}_k, \tilde{h}_k, \tilde{h}_k)$, $k = 1, 2, 3$. For strategy profiles of this type the first integer q such that some players have a different Stage 1 strategy for the time interval $[q/2, (q+1)/2]$ satisfies $q \geq 4$.

Assume that all three players start facing up. Then at time $T_0 = q/2$ they are back at their original positions, and the top and bottom players are at distance 2. By the definition of q, there are two players who move in opposite directions in the interval $[q/2, (q+1)/2]$. So for some initial configuration the player starting at 3 will move up throughout this interval and the player starting at 1 will move down. Hence at time $(q+1)/2$ the players at the ends will be at distance 3. Therefore the earliest these two player can meet is at time $(q+1)/2 + 3/2 \geq (4+1)/2 + 3/2 = 4$.

Type 2: The strategy $(\tilde{h}_1, \tilde{h}_2, \tilde{h}_2)$ is used. Consider two initial configurations $c' = \{3, 2, 1, +1, +1, +1\}$ and $c'' = \{3, 2, 1, +1, -1, -1\}$. In each of these configurations player 1, who we now call player I, starts at 3 and is back at 3 at time $T_0 = 2$. Player 3, who we now call player II, starts at 1 and is back at 1 at time 2. In both cases no players have met, so they

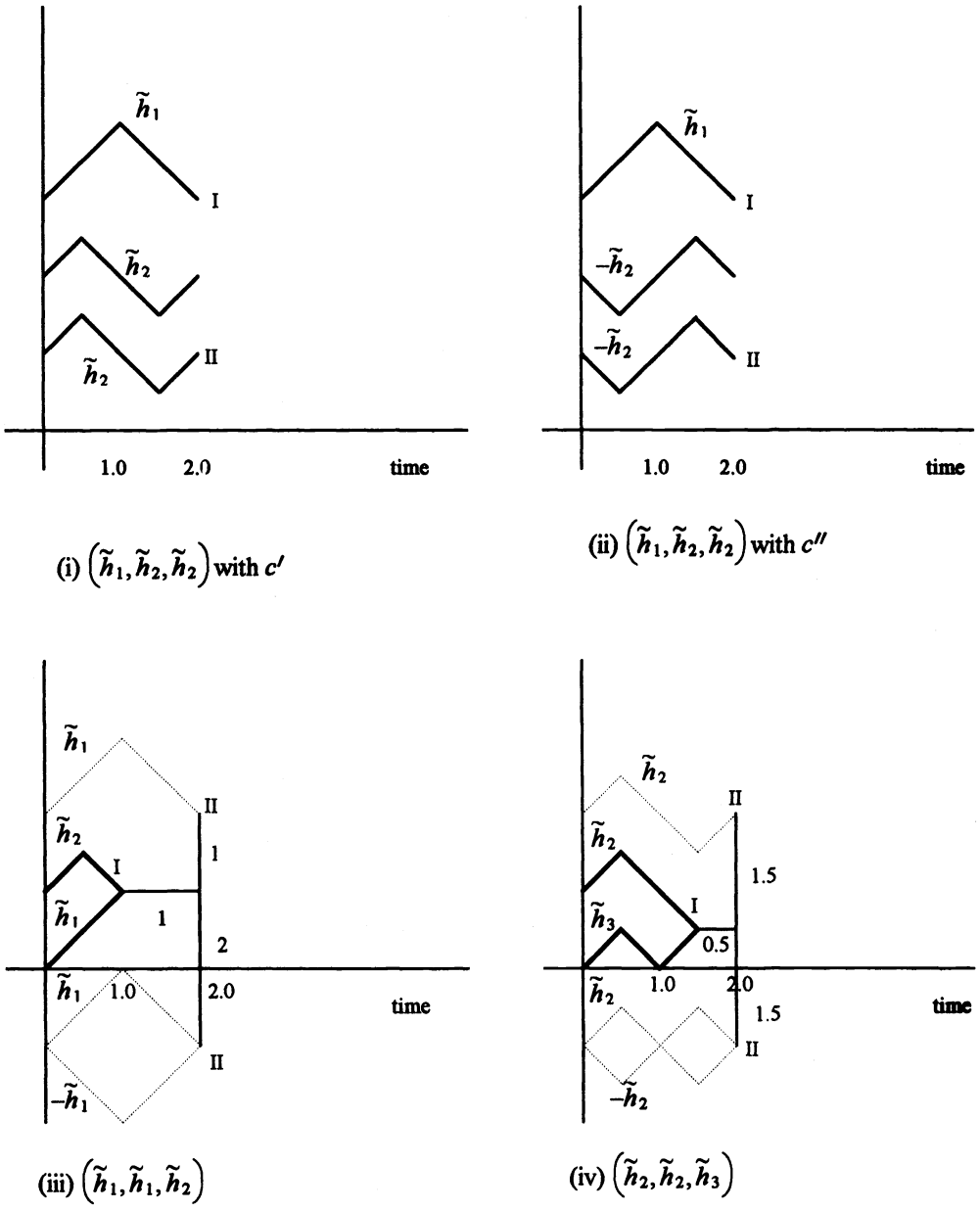


FIG. 4. *Strategies with* $s_j \in \{\tilde{h}_1, \tilde{h}_2, \tilde{h}_3\}$ *have* $T \geq 4$.

cannot determine by time 2 which of these (c' or c'') is the actual initial configuration. In case c', player II is pointing up at time T_0, while in case c'' he is pointing down. Hence by Lemma 2 with $\beta = 2$, players cannot meet before time $T_0 + \beta = 2 + 2 = 4$. This case is illustrated in Figures 4(i) and 4(ii).

For the remaining Types 3 and 4, the analysis will be as follows. For each of the two types we give a set of initial configurations. Then for each strategy profile of that type, we give a time T_0 at which two players meet (who we then call player I) but cannot distinguish between the configurations in the given set. The problem of finding player II above them is called Γ_{up} and the minimax time to find him (or turn back) is denoted by T_{up}, which can be

calculated using Lemma 2 or 3, depending on the nature of Γ_{up}. If the remaining player (II) is below then the associated problem and minimax time are denoted by Γ_{down} and T_{down}. Thus the maximum meeting time T satisfies

$$T \geq T_0 + T_{up} + T_{down}. \tag{6}$$

Examples of Types 3 and 4 are illustrated in Figures 4(iv) and 4(iii), respectively. The paths of the two players who meet at time T_0 are drawn in bold up to time T_0; the three possible paths of the remaining player (each corresponding to some initial configuation in the given set) are drawn in a dashed line up to time $2 = T_0 + \delta$; the parameters β and δ of the games Γ' and Γ'' of Lemmas 2 and 3 are drawn from player I's position at time T_0 in thin lines.

Type 3: *The strategies* $(\tilde{h}_2, \tilde{h}_2, \tilde{h}_3)$ *or* $(\tilde{h}_2, \tilde{h}_3, \tilde{h}_3)$ *are used.* Consider the set of configurations

$$C_3 = \{(1, 3, 2, +1, +1, +1), (3, 1, 2, +1, -1, -1), (3, 2, 1, +1, +1, +1)\}.$$

For these two strategy profiles and any of these configurations, a player using $\tilde{h}_2$ will meet a player using $\tilde{h}_3$ at time $T_0 = 1.5$, and they will then be unable to distinguish between these three initial configurations. Call this pair player I. For either profile we can take Γ_{up} to be $\Gamma''(1.5, .5)$ and so $T_{up} = 1$ by Lemma 3. Similarly for either profile $\Gamma_{down} = \Gamma'(1.5)$ so $T_{down} = 1.5$. Thus by (6) we have $T \geq 1.5 + 1 + 1.5 = 4$. To aid the reader we give the full analysis for the profile $(\tilde{h}_2, \tilde{h}_2, \tilde{h}_3)$, which is illustrated in Figure 4(iv). At time T_0 we normalize the time back to zero and let the meeting point be the new origin. With respect to this framework, the position of player II if above is 1.5 units above player I (i.e., the two players who met) at time .5 (at actual time 2) and facing up. Hence as claimed above, $\Gamma_{up} = \Gamma''(1.5, .5)$. If player II is below player I, then he is 1.5 units below in additional time .5, and can be facing either way, depending on the configuration. Hence, as claimed, $\Gamma_{down} = \Gamma'(1.5)$.

Type 4: *One of the strategies* $(\tilde{h}_1, \tilde{h}_1, \tilde{h}_2)$, $(\tilde{h}_1, \tilde{h}_1, \tilde{h}_3)$, $(\tilde{h}_1, \tilde{h}_3, \tilde{h}_3)$, $(\tilde{h}_1, \tilde{h}_2, \tilde{h}_3)$ *is used.* For these strategy profiles, consider the following set of configurations:

$$C_4 = \{(1, 3, 2, +1, +1, +1), (2, 1, 3, +1, +1, +1), (2, 1, 3, +1, -1, +1)\}.$$

In each of these profiles, two players meet at time $T_0 = 1$, $\Gamma_{up} = \Gamma''(1, 1)$ so $T_{up} = 1$ by Lemma 3, and $\Gamma_{down} = \Gamma'(2)$, so $T_{down} = 2$ by Lemma 2. Hence in all these cases, $T \geq 1 + 1 + 2 = 4$. □

THEOREM 3. $M_3 = 4$. *That is, the minimax rendezvous time for three players placed in random directions on consecutive integers is* 4.

Proof. Lemma 4 says that $M_3 \leq 4$, so we need only show that $M_3 \geq 4$. Since we are assuming that all paths belong to P^*, it is sufficient to show that $M_3 > 3.5$. Lemma 5 says that any strategy with maximum rendezvous time ≤ 3.5 must have all of its Stage 1 paths belonging to the set $\{\tilde{h}_1, \tilde{h}_2, \tilde{h}_3\}$. However, Lemma 6 says that any strategy triple with this property must have maximum rendezvous time of at least 4. □

6. Asymptotic value of M_n. In this section we estimate the value of the minimax rendezvous time M_n for the n-player game Γ_n when n is large. Clearly a lower bound for M_n is $(n-1)/2$, since the distance between the players initially placed at 1 and n is $n-1$. The main work of this section is the presentation and analysis of a class of strategy profiles $S(n, m)$ for the games Γ_n which have maximum rendezvous times asymptotic to $n/2$. This analysis thus gives the main result of this section (Theorem 4), that M_n is asymptotic to $n/2$.

We now define the strategy profile $S(n, m)$ for the games Γ_n. Up to time $T_0 = 3m + 1$ the players adopt *adjacency search* paths called g_k. (The paths g_1, g_2, g_3 are drawn in Figure 5, for players initially pointed to the right.) These paths remain at a player's starting point

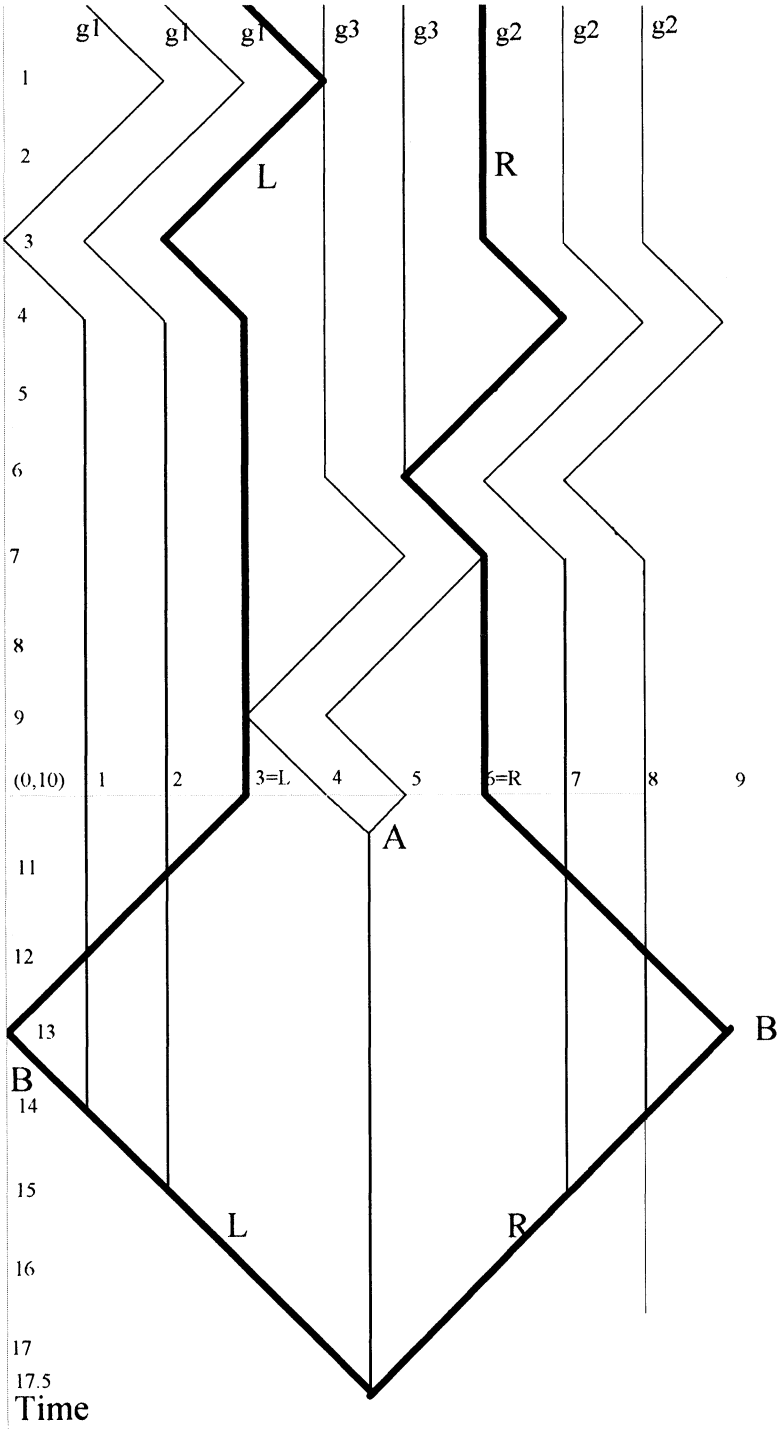


FIG. 5. $T(S(8,3)) = 17.5$.

except during the time interval $[3(k-1), 3k+1]$ of length 4, when they search first forward 1 unit, then backwards 2 units, and then forward again to return to the starting point. This path will meet any adjacent player who is stationary at their starting point during this period

(in particular at times $3(k-1)+1$ and $3(k-1)+3$). More formally these *adjacency search* paths are defined as

$$g_k(j) = \begin{cases} 1 & \text{if } j = 3(k-1)+1, \\ -1 & \text{if } j = 3(k-1)+3, \\ 0 & \text{otherwise.} \end{cases}$$

Observe that if $k < k'$ and two adjacent players are using *adjacency search* paths g_k and $g_{k'}$, then they will meet at time $3(k-1)+1$ or $3(k-1)+3$, and in any case by time $3(k-1)+3$.

In the strategy profile $S(n, m)$ the players use the first m *adjacency search* paths $g_1, \ldots, g_n$, in as equal numbers as possible. We will take as an example $S(8, 3)$, which is illustrated in Figure 5. Let $n = am + b$, $0 \le b < m$, and let exactly $a+1$ players use each strategy g_k, $k = 1, \ldots, b$, and let exactly a players use each *adjacency search* path g_k, $k = b+1, \ldots, m$, for times $0 \le t \le T_0 = 3m + 1$, disregarding (for the time being) any players they may meet. Note that $a = \text{Int}(n/m)$. (In the example, $a = 2$ and three players use g_1, three use g_2, and two use g_3.) Observe that at time $T_0 = 3m + 1$ (10, in the example) all the players are back at their starting points and that any pair of adjacent players who are using distinct strategies g_k will have already met each other, regardless of the directions in which they are initially pointed. Once the players have been placed on the integers $1, \ldots, n$, name them according to the integer where they start. We use a horizontal description of the line on which the players are placed. Let L denote the leftmost player for which the adjacent player on the right is using a different initial strategy g_k. Let R denote the rightmost player such that the player on his left is using a different strategy. (In the example $L = 3$ and $R = 6$.) Since there are at most a players to the left of L (who are using the same strategy g_k as player L) and similarly at most a players to the right of player R, we have

$$L \le a + 1 \text{ and } R \ge n - a. \tag{7}$$

Note that equality holds in the above if and only if the first $a+1$ players are all using the same strategy and the last $a+1$ players are all using the same strategy, and these end groups are initially pointed in a common direction. This configuration (shown in Figure 5) produces the maximum meeting time T.

We now describe the strategies the players adopt from time $T_0 = 3m + 1$. At this time the players have met either no adjacent players, two adjacent players, or exactly one adjacent player. (In the example of Figure 5, players 1, 2, 7, and 8 are of the first type, nobody is of the second type, and players 3, 4, 5, and 6 are of the third type.) Players of the first two types should remain still at their starting points until they meet a player who says, "follow me." Players of the third type, who have met an adjacent player in only one direction, should go in the opposite direction (at speed one) until they either (A) meet another moving player or (B) reach an unoccupied integer location (relative to their starting point). In case (A) they stop and remain still until someone says, "follow me." In case (B) they can conclude they are at position 0 or $n+1$ and hence they reverse direction and go at speed one, telling anyone they meet to follow them, until the game ends. (In the example of Figure 5, players L = 3 and R = 6 reach situation (B) at time 13, while players 4 and 5 reach situation (A) at time 10.5.) In general, it follows from the inequalities (7) that players L and R reach positions 0 and $n+1$, respectively (situation (B)), by time $(3m+1)+(a+1)$. They then meet each other, together with everyone else, by maximum T, where

$$T = T_n \le (3m+1) + (a+1) + (n+1)/2. \tag{8}$$

(In the example of Figure 5, this gives a worst case of 17.5.)

Suppose we define $m = \text{Int}(\log\ n)$ so that $a = \text{Int}(n/\text{Int}(\log\ n))$. It follows that

$$\frac{T_n}{n} \leq \frac{3\,\text{Int}(\log\ n) + 2 + \text{Int}(n/\text{Int}(\log\ n)) + 1/2}{n} + \frac{1}{2} \underset{n\to\infty}{\longrightarrow} \frac{1}{2}.$$

Consequently the minimax rendezvous time M_n satisfies

$$\frac{1}{2} = \lim_{n\to\infty} \frac{(n-1)/2}{n} \leq \lim_{n\to\infty} \frac{M_n}{n} \leq \lim_{n\to\infty} \frac{T_n}{n} = \frac{1}{2}, \text{ or } \lim_{n\to\infty} \frac{M_n}{n} = \frac{1}{2}.$$

Thus we have proved our final result.

THEOREM 4. *The minimax rendezvous time M_n, required for n players placed on adjacent integers to meet together at a single point, is asymptotic to $n/2$.*

REFERENCES

[1] S. ALPERN, *The rendezvous search problem*, SIAM J. Control Optim., 33 (1995), pp. 673–683.

[2] S. ALPERN AND S. GAL, *Rendezvous search on the line with distinguishable players*, SIAM J. Control Optim., 33 (1995), pp. 1270–1276.

[3] E. J. ANDERSON AND S. ESSEGAIER, *Rendezvous search on the line with indistinguishable players*, SIAM J. Control Optim., 33 (1995), pp. 1637–1642.

[4] V. J. BASTON AND F. A. BOSTOCK, *A high-low search game on the unit interval*, Math. Proc. Cambridge Philos. Soc., 97 (1985), pp. 345–348.

[5] S. GAL, *Search Games*, Academic Press, New York, 1980.

[6] W. S. LIM, S. ALPERN, AND A. BECK, *Rendezvous search on the line with more than two players*, Oper. Res., in press.

[7] D. J. REYNIERS, *Coordinating two searchers for an object hidden on an interval*, Journal of the Operational Research Society, 46 (1995), pp. 1386–1392.

[8] ———, *Coordinated search for an object hidden exponentially on the line*, European Journal of Operational Research, (1995), to appear.

[9] W. H. RUCKLE, *Geometric Games and their Applications*, Pitman, Boston, MA, 1983.

[10] T. SCHELLING, *The Strategy of Conflict*, Harvard University Press, Cambridge, MA, 1960.

[11] L. C. THOMAS, *Finding your kids when they are lost*, Journal of the Operational Research Society, 43 (1992), pp. 637–639.

[12] L. C. THOMAS AND P. B. HULME, *Searching for targets who want to be found*, preprint, University of Edinburgh, 1993.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1666–1690, September 1996

010

POLYNOMIAL FILTERING FOR LINEAR DISCRETE TIME NON-GAUSSIAN SYSTEMS*

FRANCESCO CARRAVETTA†, ALFREDO GERMANI‡, AND MASSIMO RAIMONDI§

Abstract. In this work we propose a new filtering approach for linear discrete time non-Gaussian systems that generalizes a previous result concerning quadratic filtering [A. De Santis, A. Germani, and M. Raimondi, *IEEE Trans. Automat. Control*, 40 (1995) pp. 1274–1278]. A recursive νth-order polynomial estimate of finite memory Δ is achieved by defining a suitable extended state which allows one to solve the filtering problem via the classical Kalman linear scheme. The resulting estimate will be the mean square optimal one among those estimators that take into account ν-polynomials of the last Δ observations. Numerical simulations show the effectiveness of the proposed method.

Key words. nonlinear filtering, polynomial estimates, recursive estimates, non-Gaussian systems

AMS subject classifications. 93E10, 93E11

1. Introduction.

In this paper the state estimation problem for linear non-Gaussian systems is considered. In many important technical areas the widely used Gaussian assumption cannot be accepted as a realistic statistical description of the random quantities involved. As shown in various papers (see for instance [1, 2]), increasing attention has been paid in control engineering to non-Gaussian systems, and the importance of parameters and state estimation problems is plainly evidenced. In these cases the conditional expectation, which gives the optimal minimum variance estimate, cannot be generally computed, so that it is necessary to look for suboptimal estimates that are easier to achieve, such as the optimal linear one. In recent years, the signal filtering and detection problems in the presence of non-Gaussian noise have been widely investigated with different signal models and statistical settings. Non-Gaussian problems often arise in digital communications when the noise interference includes noise components that are essentially non-Gaussian (this is a common situation below 100 MHz) [6]. Neglecting these components is a major source of error in communication system design. In [3, 4] the existence of stable filters for a class of nonlinear stochastic systems is studied, where the nonlinearity is defined not by its deterministic structure but by its statistical properties. In [5] the Bayesian approach to nonlinear parameter estimation is considered and the cost of computing the posterior density description is investigated when the Bayes formula is recursively applied. In telecommunication systems the detection problem in the presence of non-Gaussian noises is extensively addressed in [6]–[12], while in [13] a general abstract setting is considered for high-order statistical processing (Volterra filters). A first attempt in the definition of a polynomial filter, which in some sense generalizes the Kalman approach, is described in [14], where, in particular, an instantaneous polynomial function of the innovation process constitutes the forcing term for the linear dynamic of the filter. The computation of the polynomial coefficients, which generalizes the Kalman gain to the non-Gaussian case, remains the main problem. In [15] the linear recursive estimation is dealt with for stochastic signals having multiplicative noise and in [16] for linear discrete time systems with stochastic parameters. In [17] an asymptotic minimum variance algorithm is described for parameter estimation in non-Gaussian moving average (MA) and autoregressive moving average (ARMA)

*Received by the editors July 28, 1993; accepted for publication (in revised form) May 31, 1995. This work was partially supported by MURST.

†Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza," Via Eudossiana 18, 00184 Roma, Italy.

‡Dipartimento di Ingegneria Elettrica, Università dell'Aquila, 67100 Monteluco (L'Aquila), Italy, and Istituto di Analisi dei Sistemi ed Informatica del CNR, Viale Manzoni 30, 00185 Roma, Italy.

§Dipartimento di Matematica G. Castelnuovo, Università di Roma "La Sapienza," Piazzale Aldo Moro 2, 00185 Roma, Italy.

processes, using sample high-order statistics. The same problem is studied in [18] by using a fixed set of output cumulants. In [19], on the basis of the knowledge of the output process together with its Kronecker square products, a linear filter with respect to such information process is defined.

In this paper we consider the more general polynomial case, where past values of the output process are also considered.

The paper is organized as follows: in §2 we recall some definitions and properties on the estimation theory in a geometric framework. Moreover, some results on the Kronecker algebra are given. In §3 the non-Gaussian filtering problem is formulated with reference to a linear discrete time system. The augmented state and the corresponding dynamical model generating process are defined. In §4 some theoretical results useful for the practical implementation of the proposed algorithm are reported. Finally in §5 some numerical examples of application are presented, showing high performance of the proposed filter with respect to the Kalman one. The paper ends with a concluding remark in §6.

2. Preliminaries.

2.1. Estimates as projections. In this section, we will consider the mean square optimal (and suboptimal) estimate of a partially observed random variable as a projection onto a suitable L^2-subspace.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. For any given sub σ-algebra $\mathcal{G}$ of $\mathcal{F}$ let us denote by $L^2(\mathcal{G}, n)$ the Hilbert space of the n-dimensional, $\mathcal{G}$-measurable, random variables with finite second moment as

$$L^2(\mathcal{G}, n) = \left\{ X : \Omega \to \mathbb{R}^n,\ \mathcal{G}\text{-measurable},\ \int_\Omega \|X(\omega)\|^2 dP(\omega) < +\infty \right\},$$

where $\|\cdot\|$ is the euclidean norm in $\mathbb{R}^n$. Moreover, when $\mathcal{G}$ is the σ-algebra generated by a random variable $Y : \Omega \to \mathbb{R}^m$, that is, $\mathcal{G} = \sigma(Y)$, we will use the notation $L^2(Y, n)$ to indicate $L^2(\sigma(Y), n)$. Finally if M is a closed subspace of $L^2(\mathcal{F}, n)$, we will use the symbol $\Pi(X/M)$ to indicate the orthogonal projection of $X \in L^2(\mathcal{F}, n)$ onto M.

As is well known, the optimal minimum variance estimate of a random variable $X \in L^2(\mathcal{F}, n)$ with respect to a random variable Y, that is, $\Pi(X/L^2(Y, n))$, is given by the conditional expectation (C.E.) $E(X/Y)$. If X and Y are jointly Gaussian, then the C.E. is the following affine transformation of Y:

$$E(X/Y) = E(X) + E(X\tilde{Y}^T)E(\tilde{Y}\tilde{Y}^T)^{-1}\tilde{Y}, \tag{2.1.1}$$

where $\tilde{Y} = Y - E(Y)$.

Moreover, defining

$$Y' = \begin{bmatrix} 1 \\ Y \end{bmatrix},$$

(2.1.1) can be also interpreted as the projection on the subspace

$$\mathcal{L}(Y', n) = \{Z : \Omega \to \mathbb{R}^n \,/\, \exists\ A \in \mathbb{R}^{n\times(m+1)} \text{ such that } Z = AY'\} \subset L^2(Y', n) = L^2(Y, n).$$

Unfortunately in the non-Gaussian case, no simple characterization of the C.E. can be achieved. Consequently it is worthwhile to consider suboptimal estimates which have a simpler mathematical structure that allows the treatment of real data. The simplest suboptimal estimate is the optimal affine one, that is, $\Pi(X/\mathcal{L}(Y', n))$, which is given again by the right-hand side (R.H.S.) of (2.1.1). In the following discussion such an estimate will be denoted with $\hat{X}$ and

shortly called an optimal linear one. Intermediate estimates between the optimal linear and the C.E. can be considered by projecting onto subspaces greater than $\mathcal{L}(Y', n)$, like subspaces of polynomial transformations of Y. In order to proceed this way, we need to state some results on the Kronecker products [20] that constitute a powerful tool in treating vector polynomials.

2.2. The Kronecker algebra.

DEFINITION 2.2.1. *Let M and N be matrices of dimensions $r \times s$ and $p \times q$ respectively. Then the Kronecker product $M \otimes N$ is defined as the $(r \cdot p) \times (s \cdot q)$ matrix*

$$M \otimes N = \begin{bmatrix} m_{11}N & \dots & m_{1s}N \\ \dots\dots\dots\dots & & \\ m_{r1}N & \dots & m_{rs}N \end{bmatrix},$$

where the m_{ij} are the entries of M. Of course this kind of product is not commutative.

DEFINITION 2.2.2. *Let M be the $r \times s$ matrix*

$$M = [\, m_1 \quad m_2 \quad \dots \quad m_s \,], \tag{2.2.1}$$

where m_i denotes the ith column of M. Then the stack of M is the $r \cdot s$ vector

$$st(M) = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_s \end{bmatrix}. \tag{2.2.2}$$

Observe that a vector as in (2.2.2) can be reduced to a matrix M as in (2.2.1) by considering the inverse operation of the stack denoted by st^{-1}. We refer to [20, Chap. 12] for the main properties of the Kronecker product and stack operation. It is easy to verify that for $u \in \mathbb{R}^r$, $v \in \mathbb{R}^s$, the ith entry of $u \otimes v$ is given by

$$(u \otimes v)_i = u_l \cdot v_m, \quad l = \left[\frac{i-1}{s}\right] + 1, \quad m = |i-1|_s + 1, \tag{2.2.3}$$

where $[\cdot]$ and $|\cdot|_s$ denote integer part and s-modulo, respectively. Moreover, the Kronecker power of M is defined as

$$M^{[0]} = 1 \in \mathbb{R},$$
$$M^{[l]} = M \otimes M^{[l-1]}, \quad l \geq 1.$$

Even if the Kronecker product is not commutative in general, the following result holds [24].

THEOREM 2.2.3. *For any given pair of matrices $A \in \mathbb{R}^{r \times s}$, $B \in \mathbb{R}^{n \times m}$, we have*

$$B \otimes A = C_{r,n}^T (A \otimes B) C_{s,m}, \tag{2.2.4}$$

where $C_{r,n}$, $C_{s,m}$ are suitable $0-1$ matrices.

It is possible to show that $C_{u,v}$ is the $(u \cdot v) \times (u \cdot v)$ matrix such that its (h, l) entry is given by

$$\{C_{u,v}\}_{h,l} = \begin{cases} 1 & \text{if } l = (|h-1|_v)u + \left(\left[\dfrac{h-1}{v}\right] + 1\right); \\ 0 & \text{otherwise.} \end{cases} \tag{2.2.5}$$

Observe that $C_{1,1} = 1$, hence in the vector case when $a \in \mathbb{R}^r$ and $b \in \mathbb{R}^n$, (2.2.4) becomes

$$b \otimes a = C_{r,n}^T (a \otimes b). \tag{2.2.6}$$

Moreover, in the vector case the commutation matrices also satisfy the following recursive formula.

LEMMA 2.2.4. *For any $a, b \in \mathbb{R}^n$ and for any $l = 1, 2 \ldots$, let $G_l = C^T_{n,n^l}$ so that*

$$b^{[l]} \otimes a = G_l(a \otimes b^{[l]}). \tag{2.2.7}$$

Then the sequence $\{G_l\}$ satisfies the following equations:

$$G_1 = C^T_{n,n}, \tag{2.2.8}$$

$$G_l = (I_1 \otimes G_{l-1}) \cdot (G_1 \otimes I_{l-1}), \quad l > 1, \tag{2.2.9}$$

where I_r is the identity matrix in $\mathbb{R}^{n^r}$.

Proof. Equation (2.2.6) assures the existence of the G_l's and implies (2.2.8). Moreover, using the associative property of the Kronecker product and recalling the identity

$$(A \cdot C) \otimes (B \cdot D) = (A \otimes B) \cdot (C \otimes D)$$

with $A = I_1$, we have

$$\begin{aligned} b^{[l]} \otimes a &= b \otimes b^{[l-1]} \otimes a \\ &= b \otimes \left(G_{l-1}(a \otimes b^{[l-1]})\right) \\ &= (I_1 \otimes G_{l-1}) \cdot (b \otimes a \otimes b^{[l-1]}) \\ &= (I_1 \otimes G_{l-1}) \cdot \left((G_1(a \otimes b)) \otimes b^{[l-1]}\right) \\ &= (I_1 \otimes G_{l-1}) \cdot (G_1 \otimes I_{l-1}) \cdot (a \otimes b^{[l]}). \end{aligned}$$

Then equation (2.2.9) follows immediately by using (2.2.7). □

We can also find a binomial formula for the Kronecker power which generalizes the classical Newton one.

THEOREM 2.2.5. *For any integer $h \geq 0$ the matrix coefficients of the binomial power formula*

$$(a + b)^{[h]} = \sum_{k=0}^{h} M_k^h (a^{[k]} \otimes b^{[h-k]}) \tag{2.2.10}$$

constitute a set of matrices $\{M_0^h, \ldots, M_h^h\}$ such that

$$M_h^h = M_0^h = I_h, \tag{2.2.11}$$

$$M_j^h = (M_j^{h-1} \otimes I_1) + (M_{j-1}^{h-1} \otimes I_1) \cdot (I_{j-1} \otimes G_{h-j}), \quad 1 \leq j \leq h-1, \tag{2.2.12}$$

where G_l and I_l are as in Lemma 2.2.4.

Proof. Equation (2.2.11) is obviously true for any h.

We will prove (2.2.12) by induction for $h \geq 2$. For $h = 2$ it results in

$$\begin{aligned} (a + b)^{[2]} &= a^{[2]} + a \otimes b + b \otimes a + b^{[2]} \\ &= a^{[2]} + (I_2 + G_1)(a \otimes b) + b^{[2]}, \end{aligned} \tag{2.2.13}$$

where (2.2.7) has been used. Moreover, using (2.2.12) we obtain

$$M_1^2 = (M_1^1 \otimes I_1) + (M_0^1 \otimes I_1)(I_0 \otimes G_1) = I_2 + I_2 G_1 = I_2 + G_1$$

so that the matrix coefficient of $a \otimes b$ in (2.2.10) (which is equal to $I_2 + G_1$ by (2.2.13)) agrees with the matrix M_1^2 computed by using (2.2.12). Now suppose that (2.2.12) is true for $h \geq 2$. Then we will prove that it is true for $h+1$. We have

$$\begin{aligned}
(a+b)^{[h+1]} &= (a+b)^{[h]} \otimes (a+b) \\
&= \left(\sum_{k=0}^{h} M_k^h (a^{[k]} \otimes b^{[h-k]}) \right) \otimes (a+b) \\
&= \sum_{k=0}^{h} \left((M_k^h \otimes I_1) \cdot (a^{[k]} \otimes b^{[h-k]} \otimes a) + (M_k^h \otimes I_1) \cdot (a^{[k]} \otimes b^{[h+1-k]}) \right) \\
&= \sum_{k=0}^{h} \Big((M_k^h \otimes I_1) \cdot \big(a^{[k]} \otimes \big(G_{h-k}(a \otimes b^{[h-k]})\big)\big) \\
&\qquad + (M_k^h \otimes I_1) \cdot (a^{[k]} \otimes b^{[h+1-k]}) \Big) \\
&= \sum_{k=0}^{h} (M_k^h \otimes I_1) \cdot (I_k \otimes G_{h-k}) \cdot (a^{[k+1]} \otimes b^{[h-k]}) \\
&\qquad + \sum_{k=0}^{h} (M_k^h \otimes I_1) \cdot (a^{[k]} \otimes b^{[h+1-k]}).
\end{aligned}$$

Hence, taking into account (2.2.10) we have

$$M_j^{h+1} = (M_j^h \otimes I_1) + (M_{j-1}^h \otimes I_1) \cdot (I_{j-1} \otimes G_{h+1-j}), \quad 1 \leq j \leq h. \qquad \square$$

2.3. Polynomial estimates. Let $X \in L^2(\mathcal{F}, n)$, $Y \in L^2(\mathcal{F}, m)$ be random variables and, moreover, suppose that for some integer i, $\int_\Omega \|Y\|^{2i} dP < +\infty$. Then we can define the ith-order polynomial estimate of X as $\Pi(X/\mathcal{L}(\mathcal{Y}_i, n))$, where $\mathcal{Y}_i \in L^2(\mathcal{F}, 1 + m + \cdots + m^i)$ is given by

$$\mathcal{Y}_i = \begin{bmatrix} 1 \\ Y \\ \vdots \\ Y^{[i]} \end{bmatrix}.$$

Note that $\mathcal{L}([\begin{smallmatrix} 1 \\ Y \end{smallmatrix}], n) = \mathcal{L}(\mathcal{Y}_1, n) \subset \cdots \subset \mathcal{L}(\mathcal{Y}_{i-1}, n) \subset \mathcal{L}(\mathcal{Y}_i, n)$ so that a polynomial estimate improves (in terms of error variance) the performance of the linear one. Observe, moreover, that the previous estimate has the form

$$\sum_{l=0}^{i} c_l Y^{[l]}, \quad c_l \in \mathbb{R}^{n \times m^l}, \tag{2.3.1}$$

which justifies the term polynomial used in this paper. If $\int_\Omega \|Y\|^{2i} dP < +\infty \ \forall i \in \mathbb{N}$, let $\mathcal{H}$ be defined as the L^2-closure of $\cup_{i=0}^{+\infty} \mathcal{L}(\mathcal{Y}_i, n)$. Then the C.E. can be decomposed as

$$E(X/Y) = \Pi(X/\mathcal{H}) + \Pi(X/\mathcal{H}^\perp), \tag{2.3.2}$$

where the first term of the R.H.S. of (2.3.2) is the L^2-limit of a sequence of polynomials of Y. In particular such a sequence can be obtained by projecting X on the subspaces $\mathcal{L}(\mathcal{Y}_i, n)$,

so that the difficulty in computing the C.E. is moved to the second part of the R.H.S. of (2.3.2). In any case we can compute the coefficients in (2.3.1) of any finite-rank polynomial approximation of the term $\Pi(X/\mathcal{H})$ by using the linear estimate formula given by the R.H.S. of (2.1.1).

3. Problem formulation.

3.1. The system to be filtered. Let us consider the filtering problem for the following class of linear discrete time systems:

$$x(k+1) = Ax(k) + FN(k), \qquad x(0) = \bar{x}, \tag{3.1.1}$$

$$y(k) = Cx(k) + GN(k), \tag{3.1.2}$$

where $x(k) \in \mathbb{R}^n$, $y(k) \in \mathbb{R}^m$, $N(k) \in \mathbb{R}^u$, $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{m\times n}$, $F \in \mathbb{R}^{n\times u}$, $G \in \mathbb{R}^{m\times u}$. The random variable $\bar{x}$ (the initial condition) and the random sequence $\{N(k)\}$ satisfy the following conditions for $k \geq 0$:

i) $E\{\bar{x}\} = 0$, $E\{N(k)\} = 0$;

ii) there exists an integer $\nu \geq 1$ such that for any given multiindex $i_1, \ldots, i_L \in \{1, \ldots, u\}$, $j_1, \ldots, j_L \in \{1, \ldots, n\}$, $1 \leq L \leq 2\nu$ we have

$$F(i_1, \ldots, i_L) \triangleq E\{|N_{i_1}(k)N_{i_2}(k) \cdot \cdots \cdot N_{i_L}(k)|\} < \infty, \tag{3.1.3}$$

$$X(j_1, \ldots, j_L) \triangleq E\left\{|\bar{x}_{j_1}\bar{x}_{j_2} \cdot \cdots \cdot \bar{x}_{j_L}|\right\} < \infty; \tag{3.1.4}$$

iii) the sequence $\{N(k)\}$ forms with $\bar{x}$ a family of independent random variables.

3.2. Recursive estimates. It is well known that the optimal mean square state estimate for the state $x(k)$ of the linear system (3.1.1), (3.1.2) with respect to the observations up to the time k is given by the conditional expectation

$$\hat{x}(k) = E(x(k)/\mathcal{F}_k^y), \tag{3.2.1}$$

where $\mathcal{F}_k^y$ is the σ-algebra generated by $\{y(\tau),\ \tau \leq k\}$. Hence there exists a Borel function F such that

$$\hat{x}(k) = F(y(\tau),\ \tau \leq k).$$

As we have already seen in §2, the computation of F could be very difficult and, in general, does not produce a recursive algorithm, so it does not turn out to be very useful from an application point of view. If we are interested only in an optimal linear estimate, then we can also express the above estimate in the general recursive form,

$$\hat{x}(k) = F(k, \hat{x}(k-1), y(k)). \tag{3.2.2}$$

In fact the well-known Kalman filter, which gives the optimal linear estimate of the state, is expressed as in (3.2.2) with a linear transformation F. More generally we can consider the set of the recursive Borel transformations of finite memory Δ, that is,

$$\hat{x}(k) = \rho(k, \hat{x}(k-1), y(k), y(k-1), \ldots, y(k-\Delta)). \tag{3.2.3}$$

In order to realize (3.2.3) we will adopt the larger class of recursive functions

$$\begin{aligned} &\hat{x}(k) = \mathcal{T}\xi(k), \\ &\xi(k) = \varrho(k, \xi(k-1), y(k), y(k-1), \ldots, y(k-\Delta)), \end{aligned} \tag{3.2.4}$$

where $\xi(k) \in L^2(\mathcal{F}, n')$, $n' \geq n$, and $\mathcal{T}$ is the (linear) operator that extracts the first n components of $\xi(k)$. In particular the method that will be proposed will allow us to obtain an estimate in the form

$$\begin{aligned} &\hat{x}(k) = \mathcal{T}\xi(k), \\ &\xi(k) = L(k)\xi(k-1) + \mathcal{P}(y(k), y(k-1), \ldots, y(k-\Delta)), \end{aligned} \tag{3.2.5}$$

where $L(k) \in \mathbb{R}^{n' \times n'}$ and $\mathcal{P}$ is a polynomial transformation. One way to justify (3.2.5) is that a similar form is optimal in some interesting cases [25].

3.3. The extended system. In order to obtain a recursive estimate like in (3.2.5), as a first step we introduce the following extended vectors:

$$x_e(k) = \begin{bmatrix} x(k) \\ y(k-1) \\ \vdots \\ y(k-\Delta) \end{bmatrix} \in \mathbb{R}^q, \qquad y_e(k) = \begin{bmatrix} y(k) \\ y(k-1) \\ \vdots \\ y(k-\Delta) \end{bmatrix} \in \mathbb{R}^p, \tag{3.3.1}$$

with $q = n + m\Delta$ and $p = (\Delta + 1)m$. The model equations (3.1.1), (3.1.2) become

$$x_e(k+1) = A_e x_e(k) + F_e N(k), \quad x_e(0) = \bar{x}_e, \tag{3.3.2}$$

$$y_e(k) = C_e x_e(k) + G_e N(k), \tag{3.3.3}$$

where

$$A_e = \begin{bmatrix} A & 0 & \ldots & \ldots & 0 \\ C & 0 & \ldots & \ldots & 0 \\ 0 & I & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & I & 0 \end{bmatrix}, \quad F_e = \begin{bmatrix} F \\ G \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \bar{x}_e = \begin{bmatrix} \bar{x} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{3.3.4}$$

$$C_e = \begin{bmatrix} C & 0 & \ldots & 0 \\ 0 & I & \ddots & \\ \vdots & & \ddots & \vdots \\ 0 & & \ldots & I \end{bmatrix}, \quad G_e = \begin{bmatrix} G \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{3.3.5}$$

Moreover, let us define the generalized νth-degree polynomial observation as the vector $\mathcal{Y} \in \mathbb{R}^\mu$, $\mu = p + p^2 + \cdots + p^\nu$:

$$\mathcal{Y}(k) = \begin{bmatrix} y_e(k) \\ y_e^{[2]}(k) \\ \vdots \\ y_e^{[\nu]}(k) \end{bmatrix}. \tag{3.3.6}$$

Finally let us introduce the extended state $\mathcal{X} \in \mathbb{R}^\chi$, $\chi = q + q^2 + \cdots + q^\nu$:

$$\mathcal{X}(k) = \begin{bmatrix} x_e(k) \\ x_e^{[2]}(k) \\ \vdots \\ x_e^{[\nu]}(k) \end{bmatrix}. \tag{3.3.7}$$

In the following discussion we will denote with $M_i^j(l)$ the binomial matrices (2.2.11), (2.2.12) highlighting the dependence by the dimension l of the vectors, and the symbol $I_{i,j}$ will denote the identity in $\mathbb{R}^{i^j}$. In order to obtain a recursive filter we need to write an evolution equation for the extended state $\mathcal{X}(k)$ and another one that links it to $\mathcal{Y}(k)$. For this purpose we can prove the following important result.

LEMMA 3.3.1. *Let, on the same probability space,* $\{z(k),\ k \geq 0\}$ *and* $\{N(k),\ k \geq 0\}$ *be random sequences in* $\mathbb{R}^\alpha$ *and* $\mathbb{R}^\beta$, *respectively, such that* $\forall k\ N(k)$ *is independent by* $\{z(k), z(j), N(j),\ j < k\}$. *Moreover, let us assume*

$$w(k) = \Gamma z(k) + \Psi N(k), \tag{3.3.8}$$

where $w(k) \in \mathbb{R}^\gamma$ *and* $\Gamma,\ \Psi$ *are subsequently dimensioned deterministic matrices. Consider the Kronecker powers of* $w(k)$ *and* $z(k)$ *up to the* ν*th order aggregated in the vectors*

$$\mathcal{W}(k) = \begin{bmatrix} w(k) \\ w^{[2]}(k) \\ \vdots \\ w^{[\nu]}(k) \end{bmatrix}, \quad \mathcal{Z}(k) = \begin{bmatrix} z(k) \\ z^{[2]}(k) \\ \vdots \\ z^{[\nu]}(k) \end{bmatrix},$$

and

$$\mathcal{O} = \begin{bmatrix} \Gamma & 0 & \dots & 0 \\ O_{2,1} & \Gamma^{[2]} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ O_{\nu,1} & O_{\nu,2} & \dots & \Gamma^{[\nu]} \end{bmatrix}, \quad \mathcal{T} = \begin{bmatrix} \Psi E(N(k)) \\ \Psi^{[2]} E(N^{[2]}(k)) \\ \vdots \\ \Psi^{[\nu]} E(N^{[\nu]}(k)) \end{bmatrix},$$

where

$$O_{i,l} = M_{i-l}^i(\gamma)(\Psi^{[i-l]} \otimes \Gamma^{[l]}) \left(E(N^{[i-l]}(k)) \otimes I_{\alpha,l}\right). \tag{3.3.9}$$

Then there exists the representation

$$\mathcal{W}(k) = \mathcal{O}\mathcal{Z}(k) + \mathcal{T} + \mathcal{N}(k), \tag{3.3.10}$$

where

$$\mathcal{N}(k) = \begin{bmatrix} h_1(k) \\ h_2(k) \\ \vdots \\ h_\nu(k) \end{bmatrix} \tag{3.3.11}$$

and

$$h_i(k) = \sum_{l=0}^{i-1} M_{i-l}^i(\gamma)(\Psi^{[i-l]} \otimes \Gamma^{[l]}) \left(\left(N^{[i-l]}(k) - E(N^{[i-l]}(k))\right) \otimes I_{\alpha,l}\right) z^{[l]}(k).$$

Moreover, $\{\mathcal{N}(k)\}$ *is a zero-mean white sequence such that* $\forall k\ \mathcal{N}(k)$ *is uncorrelated with* $\{\mathcal{Z}(j),\ j \leq k\}$, *with covariance* $\mathcal{S}(k)$ *such that its* (r, s)*-block is given by*

$$\begin{aligned} \mathcal{S}_{r,s} &\triangleq E(h_s(k)h_r(k)^T) \\ &= \sum_{l=0}^{r-1} \sum_{m=0}^{s-1} M_{r-l}^r(\gamma)(\Psi^{[r-l]} \otimes \Gamma^{[l]}) \cdot st^{-1}\left((I_{\beta,s-m} \otimes C_{\beta^{r-l},\alpha^m}^T \otimes I_{\alpha,l})\right. \\ &\quad \cdot \left((E(N^{[s+r-m-l]}(k)) - E(N^{[s-m]}(k)) \otimes E(N^{[r-l]}(k)))\right. \\ &\quad \left.\left. \otimes C_{1,\alpha^m} \otimes I_{\alpha,l}\right) \cdot E(z^{[l+m]}(k))\right)(\Psi^{[s-m]} \otimes \Gamma^{[m]})^T (M_{s-m}^s(\gamma))^T, \end{aligned} \tag{3.3.12}$$

provided that there exist finite all the moments involved.

Proof. Taking the ith Kronecker power of both members in (3.3.8), we have

$$w^{[i]}(k) = (\Gamma z(k) + \Psi N(k))^{[i]}, \tag{3.3.13}$$

which can be exploited by using Theorem 2.2.5 so that

$$\begin{aligned} w^{[i]}(k) &= (\Gamma z(k))^{[i]} + \sum_{j=1}^{i} M_j^i(\gamma)\left((\Psi N(k))^{[j]} \otimes (\Gamma z(k))^{[i-j]}\right) \\ &= \Gamma^{[i]} z^{[i]}(k) + \sum_{j=1}^{i} M_j^i(\gamma)(\Psi^{[j]} \otimes \Gamma^{[i-j]})(N^{[j]}(k) \otimes z^{[i-j]}(k)) \\ &= \Gamma^{[i]} z^{[i]}(k) + \sum_{l=0}^{i-1} M_{i-l}^i(\gamma)(\Psi^{[i-l]} \otimes \Gamma^{[l]})(N^{[i-l]}(k) \otimes I_{\alpha,l}) z^{[l]}(k), \end{aligned} \tag{3.3.14}$$

from which (3.3.10) follows.

Now, let us consider the above-defined "augmented noise" $\mathcal{N}(k)$. From the independence of $z(k)$ and $N(k)$ (and hence, the independence of $N^{[i-l]}(k) - E(N^{[i-l]}(k))$ and $z^{[l]}(k)$ $\forall l = 0, \ldots, i-1$) the zero mean property for $\mathcal{N}(k)$ follows, as can be readily verified. To prove the whiteness property, suppose $k > j$. First of all, observe that because $N(k)$, by the hypotheses, is independent of $\{z(k), z(j), N(j),\ j < k\}$, it follows that

$$N^{[r-l]}(k) - E(N^{[r-l]}(k))$$

is independent of

$$z^{[l]}(k) z^{[m]^T}(j)\left((N^{[s-m]}(j) - E(N^{[s-m]}(j))) \otimes I_{\alpha,m}\right)^T;$$

then for the (r, s)-block of the covariance matrix we have

$$\begin{aligned} &\left(E(\mathcal{N}(k)\mathcal{N}(j)^T)\right)_{r,s} = E(h_r(k) h_s(j)^T) \\ &= \sum_{l=0}^{r-1} \sum_{m=0}^{s-1} M_{r-l}^r(\gamma)\left(\Psi^{[r-l]} \otimes \Gamma^{[l]}\right) \\ &\quad \cdot E\Big(((N^{[r-l]}(k) - E(N^{[r-l]}(k))) \otimes I_{\alpha,l}) z^{[l]}(k) z^{[m]}(j)^T \\ &\qquad \cdot \left(\left(N^{[s-m]}(j) - E(N^{[s-m]}(j))\right) \otimes I_{\alpha,m}\right)^T\Big)(\Psi^{[s-m]} \otimes \Gamma^{[m]})^T (M_{s-m}^s(\gamma))^T \\ &= 0, \end{aligned}$$

because $N^{[r-l]}(k) - E(N^{[r-l]}(k))$ is a zero-mean random variable. Moreover, for $j \leq k$,

$$\begin{aligned} &\left(E(\mathcal{N}(k)\mathcal{Z}(j)^T)\right)_{r,s} = E(h_r(k) z^{[s]}(j)^T) \\ &= \sum_{l=0}^{r-1} M_{r-l}^r(\gamma)\left(\Psi^{[r-l]} \otimes \Gamma^{[l]}\right) \\ &\qquad \cdot E\left(\left((N^{[r-l]}(k) - E(N^{[r-l]}(k))) \otimes I_{\alpha,l}\right) z^{[l]}(k) z^{[s]}(j)^T\right) \\ &= 0, \end{aligned}$$

which follows, as before, by the independence of the random variables involved.

In order to simplify the notation, let us introduce the following symbols for the calculation of the (r, s)-block of the covariance matrix:

$$
\begin{aligned}
M_{u,v} &= M^{u}_{u-v}(\gamma)(\Psi^{[u-v]} \otimes \Gamma^{[v]}),\\
N_{u,v} &= N^{[u-v]}(k) - E\big(N^{[u-v]}(k)\big),\\
z_u &= z^{[u]}(k),
\end{aligned}
\tag{3.3.15}
$$

where $(u, v) \in \{(r, l), (s, m)\}$. Then we have

$$
E\big(h_r(k)h_s(k)^T\big) = \sum_{l=0}^{r-1}\sum_{m=0}^{s-1} M_{r,l} E\left((N_{r,l} \otimes I_{\alpha,l}) z_l z_m^T (N_{s,m} \otimes I_{\alpha,m})^T\right) M_{s,m}^T. \tag{3.3.16}
$$

Let us now consider the argument of the expected value in (3.3.16):

$$
(N_{r,l} \otimes I_{\alpha,l}) z_l z_m^T (N_{s,m} \otimes I_{\alpha,m})^T = st^{-1}\left(st\left((N_{r,l} \otimes I_{\alpha,l}) z_l z_m^T (N_{s,m} \otimes I_{\alpha,m})^T\right)\right). \tag{3.3.17}
$$

Moreover,

(3.3.18)

$$
\begin{aligned}
&st\big((N_{r,l} \otimes I_{\alpha,l}) z_l z_m^T (N_{s,m} \otimes I_{\alpha,m})^T\big)\\
&\quad = \big((N_{s,m} \otimes I_{\alpha,m}) \otimes (N_{r,l} \otimes I_{\alpha,l})\big) \cdot st(z_l z_m^T)\\
&\quad = \big(N_{s,m} \otimes \big(C^T_{\beta^{r-l},\alpha^m}(N_{r,l} \otimes I_{\alpha,m}) C_{1,\alpha^m}\big) \otimes I_{\alpha,l}\big) \cdot (z_m \otimes z_l)\\
&\quad = \big(\big(\big(I_{\beta,s-m} \otimes C^T_{\beta^{r-l},\alpha^m}\big) \cdot \big(N_{s,m} \otimes \big((N_{r,l} \otimes I_{\alpha,m}) C_{1,\alpha^m}\big)\big)\big) \otimes I_{\alpha,l}\big) \cdot z_{l+m}\\
&\quad = \big(\big(\big(I_{\beta,s-m} \otimes C^T_{\beta^{r-l},\alpha^m}\big) \cdot \big(N_{s,m} \otimes N_{r,l} \otimes I_{\alpha,m}\big) \cdot (1 \otimes C_{1,\alpha^m})\big) \otimes I_{\alpha,l}\big) \cdot z_{l+m}\\
&\quad = \big(\big(\big(I_{\beta,s-m} \otimes C^T_{\beta^{r-l},\alpha^m}\big) \cdot \big(\big((N_{s,m} \otimes N_{r,l}) \cdot 1\big) \otimes (I_{\alpha,m} \cdot C_{1,\alpha^m})\big)\big) \otimes I_{\alpha,l}\big) \cdot z_{l+m}\\
&\quad = \big(\big(I_{\beta,s-m} \otimes C^T_{\beta^{r-l},\alpha^m} \otimes I_{\alpha,l}\big) \cdot \big(N_{s,m} \otimes N_{r,l} \otimes C_{1,\alpha^m} \otimes I_{\alpha,l}\big)\big) \cdot z_{l+m};
\end{aligned}
$$

by substituting the previous expression in (3.3.17) and then in (3.3.16), taking into account (3.3.15) we obtain formula (3.3.12). □

Now, we are able to find the "augmented" linear stochastic system that generates the observation powers, as stated in the following theorem.

THEOREM 3.3.2. *The processes* $\{\mathcal{Y}(k)\}$ *and* $\{\mathcal{X}(k)\}$ *defined in* (3.3.6), (3.3.7) *satisfy the following equations:*

$$
\begin{aligned}
\mathcal{X}(k+1) &= \mathcal{A}\mathcal{X}(k) + \mathcal{U} + \mathcal{F}(k), \quad \mathcal{X}(0) = \bar{\mathcal{X}},\\
\mathcal{Y}(k) &= \mathcal{C}\mathcal{X}(k) + \mathcal{V} + \mathcal{G}(k),
\end{aligned}
\tag{3.3.19}
$$

where

$$
\mathcal{A} = \begin{bmatrix} A_e & 0 & \dots & 0\\ H_{2,1} & A_e^{[2]} & \dots & 0\\ \dots & \dots & \dots & \dots\\ H_{\nu,1} & H_{\nu,2} & \dots & A_e^{[\nu]} \end{bmatrix}, \quad \mathcal{U} = \begin{bmatrix} 0\\ F_e^{[2]} E(N^{[2]}(k))\\ \vdots\\ F_e^{[\nu]} E(N^{[\nu]}(k)) \end{bmatrix}, \quad \bar{\mathcal{X}} = \begin{bmatrix} \bar{x}_e\\ \bar{x}_e^{[2]}\\ \vdots\\ \bar{x}_e^{[\nu]} \end{bmatrix},
$$

$$
\mathcal{C} = \begin{bmatrix} C_e & 0 & \dots & 0\\ L_{2,1} & C_e^{[2]} & \dots & 0\\ \dots & \dots & \dots & \dots\\ L_{\nu,1} & L_{\nu,2} & \dots & C_e^{[\nu]} \end{bmatrix}, \quad \mathcal{V} = \begin{bmatrix} 0\\ G_e^{[2]} E(N^{[2]}(k))\\ \vdots\\ G_e^{[\nu]} E(N^{[\nu]}(k)) \end{bmatrix},
$$

$$H_{i,l} = M^i_{i-l}(q)(F_e^{[i-l]} \otimes A_e^{[l]}) \left(E(N^{[i-l]}(k)) \otimes I_{q,l}\right),$$

$$L_{i,l} = M^i_{i-l}(p)(G_e^{[i-l]} \otimes C_e^{[l]}) \left(E(N^{[i-l]}(k)) \otimes I_{q,l}\right),$$

$$\mathcal{F}(k) = \begin{bmatrix} f_1(k) \\ f_2(k) \\ \vdots \\ f_\nu(k) \end{bmatrix}, \quad \mathcal{G}(k) = \begin{bmatrix} g_1(k) \\ g_2(k) \\ \vdots \\ g_\nu(k) \end{bmatrix}, \tag{3.3.20}$$

$$f_i(k) = \sum_{l=0}^{i-1} M^i_{i-l}(q)(F_e^{[i-l]} \otimes A_e^{[l]})\big(\big(N^{[i-l]}(k) - E(N^{[i-l]}(k))\big) \otimes I_{q,l}\big)x_e^{[l]}(k),$$

$$g_i(k) = \sum_{l=0}^{i-1} M^i_{i-l}(p)(G_e^{[i-l]} \otimes C_e^{[l]})\big(\big(N^{[i-l]}(k) - E(N^{[i-l]}(k))\big) \otimes I_{q,l}\big)x_e^{[l]}(k),$$

and $\{\mathcal{F}(k)\}$, $\{\mathcal{G}(k)\}$ *are zero-mean white sequences such that*

$$E(\mathcal{F}(k)\mathcal{G}^T(j)) = 0, \quad k \neq j. \tag{3.3.21}$$

Moreover, defining

$$\begin{aligned} P^{r,s}_{l,m}(k) = st^{-1}\big((I_{u,s-m} \otimes C^T_{u^{r-l},q^m} \otimes I_{q,l}) \\ \cdot\big(\big(E\big(N^{[s+r-m-l]}(k)\big) - E\big(N^{[s-m]}(k)\big) \otimes E\big(N^{[r-l]}(k)\big)\big) \\ \otimes C_{1,q^m} \otimes I_{q,l}\big) \cdot E\big(x_e^{[l+m]}(k)\big)\big), \end{aligned}$$

we have, for the auto-covariances $\mathcal{Q}(k)$, $\mathcal{R}(k)$ *of the noises* $\{\mathcal{F}(k)\}$, $\{\mathcal{G}(k)\}$, *respectively, and for the cross-covariance*

$$\mathcal{J}(k) \triangleq E(\mathcal{F}(k)\mathcal{G}(k)^T),$$

the following formulas:

$$\mathcal{Q}_{r,s}(k) = \sum_{l=0}^{r-1}\sum_{m=0}^{s-1} M^r_{r-l}(q)(F_e^{[r-l]} \otimes A_e^{[l]})P^{r,s}_{l,m}(k)(F_e^{[s-m]} \otimes A_e^{[m]})^T(M^s_{s-m}(q))^T, \tag{3.3.22}$$

$$\mathcal{R}_{r,s}(k) = \sum_{l=0}^{r-1}\sum_{m=0}^{s-1} M^r_{r-l}(p)(G_e^{[r-l]} \otimes C_e^{[l]})P^{r,s}_{l,m}(k)(G_e^{[s-m]} \otimes C_e^{[m]})^T(M^s_{s-m}(p))^T, \tag{3.3.23}$$

$$\mathcal{J}_{r,s}(k) = \sum_{l=0}^{r-1}\sum_{m=0}^{s-1} M^r_{r-l}(q)(F_e^{[r-l]} \otimes A_e^{[l]})P^{r,s}_{l,m}(k)(G_e^{[s-m]} \otimes C_e^{[m]})^T(M^s_{s-m}(p))^T, \tag{3.3.24}$$

where

$$\mathcal{Q}_{r,s}(k) \triangleq E(f_r(k)f_s(k)^T), \quad \mathcal{R}_{r,s}(k) \triangleq E(g_r(k)g_s(k)^T), \quad \mathcal{J}_{r,s}(k) \triangleq E(f_r(k)g_s(k)^T).$$

Proof. Equations (3.3.19) and formulas (3.3.22), (3.3.23) follow immediately by applying Lemma 3.3.1 to (3.3.2) and (3.3.3). Taking into account the structure (3.3.20) of the noises $\mathcal{F}(k)$, $\mathcal{G}(k)$, it follows that $\left(E(\mathcal{F}(k)\mathcal{G}(j)^T)\right)_{r,s}$ is the mean value of a product of terms in the

form (3.3.11) (obtained by means of a suitable substitution of Ψ, Γ, γ, α), so that (3.3.21) is easily shown, and with some manipulations similar to (3.3.18) we obtain (3.3.24). □

Given a stochastic process $\{\xi(k),\ k \geq 0\}$, $\xi(k) \in \mathbb{R}^{\alpha}$, here we say that it is an hth-order asymptotically stationary process if $\forall i,\ 1 \leq i \leq h$, there exists a constant vector $m_i \in \mathbb{R}^{\alpha^i}$ such that

$$\lim_{k \to +\infty} E(\xi^{[i]}(k)) = m_i.$$

For the sequence $\{\mathcal{F}(k)\}$ and $\{\mathcal{G}(k)\}$ in (3.3.20), we can show their second-order asymptotic stationarity, provided that the originary system (3.1.1), (3.1.2) is asymptotically stable, i.e., all the eigenvalues of the matrix A are in the open unit circle of the complex plane. For now, let us prove the following lemma.

LEMMA 3.3.3. *Let us assume the matrix A in* (3.1.1) *to be asymptotically stable. Then the sequence $\{x_e(k)\}$ is a $2v$th-order asymptotically stationary sequence.*

Proof. Let

$$m_i(k) \triangleq E(x_e^{[i]}(k)), \quad i = 1, 2, \ldots, 2v.$$

Taking the ith block in the first equation (3.3.19) we have

$$x_e^{[i]}(k+1) = A_e^{[i]} x_e^{[i]}(k) + \sum_{l=1}^{i-1} H_{i,l} x_e^{[l]}(k) + H_{i,0} + f_i(k),$$

with the $H_{i,l}$'s as in Theorem 3.3.3. Now taking the expected values of the previous equation we obtain

$$m_i(k+1) = A_e^{[i]} m_i(k) + \sum_{l=1}^{i-1} H_{i,e} m_l(k) + H_{i,0},$$

and by defining the vectors $m(k)$ and $\mathcal{U}_{2v}$ as

$$m(k) = \begin{bmatrix} m_1(k) \\ m_2(k) \\ \vdots \\ m_{2v}(k) \end{bmatrix}, \quad \mathcal{U}_{2v} = \begin{bmatrix} 0 \\ H_{2,0} \\ \vdots \\ H_{2v,0} \end{bmatrix},$$

we can write the recursive equation

$$m(k+1) = \mathcal{A}_{2v} m(k) + \mathcal{U}_{2v}, \tag{3.3.25}$$

where $\mathcal{A}_{2v}$ is defined as

$$\mathcal{A}_{2v} = \begin{bmatrix} A_e & 0 & \ldots & 0 \\ H_{2,1} & A_e^{[2]} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ H_{2v,1} & H_{2v,2} & \ldots & A_e^{[2v]} \end{bmatrix}.$$

Equation (3.3.25) is a recursive asymptotically stable equation. Actually, the asymptotic stability of A and the block-triangular structure of A_e imply the asymptotic stability of A_e itself and, hence, of all its Kronecker powers [20]. This in turn implies the asymptotic stability of $\mathcal{A}_{2v}$. The lemma is proven by observing that $\mathcal{U}_{2v}$ is a constant input. □

THEOREM 3.3.4. *The stochastic white sequences* $\{\mathcal{F}(k)\}$ *and* $\{\mathcal{G}(k)\}$ *in* (3.3.20) *are second-order asymptotically stationary processes, provided that the matrix* A *in* (3.3.1) *is asymptotically stable.*

Proof. The thesis immediately follows by using Lemma 3.3.3 and recalling that $\{\mathcal{F}(k)\}$ and $\{\mathcal{G}(k)\}$ are zero mean sequences and observing that their covariances, which are given by (3.3.22), (3.3.23), attain to a finite limit if the first 2ν moments of $x_e(k)$ are convergent. □

Note also that, under the hypotheses of Theorem 3.3.4, the cross-covariance matrix between the augmented noises, given by (3.3.24), is convergent for $k \to +\infty$.

Equation (3.3.19) is a linear model with both deterministic and stochastic forcing terms. Note that each noise is white, but they are correlated with each other at the same instant of time. Moreover, for any k, $\mathcal{F}(k)$ and $\mathcal{G}(k)$ are uncorrelated with the initial augmented state $\bar{\mathcal{X}}$, as easily follows by direct calculation. Then for this model it is possible to determine the optimal linear estimate of the extended state $\mathcal{X}(k)$ with respect to the extended observations $\mathcal{Y}(0), \mathcal{Y}(1), \ldots, \mathcal{Y}(k)$, by using the Kalman filter in the form which takes into account the cross-correlation between noises [23]. We can obtain the optimal linear estimate of the original state $x(k)$ with respect to the same set of augmented observations by extracting in $\hat{\mathcal{X}}(k)$ the first n components (as can be readily verified by observing the structure of the vectors x_e and $\mathcal{X}$). Clearly this operation produces an estimate in the generalized recursive form (3.2.5). In the following we will denote this estimate with $\hat{x}^{(\nu,\Delta)}(k)$. Observe that $\hat{x}^{(\nu,\Delta)}(k)$ agrees with the optimal mean square estimate in the (finite-dimensional) Hilbert space $\mathcal{H}_{\nu,\Delta}$ generated by objects as

$$\prod_{l=1}^{s} y(i_l), \qquad 0 \le s \le \nu, \quad 0 \le i \le k - \Delta, \quad i \le i_l \le i + \Delta,$$

which is a subspace of $\mathcal{L}(Y_{k,\nu}, n)$, where

$$Y_{k,\nu} = \begin{bmatrix} 1 \\ Y_k \\ \vdots \\ Y_k^{[\nu]} \end{bmatrix}; \quad Y_k = \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(k) \end{bmatrix}.$$

Roughly speaking we can say that the so-defined estimate approximates the projection of $x(k)$ onto $\mathcal{L}(Y_{k,\nu}, n)$, which is the most general mean square optimal polynomial estimate of fixed degree ν.

Note that the relations

$$\mathcal{H}_{\nu,\Delta} \subset \mathcal{H}_{\nu+1,\Delta},$$
$$\mathcal{H}_{\nu,\Delta} \subset \mathcal{H}_{\nu,\Delta+1}$$

hold $\forall \nu$, Δ and, hence, since $\hat{x}^{(\nu,\Delta)}(k) = \Pi(x(k)/\mathcal{H}_{\nu,\Delta})$, we have that the error variance

$$E(\|\hat{x}^{(\nu,\Delta)}(k) - x(k)\|^2)$$

decreases when ν or δ increases. Moreover, because

$$\hat{x}^{(\nu,\Delta)}(k) = \Pi(x(k)/\mathcal{H}_{\nu,\Delta}) = \Pi(\Pi(x(k)/L^2(Y_k, n))/\mathcal{H}_{\nu,\Delta}) = \Pi(E(x(k)/Y_k)/\mathcal{H}_{\nu,\Delta}),$$

we have also that the expression

$$E(\|\hat{x}^{(\nu,\Delta)}(k) - E(x(k)/Y_k)\|^2)$$

decreases when ν or Δ increases. To conclude, we say that the polynomial filter produces an estimate of the state $x(k)$ which is as "nearer" to the optimal one as the parameters ν or Δ are chosen large.

4. Implementation of the filter. For computation purposes we need to establish the following result.

THEOREM 4.1. *Let $z \in \mathbb{R}^n$. Then, $\forall k$, the ith entry of $z^{[k]}$ is*

$$(z^{[k]})_i = z_{l_1} z_{l_2} \cdot \cdots \cdot z_{l_k}, \tag{4.1}$$

where

$$l_j = \left| \left[\frac{i-1}{n^{k-j}} \right] \right|_n + 1, \quad j = 1, 2, \ldots, k-1, \tag{4.2}$$

$$l_k = |i-1|_n + 1. \tag{4.3}$$

Proof. For $k = 1$ the theorem is true. Proceeding by induction, from (2.2.3) we obtain

$$(z^{[k+1]})_i = (z \otimes z^{[k]})_i = z_{l_1} \cdot (z^{[k]})_{m_1}$$

with

$$\begin{aligned} l_1 &= \left[\tfrac{i-1}{n^k} \right] + 1 = \left| \left[\tfrac{i-1}{n^{k+1-1}} \right] \right|_n + 1, \\ m_1 &= |i-1|_{n^k} + 1 \end{aligned}$$

as in (4.2) for $k+1$. Moreover, by (4.1), (4.2), and (4.3),

$$(z^{[k]})_{m_1} = z_{\hat{l}_1} z_{\hat{l}_2} \cdot \cdots \cdot z_{\hat{l}_k}$$

with

$$\begin{aligned} \hat{l}_j &= \left| \left[\tfrac{m_1-1}{n^{k-j}} \right] \right|_n + 1, \\ \hat{l}_k &= |m_1 - 1|_n + 1 . \end{aligned}$$

Finally by denoting $l_j = \hat{l}_{j-1}$ we have $\forall j = 2, \ldots, k$

$$l_j = \left| \left[\frac{|i-1|_{n^k}}{n^{k-(j-1)}} \right] \right|_n + 1 = \left| \left[\frac{i-1}{n^{k+1-j}} \right] \right|_n + 1,$$

whereas

$$l_{k+1} = ||i-1|_{n^k} + 1 - 1|_n + 1 = |i-1|_n + 1,$$

which proves the theorem. □

Note in (3.3.22), (3.3.23), (3.3.24) that we can evaluate the covariance matrices of the noises $\mathcal{F}(k)$, $\mathcal{G}(k)$ and their cross-covariance from the moments $E(x_e^{[h]}(k))$ and $E(N^{[h']}(k))$, where $h = 1, 2, \ldots, 2(\nu-1)$ and $h' = 1, 2, \ldots, 2\nu$. From Theorem 4.1 it follows for the ith entry of $E(N^{[h']}(k))$ that

$$\left(E(N^{[h']}(k)) \right)_i = E\left(\left(N(k) \right)_{l_1} \left(N(k) \right)_{l_2} \cdot \cdots \cdot \left(N(k) \right)_{l_{h'}} \right) = F(l_1, l_2, \ldots, l_{h'}) .$$

In order to evaluate $E(x_e^{[h]}(k))$, noting that from (3.3.25) it results in

$$m(k) = \mathcal{A}_{2(\nu-1)} m(0) + \left(\sum_{i=0}^{k-1} \mathcal{A}_{2(\nu-1)}^i \right) \mathcal{U}_{2(\nu-1)},$$

we need only evaluate $m(0)$. Taking the hth block $m_h(0)$ of $m(0)$, for $1 \leq h \leq 2(\nu-1)$, we have by definition that $m_h(0) = E(x_e^{[h]}(0))$. Next, for the ith entry of $m_h(0)$, defining the h-tuple $l_1, \ldots, l_h$, which corresponds to i by Theorem 4.1, we have that

$$(m_h(0))_i = E(x_e^{[h]}(0))_i = X_e(l_1, \ldots, l_h), \tag{4.4}$$

where

$$X_e(l_1, \ldots, l_h) = \begin{cases} X(l_1, \ldots, l_h) & \text{if } 1 \leq l_i \leq n \quad \forall i = 1, \ldots, h, \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

Equations (3.3.19) are a state-space model driven by the white noise $\mathcal{F}(k)$ and with white observation noise $\mathcal{G}(k)$. Then we can obtain the optimal mean square linear estimate of the state $\mathcal{X}(k)$ defined in (3.3.7), by using the following Kalman filter equations, which take into account the correlation between noises [21], [22], [23]:

$$\hat{\mathcal{X}}(k) = \hat{\mathcal{X}}(k/k-1) + \mathcal{K}(k)\left(\mathcal{Y}(k) - \mathcal{C}\hat{\mathcal{X}}(k/k-1) - \mathcal{V}\right), \tag{4.6}$$

$$\mathcal{Z}(k) = \mathcal{J}(k)\left(\mathcal{C}\mathcal{P}(k/k-1)\mathcal{C}^T + \mathcal{R}(k)\right)^{-1}, \tag{4.7}$$

$$\begin{aligned} \hat{\mathcal{X}}(k+1/k) = {} & \left(\mathcal{A} - \left(\mathcal{A}\mathcal{K}(k) + \mathcal{Z}(k)\right)\mathcal{C}\right)\hat{\mathcal{X}}(k/k-1) \\ & + \left(\mathcal{A}\mathcal{K}(k) + \mathcal{Z}(k)\right)\left(\mathcal{Y}(k) - \mathcal{V}\right) + \mathcal{U}, \end{aligned} \tag{4.8}$$

$$\begin{aligned} \mathcal{P}(k+1/k) = {} & \mathcal{A}\mathcal{P}(k)\mathcal{A}^T + \mathcal{Q}(k) - \mathcal{Z}(k)\mathcal{J}^T(k) - \mathcal{A}\mathcal{K}(k)\mathcal{J}^T(k) \\ & - \mathcal{J}(k)\mathcal{K}^T(k)\mathcal{A}^T, \end{aligned} \tag{4.9}$$

$$\mathcal{P}(k) = \mathcal{P}(k/k-1) - \mathcal{K}(k)\mathcal{C}\mathcal{P}(k/k-1), \tag{4.10}$$

$$\mathcal{K}(k) = \mathcal{P}(k/k-1)\mathcal{C}^T\left(\mathcal{C}\mathcal{P}(k/k-1)\mathcal{C}^T + \mathcal{R}(k)\right)^{-1}, \tag{4.11}$$

where $\mathcal{K}(k)$ is the filter gain, $\mathcal{P}(k)$, $\mathcal{P}(k/k-1)$ are the filtering and prediction error covariances, respectively, and the other symbols are defined as in Theorem 3.3.2. If the matrix $\mathcal{C}\mathcal{P}(k/k-1)\mathcal{C}^T + \mathcal{R}(k)$ is singular we can use the Moore–Penrose pseudoinverse. The initial condition for (4.6) is

$$\hat{\mathcal{X}}(0/-1) = E(\bar{\mathcal{X}}),$$

and for (4.7) it is

$$\mathcal{P}(0/-1) = E\left((\bar{\mathcal{X}} - E(\bar{\mathcal{X}}))(\bar{\mathcal{X}} - E(\bar{\mathcal{X}}))^T\right),$$

which can be easily calculated by using (4.4), (4.5). By noting that the optimal linear estimate of each entry of the augmented state process $\mathcal{X}(k)$ with respect to the augmented observations $\mathcal{Y}(k)$ agrees with its optimal polynomial estimate with respect to the original observations $y(k)$, in the sense of taking into account all of the powers, up to the νth order, of $y(j)$, $j = 0, \ldots, k$, and all of the cross-products as $y^{[l_1]}(i-\Delta)\, y^{[l_2]}(i-\Delta-1) \cdot \cdots \cdot y^{[l_{\Delta+1}]}(i)$ for $\Delta \leq i \leq k$; $0 \leq l_s \leq \nu$; $\sum_{s=1}^{\Delta+1} l_s \leq \nu$, the method proposed yields the optimal polynomial (as specified before) estimate of the system (3.1.1), (3.1.2), and this estimate can be obtained by extracting the first n entries of the estimated extended state $Xa\hat{\mathcal{X}}(k)$ given by the Kalman filter. Note that in this manner we have obtained a recursive form as (3.2.5).

As we have already observed, if the dynamical matrix of system (3.1.1), (3.1.2) is asymptotically stable, the covariance matrices $\mathcal{Q}(k)$, $\mathcal{R}(k)$, and $\mathcal{J}(k)$ tend to finite limits as time goes to infinity. In this case we certainly can utilize the well-known steady-state form of the Kalman filter, and then much of the heavier calculations (such as the gain computation) can be performed before data processing.

4.1. Reduced-order filter. A considerable reduction of the filter state-space dimension can be obtained by eliminating the redundancy contained in the vector $\hat{\mathcal{X}}(k)$. In fact, the block entries of $\hat{\mathcal{X}}(k)$ are the (polynomial) estimates of monomials in the form: $(x_e)_{l_1} \cdot \cdots \cdot (x_e)_{l_h}$, $1 \le l_1, \ldots, l_h \le q$, $1 \le h \le v$. These terms do not change their values with a permutation of the indices $l_1, \ldots, l_h$, so that the same value can be repeated many times. We can avoid this by using a suitable definition of Kronecker power, instead of the classical one, which eliminates all redundancies, as suggested in [20]. This helps in reducing both memory space and computation time.

Let $X = [x_1 \ldots x_n]^T$. We will call the reduced Kronecker power of hth order the following vector:

$$X_{[h]} = \begin{bmatrix} x_1 \cdot \cdots \cdot x_1 \cdot x_1 \\ \vdots \\ x_1 \cdot \cdots \cdot x_1 \cdot x_n \\ \vdots \\ x_{l_1} \cdot \cdots \cdot x_{l_{h-1}} \cdot x_{l_h} \\ \vdots \\ x_n \cdot \cdots \cdot x_n \cdot x_n \end{bmatrix},$$

where $1 \le l_1 \le \cdots \le l_h \le n$. Note that the entries of $X_{[h]}$, are those of $X^{[h]}$, where all the monomials $x_{i_1} \cdot \cdots \cdot x_{i_h}$ which differ from each other for a permutation of the indices $i_1, \ldots, i_h$ are considered once. Let $X_{[h]}^{(n)}$ denote the hth reduced Kronecker power of X, where we highlight the dimension n of the vector X, and $d(Y)$ is the length of a vector Y. Then it is easy to find the following formulas, both giving the dimension of the vector $X_{[h]}^{(n)}$:

$$d(X_{[h]}^{(n)}) = \sum_{k=1}^{n} d(X_{[h-1]}^{(k)}),$$

$$d(X_{[h]}^{(n)}) = \sum_{i_1=0}^{h} \sum_{i_2=0}^{h-i_1} \cdots \sum_{i_{n-1}=0}^{h-(i_1+\cdots+i_{n-2})} 1.$$

Let $T_h^{(n)} \in \mathbb{R}^{n^h \times d(X_{[h]}^{(n)})}$ and $\tilde{T}_h^{(n)} \in \mathbb{R}^{d(X_{[h]}^{(n)}) \times n^h}$, with the matrices such that

$$X^{[h]} = T_h^{(n)} X_{[h]}; \quad X_{[h]} = \tilde{T}_h^{(n)} X^{[h]}.$$

Note that the following identities hold:

$$T_h^{(n)} \tilde{T}_h^{(n)} X^{[h]} = X^{[h]}; \quad \tilde{T}_h^{(n)} T_h^{(n)} X_{[h]} = X_{[h]}. \tag{4.1.1}$$

In order to obtain an expression for $T_h^{(n)}$ and $\tilde{T}_h^{(n)}$, let us consider the ith and i'th entries in the vectors $X^{[h]}$, $X_{[h]}$, respectively:

$$\{X^{[h]}\}_i = x_{l_1} \cdot \cdots \cdot x_{l_h},$$
$$\{X_{[h]}\}_{i'} = x_{l_1'} \cdot \cdots \cdot x_{l_h'},$$

where $1 \le l_1' \le \cdots \le l_h' \le n$. We shall indicate with ζ, η the functions such that

$$\zeta(i) = (l_1, \ldots, l_h),$$
$$\eta(i') = (l_1', \ldots, l_h').$$

Of course the inverse functions ζ^{-1}, η^{-1} are well defined. Moreover, let $o(l_1, \ldots, l_h)$ be the ordering function acting on a h-tuple $(l_1, \ldots, l_h)$. Then the following expressions hold:

$$\{T_h^{(n)}\}_{i,j} = \begin{cases} 1, & \text{if } j = \eta^{-1}(o(\zeta(i))); \\ 0, & \text{otherwise;} \end{cases}$$

$$\{\tilde{T}_h^{(n)}\}_{i,j} = \begin{cases} 1, & \text{if } j = \zeta^{-1}(\eta(i)); \\ 0, & \text{otherwise.} \end{cases}$$

Note that the function ζ is easily obtained by applying Theorem 4.1. Moreover, it is easy to show that

$$i = \zeta^{-1}(l_1, \ldots, l_h) = \sum_{j=1}^{h-1} (l_j - 1) n^{h-j} + l_h.$$

Now, if we define for a fixed ν

$$\mathcal{X}_r = \begin{bmatrix} X \\ X_{[1]} \\ \vdots \\ X_{[\nu]} \end{bmatrix}; \quad \mathcal{X}_{nr} = \begin{bmatrix} X \\ X^{[1]} \\ \vdots \\ X^{[\nu]} \end{bmatrix},$$

then we have

$$\mathcal{X}_{nr} = T^{(n)} \mathcal{X}_r, \quad \mathcal{X}_r = \tilde{T}^{(n)} \mathcal{X}_{nr},$$

where

$$T^{(n)} = \begin{bmatrix} T_1^{(n)} & 0 & \cdots & \\ 0 & T_2^{(n)} & & \\ \vdots & & \ddots & \\ & & \cdots & T_\nu^{(n)} \end{bmatrix}; \quad \tilde{T}^{(n)} = \begin{bmatrix} \tilde{T}_1^{(n)} & 0 & \cdots & \\ 0 & \tilde{T}_2^{(n)} & & \\ \vdots & & \ddots & \\ & & \cdots & \tilde{T}_\nu^{(n)} \end{bmatrix}.$$

We are now able to write down the reduced-order filter equations.

Let $\mathcal{X}_r(k)$, $\mathcal{Y}_r(k)$ be defined as

$$\mathcal{X}_r(k) = \begin{bmatrix} x_e(k) \\ x_{e[2]}(k) \\ \vdots \\ x_{e[\nu]}(k) \end{bmatrix}, \quad \mathcal{Y}_r(k) = \begin{bmatrix} y_e(k) \\ y_{e[2]}(k) \\ \vdots \\ y_{e[\nu]}(k) \end{bmatrix},$$

where $x_e(k)$, $y_e(k)$ are still given by (3.3.1). Then we have for any k

$$\mathcal{X}(k) = T^{(q)} \mathcal{X}_r(k), \quad \mathcal{X}_r(k) = \tilde{T}^{(q)} \mathcal{X}(k),$$

$$\mathcal{Y}(k) = T^{(p)} \mathcal{Y}_r(k), \quad \mathcal{Y}_r(k) = \tilde{T}^{(p)} \mathcal{Y}(k), \tag{4.1.2}$$

where $\mathcal{Y}(k)$, $\mathcal{X}(k)$ are given by (3.3.6), (3.3.7) and q and p are the same as in (3.3.1). Moreover, we have that the same relations link the vectors $\hat{\mathcal{X}}(k)$, $\hat{\mathcal{X}}(k/k-1)$ in (4.4), (4.5) to their reduced counterparts $\hat{\mathcal{X}}_r(k)$, $\hat{\mathcal{X}}_r(k/k-1)$:

$$\hat{\mathcal{X}}(k) = T^{(q)} \hat{\mathcal{X}}_r(k), \quad \hat{\mathcal{X}}_r(k) = \tilde{T}^{(q)} \hat{\mathcal{X}}(k), \tag{4.1.3}$$

$$\hat{\mathcal{X}}(k/k-1) = T^{(q)}\hat{\mathcal{X}}_r(k/k-1), \quad \hat{\mathcal{X}}_r(k/k-1) = \tilde{T}^{(q)}\hat{\mathcal{X}}(k/k-1). \tag{4.1.4}$$

By using (4.1.2), (4.1.3), (4.1.4) in (4.6), (4.8) and taking into account (4.1.1), we obtain

$$\hat{\mathcal{X}}_r(k) = \hat{\mathcal{X}}_r(k/k-1) - \mathcal{A}_1(k)Xa\hat{\mathcal{X}}_r(k/k-1) + \mathcal{B}_1(k)\mathcal{Y}_r(k) - \mathcal{V}_1(k),$$
$$\hat{\mathcal{X}}_r(k+1/k) = \mathcal{A}_2(k)\hat{\mathcal{X}}_r(k/k-1) + \mathcal{B}_2(k)\mathcal{Y}_r(k) - \mathcal{V}_2(k) + \mathcal{U}_1,$$

where

$$\mathcal{A}_1(k) = \tilde{T}^{(q)}\mathcal{K}(k)\mathcal{C}T^{(q)}, \quad \mathcal{B}_1(k) = \tilde{T}^{(q)}\mathcal{K}(k)T^{(p)}, \quad \mathcal{V}_1(k) = \tilde{T}^{(q)}\mathcal{K}(k)\mathcal{V},$$
$$\mathcal{A}_2(k) = \tilde{T}^{(q)}(\mathcal{A} - (\mathcal{A}\mathcal{K}(k) + \mathcal{Z}(k))\mathcal{C})T^{(q)}, \quad \mathcal{B}_2(k) = \tilde{T}^{(q)}(\mathcal{A}\mathcal{K}(k) + \mathcal{Z}(k))T^{(p)},$$
$$\mathcal{V}_2(k) = \tilde{T}^{(q)}(\mathcal{A}\mathcal{K}(k) + \mathcal{Z}(k))\mathcal{V}, \quad \mathcal{U}_1 = \tilde{T}^{(q)}\mathcal{U},$$

which is the reduced-order filter.

5. Numerical results. Numerical simulations on an IBM Risk 6000 endowed with "Mathematica" have been performed for two examples in order to test the method. In both of them, we consider the problems of signal and state filtering for the following linear discrete time system, where the state and the output noises are non-Gaussian:

$$\begin{aligned} x(k+1) &= Ax(k) + f(k), \qquad x(0) = 0, \\ s(k) &= Cx(k), \\ y(k) &= s(k) + g(k). \end{aligned} \tag{5.1}$$

In the first example it is assumed

$$A = \begin{bmatrix} 0.1 & 0.3 \\ 0.12 & 0.1 \end{bmatrix}, \qquad C = [0.7 \quad 0.3],$$
$$f(k) = \begin{bmatrix} f_1(k) \\ f_2(k) \end{bmatrix} \in \mathbb{R}^2, \qquad g(k) \in \mathbb{R},$$

$\{f(k)\}$ and $\{g(k)\}$ are independent, zero-mean random sequences in $(\Omega, \mathcal{F}, P)$ defined as

$$f_1(k)(\omega) = -0.4\chi_{F_1}(\omega) + 0.1\chi_{F_2}(\omega), \qquad f_2(k)(\omega) = -0.02\chi_{F_3}(\omega) + 0.18\chi_{F_4}(\omega),$$
$$g(k)(\omega) = -0.28\chi_{G_1}(\omega) + 0.62\chi_{G_2}(\omega) + 1.62\chi_{G_3}(\omega),$$

where χ_Q, $Q \in \mathcal{F}$ denotes the characteristic function of Q and the disjoint events (F_1, F_2), (F_3, F_4) and (G_1, G_2, G_3) have probability

$$P(F_1) = 0.2, \quad P(F_2) = 0.8,$$
$$P(F_3) = 0.9, \quad P(F_4) = 0.1,$$
$$P(G_1) = 0.8, \quad P(G_2) = 0.1, \quad P(G_3) = 0.1.$$

The optimal linear, quadratic, and cubic algorithms, without memory ($\Delta = 0$), and the quadratic with $\Delta = 1$ have been implemented. In order to simplify the computations, we have used the steady-state Kalman filter, starting from initial conditions $x(0) = \hat{x}(0) = 0$. The results are displayed in Figs. 5.1–5.5 for 30 iterations with reference to the signal $s(k)$. It can be seen that the quadratic filter follows the true state evolution better than the linear filter, although the quadratic one with $\Delta = 1$ does not give in this case a meaningful improvement. A further remarkable improvement is indeed obtained with the cubic filter. All the mentioned results agree with the steady-state error estimate covariance values obtained by solving the

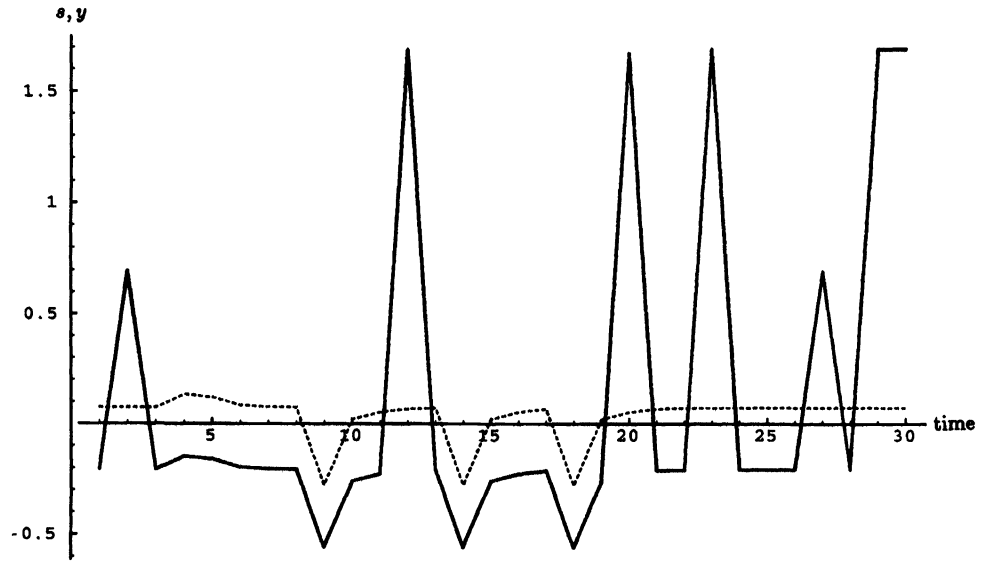


FIG. 5.1. *s=signal (dashed line), y=observation (solid line).*

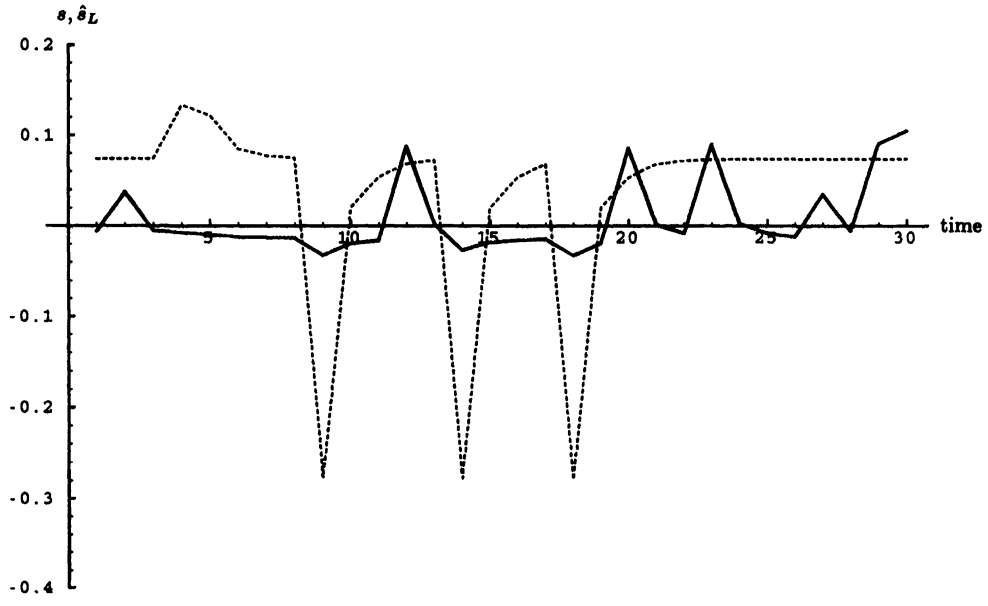


FIG. 5.2. *s=signal (dashed line), $\hat{s}_L$=optimal linear estimate with $\Delta = 0$ (solid line).*

Riccati equation for the linear, quadratic, and cubic cases with $\Delta = 0$, and for the quadratic one with $\Delta = 1$. In particular the above error covariance matrices, namely S_L, S_Q, S_C, and $S_{Q\Delta}$ (which are 2×2, 6×6, 20×20, and 12×12, respectively), have the form

$$S_L = \begin{bmatrix} 0.03864 & 0.00049 \\ 0.00049 & 0.00420 \end{bmatrix}, \qquad S_Q = \begin{bmatrix} 0.02773 & -0.00014 & \cdots \\ -0.00014 & 0.00401 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix},$$

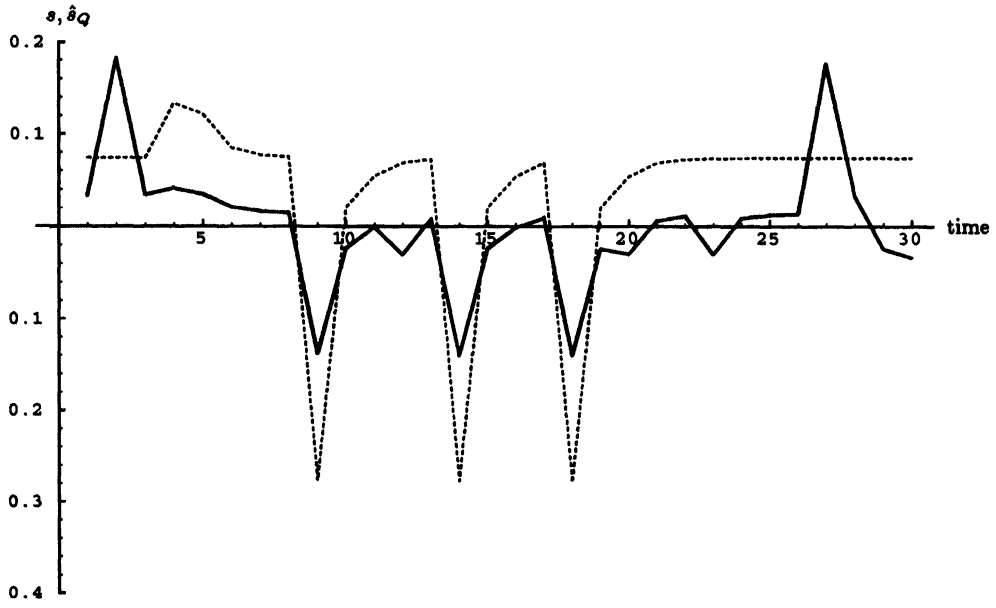


FIG. 5.3. *s=signal (dashed line), $\hat{s}_Q$=optimal quadratic estimate with $\Delta = 0$ (solid line).*

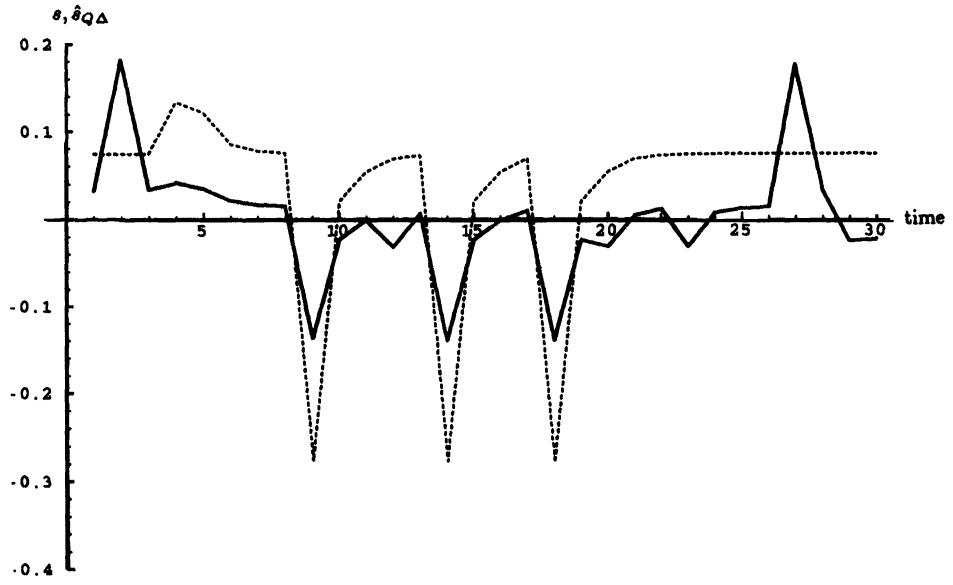


FIG. 5.4. *s=signal (dashed line), $\hat{s}_{Q\Delta}$=optimal quadratic estimate with $\Delta = 1$ (solid line).*

$$S_C = \begin{bmatrix} 0.00898 & -0.00085 & \dots \\ -0.00085 & 0.003711 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}, \qquad S_{Q\Delta} = \begin{bmatrix} 0.02773 & -0.00014 & \dots \\ -0.00014 & 0.00401 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix},$$

where we have remarked only the 2×2 matrix blocks in the top left side for the matrices S_Q, S_C, and $S_{Q\Delta}$ because they contain in the main diagonal the steady-state estimate error covariance

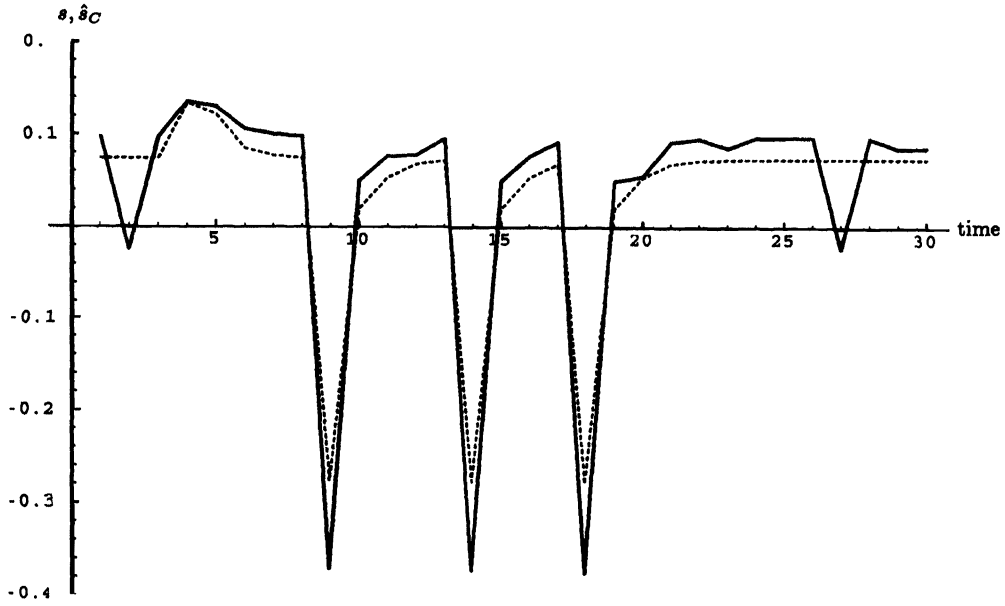


FIG. 5.5. *s=signal (dashed line), $\hat{s}_C$=optimal cubic estimate with $\Delta = 0$ (solid line).*

TABLE 5.1.

	Linear $\Delta = 0$	Quadratic $\Delta = 0$	Quadratic $\Delta = 1$	Cubic $\Delta = 0$
$\sigma^2_{(x-\hat{x})_1}$, $N = 30$	0.02188	0.01377	0.01373	0.00145
$\sigma^2_{(x-\hat{x})_2}$, $N = 30$	0.00181	0.00156	0.00156	0.00145
$\sigma^2_{(s-\hat{s})}$, $N = 30$	0.01084	0.00662	0.00660	0.00060
$\sigma^2_{(x-\hat{x})_1}$, $N = 5000$	0.03809	0.02800	0.02800	0.00913
$\sigma^2_{(x-\hat{x})_2}$, $N = 5000$	0.00429	0.00409	0.00409	0.00378
$\sigma^2_{(s-\hat{s})}$, $N = 5000$	0.01947	0.01422	0.01422	0.00447

of each component of the state. It results in $S_C(1,1) < S_Q(1,1) = S_{Q\Delta}(1,1) < S_L(1,1)$ and $S_C(2,2) < S_Q(2,2) = S_{Q\Delta}(2,2) < S_L(2,2)$. By executing the product CSC^T for all the filters implemented, where S is the block of interest in S_L, S_Q, $S_{Q\Delta}$, S_C, we have the corresponding values v_L, v_Q, $v_{Q\Delta}$, and v_C for the steady-state signal error variances

$$v_L = 0.01952,$$
$$v_Q = 0.01389,$$
$$v_{Q\Delta} = 0.01389,$$
$$v_C = 0.00438.$$

As expected, these values are close to the sampled ones obtained via numerical simulation. In Table 5.1 the sampled variances of the state and signal, obtained with a number of $N = 30$ and $N = 5000$ trials, are reported.

In the second example we will see a case where the quadratic filter with $\Delta = 1$ improves not only the simpler quadratic ($\Delta = 0$) one but also the cubic ($\Delta = 0$) one, showing the

TABLE 5.2.

	Linear $\Delta = 0$	Quadratic $\Delta = 0$	Quadratic $\Delta = 1$	Cubic $\Delta = 0$
$\sigma^2_{(x-\hat{x})}$, $N = 5000$	4.41906	1.88443	1.5012	1.84051
$\sigma^2_{(s-\hat{s})}$, $N = 5000$	2.8282	1.20604	0.96077	1.17793

nonexistence of any relation between ν and Δ. Let us consider a scalar system with

$$A = 0.6, \qquad C = 0.8,$$

and the following noises:

$$f(k)(\omega) = -\chi_{F_1}(\omega) + 3\chi_{F_2}(\omega) + 9\chi_{F_3}(\omega),$$
$$g(k)(\omega) = -9\chi_{G_1}(\omega) - 3\chi_{G_2}(\omega) + \chi_{G_3}(\omega),$$

$$P(F_1) = 15/18, \quad P(F_2) = 2/18, \quad P(F_3) = 1/18,$$
$$P(G_1) = 1/18, \quad P(G_2) = 2/18, \quad P(G_3) = 15/18,$$

where the two systems of disjoint events (F_1, F_2, F_3) and (G_1, G_2, G_3) are independent.

For this system the linear, quadratic, and cubic filters with $\Delta = 0$ and, moreover, the quadratic filter with $\Delta = 1$ have been implemented. Similar to the first example, we report the steady-state error covariance matrices S_L, S_Q, S_C, and $S_{Q\Delta}$ (which are 1×1, 2×2, 3×3, and 6×6, respectively) remarking the entry (1,1):

$$S_L = \begin{bmatrix} 4.39815 & * \\ * & * \end{bmatrix}, \qquad S_Q = \begin{bmatrix} 1.7734 & \dots \\ \vdots & \ddots \end{bmatrix},$$

$$S_C = \begin{bmatrix} 1.75661 & \dots \\ \vdots & \ddots \end{bmatrix}, \qquad S_{Q\Delta} = \begin{bmatrix} 1.53214 & \dots \\ \vdots & \ddots \end{bmatrix}.$$

The corresponding signal error variances are

$$v_L = 2.81482,$$
$$v_Q = 1.13498,$$
$$v_{Q\Delta} = 0.98057,$$
$$v_C = 1.12423.$$

Moreover, in Table 5.2 the error variance results for a numerical simulation with $N = 5000$ trials are reported.

Simulations of higher-order polynomial filters should require a more sophisticated numerical implementation, which is not the aim of this paper.

5.1. An example of polynomial estimate converging to the C.E. It would be of real practical interest (and also add further theoretic insight) to have some idea about the performance of the polynomial approximation, i.e., about the distance between the ideal and approximate estimate (given a particular model). For this purpose, let us consider the following model:

$$x(k) = f(k), \tag{5.1.1}$$
$$y(k) = x(k) + g(k),$$

where $\{f(k)\}, \{g(k)\}$ are scalar white sequences defined on some probability space $(\Omega, \mathcal{F}, P)$, independent of each other, defined as

$$f(k)(\omega) = -2\chi_{F_1}(\omega) + \chi_{F_2}(\omega),$$

$$g(k)(\omega) = -\chi_{G_1}(\omega) + \chi_{G_2}(\omega) + 2\chi_{G_3}(\omega),$$

where (F_1, F_2, F_3) and (G_1, G_2, G_3) are disjoint events having the following probabilities:

$$P(F_1) = 1/4, \quad P(F_2) = 1/2, \quad P(F_3) = 1/4,$$

$$P(G_1) = 4/7, \quad P(G_2) = 2/7, \quad P(G_3) = 1/7.$$

Because (5.1.1) is an instantaneous system it results in

$$\hat{x}(k) = E(x(k)/y(0), y(1), \ldots, y(k)) = E(x(k)/y(k)).$$

Moreover, $x(k)$ and $y(k)$ assume for any k only a finite number of values. Hence we have

$$\hat{x}(k)(\omega) = \sum_{i=1}^{6} \left(\sum_{j=1}^{3} x_j P(x(k)/y = y_i) \right) \chi_{\{y=y_i\}}(\omega). \tag{5.1.2}$$

Moreoever, being

$$y(k)(\omega) = -3\chi_{F_1 \cap G_1}(\omega) - \chi_{(F_1 \cap G_2) \cup (F_3 \cap G_1)}(\omega) + \chi_{F_3 \cap G_2}(\omega) + 2\chi_{(F_3 \cap G_3) \cup (F_2 \cap G_2)}(\omega) + 3\chi_{F_2 \cap G_3}(\omega),$$

by direct calculation and taking into account (5.1.2), we obtain

$$\hat{x}(k)(\omega) = -2\chi_{F_1 \cap G_1}(\omega) - (2/3)\chi_{(F_1 \cap G_2) \cup (F_3 \cap G_1)}(\omega) + (2/3)\chi_{(F_1 \cap G_3) \cup (F_2 \cap G_1)(\omega)} + (4/5)\chi_{(F_3 \cap G_3) \cup (F_2 \cap G_2)}(\omega) + \chi_{F_2 \cap G_3}(\omega),$$

and using this we can calculate the error variance

$$v_o = E((x(k) - \hat{x}(k))^2) = 0.504762.$$

By denoting with v_i, $i = 1, \ldots, 5$, the a priori error variances in the polynomial estimates of degree i, respectively, obtained by applying the polynomial filter to (5.1.1) it results in

$$\begin{aligned} v_1 &= 0.731707, \\ v_2 &= 0.615380, \\ v_3 &= 0.614835, \\ v_4 &= 0.567688, \\ v_5 &= 0.504762. \end{aligned}$$

Observe that $v_o = v_5$. This is not surprising because, from the fact that the observation takes values on a finite set of six numbers, it follows that an at most 5th-degree polynomial is the exact interpolator of $\hat{x}(k)$ versus $y(k)$.

In this example we can see how the polynomial estimates converge (as the polynomial degree increases) to the C.E., even in a finite number of steps.

6. Concluding remark. The method proposed allows us to obtain recursively a νth-order polynomial state estimate of the stochastic linear non-Gaussian system (3.1.1), (3.1.2). For this purpose, we have defined a new linear system in which the state and the observation are obtained in two steps: first of all by augmenting the original ones with the past values of the observations taken over a time window of fixed length Δ, then by aggregating the previous augmented vectors with their powers up to the νth order. The optimal linear estimate of the extended state with respect to the extended observations agrees with the optimal polynomial estimate (of finite memory Δ) with respect to the original observations, so that it can be obtained via the well-known Kalman filter.

It should be noted that denoting by $\sigma^2(\nu, \Delta)$ the signal error covariance (highlighting the dependence from the polynomial order ν and the memory Δ), from the above-developed theory it follows that for any ν,Δ

$$\sigma^2(\nu + 1, \Delta) \leq \sigma^2(\nu, \Delta),$$
$$\sigma^2(\nu, \Delta + 1) \leq \sigma^2(\nu, \Delta).$$

Moreover, $\sigma^2(\nu + 1, \Delta)$ and $\sigma^2(\nu, \Delta + 1)$ are not in any reciprocal relation, and this agrees with the results shown in the numerical simulations where the quadratic filter with $\Delta = 1$ gives, with respect to the cubic one with $\Delta = 0$, a worse and a better result in the first and second case, respectively. Finally, numerical simulations show the heavy inadequacy of optimal linear filtering in a non-Gaussian environment together with the high performance of polynomial filters. Of course this nice behavior is at the expense of growing computational complexity. Nevertheless, it should be stressed that this larger amount of calculations can be performed before the real time data processing, because they are mostly concerned with the computation of the covariance error matrix. Moreover, a further reduction of the filter dimensions can be obtained by using the reduced-order Kronecker powers.

Acknowledgment. The authors thank Prof. S. I. Marcus, the Associate Editor, and the anonymous referees for their critical reading and useful comments.

REFERENCES

[1] E. YAZ, *Relationship between several novel control schemes proposed for a class of nonlinear stochastic systems*, Internat. J. Control, 45 (1987), pp. 1447–1454.

[2] ———, *A control scheme for a class of discrete nonlinear stochastic systems*, IEEE Trans. Automat. Control, 32 (1987), pp. 77–80.

[3] ———, *Linear state estimators for nonlinear stochastc systems with noisy nonlinear observation*, Internat J. Control, 48 (1988), pp. 2465–2475.

[4] ———, *On the optimal state estimation of a class of discrete-time nonlinear systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 34 (1987), pp. 1127–1129.

[5] R. KULHAVY, *Differential geometry of recursive nonlinear estimation*, in Proc. of IFAC, 3, Tallinn, USSR, 1990, pp. 113–118.

[6] S. S. RAPPAPORT AND L. KURTZ, *An optimal nonlinear detector for digital data transmission through non-Gaussian channels*, IEEE Trans. Comm. Tech., COM-14 (1966), pp. 266–274.

[7] B. PICINBONO AND G. VEZZOSI, *Detection d'un signal certain dans un bruit non stationnaire et non gaussien*, Ann. des Telecomm., 25 (1970), pp. 433–439.

[8] R. D. MARTIN AND S. C. SCHWARTZ, *Robust detection of a known signal in nearly Gaussian noise*, IEEE Trans. Inform. Theory, 17 (1971), pp. 50–56.

[9] J. H. MILLER AND J. B. THOMAS, *Detectors for discrete-time signals in non-Gaussian noise*, IEEE Trans. Inform. Theory, 18 (1972), pp. 241–250.

[10] N. H. LU AND B. A. EISENSTEIN, *Detection of weak signals in non-Gaussian noise*, IEEE Trans. Inform. Theory, 28 (1982), pp. 84–91.

[11] S. A. KASSAM, G. MOUSTAKIDES, AND J. G. SHIN, *Robust detection of known signals in asymmetric noise*, IEEE Trans. Inform. Theory, 28 (1982), pp. 84–91.

[12] B. PICINBONO AND P. DUVAUT, *Optimal linear quadratic system for detection and estimation*, IEEE Trans. Inform. Theroy, 34 (1988), pp. 304–311.

[13] B. PICINBONO, *Geometrical properties of optimal Volterra filters for signal detection*, IEEE Trans. Inform. Theory, 36 (1990), pp. 1061–1068.

[14] T. SUBBA RAO AND M. YAR, *Linear and non-linear filters for linear, but not Gaussian processes*, Internat. J. Control, 39 (1984), pp. 235–246.

[15] P. K. RAJESEKARAN, N. SATYANARAYANA, AND M. D. SRINATH, *Optimum linear estimation of stochastic signals in the presence of multiplicative noise*, IEEE Trans. Aerospace Electron. Systems, AES-7 (1971), pp. 462–468.

[16] W. I. DE KONIG, *Optimal estimation of linear discrete-time systems with stochastic parameters*, Automatica J. IFAC, 20 (1984), pp. 113–115.

[17] B. FRIEDLANDER AND B. PORAT, *Asymptotically optimal estimation of MA and ARMA parameters*, IEEE Trans. Automat. Control, 35 (1990).

[18] G. B. GIANNAKIS, *On the identifiability of non-Gaussian ARMA models using cumulants*, IEEE Trans. Automat. Control, 35 (1990), pp. 18–26, pp. 27–35.

[19] A. DE SANTIS, A. GERMANI, AND M. RAIMONDI, *Optimal quadratic filtering of linear discrete time non-Gaussian systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 1274–1278.

[20] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw–Hill, 1970.

[21] G. C. GOODWIN AND R. L. PAYNE, *Dynamical System Identification: Experiment Design and Data Analysis*, Math. Sci. Engrg., 136, R. Bellman, ed., Academic Press, New York, 1977, pp. 77–78.

[22] T. KAILATH, *An innovation approach to least square estimation Part* I: *Linear filtering in additive white noise*, IEEE Trans. Automat. Control, 13 (1968), pp. 646–655.

[23] A. V. BALAKRISHNAN, *Kalman Filtering Theory*, Optimization Software, Inc., Publication Division, New York, 1984.

[24] G. S. RODGERS, *Matrix Derivatives*, Marcel Dekker, New York, Basel, 1980.

[25] S. I. MARCUS, *Optimal nonlinear estimation for a class of discrete-time stochastic systems*, IEEE Trans. Automat. Control, 24 (1979), pp. 297–302.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1691–1706, September 1996

011

AN ENTROPY FORMULA FOR TIME-VARYING DISCRETE-TIME CONTROL SYSTEMS*

PABLO A. IGLESIAS†

Abstract. The results of this paper generalize the formula for the entropy of a transfer function to time-varying systems. This is done through the use of some results on spectral factorizations due to Arveson and properties of the $\mathcal{W}$-transform which generalizes the usual $\mathcal{Z}$-transform for time-varying systems. Using the formula defined, it is shown that for linear fractional transformations like those that arise in time-varying $\mathcal{H}_\infty$ control, there exists a unique, bounded contraction which maximizes the entropy. This generalizes known results in the time-invariant case. Possible extensions are discussed, along with state-space formulae.

Key words. time-varying systems, optimal control, spectral factorizations

AMS subject classifications. 93B36, 93C55, 93C50

1. Introduction.

Since the pioneering work of Zames [27], there has been much interest in finding stabilizing controllers which ensure that the $\mathcal{H}_\infty$ norm of a closed-loop transfer function is below a given number $\gamma > 0$. In particular, consider the system depicted in Figure 1.1, and suppose that the open-loop system is given by

$$\begin{bmatrix} z \\ y \end{bmatrix} = G \begin{bmatrix} w \\ u \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} w \\ u \end{bmatrix},$$

where the signals w, z, y, and u are the external disturbance, output error, measured output, and control input, respectively. The goal of $\mathcal{H}_\infty$ control theory is to find a control law $u = Ky$ such that the closed-loop system, denoted

$$\mathcal{F}_\ell(G, K) = G_{11} + G_{12}K(I - G_{22}K)^{-1}G_{21},$$

has $\mathcal{H}_\infty$ norm less than γ, assuming that such a controller exists.

While early developments relied on transfer function and operator methods, a recent emphasis, based on the work of Glover and co-workers [5, 10], has been to approach the $\mathcal{H}_\infty$ control problem using state-space methods. Glover et al. have shown that the existence of stabilizing controllers achieving the required norm bound is equivalent to the existence of positive semidefinite, stabilizing solutions to two indefinite algebraic Riccati equations. Because of this connection with algebraic Riccati equations, these results are reminiscent of earlier work on linear quadratic Gaussian (LQG) control.

A complete characterization of all controllers achieving the closed-loop $\mathcal{H}_\infty$ norm bound is given in [10]. Assuming that a controller exists such that $\|\mathcal{F}_\ell(G, K)\|_\infty < \gamma$, it can be shown that the set of all such controllers can also be parameterized by a linear fractional transformation of a specific controller K_a and a stable contraction Q:

$$K = \mathcal{F}_\ell(K_a, Q), \qquad \|Q\|_\infty < 1.$$

Given this parameterization, it is easy to see that all possible closed-loop transfer functions satisfying $\|\mathcal{F}_\ell(G, K)\|_\infty < \gamma$ are given by

$$\mathcal{F}_\ell(G, \mathcal{F}_\ell(K_a, Q)) = \mathcal{F}_\ell(J, Q) =: H(Q)$$

*Received by the editors March 9, 1994; accepted for publication (in revised form) June 2, 1995. This research was supported in part by National Science Foundation contract ECS-9309387.

†Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD 21218 (pi@jhu.edu).

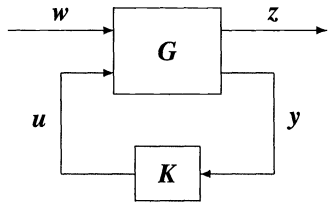


FIG. 1.1. *Closed-loop system.*

for some J. To choose among this "ball" of solutions, it has been proposed that the controller selected be chosen so as to maximize the following "entropy" integral (see [2, 6, 11, 9]):

$$I_d(H(Q),\gamma,\lambda_0) := \frac{\gamma^2}{2\pi}\int_{-\pi}^{\pi} \ln\left|\det\left(I-\gamma^{-2}H^{\sim}(e^{i\omega})H(e^{i\omega})\right)\right| \frac{1-|\lambda_0|^2}{|e^{i\omega}-\lambda_0|^2}d\omega. \tag{1.1}$$

The benefits of using controllers which maximize this entropy integral are outlined in the monograph [21], which treats the continuous-time case, and [16, 15] for the discrete-time case. As shown in [21], these controllers can be thought of as lying between $\mathcal{H}_\infty$ optimal controllers and LQG optimal controllers. Specifically, I_d exhibits some normlike properties; it is monotonically decreasing with respect to γ, and it bounds the LQG cost of the closed-loop system. It has the added property that controllers which maximize the entropy are also optimal with respect to the risk-sensitive control problem of stochastic control theory [25].

Owing to the similarities between $\mathcal{H}_\infty$ control and classical LQG control, which were highlighted in [5], many straightforward extensions have now appeared. In this paper we are particularly interested in controllers for time-varying systems as have been considered in [20, 19, 22]. While the controllers achieving a norm bound can also be written as a linear fractional transformation of an operator K_a and a stable contractive operator Q, it is not clear how to choose among the possible controllers, since the entropy integral (1.1) is given in terms of the transfer functions and is therefore not amenable to time-varying systems. In this paper we give a generalization of the entropy integral for discrete-time, time-varying systems. This generalization will be based on the $\mathcal{W}$-transform, introduced by Alpay, Dewilde, and Dym [1] in the context of interpolation problems for nonstationary processes. As in the stationary case, the entropy defined here for control systems will be related to the entropy used in [12] in the context of interpolants for band extension problems.

In the study of linear time-invariant controllers, the entropy evaluated at $\lambda_0 = 0$ is of particular importance. In this case, it can be shown that the integral (1.1) is an upper bound for the $\mathcal{H}_2$ norm of the transfer function. For our time-varying systems, our entropy definition will deal only with the analogous evaluation at the origin. We will outline the difficulties that arise in generalizing this definition.

The rest of the paper will be organized as follows. We begin by introducing the $\mathcal{W}$-transform in §2 and giving some of its properties. In §3, the definition of the entropy for time-varying systems is given. In §4 we show that for systems that can be expressed as linear fractional transformations of linear, causal, contractive operators, the entropy defined has a maximum. Some possible extensions of the theory are discussed in §5, and, finally, in §6 we give some conclusions.

2. Preliminaries. In this section we introduce the notation and present some preliminary results concerning operators and the $\mathcal{W}$-transform that will be needed in the rest of the paper. Most of these results are taken from [1]. Note that the presentation in [1] assumes that causal

operators are represented as upper triangular matrices. In our presentation, we will use the more common representation of causal operators as lower triangular operators.

2.1. Notation. Let $\boldsymbol{x} = \{x(k) : k \in \mathbb{Z}\}$ denote a sequence of vectors $x(k) \in \mathbb{C}^n$. The set of all such sequences is denoted $\mathcal{S}^n$. The subset of $\mathcal{S}^n$ of square-summable sequences is ℓ_2^n. The space ℓ_2^n is a Hilbert space with inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \sum_{-\infty}^{\infty} x^*(k) y(k)$$

and induced norm $\|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$.

Let $\boldsymbol{G}$ represent a linear operator from ℓ_2^m to ℓ_2^p. Then $\boldsymbol{G}$ has a natural representation as a doubly infinite matrix $\{G(i, j)\}$, $i, j \in \mathbb{Z}$, $G(i, j) \in \mathbb{C}^{p \times m}$. We will denote the operation $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{u}$ as follows:

$$\begin{bmatrix} \vdots \\ y(-1) \\ \boxed{y(0)} \\ y(1) \\ \vdots \end{bmatrix} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \\ \cdots & G(-1,-1) & G(-1,0) & G(-1,1) & \cdots \\ \cdots & G(0,-1) & \boxed{G(0,0)} & G(0,1) & \cdots \\ \cdots & G(1,-1) & G(1,0) & G(1,1) & \cdots \\ & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \vdots \\ u(-1) \\ \boxed{u(0)} \\ u(1) \\ \vdots \end{bmatrix}.$$

The box around the elements in the vectors (resp., matrix) denotes the element with index 0 (resp., 0, 0).

Let $\mathfrak{X}^{p \times m}$ denote the space of bounded linear operators from the space ℓ_2^m to ℓ_2^p. The subspace of $\mathfrak{X}^{p \times m}$ consisting of causal (resp., diagonal) operators is denoted $\mathfrak{L}^{p \times m}$ (resp., $\mathfrak{D}^{p \times m}$).

We will usually drop the superscript on these spaces; the Hilbert spaces on which the operators act should be clear from the context. We write $\mathfrak{X}^{-1}$ to mean the space of operators whose inverses are in $\mathfrak{X}$. Similar expressions are used for $\mathfrak{L}$ and $\mathfrak{D}$.

For operators in $\mathfrak{X}$, the following two facts will be useful. Proofs may be found in [1].

LEMMA 2.1. *For an operator $\boldsymbol{X} \in \mathfrak{X}$, the elements $X(i, j)$ satisfy*

$$\|\boldsymbol{X}(i, j)\| \leq \|\boldsymbol{X}\| \qquad (\forall i, j).$$

LEMMA 2.2. *If $\boldsymbol{D} \in \mathfrak{D}$ is a diagonal operator, then*

$$\|\boldsymbol{D}\| = \sup_i \|D(i)\|.$$

In subsequent discussion, the forward shift operator will play a prominent role. This is the operator $\boldsymbol{Z} \in \mathfrak{X}^{m \times m}$:

$$\boldsymbol{Z} \begin{bmatrix} \vdots \\ y(-1) \\ \boxed{y(0)} \\ y(1) \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ y(0) \\ \boxed{y(1)} \\ y(2) \\ \vdots \end{bmatrix}.$$

This operator has a matrix representation

$$\boldsymbol{Z} = \{Z(i, j)\} = \begin{cases} I_m, & j - i = 1, \\ 0 & \text{otherwise.} \end{cases}$$

One other useful operator is the projection operator $\boldsymbol{P}_k \in \mathfrak{D}^{m\times m}$, which has a matrix representation $\boldsymbol{P}_k = \{P(i,i)\}$, where

$$P(i,i) = \begin{cases} I_m & \text{if } i < k, \\ 0 & \text{otherwise.} \end{cases}$$

2.2. The $\mathcal{W}$-transform. In order to consider interpolation problems for non-Toeplitz operators, Alpay, Dewilde, and Dym introduced a generalization of the usual Fourier transform on sequences, known as the $\mathcal{W}$-transform. In this section we provide an introduction to this transform as well as some of its properties. Details may be found in [1].

Given an operator $\boldsymbol{G} \in \mathfrak{X}^{p\times m}$, $\boldsymbol{G} = \{G(i,j)\}$, we define the set of diagonal operators $\boldsymbol{G}_{[k]}$ corresponding to the kth subdiagonal shifted up to the diagonal:

$$\boldsymbol{G}_{[k]} = \operatorname{diag}\{\ldots, G(-1,-1-k), \boxed{G(0,-k)}, G(1,1-k), \ldots\}.$$

From Lemmas 2.1 and 2.2,

$$\left\|\boldsymbol{G}_{[k]}\right\| = \sup_i \|G(i,i-k)\| \le \|\boldsymbol{G}\|.$$

An operator $\boldsymbol{G} \in \mathfrak{L}$ has a unique representation as a series in terms of the $\boldsymbol{G}_{[k]}$ as follows:

$$\boldsymbol{G} = \sum_{k=0}^{\infty} \boldsymbol{G}_{[k]} (\boldsymbol{Z}^*)^k,$$

where the sum converges weakly [28].

Let $\boldsymbol{X} \in \mathfrak{X}$. We denote by $\sigma(\boldsymbol{X})$ the spectrum of $\boldsymbol{X}$ and by $\rho(\boldsymbol{X})$ the spectral radius of $\boldsymbol{X}$. It is well known that

$$\begin{aligned} \rho(\boldsymbol{X}) &:= \max\{|\lambda| : \lambda \in \sigma(\boldsymbol{X})\} \\ &= \lim_{n\to\infty} \left\|\boldsymbol{X}^n\right\|^{1/n}. \end{aligned}$$

Finally, for $\boldsymbol{W} \in \mathfrak{D}$, we define

$$\ell(\boldsymbol{W}) := \rho(\boldsymbol{W}\boldsymbol{Z}^*).$$

We are now ready to introduce the $\mathcal{W}$-transform.

DEFINITION 2.3. *Let $\boldsymbol{G} \in \mathfrak{L}$ and $\boldsymbol{W} \in \mathfrak{D}$, with $\ell(\boldsymbol{W}) < 1$. We define*

$$\widehat{\boldsymbol{G}}(\boldsymbol{W}) := \sum_{k=0}^{\infty} \boldsymbol{G}_{[k]} (\boldsymbol{Z}^*)^k (\boldsymbol{Z}\boldsymbol{W})^k. \tag{2.1}$$

This series will converge in norm, provided that $\ell(\boldsymbol{W}) < 1$. The transform (2.1) acts like the λ-transform[1] of a sequence in $\mathbb{C}^n$. Note that in this case, the transform is taken of a sequence $\{\boldsymbol{G}_{[k]}\}$ of operators in $\mathfrak{D}$. In terms of the doubly infinite matrix representation $\boldsymbol{G} = \{G(i,j)\}$ and $\boldsymbol{W} = \operatorname{diag}\{W(i)\}$,

$$(\boldsymbol{Z}^*)^k (\boldsymbol{Z}\boldsymbol{W})^k = \operatorname{diag}\{W(i)W(i+1)\cdots W(i+k-1)\}, \qquad k \ge 1,$$

and hence the $\mathcal{W}$-transform can be written as

$$\widehat{\boldsymbol{G}}(\boldsymbol{W}) = \operatorname{diag}\left\{\sum_{k=0}^{\infty} G(i,i-k)W(i-k+1)W(i-k+2)\cdots W(i-1)W(i)\right\}. \tag{2.2}$$

[1]The λ-transform is just the $\mathcal{Z}$-transform with z^{-1} replaced by λ.

In order to illustrate some of the properties of the $\mathcal{W}$-transform, we provide some examples.

Example 2.4 (time-invariant systems). Suppose that $\boldsymbol{G} \in \mathfrak{L}$ represents a time-invariant operator. Then $\boldsymbol{G}$ has a characterization as a block Toeplitz matrix $\boldsymbol{G} = \{G(i, i-k)\}$ with $G(i, i-k) = g_k$. And thus $\boldsymbol{G}_{[k]} = \operatorname{diag}\{\ldots, g_k, \boxed{g_k}, g_k, \ldots\}$, $k \in \mathbb{Z}_+$. We wish to evaluate this at $W = \lambda \boldsymbol{I}$, where $\lambda \in \mathbb{C}$, $|\lambda| < 1$, and $\boldsymbol{I}$ is the identity operator in $\mathfrak{X}$. Then

$$\widehat{\boldsymbol{G}}(\boldsymbol{W}) = \sum_{k=0}^{\infty} \boldsymbol{G}_{[k]} (\boldsymbol{Z}^*)^k (\lambda \boldsymbol{Z})^k = \sum_{k=0}^{\infty} \lambda^k \boldsymbol{G}_{[k]} = \operatorname{diag}\{G(\lambda)\},$$

where $G(\lambda)$ is the usual λ-transform of the sequence $\{g_k\}$.

Example 2.5 (frozen-time systems). Consider a general causal time-varying operator $\boldsymbol{G}$, but evaluate this as in the previous example at $\boldsymbol{W} = \lambda \boldsymbol{I}$. This gives

$$\widehat{\boldsymbol{G}}(\boldsymbol{W}) = \operatorname{diag}\{\ldots, G_{-1}(\lambda), \boxed{G_0(\lambda)}, G_1(\lambda), \ldots\},$$

where

$$G_i(\lambda) = \sum_{k=0}^{\infty} G(i, i-k) \lambda^k, \qquad i \in \mathbb{Z},$$

is the λ-transform of the frozen-time system at time i. These frozen-time systems have received considerable interest recently in the study of slowly time-varying systems [23, 24, 28].

In what follows we will be interested primarily in the evaluation of $\widehat{\boldsymbol{G}}(\boldsymbol{W})$ at $\boldsymbol{W} = \boldsymbol{0}$. In this case it is easy to see that $\widehat{\boldsymbol{G}}(\boldsymbol{W}) = \boldsymbol{G}_{[0]}$, that is, the diagonal element of the operator $\boldsymbol{G}$.

2.3. $\mathcal{W}$-transform of a state-space system. Consider the system

$$\Sigma_G := \left\{ \begin{array}{rcl} x(k+1) & = & A(k)x(k) + B(k)u(k), \quad k \in \mathbb{Z}^+, \\ y(k) & = & C(k)x(k) + D(k)u(k). \end{array} \right. \tag{2.3}$$

Define the following operator in $\mathfrak{D}$:

$$\boldsymbol{A} := \operatorname{diag}(\ldots, 0, 0, \boxed{A(0)}, A(1), A(2), \ldots),$$

with similar representations for $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$. Let $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ represent the elements of $\mathcal{S}^n$, $\mathcal{S}^p$, and $\mathcal{S}^m$ corresponding to the $x(k)$, $y(k)$, and $u(k)$. We can express the state-space equations (2.3) as

$$\begin{aligned} \boldsymbol{Z}\boldsymbol{x} &= \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}, \\ \boldsymbol{y} &= \boldsymbol{C}\boldsymbol{x} + \boldsymbol{D}\boldsymbol{u}. \end{aligned} \tag{2.4}$$

The operator mapping $\boldsymbol{u}$ to $\boldsymbol{y}$ is the $\mathfrak{L}$ operator:

$$\begin{aligned} \boldsymbol{G} &:= \boldsymbol{C}(\boldsymbol{Z} - \boldsymbol{A})^{-1}\boldsymbol{B} + \boldsymbol{D} \\ &= \boldsymbol{C}\left(\sum_{k=0}^{\infty} (\boldsymbol{Z}^*\boldsymbol{A})^k\right)\boldsymbol{Z}^*\boldsymbol{B} + \boldsymbol{D} \\ &=: \left[\begin{array}{c|c} \boldsymbol{A} & \boldsymbol{B} \\ \hline \boldsymbol{C} & \boldsymbol{D} \end{array}\right]. \end{aligned} \tag{2.5}$$

The series in (2.5) converges, provided that $\rho(\boldsymbol{Z}^*\boldsymbol{A}) < 1$. It can be shown that this condition is equivalent to uniform exponential stability of the autonomous system in (2.3) [18, 13].

For this linear time-varying system, we wish to calculate the $\mathcal{W}$-transform of $\boldsymbol{G}$ for $\boldsymbol{W} = \operatorname{diag}\{W(i)\}$. First of all,

$$\boldsymbol{G}_{[k]} := \begin{cases} \boldsymbol{CZ}^*(\boldsymbol{AZ}^*)^{k-1}\boldsymbol{BZ}^k & \text{for } k > 0, \\ \boldsymbol{D}, & k = 0, \\ \boldsymbol{0}, & k < 0. \end{cases}$$

Thus

$$\begin{aligned} \widehat{\boldsymbol{G}}(\boldsymbol{W}) &= \boldsymbol{D} + \boldsymbol{CZ}^* \sum_{k=1}^{\infty} (\boldsymbol{AZ}^*)^{k-1}\boldsymbol{B}(\boldsymbol{ZW})^k \\ &= \operatorname{diag}\left\{ D(i) + \sum_{k=0}^{\infty} C(i)\Phi_A(i, i-k+1)B(i-k)\Phi_W(i-k, i) \right\}. \end{aligned}$$

Here, Φ represents the transition matrix of the sequences $\boldsymbol{A}$ and $\boldsymbol{W}$, that is,

$$\Phi_A(i, j) = \begin{cases} I & \text{for } i = j, \\ A(i-1)A(i-2)\cdots A(j) & \text{for } i > j, \\ A(i+1)A(i+2)\cdots A(j) & \text{for } i < j. \end{cases}$$

2.4. Properties of the $\mathcal{W}$-transform. In this section we outline some properties of the $\mathcal{W}$-transform.

LEMMA 2.6. *Let $\boldsymbol{G} \in \mathfrak{X}$ and $\boldsymbol{D}, \boldsymbol{W} \in \mathfrak{D}$ with $\ell(\boldsymbol{W}) < 1$; then*

$$\widehat{\boldsymbol{DG}}(\boldsymbol{W}) = \boldsymbol{D}\widehat{\boldsymbol{G}}(\boldsymbol{W}).$$

Proof. Using the identity

$$(\boldsymbol{DG})_{[k]} = \boldsymbol{DG}_{[k]},$$

the proof is straightforward. □

For the next property we need a special operator. Let $\boldsymbol{D} \in \mathfrak{D}$. We define

$$\boldsymbol{D}^{(k)} := (\boldsymbol{Z}^*)^k \boldsymbol{D}(\boldsymbol{Z})^k \in \mathfrak{D}.$$

This operator has the effect of moving the elements of the diagonal $\boldsymbol{D}$ "down" k steps.

LEMMA 2.7. *Let $\boldsymbol{G} \in \mathfrak{X}$, $\boldsymbol{D} \in \mathfrak{D} \cap \mathfrak{D}^{-1}$, and $\boldsymbol{W} \in \mathfrak{D}$ with $\ell(\boldsymbol{W}) < 1$; then*

$$\widehat{\boldsymbol{GD}}(\boldsymbol{W}) = \hat{\boldsymbol{G}}(\boldsymbol{D}^{(1)}\boldsymbol{W}\boldsymbol{D}^{-1})\boldsymbol{D}.$$

Proof. First, note that

$$\boldsymbol{G}_{[k]}(\boldsymbol{Z}^*)^k\boldsymbol{D} = [\boldsymbol{GD}]_{[k]}(\boldsymbol{Z}^*)^k. \tag{2.6}$$

It follows that

$$\begin{aligned} \hat{\boldsymbol{G}}(\boldsymbol{D}^{(1)}\boldsymbol{W}\boldsymbol{D}^{-1})\boldsymbol{D} &= \left(\sum_{k=0}^{\infty} \boldsymbol{G}_{[k]}(\boldsymbol{Z}^*)^k \left(\boldsymbol{Z}\left[\boldsymbol{Z}^*\boldsymbol{DZWD}^{-1}\right]\right)^k \right) \boldsymbol{D} \\ &= \sum_{k=0}^{\infty} \boldsymbol{G}_{[k]}(\boldsymbol{Z}^*)^k \boldsymbol{D}(\boldsymbol{ZW})^k \\ &= \sum_{k=0}^{\infty} [\boldsymbol{GD}]_{[k]}(\boldsymbol{Z}^*)^k(\boldsymbol{ZW})^k \end{aligned}$$

as required. □

The following corollary is straightforward.

COROLLARY 2.8. *Let* $\boldsymbol{G} \in \mathfrak{X}$ *and* $\boldsymbol{D} \in \mathfrak{D}$; *then*

$$\widehat{\boldsymbol{GD}}(\boldsymbol{0}) = \widehat{\boldsymbol{G}}(\boldsymbol{0})\boldsymbol{D}.$$

The following result and, more important, its corollary will be crucial to the results that follow.

LEMMA 2.9. *Let* $\boldsymbol{G}, \boldsymbol{H} \in \mathfrak{X}$, *and* $\boldsymbol{W} \in \mathfrak{D}$ *with* $\ell(\boldsymbol{W}) < 1$; *then*

$$\widehat{\boldsymbol{GH}}(\boldsymbol{W}) = \widehat{\boldsymbol{G}}\widehat{\boldsymbol{H}}(\boldsymbol{W})(\boldsymbol{W}).$$

Proof. See [1, Lem. 3.7]. □

Finally, the following corollary combines the results of Lemma 2.9 and Corollary 2.8.

COROLLARY 2.10. *Let* $\boldsymbol{G}, \boldsymbol{H} \in \mathfrak{X}$; *then*

$$\widehat{\boldsymbol{GH}}(\boldsymbol{0}) = \widehat{\boldsymbol{G}}(\boldsymbol{0})\widehat{\boldsymbol{H}}(\boldsymbol{0}).$$

2.5. Spectral factorizations. For our definition of entropy, we require spectral factorizations of operators. The following lemma, due to Arveson, guarantees the existence of a spectral factor for positive self-adjoint operators.

LEMMA 2.11 (see [3]). *Suppose that* $\boldsymbol{G} \in \mathfrak{X} \cap \mathfrak{X}^{-1}$ *is a positive, self-adjoint operator. There exist operators* $\boldsymbol{A}, \boldsymbol{B} \in \mathfrak{L} \cap \mathfrak{L}^{-1}$ *such that*

$$\boldsymbol{G} = \boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{B}^*\boldsymbol{B}.$$

Moreover, $\boldsymbol{A} = \boldsymbol{DB}$ *where* $\boldsymbol{D} \in \mathfrak{D} \cap \mathfrak{D}^{-1}$ *and* $\boldsymbol{D}^*\boldsymbol{D} = \boldsymbol{I}$.

2.5.1. State-space formulae. We are interested in computing spectral factorizations for operators of the form $\boldsymbol{I} - \boldsymbol{G}^*\boldsymbol{G}$, where $\|\boldsymbol{G}\| < 1$ and $\boldsymbol{G}$ is given by (2.4). The following result provides a state-space equation for a particular spectral factorization.

THEOREM 2.12 (see [17]). *Suppose that* $\boldsymbol{G}$ *is given by* (2.4) *with* $\rho(\boldsymbol{Z}^*\boldsymbol{A}) < 1$. *The following statements are equivalent:*

1. $\boldsymbol{I} - \boldsymbol{G}^*\boldsymbol{G} > 0$.
2. *There exists a uniformly bounded solution* $\boldsymbol{X} = \boldsymbol{X}^* \geq 0$ *to the operator algebraic Riccati equation*

$$\boldsymbol{X} = \boldsymbol{A}^*\boldsymbol{ZXZ}^*\boldsymbol{A} + \boldsymbol{C}^*\boldsymbol{C} + \left(\boldsymbol{A}^*\boldsymbol{ZXZ}^*\boldsymbol{B} + \boldsymbol{C}^*\boldsymbol{D}\right)\boldsymbol{V}^{-1}\left(\boldsymbol{B}^*\boldsymbol{ZXZ}^*\boldsymbol{A} + \boldsymbol{D}^*\boldsymbol{C}\right)$$

with $\boldsymbol{V} := \boldsymbol{I} - \boldsymbol{D}^*\boldsymbol{D} - \boldsymbol{B}^*\boldsymbol{ZXZ}^*\boldsymbol{B} > \boldsymbol{0}$ *and* $\boldsymbol{A} + \boldsymbol{BV}^{-1}\left(\boldsymbol{B}^*\boldsymbol{ZXZ}^*\boldsymbol{A} + \boldsymbol{D}^*\boldsymbol{C}\right)$ *uniformly exponentially stable.*

3. *The operator* $\boldsymbol{I} - \boldsymbol{G}^*\boldsymbol{G}$ *has a spectral factorization, i.e.,*

$$\boldsymbol{I} - \boldsymbol{G}^*\boldsymbol{G} = \boldsymbol{M}^*\boldsymbol{M}$$

with

$$\boldsymbol{M} = \left[\begin{array}{c|c} \boldsymbol{A} & \boldsymbol{B} \\ \hline -\boldsymbol{V}^{-1/2}\left(\boldsymbol{B}^*\boldsymbol{ZXZ}^*\boldsymbol{A} + \boldsymbol{D}^*\boldsymbol{C}\right) & \boldsymbol{V}^{1/2} \end{array}\right].$$

For conciseness we have chosen to write the Riccati equation as an operator algebraic Riccati equation in this theorem. An equivalent representation of this equation is in terms of a recursive Riccati difference equation:

$$\begin{aligned} X_k = {} & A_k^T X_{k+1} A_k + C_k^T C_k \\ & + (A_k^T X_{k+1} B_k + C_k^T D_k) V_k^{-1} (B_k^T X_{k+1} A_k + D_k^T C_k), \end{aligned}$$

where $V_k := I - D_k^T D_k - B_k^T X_{k+1} B_k$. Note that this recursion has no terminal condition.

3. Time-varying entropy.

3.1. Entropy operator. In this section we present our definition of the entropy for a linear time-varying operator $\boldsymbol{G}$.

Suppose that $\boldsymbol{G} \in \mathfrak{X}$ has operator norm $\|\boldsymbol{G}\| < \gamma$. It follows that the self-adjoint operator $\boldsymbol{I} - \gamma^{-2}\boldsymbol{G}^*\boldsymbol{G}$ is positive. By Lemma 2.11, it has a spectral factor $\boldsymbol{M}$. Using this spectral factor we begin by defining an entropy *operator*.

DEFINITION 3.1. *Suppose that* $\boldsymbol{G} \in \mathfrak{X}$ *and* $\|\boldsymbol{G}\| < \gamma$. *Let* $\boldsymbol{M} \in \mathfrak{L} \cap \mathfrak{L}^{-1}$ *be a spectral factor of the positive operator* $\boldsymbol{I} - \gamma^{-2}\boldsymbol{G}^*\boldsymbol{G}$ *and* $\boldsymbol{W} \in \mathfrak{D}$ *be a diagonal operator with* $\ell(\boldsymbol{W}) < 1$. *We define*

$$\boldsymbol{E}(\boldsymbol{G}, \gamma, \boldsymbol{W}) := \widehat{\boldsymbol{M}}^*(\boldsymbol{W})\widehat{\boldsymbol{M}}(\boldsymbol{W}). \tag{3.1}$$

Since spectral factors are not unique, in order for the expression in (3.1) to make sense, we must show that it does not depend on the particular spectral factor chosen. Suppose that

$$\boldsymbol{I} - \gamma^{-2}\boldsymbol{G}^*\boldsymbol{G} = \boldsymbol{M}^*\boldsymbol{M} = \boldsymbol{N}^*\boldsymbol{N},$$

where, from Lemma 2.11, we know that $\boldsymbol{M} = \boldsymbol{D}\boldsymbol{N}$ for some $\boldsymbol{D} \in \mathfrak{D}$ such that $\boldsymbol{D}^*\boldsymbol{D} = \boldsymbol{I}$. Let

$$\boldsymbol{E}_1(\boldsymbol{G}, \gamma, \boldsymbol{W}) := \widehat{\boldsymbol{M}}^*(\boldsymbol{W})\widehat{\boldsymbol{M}}(\boldsymbol{W}), \qquad \boldsymbol{E}_2(\boldsymbol{G}, \gamma, \boldsymbol{W}) := \widehat{\boldsymbol{N}}^*(\boldsymbol{W})\widehat{\boldsymbol{N}}(\boldsymbol{W}).$$

Now, from Lemma 2.6

$$\begin{aligned}\widehat{\boldsymbol{M}}(\boldsymbol{W}) &= \widehat{\boldsymbol{D}\boldsymbol{N}}(\boldsymbol{W}) \\ &= \boldsymbol{D}\widehat{\boldsymbol{N}}(\boldsymbol{W}).\end{aligned}$$

Thus,

$$\begin{aligned}\boldsymbol{E}_1(\boldsymbol{G}, \gamma, \boldsymbol{W}) &= \widehat{\boldsymbol{M}}^*(\boldsymbol{W})\widehat{\boldsymbol{M}}(\boldsymbol{W}) \\ &= \widehat{\boldsymbol{N}}^*(\boldsymbol{W})\boldsymbol{D}^*\boldsymbol{D}\widehat{\boldsymbol{N}}(\boldsymbol{W}) \\ &= \widehat{\boldsymbol{N}}^*(\boldsymbol{W})\widehat{\boldsymbol{N}}(\boldsymbol{W}) \\ &= \boldsymbol{E}_2(\boldsymbol{G}, \gamma, \boldsymbol{W}).\end{aligned}$$

The entropy operator (3.1) has many of the properties that the integral (1.1) exhibits for time-invariant systems. In the next lemma we outline some of these properties. Since we are primarily interested in the $\boldsymbol{W} = \boldsymbol{0}$ case, we abbreviate the notation for this special case: $\boldsymbol{E}(\boldsymbol{G}, \gamma) := \boldsymbol{E}(\boldsymbol{G}, \gamma, \boldsymbol{0})$.

LEMMA 3.2. *With the notation of Definition* 3.1 *we have*

(i) $\boldsymbol{E}(\boldsymbol{G}, \gamma) \geq \boldsymbol{0}$.

(ii) $\boldsymbol{E}(\boldsymbol{G}, \gamma) \leq \boldsymbol{I}$ *with equality iff* $\boldsymbol{G} \equiv \boldsymbol{0}$.

(iii) *If* $\boldsymbol{U} \in \mathfrak{X}$, $\boldsymbol{V} \in \mathfrak{L}$ *with* $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$ *and* $\boldsymbol{V}^*\boldsymbol{V} = \boldsymbol{I}$, *then*

$$\boldsymbol{E}(\boldsymbol{U}\boldsymbol{G}\boldsymbol{V}, \gamma) = \widehat{\boldsymbol{V}}^*(\boldsymbol{0})\boldsymbol{E}(\boldsymbol{G}, \gamma)\widehat{\boldsymbol{V}}(\boldsymbol{0}).$$

Proof. Property (i) is straightforward. For property (ii), note from Lemmas 2.1 and 2.2 that since $\widehat{\boldsymbol{M}}(\boldsymbol{0}) = \boldsymbol{M}_{[0]}$,

$$\left\|\widehat{\boldsymbol{M}}(\boldsymbol{0})\right\| = \sup_i \|\boldsymbol{M}(i, i)\| \leq \|\boldsymbol{M}\| \leq 1$$

with equality only when $\boldsymbol{G} \equiv \boldsymbol{0}$. Thus

$$\boldsymbol{I} - \widehat{\boldsymbol{M}}^*(\boldsymbol{0})\widehat{\boldsymbol{M}}(\boldsymbol{0}) \geq \boldsymbol{0}$$

and hence $\boldsymbol{E}(\boldsymbol{G}, \gamma) \leq \boldsymbol{I}$.

To show (iii), first note that

$$\|\boldsymbol{UGV}\| \leq \|\boldsymbol{U}\| \, \|\boldsymbol{G}\| \, \|\boldsymbol{V}\| = \|\boldsymbol{G}\| < \gamma.$$

Thus

$$\begin{aligned} \boldsymbol{I} - \gamma^{-2}(\boldsymbol{UGV})^*(\boldsymbol{UGV}) &= \boldsymbol{V}^* \left(\boldsymbol{I} - \gamma^{-2}\boldsymbol{G}^*\boldsymbol{G}\right) \boldsymbol{V} \\ &= (\boldsymbol{MV})^*(\boldsymbol{MV}). \end{aligned}$$

Since $\boldsymbol{M}$ and $\boldsymbol{V} \in \mathfrak{L}$, the product $\boldsymbol{MV} \in \mathfrak{L}$. Thus $\boldsymbol{MV}$ is a spectral factor for

$$\boldsymbol{I} - \gamma^{-2}(\boldsymbol{UGV})^*(\boldsymbol{UGV}).$$

Using Corollary 2.10, we have

$$\widehat{\boldsymbol{MV}}(\boldsymbol{0}) = \widehat{\boldsymbol{M}}(\boldsymbol{0})\widehat{\boldsymbol{V}}(\boldsymbol{0}),$$

which proves property (iii). □

We now show that this definition of the entropy is a natural generalization of the entropy integral (1.1) by showing that, for time-invariant systems, the two entropies are strongly related.

PROPOSITION 3.3. *Suppose that $\boldsymbol{G} \in \mathfrak{L}$ and that $\boldsymbol{G}$ represents a time-invariant system (that is, $\boldsymbol{G}$ commutes with the shift $\boldsymbol{Z}$). Then $\boldsymbol{E}(\boldsymbol{G}, \gamma, \lambda \boldsymbol{I})$ is a diagonal operator with constant matrix elements along the diagonal, i.e.,*

$$\boldsymbol{E}(\boldsymbol{G}, \gamma, \lambda \boldsymbol{I}) = \operatorname{diag}\{\ldots, E_0, \boxed{E_0}, E_0, \cdots\}. \tag{3.2}$$

Moreover, if $G(\lambda)$ is the transfer function associated with this operator, the entropies $\boldsymbol{E}(\boldsymbol{G}, \gamma, \lambda_0 \boldsymbol{I})$ and $I_d(G, \gamma, \lambda_0)$ are related by

$$I_d(G, \gamma, \lambda_0) = \gamma^2 \ln |\det E_0|. \tag{3.3}$$

Proof. We first evaluate $I_d(G, \gamma, \lambda_0)$. Suppose that

$$I - \gamma^{-2} G^{\sim}(\lambda) G(\lambda) = M^{\sim}(\lambda) M(\lambda)$$

is a spectral factorization. It follows that

$$\begin{aligned} I_d(G, \gamma, \lambda_0) &= \frac{\gamma^2}{\pi} \int_{-\pi}^{\pi} \ln \left|\det \left(M(e^{i\omega})\right)\right| \frac{1 - |\lambda_0|^2}{|e^{i\omega} - \lambda_0|^2} d\omega \\ &= 2\gamma^2 \ln |\det (M(\lambda_0))| \\ &= \gamma^2 \ln \left|\det \left(M^T(\lambda_0) M(\lambda_0)\right)\right|, \end{aligned}$$

where we have used the Poisson integral formula in the second line and the fact that $\det(X) = \det(X^T) = \det(X^T X)$ in the third. Now, let $\boldsymbol{M}$ be the Toeplitz operator with symbol $M(\lambda)$. It follows that this operator is a spectral factor of $\boldsymbol{I} - \gamma^{-2}\boldsymbol{G}^*\boldsymbol{G}$ and

$$\boldsymbol{E}(\boldsymbol{G}, \gamma, \lambda \boldsymbol{I}) = \widehat{\boldsymbol{M}}^*(\lambda_0 \boldsymbol{I}) \widehat{\boldsymbol{M}}(\lambda_0 \boldsymbol{I}).$$

Hence, from Example 2.4 we know that

$$\widehat{\boldsymbol{M}}(\lambda_0 \boldsymbol{I}) = \operatorname{diag}\{M(\lambda_0)\}.$$

Thus (3.2) and (3.3) hold with $E_0 := M^T(\lambda_0) M(\lambda_0)$. □

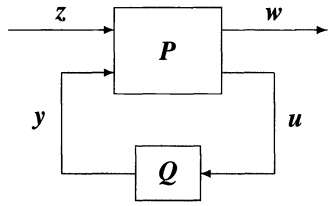


FIG. 4.1. *Closed-loop system.*

3.2. State-space formulae. For systems defined by the state-space representation (2.3), it is possible to give a state-space formula for the entropy. For notational simplicity, we assume that $\gamma = 1$.

It follows from Theorem 2.12 that a spectral factor for $\boldsymbol{M}$ for $\boldsymbol{I} - \boldsymbol{G}^*\boldsymbol{G}$ is given by

$$\boldsymbol{M} = \left[\begin{array}{c|c} \boldsymbol{A} & \boldsymbol{B} \\ \hline -\boldsymbol{V}^{-1/2}(\boldsymbol{B}^*\boldsymbol{Z}\boldsymbol{X}\boldsymbol{Z}^*\boldsymbol{A} + \boldsymbol{D}^*\boldsymbol{C}) & \boldsymbol{V}^{1/2} \end{array}\right].$$

The diagonal component of this spectral factor is $\boldsymbol{V}^{1/2}$, and thus

$$\boldsymbol{E}(\boldsymbol{G}, 1) = \boldsymbol{V} = \boldsymbol{I} - \boldsymbol{D}^*\boldsymbol{D} - \boldsymbol{B}^*\boldsymbol{Z}\boldsymbol{X}\boldsymbol{Z}^*\boldsymbol{B}.$$

In the study of $\mathcal{H}_\infty$ control theory, we find that controllers, and thus closed-loop systems, can often be written as linear fractional transformations of an inner transfer function P and a stable contraction Q. In the next section we show that, as in the case of time-invariant systems, the entropy operator can be used to choose from among a set of controllers.

4. Maximizing the entropy operator. For linear time-invariant systems, the set of closed-loop systems can be characterized in the form of a linear fractional transformations of an inner transfer function P and a stable contraction Q [10, 5, 14]. For time-varying systems, a similar characterization of stabilizing controllers exists [22].

In the time-invariant case, the integral (1.1) can be used to select from among the possible closed-loop systems. Partition the transfer function P as $\left(\begin{smallmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{smallmatrix}\right)$ with respect to the stable contraction Q. It is known that controller which maximizes the entropy integral (1.1) for z_0 is given by the choice $Q_{\max} = [P_{22}(z_0)]^\sim$ [11, 16]. In the time-invariant case this block of the transfer function P is strictly proper, and thus when $z_0 = 0$, the entropy maximizing contraction is $Q = 0$. This is known as the *central* controller of $\mathcal{H}_\infty$ control, which happens to coincide with with the optimal risk-sensitive controller of stochastic control [10].

In this section we will show that the entropy operator $\boldsymbol{E}(\boldsymbol{G}, \gamma)$ plays the corresponding role in time-varying optimization problems. Before doing so, we must show that a linear fractional transformation of the corresponding operators is well posed. In order to do this, we require a time-varying version of Redheffer's lemma.

LEMMA 4.1. *Suppose that, in Figure* 4.1, $\boldsymbol{P} = \left(\begin{smallmatrix} \boldsymbol{P}_{11} & \boldsymbol{P}_{12} \\ \boldsymbol{P}_{21} & \boldsymbol{P}_{22} \end{smallmatrix}\right)$, *with* $\boldsymbol{P}_{11}, \boldsymbol{P}_{12}, \boldsymbol{P}_{22} \in \mathfrak{L}$, $\boldsymbol{P}_{21} \in \mathfrak{L} \cap \mathfrak{L}^{-1}$, *is an isometry that admits a doubly coprime factorization. Furthermore, assume that* $\boldsymbol{Q}$ *is a causal (not necessarily bounded) operator also admitting a doubly coprime factorization. The following two statements are equivalent.*

(i) *The system is internally stable and well posed with* $\|\mathcal{F}_\ell(\boldsymbol{P}, \boldsymbol{Q})\| < 1$.

(ii) $\boldsymbol{Q} \in \mathfrak{L}$ *and* $\|\boldsymbol{Q}\| < 1$.

Proof. See §A.1. □

Suppose that all closed-loop systems $\boldsymbol{H}$ are characterized as a lower linear fractional transformation

$$\boldsymbol{H} = \gamma\mathcal{F}_\ell(\boldsymbol{P}, \boldsymbol{Q}),$$

where P is as in Lemma 4.1 and Q is a causal contraction. By Lemma 4.1, $H \in \mathfrak{L}$ and contractive. Thus, the entropy operator $E(H, \gamma)$ is well defined for all allowable Q. In the following proposition, we show that as Q varies over the set of all causal contractive operators, $E(H, \gamma)$ has a maximum.

PROPOSITION 4.2. *Suppose that* $H = \gamma \mathcal{F}_\ell(P, Q)$ *denotes the set of all closed-loop systems, where* P *is as in Lemma* 4.1 *with* $\|P_{22}(0)\| < 1$ *and* Q *is a causal, bounded contractive operator. Then*

(a) $E(H, \gamma)$ *is maximized by the unique choice*

$$Q = Q_{\max} := \hat{P}_{22}^*(0);$$

(b) *the maximum value of the entropy is given by*

$$E(\mathcal{F}_\ell(P, Q_{\max}), \gamma) = \hat{P}_{21}^*(0)\big(I - \hat{P}_{22}(0)\hat{P}_{22}^*(0)\big)^{-1}\hat{P}_{21}(0).$$

Proof. Since Q is a contractive operator, the bounded, Hermitian operator $I - Q^*Q$ has a spectral factorization. Denote the spectral factor by L. Now,

$$\begin{aligned} I - \gamma^{-2}H^*H &= I - \mathcal{F}_\ell(P, Q)^*\mathcal{F}_\ell(P, Q) \\ &= P_{21}^*\left(I - Q^*P_{22}^*\right)^{-1}\left(I - Q^*Q\right)\left(I - P_{22}Q\right)^{-1}P_{21} \\ &= \left[L\left(I - P_{22}Q\right)^{-1}P_{21}\right]^*\left[L\left(I - P_{22}Q\right)^{-1}P_{21}\right] \\ &=: N^*N. \end{aligned}$$

By assumption, $P_{21} \in \mathfrak{L} \cap \mathfrak{L}^{-1}$. Moreover, $L \in \mathfrak{L} \cap \mathfrak{L}^{-1}$, since L is a spectral factor. Finally, it is shown in the proof of Lemma 4.1 that $(I - P_{22}Q) \in \mathfrak{L} \cap \mathfrak{L}^{-1}$. Thus $N \in \mathfrak{L} \cap \mathfrak{L}^{-1}$, and it is clearly a spectral factor of the operator $I - \gamma^{-2}H^*H$.

Using Corollary 2.10, we can evaluate

$$\left[L\left(I - P_{22}Q\right)^{-1}P_{21}\right]\hat{}(0) = \hat{L}(0)\left(I - P_{22}Q\right)^{-1}\hat{}(0)\hat{P}_{21}(0).$$

Writing $(I - P_{22}Q)^{-1}$ as a series and again using Corollary 2.10 we see that

$$\begin{aligned} \left(I - P_{22}Q\right)^{-1}\hat{}(0) &= \left[\sum_{k=0}^{\infty}(P_{22}Q)^k\right]\hat{}(0) \\ &= \sum_{k=0}^{\infty}\big(\hat{P}_{22}(0)\hat{Q}(0)\big)^k \\ &= \big(I - \hat{P}_{22}(0)\hat{Q}(0)\big)^{-1}. \end{aligned} \tag{4.1}$$

The sum in (4.1) converges, since $\|\hat{P}_{22}(0)\hat{Q}(0)\| \le \|P_{22}Q\| < 1$. It follows that[2]

$$E(H, \gamma) = \hat{P}_{21}^*(0)\big(I - \hat{P}_{22}(0)\hat{Q}(0)\big)^{-*}E(Q, 1)\big(I - \hat{P}_{22}(0)\hat{Q}(0)\big)^{-1}\hat{P}_{21}(0). \tag{4.2}$$

Note that since $\|\hat{P}_{22}(0)\| < 1$, the operators

$$I - \hat{P}_{22}(0)\hat{P}_{22}^*(0), \qquad I - \hat{P}_{22}^*(0)\hat{P}_{22}(0)$$

admit unique positive-definite Hermitian square roots; see [7, p. 88].

[2]We use the notation $A^{-*} = (A^*)^{-1}$.

We proceed now as in [2] and define the following Julia $\mathfrak{D}^{2\times 2}$ operator (see [26, p. 148]) acting on the Hilbert space $\ell_2 \oplus \ell_2$:

$$X := \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$$
$$:= \begin{bmatrix} -\widehat{P}_{22}(0) & \left(I - \widehat{P}_{22}^*(0)\widehat{P}_{22}(0)\right)^{1/2} \\ \left(I - \widehat{P}_{22}(0)\widehat{P}_{22}^*(0)\right)^{1/2} & \widehat{P}_{22}(0) \end{bmatrix}.$$

It is straightforward to check that X is an isometry and thus, by Lemma 4.1, the linear fractional transformation

$$T := \gamma \mathcal{F}_\ell(X, Q)$$

is well defined. Moreover, note that $T = 0 \iff Q = \widehat{P}_{22}^*(0)$.

Since $X_{21} \in \mathfrak{D}$, then $\hat{X}_{21}(0) = X_{21}$. Proceeding as above, we can evaluate

$$E(T, \gamma) = X_{21}^*\left(I - \widehat{P}_{22}(0)\widehat{Q}(0)\right)^{-*} E(Q, 1)\left(I - \widehat{P}_{22}(0)\widehat{Q}(0)\right)^{-1} X_{21}. \tag{4.3}$$

Comparing (4.2) and (4.3), we see that

$$\begin{aligned} E(Q, 1) &= \left(I - \widehat{P}_{22}(0)\widehat{Q}(0)\right)^* \widehat{P}_{21}^{-*}(0) E(H, \gamma) \widehat{P}_{21}^{-1}(0)\left(I - \widehat{P}_{22}(0)\widehat{Q}(0)\right) \\ &= \left(I - \widehat{P}_{22}(0)\widehat{Q}(0)\right)^* X_{21}^{-*} E(T, \gamma) X_{21}^{-1}\left(I - \widehat{P}_{22}(0)\widehat{Q}(0)\right). \end{aligned}$$

Thus

$$E(H, \gamma) = \widehat{P}_{21}^*(0) X_{21}^{-*} E(T, \gamma) X_{21}^{-1} \widehat{P}_{21}(0). \tag{4.4}$$

Since X_{21} and $\widehat{P}_{21}(0)$ are both independent of Q, the maximum in (4.4) is obtained whenever $E(T, \gamma)$ is maximized. By property (ii) of Lemma 3.2, we know that this achieved uniquely for $T \equiv 0 \Rightarrow Q = \widehat{P}_{22}^*(0)$, from which the result of part (a) follows. Part (b) follows immediately upon substitution. □

5. Extensions. In this section we outline some possible extensions of the entropy operator defined here and some difficulties that arise with each.

5.1. Finite-time-horizon systems. The entropy defined in this paper, despite having these many desirable properties, differs significantly from the usual entropy in that the expression in (3.1) defines entropy to be an operator and not a real number as in (1.1). For operators associated with finite-time-horizon systems, we may define an entropy number analogous to that of (3.1); this is now done.

Let $M = \{M(i, j)\}$ represent a bounded, causal operator. Define the operator

$$M_N := P_N M(I - P_0).$$

With respect to the direct sum decomposition $\ell_2 = P_0\ell_2 \oplus (P_N - P_0)\ell_2 \oplus (I - P_N)\ell_2$, this operator has a natural partition as

$$M_N = \begin{bmatrix} 0 & 0 & 0 \\ 0 & H^N & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where

$$H^N = \begin{cases} \{M(i, j)\}, & 0 \le i, j \le N - 1, \\ 0 & \text{elsewhere.} \end{cases}$$

Note that if $\boldsymbol{M} = \sum_{k=0}^{\infty} \boldsymbol{M}_{[k]}(\boldsymbol{Z}^*)^k$, then

$$\begin{aligned} \boldsymbol{M}_N &= \sum_{k=0}^{\infty} \big(\boldsymbol{P}_N \boldsymbol{M}_{[k]} (\boldsymbol{Z}^*)^k (\boldsymbol{I} - \boldsymbol{P}_0) \boldsymbol{Z}^k\big) (\boldsymbol{Z}^*)^k \\ &= \sum_{k=0}^{N-1} \big(\boldsymbol{P}_N \boldsymbol{M}_{[k]} (\boldsymbol{I} - \boldsymbol{P}_k)\big) (\boldsymbol{Z}^*)^k \qquad (5.1) \\ &=: \sum_{k=0}^{N-1} \boldsymbol{M}^N_{[k]} (\boldsymbol{Z}^*)^k, \end{aligned}$$

where, in (5.1), we have used the identity $(\boldsymbol{Z}^*)^k \boldsymbol{P}_0 \boldsymbol{Z}^k = \boldsymbol{P}_k$. We can evaluate

$$\begin{aligned} \widehat{\boldsymbol{M}}_N(\boldsymbol{0}) = \boldsymbol{M}^N_{[0]} &= \boldsymbol{P}_N \boldsymbol{M}_{[0]} (\boldsymbol{I} - \boldsymbol{P}_0) \\ &=: \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}^N_{[0]} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}. \end{aligned}$$

DEFINITION 5.1. *Suppose that $\boldsymbol{G} \in \mathfrak{X}$ and $\|\boldsymbol{G}\| < \gamma$, and let $\boldsymbol{M}$ be a spectral factor of the positive operator $\boldsymbol{I} - \gamma^{-2}\boldsymbol{G}^*\boldsymbol{G}$. Consider the finite rank operator $\boldsymbol{M}_N$ defined as above as well as its nonzero component $\boldsymbol{H}^N$. We define*

$$\mathcal{E}_N(\boldsymbol{G}, \gamma) := \frac{\gamma^2}{N} \ln \left| \det \left((\hat{\boldsymbol{H}}^N_{[0]})^* \hat{\boldsymbol{H}}^N_{[0]} \right) \right|.$$

In the next result we show that, for time-invariant systems, the entropy $\mathcal{E}_N$ coincides with the integral (1.1).

PROPOSITION 5.2. *For time-invariant operators, $\boldsymbol{G} = \{G(i, i-k)\} \in \mathfrak{L}$, with $G(i, i-k) = g_k$ and $\gamma \in \mathbb{R}$ such that $\|\boldsymbol{G}\| < \gamma$, the following holds:*

$$\mathcal{E}_N(\boldsymbol{G}, \gamma) = I_d(G, \gamma, 0),$$

where $G(\lambda)$ is the λ-transform of the sequence $\{g_k\}$.

Proof. The proof of this is very similar to that of Proposition 3.3. Following the notation in that proof, we note that, if

$$G(\lambda) = \sum_{k=0}^{\infty} g_k \lambda^k \qquad \text{and} \qquad M(\lambda) = \sum_{k=0}^{\infty} m_k \lambda^k,$$

then

$$I_d(G, \gamma, 0) = 2\gamma^2 \ln |\det(m_0)|$$

and, recalling from Example 2.2 that for Toeplitz operators, $\boldsymbol{M}_{[k]} = m_k \boldsymbol{I}$,

$$\widehat{\boldsymbol{M}}_N(\boldsymbol{0}) = \boldsymbol{P}_N \boldsymbol{M}_{[0]} (\boldsymbol{I} - \boldsymbol{P}_0) = m_0 \boldsymbol{P}_N (\boldsymbol{I} - \boldsymbol{P}_0)$$

and

$$\hat{\boldsymbol{H}}^N_{[0]} = m_0 I_N,$$

where I_N is a block $N \times N$ identity matrix. Thus

$$\mathcal{E}_N(\boldsymbol{G}, \gamma) = 2\frac{\gamma^2}{N} \ln \left| \det \left(\hat{H}^N_{[0]} \right) \right| = 2\gamma^2 \ln |\det(m_0)|. \qquad \square$$

5.2. General W. While the entropy definition in (1.1) allows one to work with an entropy with respect to any $z_0 \in \{z : |z| < 1\}$, the point of greatest interest is that with $z_0 = 0$. It is to this particular entropy that our operator entropy maximization corresponds. Nevertheless, it would be desirable to generalize Proposition 4.2 to more general "points" corresponding to operators $\boldsymbol{W} \neq \boldsymbol{0}$. Consider the general form of the entropy of Definition 3.1:

$$\boldsymbol{E}(\boldsymbol{G}, \boldsymbol{W}, \gamma) = \widehat{\boldsymbol{M}}(\boldsymbol{W})^* \widehat{\boldsymbol{M}}(\boldsymbol{W}).$$

This formula satisfies properties (i) and (ii) of Lemma 3.2; however, it does not satisfy property (iii). More importantly, it does not seem to satisfy the same maximization property of Proposition 4.2 and for this reason is of limited use. The proof of Proposition 4.2 breaks down since it relies heavily on Corollary 2.10, which does not hold for general $\boldsymbol{W}$.

5.3. Continuous-time systems. In the case of continuous-time, linear time-invariant systems, there exists an entropy integral analogous to that of (1.1); see [21]. For time-varying systems, however, it is well known that the input-output operators which are analogous to $\boldsymbol{G}$ exist in continuous resolution spaces. In general, positive, invertible Hermitian operators in these spaces *do not* have spectral factorizations; see [4, Thm. 14.2]. Since the definition of the entropy for discrete-time systems given here depends crucially on the existence of these factorizations, it is not clear how to generalize this to continuous-time systems.

6. Conclusions. State-space methods have by now become prevalent in the theory of $\mathcal{H}_\infty$ control. Apart from being advantageous in terms of the numerical computations required, they have also allowed straightforward extensions of the theory to other settings, including time-varying systems. Until now, however, it has not been possible to extend the definition of the entropy of a system to this setting, since it relied heavily on the transfer function of the system. This paper has given this extension in terms of non-Toeplitz operators. The entropy defined here, while being an operator rather than a real number, has many of the same properties as that used in the time-invariant case. Moreover, for time-varying systems with state-space realizations, state-space formulae for this entropy have been provided.

Appendix A. In this appendix we prove our time-varying version of Redheffer's lemma.

A.1. Proof of Lemma 4.1. (i)$\Leftarrow$(ii) For our proof, we modify the proof for the time-invariant case found in [5, Lem. 15].

Since $\boldsymbol{P}$ is an isometry, $\|\boldsymbol{P}_{22}\| \leq 1$. This, together with the fact that $\boldsymbol{Q}$ is a contraction, implies that $\|\boldsymbol{P}_{22}\boldsymbol{Q}\| < 1$. Thus, the series

$$\sum_{k=0}^{\infty} (\boldsymbol{P}_{22}\boldsymbol{Q})^k$$

converges in $\mathfrak{L}$ and is equal to $(\boldsymbol{I} - \boldsymbol{P}_{22}\boldsymbol{Q})^{-1}$. This implies that $\boldsymbol{Q}$ stabilizes $\boldsymbol{P}_{22}$. By the coprime assumption on $\boldsymbol{P}$ and a time-varying version of Lemma 4.2.1 in [8], it follows that $\boldsymbol{Q}$ internally stabilizes $\boldsymbol{P}$. Now to show that $\mathcal{F}_\ell(\boldsymbol{P}, \boldsymbol{Q})$ is a contraction, we use the fact that $\boldsymbol{P}$ is an isometry and a little algebra to show that

$$\begin{aligned} \mathcal{F}_\ell(\boldsymbol{P}, \boldsymbol{Q})^* \mathcal{F}_\ell(\boldsymbol{P}, \boldsymbol{Q}) &= \boldsymbol{I} - \boldsymbol{P}_{21}^*(\boldsymbol{I} - \boldsymbol{Q}^*\boldsymbol{P}_{22}^*)^{-1}(\boldsymbol{I} - \boldsymbol{Q}^*\boldsymbol{Q})(\boldsymbol{I} - \boldsymbol{P}_{22}\boldsymbol{Q})^{-1}\boldsymbol{P}_{21} \\ &\leq \boldsymbol{I}, \end{aligned}$$

where we have used the fact that $\boldsymbol{Q}$ is a contraction.

(ii)$\Leftarrow$(i) To show the converse, we first prove that $\boldsymbol{Q}$ is a bounded operator. Recall that $\boldsymbol{Q}$ has a right coprime factorization $\boldsymbol{Q} = \boldsymbol{N}\boldsymbol{D}^{-1}$, where $\boldsymbol{N}, \boldsymbol{D} \in \mathfrak{L}$. Note that $\boldsymbol{Q} \in \mathfrak{L} \iff \boldsymbol{D} \in$

$\mathfrak{L} \cap \mathfrak{L}^{-1}$; see [7, p. 182]. From the internal stability assumption, we know that

$$\boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{P}_{22}\boldsymbol{Q})^{-1} \in \mathfrak{L} \Rightarrow \boldsymbol{N}(\boldsymbol{D} - \boldsymbol{P}_{22}\boldsymbol{N})^{-1} \in \mathfrak{L}.$$

Now, since $\boldsymbol{N}$ and $\boldsymbol{D}$ are right coprime, it follows that $\boldsymbol{N}$ and $\boldsymbol{D} - \boldsymbol{P}_{22}\boldsymbol{N}$ are also right coprime. To see this, suppose that $\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}} \in \mathfrak{L}$ such that $\tilde{\boldsymbol{X}}\boldsymbol{N} + \tilde{\boldsymbol{Y}}\boldsymbol{D} = \boldsymbol{I}$. Then,

$$\boldsymbol{X}\boldsymbol{N} + \boldsymbol{Y}(\boldsymbol{D} - \boldsymbol{P}_{22}\boldsymbol{N}) = \boldsymbol{I}$$

with $\boldsymbol{X} = \tilde{\boldsymbol{X}} + \boldsymbol{Y}\boldsymbol{P}_{22} \in \mathfrak{L}$ and $\boldsymbol{Y} = \hat{\boldsymbol{Y}} \in \mathfrak{L}$, which proves coprimeness. It follows, again from [7, p. 182], that

$$\begin{aligned} \boldsymbol{N}(\boldsymbol{D} - \boldsymbol{P}_{22}\boldsymbol{N})^{-1} \in \mathfrak{L} &\Rightarrow (\boldsymbol{D} - \boldsymbol{P}_{22}\boldsymbol{N})^{-1} \in \mathfrak{L} \\ &\Rightarrow \boldsymbol{D}^{-1}(\boldsymbol{I} - \boldsymbol{P}_{22}\boldsymbol{Q})^{-1} \in \mathfrak{L} \\ &\Rightarrow \boldsymbol{D}^{-1} \in \mathfrak{L}, \end{aligned}$$

where the last line comes from that fact that $(\boldsymbol{I} - \boldsymbol{P}_{22}\boldsymbol{Q})^{-1} \in \mathfrak{L} \cap \mathfrak{L}^{-1}$. Thus, $\boldsymbol{Q} \in \mathfrak{L}$.

We now show that $\boldsymbol{Q}$ is a contraction. Assume otherwise; thus there exists a signal $\boldsymbol{y} \in \ell_2$ such that $\boldsymbol{u} = \boldsymbol{Q}\boldsymbol{y} \in \ell_2$ and $\|\boldsymbol{u}\| \geq \|\boldsymbol{y}\|$. Let $\boldsymbol{w} = \boldsymbol{P}_{21}^{-1}(\boldsymbol{I} - \boldsymbol{P}_{22}\boldsymbol{Q})\boldsymbol{y}$. This is in ℓ_2, since $\boldsymbol{P}_{21} \in \mathfrak{L} \cap \mathfrak{L}^{-1}$. Moreover, from the isometry condition we know that

$$\begin{aligned} \|\boldsymbol{z}\|^2 + \|\boldsymbol{y}\|^2 &= \|\boldsymbol{w}\|^2 + \|\boldsymbol{u}\|^2 \\ &\geq \|\boldsymbol{w}\|^2 + \|\boldsymbol{y}\|^2. \end{aligned}$$

Thus $\|\boldsymbol{z}\|^2 \geq \|\boldsymbol{w}\|^2$, which contradicts the assumption that $\mathcal{F}_\ell(\boldsymbol{P}, \boldsymbol{Q})$ is a contraction.

REFERENCES

[1] D. ALPAY, P. DEWILDE, AND H. DYM, *Lossless Inverse Scattering and Reproducing Kernels for Upper Triangular Operators*, Oper. Theory Adv. Appl. OT 47, Birkhäuser, Basel, 1990, pp. 61–135.

[2] D. AROV AND M. KREĬN, *Problem of search of the minimum entropy in indeterminate extension problems*, Functional Anal. Appl., 15 (1981), pp. 123–126.

[3] W. ARVESON, *Interpolation in nest algebras*, J. Funct. Anal., 20 (1975), pp. 208–233.

[4] K. DAVIDSON, *Nest Algebras*, Pitman Res. Notes Math. Ser. 191, Longman Scientific & Technical, Harlow, UK, 1988.

[5] J. DOYLE, K. GLOVER, P. KHARGONEKAR, AND B. FRANCIS, *State-space solutions to standard $\mathcal{H}_\in$ and $\mathcal{H}_\infty$ control problems*, IEEE Trans. Automat. Contr., 34 (1989), pp. 831–847.

[6] H. DYM AND I. GOHBERG, *A maximum entropy principle for contractive interpolants*, J. Funct. Anal., 65 (1986), pp. 83–125.

[7] A. FEINTUCH AND R. SAEKS, *System Theory: A Hilbert Space Approach*, Academic Press, New York, 1982.

[8] B. FRANCIS, *A Course in $\mathcal{H}_\infty$ Control Theory*, Lecture Notes in Control and Inform. Sci. 88, Springer-Verlag, New York, 1987.

[9] K. GLOVER, *Relations between $\mathcal{H}_\infty$ and risk sensitive controllers*, in 8th Intl. Conf. Analysis & Optimization of Systems, INRIA, Antibes, France, 1988.

[10] K. GLOVER AND J. DOYLE, *State-space formulae for all stabilizing controllers that satisfy an $\mathcal{H}_\infty$ norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[11] K. GLOVER AND D. MUSTAFA, *Derivation of the maximum entropy $\mathcal{H}_\infty$-controller and a state space formula for its entropy*, Internat. J. Control, 50 (1989), pp. 899–916.

[12] I. GOHBERG, M. KAASHOEK, AND H. WOERDEMAN, *A maximum entropy principle in the general framework of the band method*, J. Funct. Anal., 95 (1991), pp. 231–254.

[13] P. IGLESIAS, *On the Stabilization of Discrete-Time Linear Time-Varying Systems*, Tech. rep. TR 94/08, Dept. Elec. & Comp. Eng., Johns Hopkins Univ., Baltimore, MD, 1994. Submitted for publication.

[14] P. IGLESIAS AND K. GLOVER, *A state space approach to discrete-time $\mathcal{H}_\infty$ control*, Internat. J. Control, 54 (1991), pp. 1031–1074.

[15] P. IGLESIAS AND D. MUSTAFA, *A separation principle for discrete time controllers satisfying a minimum entropy criterion*, IEEE Trans. Automat. Control, 38 (1993), pp. 1525–1530.

[16] P. IGLESIAS, D. MUSTAFA, AND K. GLOVER, *Discrete-time $\mathcal{H}_\infty$ controllers satisfying a minimum entropy criterion*, Systems Control Lett., 14 (1990), pp. 275–286.

[17] P. IGLESIAS AND M. PETERS, *On the induced norms of discrete-time and hybrid time-varying systems*, International Journal of Robust and Nonlinear Control, to appear.

[18] E. KAMEN, P. KHARGONEKAR, AND K. POOLLA, *A transfer-function approach to linear time-varying discrete-time systems*, SIAM J. Control Optim., 23 (1985), pp. 550–565.

[19] P. KHARGONEKAR, R. RAVI, AND K. NAGPAL, *$\mathcal{H}_\infty$ control of linear time-varying systems: A state-space approach*, SIAM J. Control Optim., 29 (1991), pp. 1394–1413.

[20] D. LIMEBEER, M. GREEN, AND D. WALKER, *Discrete time $\mathcal{H}_\infty$ control*, in IEEE Conf. Decision and Control, Tampa Bay, FL, Dec. 1989, IEEE, Piscataway, NJ, pp. 392–396.

[21] D. MUSTAFA AND K. GLOVER, *Minimum Entropy $\mathcal{H}_\infty$ Control*, Lecture Notes in Control and Inform. Sci. 146, Springer-Verlag, Heidelberg, 1990.

[22] M. VERHAEGEN AND A.-J. VAN DER VEEN, *The bounded real lemma for discrete-time varying systems with application to robust output feedback*, in IEEE Conf. Decision and Control, San Antonio, TX, Dec. 1993, pp. 45–50.

[23] L. WANG AND G. ZAMES, *Local-global double algebras for slow $\mathcal{H}_\infty$ adaptation: Part* II*—Optimization of stable plants*, IEEE Trans. Automat. Control, 36 (1991), pp. 143–151.

[24] ———, *Local-global double algebras for slow $\mathcal{H}_\infty$ adaptation: The case of ℓ^2 disturbances*, IMA J. Math. Control Inform., 8 (1991), pp. 287–319.

[25] P. WHITTLE, *Risk-sensitive Optimal Control*, John Wiley and Sons, New York, 1990.

[26] N. YOUNG, *An Introduction to Hilbert Space*, Cambridge University Press, Cambridge, UK, 1988.

[27] G. ZAMES, *Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, 26 (1981), pp. 301–320.

[28] G. ZAMES AND L. WANG, *Local-global double algebras for slow $\mathcal{H}_\infty$ adaptation: Part* I*—Inversion and stability*, IEEE Trans. Automat. Control, 36 (1991), pp. 130–142.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1707–1733, September 1996

012

ON MARKOVIAN FRAGMENTS OF COCOLOG FOR LOGIC CONTROL SYSTEMS*

YUANJUN WEI† AND PETER E. CAINES‡

Abstract. The COCOLOG (Conditional Observer and Controller Logic) system is a partially ordered family of first-order logical theories expressed in the typed first-order languages $\{\mathcal{L}_k; k \geq 0\}$ describing the controlled evolution of the state of a given partially observed finite machine $\mathcal{M}$. The initial theory of the system, denoted Th_0, gives the theory of $\mathcal{M}$ with no data being given on the initial state. Later theories, $\{Th(o_1^k); k \geq 1\}$, depend upon the (partially ordered lists of) observed input-output trajectories $\{o_1^k; k \geq 1\}$, where new data are accepted in the form of the new axioms $AXM^{obs}(\mathcal{L}_k)$, $k \geq 1$. A feedback control input $U(k)$ is determined by the solution of a collection of control problems posed in the form of a set of conditional control rules $CCR(\mathcal{L}_k)$, such a set being paired with the theory $Th(o_1^k)$ for each $k \geq 1$. In this paper, we introduce a restricted version of COCOLOG, called a system of Markovian fragments of COCOLOG, in which a smaller amount of information is communicated from one theory to the next. Such fragment theories are associated with a restricted set of candidate control problems, denoted $CCR(\mathcal{L}_k^m)$, $k \geq 1$. It is shown that a Markovian fragment theory $MTh(o_1^k)$ contains a large subset of $Th(o_1^k)$, which includes, in particular, the state estimation theorems of the corresponding full COCOLOG system and, for the set of control rules $CCR(\mathcal{L}_k^m)$, has what is termed the same control reasoning power. Next, it is shown that proofs of theorems in the fragment systems are necessarily shorter than their proofs in the full COCOLOG systems. Finally some computer-generated examples are given, illustrating this increased theorem-proving efficiency.

Key words. discrete-event systems, finite machines, logic control

AMS subject classifications. 93, 68, 03

1. Introduction.

The COCOLOG system, introduced by P. E. Caines and S. Wang [CW95], [CW91], [W91], is a partially ordered family of first-order logical theories which describe the controlled evolution of the state of a given partially observed finite machine $\mathcal{M}$. Unlike most knowledge-based control systems, reasoning in COCOLOG is based upon the representation of the two fundamental properties of a partially observed dynamical system with inputs, namely, *controllability* (or *reachability*) and *observability*. The predicate symbols Rbl and CSE_k are introduced for this purpose. The initial theory of the system, Th_0, gives the general theory of $\mathcal{M}$ without any data being given on the initial state. Later theories, $\{Th(o_1^k); k \geq 1\}$, depend upon the (partially ordered lists of) observed input-output trajectories $\{o_1^k; k \geq 1\}$, through their axiom sets $\{\Sigma_k; k \geq 1\}$, since new data are accepted sequentially into the subsequent theories in the form of the *new axioms* $AXM^{obs}(\mathcal{L}_k)$. The inputs $U(k)$ are determined by the solution to control problems associated with each theory in the form of the *conditional control rules* $CCR(\mathcal{L}_k)$. An important class of control problems involves the aforementioned reachability predicate $Rbl(x, y, l)$, which is defined axiomatically in each theory and corresponds to the reachability of the state y from the state x in l steps. The solution to one problem of this type would be, say, the first control in a sequence of controls giving a minimal length path to the state y from the current state x.

This paper is concerned with the definition of tractable fragments of the full COCOLOG system which carry enough information to enable significant classes of control problems to be posed (through the conditional control rules) and resolved in a limited subset of the language $\mathcal{L}_k$ at each instant k. Due to the overall dynamical framework of a COCOLOG system, we are able to formulate what we call the Markovian fragments of a general COCOLOG system in such a way that they possess the required tractability properties. These are introduced in

*Received by the editors August 26, 1993; accepted for publication (in revised form) June 6, 1995. This research was supported by NSERC grant A1329 and NSERC–NCE–IRIS-I Program project B-5.

†Department of Electrical Engineering, McGill University, Montreal, PQ, H3A 2A7, Canada.

‡Department of Electrical Engineering, McGill University, Montreal, PQ, H3A 2A7, Canada, and Canadian Institute for Advanced Research.

this paper via the definition of a restricted set of languages $\{\mathcal{L}_k^m; k \geq 0\}$ and the associated set of axioms $\{M\Sigma_k(o_1^k); k \geq 0\}$ for the set of theories $\{MTh(o_1^k); k \geq 1\}$. In contrast to the evolution of a full COCOLOG system, the evolution of a system of Markovian fragments consists of a combination of axiom set expansion (adding some previously derived theorems and new axioms into the axiom set) and axiom set contraction (deleting some old axioms) at each time instant. A part of $\{M\Sigma_k(o_1^k); k \geq 0\}$ expresses only the basic dynamical properties of the machine under control plus the most recent observations, while another part expresses the state estimate generated in the most recent COCOLOG theory. In addition, an updated version of a set of control problems is carried in a corresponding set of conditional control rules; these are phrased only in terms of the predicates and axioms available in the restricted theories $MTh(o_1^k)$, $k \geq 1$. In particular, this avoids the unbounded increment in axiomatic theorem proving (ATP) complexity due to the increase in the number of formulas in the successive axiom sets.

The implementation of control reasoning in COCOLOG requires efficient ATP methodologies. The development of function evaluation (FE) resolution [WC92] and the Blitzensturm methodology [CMW93] have been steps in this direction. Both have recently been shown to be efficiently implementable. It may be seen from the analysis in this paper that the sizes of proof trees in Markovian fragment theories are necessarily smaller than or equal to the corresponding proof trees in full COCOLOG theories. At the end of this paper, we present some computer-generated examples to illustrate this comparison. These and other experiments implementing ATP in the full and the Markovian fragment COCOLOG systems result in a corresponding speed-up of the computing time required for certain COCOLOG control problems.

The paper is organized as follows. After a brief review of COCOLOG systems, §3 presents the definition of the fragment languages $\mathcal{L}_k^m$, $k \geq 0$. The axiomatizations of fragment theories and their semantics are given in §§4 and 5, respectively. Section 6 contains the first main result of the paper, stating that, for a large class of control problems, no loss in control reasoning power is incurred by restriction to the fragment theories. Section 7 presents the second main result, which deals with proof complexity. Appendix 1 consists of a complete description of an axiom set of the basic theory, Σ_0. Finally, two pairs of computer-based experiments are presented in Appendix 2 to illustrate the difference between proofs of the same theorems in full COCOLOG and in its Markovian fragment counterpart, respectively.

Earlier versions of the theory developed in this paper were presented in [WeC92] and [CWe94].

2. COCOLOG. The reader is referred to [CW91, CW95, W91, WC92] for a full exposition of all terms and expressions which are not completely explained in the summary of the formulation of a COCOLOG system given below.

2.1. Syntax of COCOLOG.

DEFINITION 2.1. *A (partially observed) finite (input-state-output) machine is a quintuple*

$$\mathcal{M} \underline{\Delta} (\mathbf{U}, \mathbf{X}, \mathbf{Y}, \mathbf{\Phi}, \eta),$$

where $\mathbf{U}$ *is a (finite) set of* inputs, $\mathbf{X}$ *is a (finite) set of* states, $\mathbf{Y}$ *is a (finite) set of* outputs, $\mathbf{\Phi} : X \times U \to X$ *is a* transition function, *and* $\eta : \mathbf{X} \to \mathbf{Y}$ *is an* output function.

In this paper, we set $|\mathbf{U}| = R$, $|\mathbf{X}| = N$, and $|\mathbf{Y}| = M$.

We always use boldface letters to distinguish semantic objects from the symbols in the first-order language that describe them. For the purpose of describing such machines in a first-order language, we define the symbol set $S(\mathcal{L}_0)$ that contains the *constant symbol set*:

$$\begin{aligned}\text{Const}(\mathcal{L}_0) &= U \cup X \cup Y \cup I_{K(N)} \\ &= \{u^1, \ldots, u^R\} \cup \{x^1, \ldots, x^N\} \cup \{y^1, \ldots, y^M\} \cup \{0, 1, \ldots, K(N)+1\};\end{aligned}$$

the *variable symbol set*:

$$\text{Var}(\mathcal{L}_0) = \{u, u', \ldots\} \cup \{x, x', \ldots\} \cup \{y, y', \ldots\} \cup \{i, j, l, \ldots\};$$

the *function symbol set*, $\text{Func}(\mathcal{L}_0)$, that contains the symbols $\{\Phi(\ ,\ ), \eta(), +_{K(N)}, -_{K(N)}\}$; the *atomic predicate symbol set*, $\text{Pre}(\mathcal{L}_0)$, that contains $\{Eq, Rbl\}$; and the *logical symbol set*, $\{\forall, \rightarrow, \bot\}$, where $\bot$ is a logical constant together with the derived symbols $\exists, \neg, \top$. The set of typed terms $\text{Term}(\mathcal{L}_0)$ includes the first two sets of symbols, together with certain finite strings of symbols, whose farthest left symbol is an $n-ary$ functional symbol followed by n terms (see [CW95]).

Any *well-formed formula* of $\mathcal{L}_0$ is given by the standard first-order grammar. The set of such formulas will be denoted $WFF(\mathcal{L}_0)$.

2.2. Axiomatic theory of Th_0. The basic axiom set, which generates the basic theory Th_0, has a set of *logical axioms*, a set of *equality axioms* for an equality predicate, a set of *arithmetic axioms*, and a set of *special axioms* which specify the facts concerning the subject that the logic describes (in at least one of its interpretations).

Finite machine axioms. The special axiom set of Th_0 corresponds exactly to the state transitions and output map relations of the given machine $\mathcal{M}$:

State transition axioms:

$$AXM^{dyn}(\mathcal{L}_0) \underline{\Delta} \{Eq(\Phi(x^i, u^l), x^j);\ x^i, x^j \in X, u^l \in U\},$$

where an entry appears in the braces if and only if, for $\mathcal{M}$, $\mathbf{\Phi(x^i, u^l) = x^j}$.

Output axioms:

$$AXM^{out}(\mathcal{L}_0) \underline{\Delta} \{Eq(\eta(x^i), y^j);\ x^i \in X, y^j \in Y\},$$

where an entry appears in the braces if and only if, for $\mathcal{M}$, $\mathbf{\eta(x^i) = y^j}$.

Example 2.1 (a seven-state machine). Some members of the axiom set AXM^{dyn} for the machine in Figure 1 are

$$\begin{array}{lll} Eq(\Phi(x^1, u^1), x^2) & Eq(\Phi(x^1, u^2), x^3) & Eq(\Phi(x^1, u^3), x^4) \\ Eq(\Phi(x^2, u^1), x^3) & Eq(\Phi(x^2, u^2), x^4) & Eq(\Phi(x^2, u^3), x^5) \\ \cdots & \cdots & \cdots, \end{array}$$

and some members of the axiom set AXM^{out} are

$$\begin{array}{lll} Eq(\eta(x^1), y^1) & Eq(\eta(x^2), y^2) & Eq(\eta(x^3), y^3) \\ \cdots & \cdots & \cdots. \end{array} \quad \square$$

Reachability axioms. Denoted by $AXM^{Rbl}(\mathcal{L}_0)$, these are recursively defined for the *reachability predicate Rbl* by the following:

0. $\forall x \forall x',\ Eq(x, x') \longleftrightarrow Rbl(x, x', 0)$,
1. $\forall x \forall x',\ (\exists u,\ Eq(\Phi(x, u), x')) \longleftrightarrow Rbl(x, x', 1)$,
2. $\forall x \forall x'' \forall l,\ Eq(l, K(N)+1)$
   $\vee [\{\exists x' \exists u,\ Rbl(x', x'', l) \wedge Eq(\Phi(x, u), x')\} \longleftrightarrow Rbl(x, x'', l +_{K(N)} 1)]$. $\square$

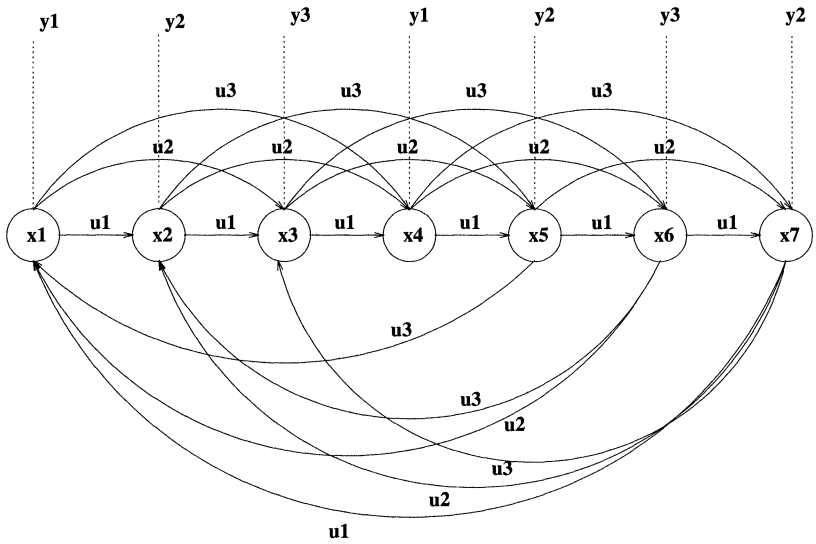


FIG. 1. *A seven-state machine.*

The reachability axioms specify the l-step reachability relation $Rbl(x, x', l)$ among any pair of states x, x'. We note that in these formulas the variables x, x', x'' range over X, the variable u ranges over U, and l ranges over the integers $0, 1, \ldots, K(N)+1$. Axiom 2 excludes consideration of the infinity case in order to characterize reachability on the finite numbers in the arithmetic.

Size axioms (see Appendix 1). Denoted by $AXM^{size}(\mathcal{L}_0)$, these specify the restriction that any model of this axiom set must have a domain that contains exactly N state objects, R input objects, M output objects, and $K(N) + 2$ integers. This restriction is natural since the controlled machine $\mathcal{M}$ is fixed.

Finite arithmetic axioms (see Appendix 1). Denoted by $AXM^{arith}(\mathcal{L}_0)$, these define the arithmetical operations $+_{K(N)}$ and $-_{K(N)}$ on the initial segment of the natural numbers $\{0, \ldots, K(N) + 1\}$.

Equality axioms (see Appendix 1). Denoted by $AXM^{Eq}(\mathcal{L}_0)$, these consist of the basic axioms for equality and the substitution axioms for every functional symbol and predicate symbol in the language.

Logical axioms (see Appendix 1). Denoted by $AXM^{logic}(\mathcal{L}_0)$, this is the standard set of axiom schemata for first-order logic.

Axiom set Σ_0. We write Σ_0 for the union of the above axiom sets of $\mathcal{L}_0$, i.e.,

$$\Sigma_0 \triangleq \{AXM^{arith}(\mathcal{L}_0), AXM^{dyn}(\mathcal{L}_0), AXM^{out}(\mathcal{L}_0), AXM^{Rbl}(\mathcal{L}_0), AXM^{Eq}(\mathcal{L}_0), AXM^{logic}(\mathcal{L}_0)\}.$$

Rules of inference (see Appendix 1). The rules of inference in Th_0 are *Modus Ponens* (MP) and *generalization.*

2.3. Axiomatic Theory of $Th(o_1^k)$. At each instant $k \geq 1$, the controlled machine $\mathcal{M}$ generates a pair of observed data $(\mathbf{u}^p, \mathbf{y}^q)$ for some $\mathbf{u}^p \in \mathbf{U}$, $\mathbf{y}^q \in \mathbf{Y}$. The axiom set is updated to express the acquisition of this new information, and new control rules are added to decide

the subsequent control action $U(k)$. In order to formalize this change the language is extended by introducing new constant symbols.

We will use the notation $\mathbf{o}_1^k$, $k \geq 1$, to denote the observed sequence received on the time interval $[1, k]$, i.e.,

$$\{(\emptyset, \mathbf{y}^{j_1}), (\mathbf{u}^{i_2}, \mathbf{y}^{j_2}), \ldots, (\mathbf{u}^{i_k}, \mathbf{y}^{j_k})\}$$

The typed language $\mathcal{L}_k \triangleq L(o_1^k)$ is an extension of the language $\mathcal{L}_0$ which is obtained by adding new constant symbols and predicates symbols in the following way:

$$S(\mathcal{L}_k) \triangleq S(L(o_1^k)) = S(\mathcal{L}_0) \bigsqcup_{j=1}^{k} \{U(j-1), Y(j)\} \bigsqcup_{j=1}^{k} \{CSE_j\}.$$

Here $U(j-1)$ and $Y(j)$ are new constant symbols, representing the observed control input at the time instant $j-1$ and output at the time instant j generated by the controlled machine $\mathcal{M}$. CSE_j is a new one-place predicate symbol which is called the current state estimation predicate at time j, $j \geq 1$.

The syntax of $\mathcal{L}_k$ is standard, and we remark only that the variables and constants are sorted and the well-formed formulas parse according to the first-order grammar of each $\mathcal{L}_k$.

Observation axioms ($AXM^{obs}(\mathcal{L}_k)$). The observation of the control action $\mathbf{u^p} \in \mathbf{U}$, taken at the instant $k-1$, and the output $\mathbf{y^q} \in \mathbf{Y}$, generated at the instant k, are expressed in the form $Eq(U(k-1), u^p)$ and $Eq(Y(k), y^q)$. These are added to the previous axiom set Σ_{k-1} as axioms to express the fact that these observations have taken place. Let

$$AXM^{obs}(\mathcal{L}_k) \triangleq \{Eq(Y(k), y^q), Eq(U(k-1), u^p)\}, \quad k \geq 1,$$

where this set of formulas is subject to the convention that the second axiom above holds only in case $k > 1$.

State estimation axioms ($AXM^{est}(\mathcal{L}_k)$). These express in axiomatic form the recursive formulas for the current state estimate sets.

In case $k = 1$:

$$Eq(\eta(x^i), Y(k)) \longleftrightarrow CSE_1(x^i), \qquad 1 \leq i \leq N. \tag{1}$$

In case $k > 1$:

$$\exists x, CSE_{k-1}(x) \wedge Eq(\Phi(x, U(k-1)), x^i) \wedge Eq(Y(k), \eta(x^i)) \longleftrightarrow CSE_k(x^i), \qquad 1 \leq i \leq N. \tag{2}$$

This axiom set will also be added to the previous axiom set. Finally we add the following.

Substitution axioms ($AXM^{subs}(\mathcal{L}_k)$ (for CSE_k, $k \geq 1$)).

$$\forall x_1 \forall x_2 (Eq(x_1, x_2) \rightarrow (CSE_k(x_1) \rightarrow CSE_k(x_2))). \tag{3}$$

Henceforth we set

$$\Sigma_k \triangleq \Sigma_k(o_1^k) = \Sigma_0 \bigsqcup_{j=1}^{k} \{AXM^{obs}(\mathcal{L}_j), AXM^{subs}(\mathcal{L}_j), AXM^{est}(\mathcal{L}_j)\}. \tag{4}$$

Rules of inference (see Appendix 1). The rules of inference in $Th(o_1^k)$ are MP and *generalization.*

We formally define the proof mechanics of Σ_k as follows.

DEFINITION 2.2. *For any $k \geq 0$, a proof sequence $\mathcal{P}$ for a formula F with respect to the axiom set Σ_k is a finite indexed list of formulas in which F is the last on the list, and any other formula in the list is either an instance of a logical axiom schemata, a member of Σ_k, or a formula deduced from previous formulas in the list through MP or generalization. We call $\mathcal{P}$ a Σ_k-*proof. *If such a proof sequence exists, we say that F is* deducible from *or* provable from *Σ_k and denote this by $\Sigma_k \vdash F$. F is called a theorem of Σ_k. Finally we use $|\mathcal{P}|$, the length of $\mathcal{P}$, to denote the number of formulas in $\mathcal{P}$.*

$Th(o_1^k)$, called the theory generated from *Σ_k, is the set of all theorems that can be deduced from Σ_k; that is to say, it is the set*

$$\{F : F \in WFF(\mathcal{L}_k), \Sigma_k \vdash F\}. \qquad \square$$

Conditional control rules ($CCR(\mathcal{L}_k)$). The following is the general form of a set of *conditional control rules* at time instant k. Let $C^j(\mathcal{L}_k)$, $1 \leq j \leq R$, be a set of conditional control formulas associated with u^j, $1 \leq j \leq R$, expressed in $WFF(\mathcal{L}_k)$, and let

$$D^p(\mathcal{L}_k) = \bigwedge_{i=1}^{p-1} \neg C^i(\mathcal{L}_k) \wedge C^p(\mathcal{L}_k). \tag{5}$$

Then the associated $CCR(\mathcal{L}_k)$ set is the set of extralogical statements

if	$D^1(\mathcal{L}_k)$,	**then**	$Eq(U(k), u^1)$;
if	$D^2(\mathcal{L}_k)$,	**then**	$Eq(U(k), u^2)$;
if	$\vdots$	**then**	$\vdots$
if	$D^R(\mathcal{L}_k))$,	**then**	$Eq(U(k), u^R)$;
if	$\bigwedge_{j=1}^{R}(\neg C^j(\mathcal{L}_k))$,	**then**	$Eq(U(k), u^*)$. $\square$

We use the notation $CCF(\mathcal{L}_k)$ (standing for the set of *conditional control formulas* in the language $\mathcal{L}_k$) to denote the conditions in the **if** parts of a given set of rules $CCR(\mathcal{L}_k)$, i.e.,

$$CCF(\mathcal{L}_k) \triangleq \left\{ D^1(\mathcal{L}_k), D^2(\mathcal{L}_k), \ldots, D^R(\mathcal{L}_k), \bigwedge_{j=1}^{R} \neg C^j(\mathcal{L}_k) \right\}.$$

The set of rules $CCR(\mathcal{L}_k)$, $k \geq 1$, is central to the construction of a COCOLOG controller. The operation of the members of such a set is as follows.

Extralogical feedback control specification. If the condition $C^1(\mathcal{L}_k)$ is provable from the theory $Th(o_1^k)$, then the first rule gives the value u^1 to the control constant $U(k)$ (i.e., the control action u^1 will take place in the controlled system); if not, but if $C^2(\mathcal{L}_k)$ is provable, then the second rule gives the value u^2 to the control constant $U(k)$, and so on. If none of the conditions $C^1(\mathcal{L}_k), C^2(\mathcal{L}_k), \ldots, C^R(\mathcal{L}_k)$ is provable, then the last rule sets the control action equal to the arbitrary constant control u^*. This procedure uniquely determines the value of $U(k)$. $\square$

When $k \to k+1$, we make the extralogical step of passing to the theory $Th(o_1^{k+1})$ by carrying along all the previous axioms and adding the axioms $AXM^{obs}(\mathcal{L}_{k+1})$ specifying the

observation of the input, e.g., $Eq(U(k), u^i)$ in case u^i was selected, and the observation of the output, e.g., $Eq(Y(k+1), y^q)$ in case y^q was generated. This is formally enforced by the above definition of the axiom set generating $Th(o_1^{k+1})$. Hence, in the new theory $Th(o_1^{k+1})$, the observed control action $U(k)$ is specified so as to be the constant value u^i determined by $Th(o_1^k)$ through $CCR(\mathcal{L}_k)$.

DEFINITION 2.3. *A* COCOLOG controller *for $\mathcal{M}$, together with the observation sequence $\mathbf{o}_1^k$ at time instant $k \geq 0$, is a pair $\langle \Sigma_k, CCR(\mathcal{L}_k) \rangle$, where Σ_k and $CCR(\mathcal{L}_k)$ are defined above.* □

2.4. Semantics of COCOLOG. For the semantics of any theory $Th(o_1^k)$ in a COCOLOG system, we adopt, in Definitions 2.4 and 2.5 below, a conventional set theoretic model interpretation (see, e.g., [RG87]). Since Σ_k contains all symbols of $\mathcal{L}_k$, we will not distinguish the prestructure of the axiom set and that of the language when the context is clear.

DEFINITION 2.4. *For all $k \geq 0$, a* prestructure *of $\mathcal{L}_k$ is a pair $\langle I_k, \mathbf{D} \rangle$, denoted by $\mathcal{I}_k$, where $\mathbf{D}$ is a nonempty set called the domain, which is the union of $\mathbf{U}$, $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{I}_{K(N)}$, and where I_k is the interpretation mapping which is defined as follows:*

(1) $I_k(c) = \mathbf{c} \in \mathbf{D}$ *for all $c \in \mathcal{L}_k$ such that* $I_k : U \to \mathbf{U}$, $I_k : X \to \mathbf{X}$, $I_k : Y \to \mathbf{Y}$, $I_k : I_{K(N)} \to \mathbf{I}_{K(N)}$.

(2) $I_K(\Phi) = \mathbf{\Phi} : \mathbf{X} \times \mathbf{U} \to \mathbf{X}$.

(3) $I_K(\eta) = \eta : \mathbf{X} \to \mathbf{Y}$.

(4) $I_k(+_{K(N)}) = +_{\mathbf{K(N)}} : \mathbf{I_{K(N)}} \times \mathbf{I_{K(N)}} \to \mathbf{I_{K(N)}}$.

(5) $I_k(-_{K(N)}) = -_{\mathbf{K(N)}} : \mathbf{I_{K(N)}} \times \mathbf{I_{K(N)}} \to \mathbf{I_{K(N)}}$.

(6) *For* $t = f(\vec{t}) \in \mathrm{Term}(\mathcal{L}_k) : I_k(f(\vec{t})) = I_k(f)(I_k(\vec{t}))$.

(7) $I_k(Eq) = \{(\mathbf{d}, \mathbf{d}) : \mathbf{d} \in \mathbf{D}\}$.

(8) $I_k(Rbl) \subset \mathbf{X} \times \mathbf{X} \times \mathbf{I_{K(N)}}$.

(9) $I_k(CSE_j) \subset \mathbf{X}$, $1 \leq j \leq k$.

(10) $I_k(U(j-1)) \in \mathbf{U}$, $1 \leq j \leq k$.

(11) $I_k(Y(j)) \in \mathbf{Y}$, $1 \leq j \leq k$.

(12) *For* $P \in \mathrm{Pre}(\mathcal{L}_k)$, $\vec{t} \in \mathrm{Term}(\mathcal{L}_k)$, $I_k(P(\vec{t})) = I_k(P)(I_k(\vec{t}))$. □

DEFINITION 2.5. *For all $k \geq 0$, a* structure *for Σ_k and an input-output string $\mathbf{o}_1^k$, $k \geq 1$, is a pair $\langle \mathcal{I}_k, V_k \rangle$, denoted by $\mathcal{H}_k$, where $\mathcal{I}_k$ is a prestructure of Σ_k and V_k is a corresponding* evaluation mapping $WFF(\mathcal{L}_k) \to \{0, 1\}$ *with corresponding type. Under this structure, each formula in $WFF(\mathcal{L}_k)$ will be assigned recursively a truth value $V_k(F) \in \{0, 1\}$ as follows:*

(1) $V_k(\perp) = 0$.

(2) *For a ground atomic formula $P(\vec{t})$, $V_k(P(\vec{t})) = 1$ iff $I_k(P(\vec{t})) \in I_k(P)$.*

(3) *For $F = F_1 \vee F_2$, $V_k(F) = 1$ iff either $V_k(F_1) = 1$ or $V_k(F_2) = 1$.*

(4) *For $F = \neg F_1$, $V_k(F) = 1$ iff $V_k(F_1) = 0$.*

(5) *For $F = F_1 \to F_2$, $V_k(F) = 1$ iff either $V_k(F_1) = 0$ or $V_k(F_2) = 1$.*

(6) *For $F = \forall x F_1$, $V_k(F) = 1$ iff $V_k(F_1(x/c)) = 1$ for all $c \in \mathrm{Const}(\mathcal{L}_k)$.*

(7) *For $F = \exists x F_1$, $V_k(F) = 1$ iff $V_k(F_1(x/c)) = 1$ for some $c \in \mathrm{Const}(\mathcal{L}_k)$.*

In cases (6) *and* (7), *the term c is called a* witness *of x.*

When $V_k(F) = 1$, we say that $\mathcal{H}_k$ satisfies *F, or $\mathcal{H}_k$ is a* model *of F, and denote this by $\mathcal{I}_k \models F$. If $\mathcal{I}_k \models F$ holds for every $F \in \Sigma_k$, then we say that this structure is a model of Σ_k and write $\mathcal{H}_k \models \Sigma_k$. If every model of Σ_k is also a model of F, then we call F a* logical consequence *of Σ_k and write $\Sigma_k \models F$.* □

It should be noted that the sets $\mathbf{U}$, $\mathbf{X}$, and $\mathbf{Y}$ in Definition 2.4 are not in general identical to the sets appearing in the definition of the machine $\mathcal{M}$ which defines the language $\mathcal{L}_0$.

Henceforth, for all $k \geq 0$, we assume that any language $\mathcal{L}_k$ and axiom system $\{\Sigma_k, k \geq 0\}$ are defined so that some machine $\mathcal{M}$, together with a given input-output sequence $\mathbf{o}_1^k$, $k \geq 0$, is a model in the sense of Definitions 2.4 and 2.5 for Σ_k. This is expressed by saying that

$\{\Sigma_k; k \geq 0\}$ is an axiom system for $\mathcal{M}$ and the observed sequence $\mathbf{o}_1^k$. Note that for $k \geq 0$ a model $\mathcal{H}_k$ is defined without any reference to an initial state for a given machine $\mathcal{M}$ and that such an entity is not defined in the languages $\{\mathcal{L}_k, k \geq 0\}$.

Some important properties of COCOLOG families of theories are given in the following theorems.

THEOREM 2.1 (see [CW95, W91]). *For all $k \geq 0$, the axiom set Σ_k for $\mathcal{M}$ together with the observed input-output sequence $\mathbf{o}_1^k$ is consistent.* □

THEOREM 2.2 (unique model property [CW95, W91]). *For all $k \geq 0$, the logical theory $Th(o_1^k)$, generated by the axiom system Σ_k for $\mathcal{M}$ together with the observed input-output sequence $\mathbf{o}_1^k$, has a unique model up to isomorphism.* □

THEOREM 2.3 (decidable theoremhood [CW95, W91]). *For all $k \geq 0$, the logical theory $Th(o_1^k)$, generated by the axiom set Σ_k for $\mathcal{M}$ together with the observed input-output sequence $\mathbf{o}_1^k$, is decidable.* □

THEOREM 2.4 (the nesting theorem [CW95, W91]). *For all $k \geq 0$ and $\mathbf{o}_1^k \subset \mathbf{o}_1^{k+1}$, the logical theory $Th(o_1^k)$, generated by the axiom set Σ_k for $\mathcal{M}$ together with the given sequence $\mathbf{o}_1^k$, is a subtheory of $Th(o_1^{k+1})$, generated by the axiom set Σ_{k+1} for $\mathcal{M}$ and the sequence $\mathbf{o}_1^{k+1}$, i.e., $Th(o_1^k) \subset Th(o_1^{k+1})$.* □

Concerning the size of the axiom set at time $k \geq 0$, we have the following lemma.

LEMMA 2.1. *The size of the axiom sets Σ_k for $\mathcal{M}$ together with the generated sequence $\mathbf{o}_1^k$, $k \geq 0$, satisfies*

$$|\Sigma_k| = |\Sigma_0| + k(N+3).$$

Proof.

$$\begin{aligned}
|\Sigma_k| &= \left|\Sigma_0 \bigcup_{j=1}^{j=k} (AXM^{est}(L_j) \cup AXM^{obs}(L_j) \cup AXM^{subs}(L_j))\right| \\
&= |\Sigma_0| + \left|\bigcup_{j=1}^{j=k} AXM^{est}(L_j)\right| + \left|\bigcup_{j=1}^{j=k} AXM^{obs}(L_j)\right| + \left|\bigcup_{j=1}^{j=k} AXM^{subs}(L_j))\right| \\
&= |\Sigma_0| + \sum_{j=1}^{j=k} |AXM^{est}(L_j)| + \sum_{j=1}^{j=k} |AXM^{obs}(L_j)| + \sum_{j=1}^{j=k} |AXM^{subs}(L_j))| \\
&= |\Sigma_0| + k \cdot N + k \cdot 2 + k \cdot 1 \\
&= |\Sigma_0| + k(N+3). \qquad \square
\end{aligned}$$

Since the results in this paper do not depend explicitly on the sequence of values of any given observed sequence $\mathbf{o}_1^k$, we will, from now on, omit the indication of the particular observation sequence, and in particular, we shall write Th_k instead of $Th(o_1^k)$.

Example 2.2. Suppose that the machine in Example 2.1 generates the observation sequence

$$(\emptyset, \mathbf{y}^1) \quad (\mathbf{u}^1, \mathbf{y}^2) \quad (\mathbf{u}^2, \mathbf{y}^1).$$

Then corresponding observation axiom sets are

$$AXM^{obs}(\mathcal{L}_1) = \{Eq(Y(1), y^1)\}, \quad AXM^{obs}(\mathcal{L}_2) = \{Eq(U(1), u^1), Eq(Y(2), y^2)\},$$
$$AXM^{obs}(\mathcal{L}_3) = \{Eq(U(2), u^2), Eq(Y(3), y^1)\}.$$

A proof sequence in Th_3 which gives a deduction of $CSE_3(x^4)$ is as follows:

(1) $Eq(\Phi(x^1, u^1), x^2)$ AXM^{dyn},
(2) $Eq(\Phi(x^2, u^2), x^4)$ AXM^{dyn},
(3) $Eq(\eta(x^1), y^1)$ AXM^{out},
(4) $Eq(\eta(x^2), y^2)$ AXM^{out},
(5) $Eq(\eta(x^4), y^1)$ AXM^{out},
(6) $Eq(Y(1), y^1)$ $AXM^{obs}(\mathcal{L}_1)$,
(7) $Eq(\eta(x^1), Y(1)) \rightarrow CSE_1(x^1)$ $AXM^{est}(\mathcal{L}_1)$,
(8) $CSE_1(x^1)$ (6), (7), MP,
(9) $Eq(U(1), u^1)$ $AXM^{obs}(\mathcal{L}_2)$,
(10) $Eq(Y(2), y^2)$ $AXM^{obs}(\mathcal{L}_2)$,
(11) $\exists x, CSE_1(x) \wedge Eq(\Phi(x, U(1)), x^2) \wedge Eq(\eta(x^2), Y(2)) \rightarrow CSE_2(x^2)$ $AXM^{est}(\mathcal{L}_2)$,
(12) $CSE_1(x^1) \wedge Eq(\Phi(x^1, U(1)), x^2) \wedge Eq(\eta(x^2), Y(2))$ (1), (4), (9), (8), (10), $\wedge$−rule,
(13) $\exists x, CSE_1(x) \wedge Eq(\Phi(x, U(1)), x^2) \wedge Eq(\eta(x^2), Y(2))$ (12), $\exists$−rule,
(14) $CSE_2(x^2)$ (11), (13), MP,
(15) $Eq(U(2), u^2)$ $AXM^{obs}(\mathcal{L}_3)$,
(16) $Eq(Y(3), y^1)$ $AXM^{obs}(\mathcal{L}_3)$,
(17) $\exists x, CSE_2(x) \wedge Eq(\Phi(x, U(2)), x^4) \wedge Eq(\eta(x^4), Y(3)) \rightarrow CSE_3(x^4)$ $AXM^{est}(\mathcal{L}_3)$,
(18) $CSE_2(x^2) \wedge Eq(\Phi(x^2, U(2)), x^4) \wedge Eq(\eta(x^4), Y(3))$ (2), (5), (14), (15), (16), $\wedge$−rule,
(19) $\exists x, CSE_2(x) \wedge Eq(\Phi(x, U(2)), x^4) \wedge Eq(\eta(x^4), Y(3))$ (18), $\exists$−rule,
(20) $CSE_3(x^4)$ (17), (19), MP. □

One may notice that in order to prove (20), $AXM^{obs}(\mathcal{L}_3)$ and (14) are used. Further, the proof of (14) invokes $AXM^{obs}(\mathcal{L}_2)$ and (8), etc. Actually this chain goes back to the first observation pair. It is obvious that the lengths of the proof sequences of theorems in COCOLOG increase with the lengths of the observation sequences. This is an obstacle to the efficiency of a COCOLOG controller. In the rest of this paper, we present a restricted fragment of the COCOLOG system in which the reasoning complexity is independent of the lengths of the observation sequences.

3. Language fragment $\mathcal{L}_k^m$. The full COCOLOG language defined in the previous section has the power to express the whole observation history of the system, and this gives rise to a monotonic evolution of the theories $\{Th_k; k \geq 0\}$(see the nesting theorem above) which, in particular, permits reference to the past. For example, one may write a formula to express the following control law: If the first control has not been invoked since the beginning of the process, then invoke it now. This control rule can be written as

$$\text{(6)} \qquad \textbf{if} \quad \bigwedge_{j=2}^{j=k} \neg Eq(U(j-1), u^1), \quad \textbf{then} \quad Eq(U(k), u^1),$$

and it can be seen to involve the whole collection of languages from $j = 2$ up to $j = k$ whose union is precisely $\mathcal{L}_k$. On the other hand, such expressive power is unnecessary for the purposes of control with respect to control criteria depending on present and future states and outputs, since the controlled dynamical system is, by definition, current state dependent.

There are two motivating ideas for the introduction of the Markovian fragment languages $\mathcal{L}_k^m$ and theories Th_k^m; first, it is well known that the control of partially observed state-space systems, with respect to current and future input-, state-, and output-dependent criteria, need

only be a function of current state estimates; second, the first fact is reflected in the complexity properties of ATP procedures implementing COCOLOG control rules.

Specifically, we notice that the evolution of COCOLOG theories takes place only by acquiring new observation information; there is no other "learning" taking place, and hence all facts which do not simply describe input and output observations must be deduced within each theory Th_k, $k \geq 1$. This will, in many cases, handicap the efficiency of the reasoning process as the length of the observation sequence increases. For instance, in Example 2.4, (14) is necessary to verify the desired theorem (20); but in order to verify (14), (8) is necessary. In a typical ATP implementation, this chaining effect goes back to the state estimation at the time instant $k = 1$. However, if we take (in an extralogical way) the formula in (14) as an axiom when we make the extralogical transition from Σ_2 to Σ_3, the above proof sequence will be shortened considerably. As a result, we introduce the fragment theories $\{MTh_k; k \geq 1\}$. These are propagated from instant to instant and have an exact analogy to the current state estimates which are propagated in classical control theory problems; their use enormously increases the efficiency of the theorem-proving procedures. The first step in defining the fragment theories is to restrict the set of symbols to that sufficient to express the current observation and the immediately previous one at each time instant.

DEFINITION 3.1. *The symbol set of the* Markovian fragment $\mathcal{L}_k^m$ *of* $\mathcal{L}_k$, $k \geq 1$, *is defined via*

$$S(\mathcal{L}_k^m) \triangleq S(\mathcal{L}_0) \bigcup \{CSE_k, CSE_{k-1}\} \bigcup \{U(k-1), Y(k)\},$$

where the constants and variables are sorted accordingly with respect to $\mathcal{L}_k$. □

It is obvious that $\mathcal{L}_k^m$ has a fixed number of symbols, which are fewer in number than those in $\mathcal{L}_k$. Specifically, compared with $\mathcal{L}_k$,

(1) $\text{Const}(\mathcal{L}_k^m) = \text{Const}(\mathcal{L}_0) \bigcup \{U(k-1), Y(k)\} \subset \text{Const}(\mathcal{L}_k)$.

(2) $\text{Pre}(\mathcal{L}_k^m) = \{Rbl, Eq, CSE_{k-1}, CSE_k\} \subset \text{Pre}(\mathcal{L}_k)$.

(3) $\mathcal{L}_k^m$ has the same set of functional symbols as $\mathcal{L}_k$, i.e., $\text{Func}(\mathcal{L}_k^m) = \text{Func}(\mathcal{L}_k)$.

By the above definition, a term in $\mathcal{L}_k^m$ can be formed only through $\text{Func}(\mathcal{L}_k^m)$, $\text{Var}(\mathcal{L}_k^m)$, and $\text{Const}(\mathcal{L}_k^m)$. Hence a term like $\Phi(x^i, U(k-3))$ is not a term of $\mathcal{L}_k^m$. This restriction also holds for the formulas of $\mathcal{L}_k^m$ defined below.

DEFINITION 3.2. *For all* $k \geq 0$, *the set of* well-formed formulas $WFF(\mathcal{L}_k^m)$ *of* $\mathcal{L}_k^m$ *is defined using the same connectives and formation rules as* $\mathcal{L}_k$ *but is subject to the restriction that the only permitted atomic formulas are instances of Rbl,* CSE_k, CSE_{k-1}, *and Eq with respect to terms of* $\mathcal{L}_k^m$. □

Due to the construction of $\mathcal{L}_k^m$,

$$WFF(\mathcal{L}_k^m) \subset WFF(\mathcal{L}_k), \qquad k \geq 1.$$

As a consequence, $\mathcal{L}_k^m$ cannot express the state estimate formulas concerning the state at time $k-2$ or earlier, and so, for example, (6) is no longer a conditional control rule with respect to $WFF(\mathcal{L}_k^m)$. Intuitively, the language fragment $\mathcal{L}_k^m$ can only express information that relates to the most recent change and the current configuration of the controlled machine. Hence $CCR(\mathcal{L}_k^m)$ can be written only with respect to this fraction of the total information concerning $\mathcal{M}$ at the instant k.

4. Construction of $M\Sigma_k$. With the language fragment in hand, we shall give, in this section, the axiom sets for the Markovian fragment system $\{MTh_k; k \geq 0\}$. We shall continue to make the restriction that the admissible control objectives within a fragment $\{MTh_k; k \geq 0\}$, expressed via the control rules $\{CCR(\mathcal{L}_k^m); k \geq 1\}$, shall refer only to the current and future state (estimate) behavior of the controlled system. Correspondingly, we limit the information

to be transferred into MTh_k at any instant k to that necessary to deduce the current state estimate at k or, more precisely, to deduce the set of states satisfying the predicate CSE_k in MTh_k.

We shall be concerned with the temporal structure of the fragment sequences $\ldots, MTh_{k-1}, MTh_k, MTh_{k+1}, \ldots$, each of them nested respectively within $\ldots, Th_{k-1}, Th_k, Th_{k+1}, \ldots$. We observe in passing that this is the logical analogue of the generation of the state estimate in a linear stochastic control problem. Furthermore, since a critical subset of the theorems of MTh_{k-1} forms a part of the axiom set for MTh_k, a certain form of learning may be said to take place, since these theorems do not have to be deduced again from more elementary information given in axiomatic form.

The definitions below specify an axiom set $M\Sigma_k$, expressed within the language $\mathcal{L}_k^m$, to be a certain combination of (i) axioms for the machine dynamics, reachability, and machine size; (ii) a set of axioms carrying the most recent state estimate theorems; (iii) the most recent observation axioms expressed via the equality predicate; and, in addition to the above, (iv) the most recent estimation axioms.

DEFINITION 4.1. *For a given machine $\mathcal{M}$ and the input-output sequence $\mathbf{o}_1^k$, the axiom set $M\Sigma_k$, $k \geq 0$, of a* Markovian fragment *of a COCOLOG system is recursively defined as follows:*

$$M\Sigma_0 = \Sigma_0, \tag{7}$$

$$M\Sigma_1 = \Sigma_1 = \Sigma_0 \cup AXM^{special}(\mathcal{L}_1^m). \tag{8}$$

Suppose that $M\Sigma_{k-1}$ is defined. Then

$$M\Sigma_k = M\Sigma_0 \cup AXM^{special}(\mathcal{L}_k^m) \cup K(M\Sigma_{k-1}), \quad k > 1, \tag{9}$$

where $AXM^{special}(\mathcal{L}_k^m)(\subset WFF(\mathcal{L}_k^m))$ denotes the following union:

$$AXM^{est}(\mathcal{L}_k^m) \cup AXM^{obs}(\mathcal{L}_k^m) \cup AXM^{subs}(CSE_{k-1}) \cup AXM^{subs}(CSE_k), \tag{10}$$

where $AXM^{obs}(\mathcal{L}_k^m) = \{Eq(U(k-1), u^p), Eq(Y(k), y^q)\}$ if and only if $(\mathbf{u}^p, \mathbf{y}^q)$ is the observation pair at time k and where $K(M\Sigma_{k-1})$ is defined as

$$K(M\Sigma_0) = \emptyset, \tag{11}$$

$$K(M\Sigma_{k-1}) \equiv \{?CSE_{k-1}(x^i) : M\Sigma_{k-1} \vdash ?CSE_{k-1}(x^i),\ ? \in \{\neg,\ \},\ x^i \in X\}, \quad k > 1,$$

where the notation $? = \neg$ indicates the negative assertion of CSE_{k-1} and the positive assertion is indicated by the lack of a symbol before CSE_{k-1}. □

The definition of CSE_{k+1} in a Markovian fragment theory is displayed in Figure 2. Those axiom sets that lie on the left of the dotted box will not appear in $M\Sigma_{k+1}$.

It is to be noted that

$$AXM^{est}(\mathcal{L}_k) = AXM^{est}(\mathcal{L}_k^m), \quad AXM^{subs}(CSE_{k-1}), AXM^{subs}(CSE_k) \subset WFF(\mathcal{L}_k^m)$$

by virtue of the definition of the Markovian fragment languages $\mathcal{L}_k^m$, $k \geq 0$. From the definitions of §2.3, we observe that $AXM^{obs}(\mathcal{L}_k) \subset WFF(\mathcal{L}_k^m)$, so the axiom set $M\Sigma_k$ in Definition 4.1 is well defined within $\mathcal{L}_k^m$.

Example 4.1. For the same observation sequence as in Example 2.2 we have

$$\begin{aligned} K(M\Sigma_1) = \{&CSE_1(x^1), CSE_1(x^4), \neg CSE_1(x^2), \neg CSE_1(x^3), \\ &\neg CSE_1(x^5), \neg CSE_1(x^6), \neg CSE_1(x^7)\}, \\ K(M\Sigma_2) = \{&CSE_2(x^2), CSE_2(x^5), \neg CSE_2(x^1), \neg CSE_2(x^4), \\ &\neg CSE_2(x^3), \neg CSE_2(x^6), \neg CSE_2(x^7)\}. \end{aligned}$$

□

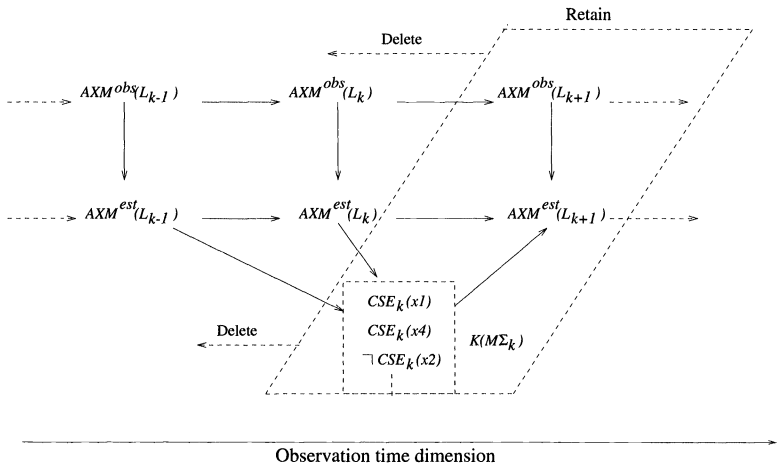


FIG. 2. *Description of definition of* CSE_{k+1}.

Informally speaking, $M\Sigma_k$ drops all the estimation axioms and observation axioms that were added to each Σ_j for $1 \leq j \leq k-1$. The loss of estimation axioms at time instant $k-1$ is compensated for by adding in $K(M\Sigma_{k-1})$, which carries the state estimate theorems from MTh_{k-1} to MTh_k in the form of axioms. Compared to Σ_k, $M\Sigma_k$ contains fewer axioms. But it is not the case that $M\Sigma_k \subset \Sigma_k$ since $K(M\Sigma_{k-1}) \not\subset \Sigma_k$. The axiom evolution for Markovian fragment theories at time instant k can be highlighted by the following equation:

$$\begin{aligned} M\Sigma_k &= M\Sigma_{k-1} \setminus (K(M\Sigma_{k-2}) \cup AXM^{special}(\mathcal{L}_{k-1}^m)) \cup (K(M\Sigma_{k-1}) \cup AXM^{special}(\mathcal{L}_k^m)) \\ &= \Sigma_0 \cup K(M\Sigma_{k-1}) \cup AXM^{special}(\mathcal{L}_k^m). \end{aligned}$$

It should be pointed out that Definition 4.1 itself states that $M\Sigma_k$ and Σ_k use the same state estimation axioms but *does not state that they receive the same observation axioms*; that this is the case requires a proof that the two theories yield the same inputs to the controlled machine $\mathcal{M}$. This is established below in our main result, Theorem 6.1.

The conditional control rules associated with the fragment theories are defined in a way similar to those associated with the full theories.

DEFINITION 4.2 ($CCR(\mathcal{L}_k^m)$). *Let* $\{C^j(\mathcal{L}_k^m), \quad 1 \leq j \leq R\}$ *be a set of formulas in* $WFF(\mathcal{L}_k^m)$, *and let*

$$D^p(\mathcal{L}_k^m) \underline{\Delta} (\neg C^1(\mathcal{L}_k^m) \wedge \cdots \wedge \neg C^{p-1}(\mathcal{L}_k^m)) \wedge C^p(\mathcal{L}_k^m) \quad \textit{for} \quad 1 \leq p \leq R.$$

Then $CCR(\mathcal{L}_k^m)$ *is defined as follows:*

$$\begin{array}{llll} \textbf{if} & D^1(\mathcal{L}_k^m), & \textbf{then} & Eq(U(k), u^1); \\ \textbf{if} & D^2(\mathcal{L}_k^m), & \textbf{then} & Eq(U(k), u^2); \\ \textbf{if} & \vdots & \textbf{then} & \vdots \\ \textbf{if} & D^R(\mathcal{L}_k^m), & \textbf{then} & Eq(U(k), u^R); \\ \textbf{if} & \bigwedge_{i=1}^{R} \neg C^i(\mathcal{L}_k^m), & \textbf{then} & Eq(U(k), u^*). \end{array} \quad \square$$

The *extralogical feedback control specification* and *fragment controller* for the fragment are exactly similar to that for a full COCOLOG system as given in §2.

In exact analogy with Definition 2.2, we define a proof sequence with respect to $M\Sigma_k$ to be a proof sequence which invokes only the members of $M\Sigma_k$. It follows that all formulas of an $M\Sigma_k$-proof sequence fall in $WFF(\mathcal{L}_k^m)$.

We define MTh_k to be the theory generated by $M\Sigma_k$, i.e.,

$$MTh_k \triangleq \{F : M\Sigma_k \vdash F, F \in WFF(\mathcal{L}_k^m)\}.$$

The following assertion concerning the size of the axiom set $M\Sigma_k$ is immediate.

LEMMA 4.1. *With the definition of a Markovian fragment given above, we have*

$$|M\Sigma_k| = |\Sigma_0| + 2N + 4, \qquad k > 1.$$

Proof. It is sufficient to observe that $K(M\Sigma_{k-1})$ contributes N axioms and the bounds are independent of the time index. □

5. Semantics of Markovian fragment theories. The fragment language $\mathcal{L}_k^m$ is a sublanguage of $\mathcal{L}_k$; hence any prestructure of $\mathcal{L}_k^m$ is a substructure of a prestructure of $\mathcal{L}_k$. We observe that the construction of $M\Sigma_k$ is ordained in such a way that a model of $M\Sigma_k$ preserves the essential properties of Σ_k; by this we mean that the properties of an interpretation of $\mathcal{L}_0$ must be preserved in an interpretation of $M\Sigma_k$. These requirements leads to the following definition of a structure for $M\Sigma_k$.

DEFINITION 5.1. *For each $k \geq 0$, the structure $\langle \mathcal{I}_k^m, V_k^m \rangle$ for $M\Sigma_k$ is defined in exactly the same way as that for Σ_k (see Definitions 2.4 and 2.5) except for the following modifications:*

(6′) *for* $t = f(\vec{t}) \in \text{Term}(\mathcal{L}_k^m) : I_k^m(f(\vec{t})) = I_k^m(f)(I_k^m(\vec{t}))$,

(9′) $I_k^m(CSE_j) \subset \mathbf{X}$, $j = k-1, k$,

(10′) $I_k^m(U(k-1)) \in \mathbf{U}$,

(11′) $I_k^m(Y(k)) \in \mathbf{Y}$,

and $V_k^m : WFF(\mathcal{L}_k^m) \longrightarrow \{0, 1\}$. □

In this definition, the model of $M\Sigma_k$ does not depend upon the whole observation sequence but upon the current observation pair and the previous state estimate set characterised by the set of x^i, $1 \leq i \leq N$, such that $CSE_k(x^i)$ holds. We now examine the evolution of full and fragment theories and define the relationship between $\mathcal{H}_{k-1}^m$ and $\mathcal{H}_k^m$.

We give below a well-known lemma for later use.

LEMMA 5.1 (coincidence lemma [EFT84]). *Let $\mathcal{H}' = \langle \mathcal{I}', V' \rangle$ and $\mathcal{H}'' = \langle \mathcal{I}'', V'' \rangle$ be structures for the languages $\mathcal{L}'$ and $\mathcal{L}''$, respectively, both with the same domain* **D**. *Let $\mathcal{L} = \mathcal{L}' \cap \mathcal{L}''$.*

(a) *For any $t \in \text{Term}(\mathcal{L})$, if $\mathcal{H}'$ and $\mathcal{H}''$ agree on the symbols of $S(\mathcal{L})$ occurring in t, then $I'(t) = I''(t)$.*

(b) *For $F \in WFF(\mathcal{L})$, if $\mathcal{H}'$ and $\mathcal{H}''$ agree on the symbols of $S(\mathcal{L})$ occurring in F, and $V'(x) = V''(x)$ for x free in F, then $\mathcal{H}' \models F$ iff $\mathcal{H}'' \models F$.* □

LEMMA 5.2. *Let $\mathcal{H}_{k-1}^m = \langle \mathcal{I}_{k-1}^m, V_{k-1}^m \rangle$ and $\mathcal{H}_k^m = \langle \mathcal{I}_k^m, V_k^m \rangle$ be models of $M\Sigma_{k-1}$ and $M\Sigma_k$, respectively. Then*

(a)

$$I_{k-1}^m|_{\mathcal{L}_{k-1}^m \cap \mathcal{L}_k^m} = I_k^m|_{\mathcal{L}_{k-1}^m \cap \mathcal{L}_k^m}. \tag{12}$$

(b) *V_{k-1}^m and V_k^m agree on the following set:*

$$\Sigma_0 \cup K(M\Sigma_{k-1}). \tag{13}$$

Proof. (a) follows from the definition of $\mathcal{H}_k^m$ and Lemma 5.1. For the proof of (b), it is enough to observe that

$$K(M\Sigma_{k-1}) \subset MTh_{k-1}, \quad K(M\Sigma_{k-1}) \subset M\Sigma_k. \qquad \square$$

DEFINITION 5.2. *Two prestructures $\mathcal{I}$ and $\mathcal{I}'$ of a language L are said to be* isomorphic *if there exists a bijection h between the domains of $\mathcal{I}$ and $\mathcal{I}'$ such that*

(1) *For every $c \in \text{Const}(L)$, $h(I(c)) = I'(c)$.*

(2) *For every $f \in \text{Func}(L)$ and $t_1, \ldots, t_l \in \text{Term}(L)$, $h(I(f(\vec{t}))) = I'(f)(I'(\vec{t}))$.*

(3) *For every $P \in \text{Pre}(L)$ and $t_1, \ldots, t_l \in \text{Term}(L)$, $I(P(\vec{t}))$ iff $I'(P)I'(\vec{t})$.* □

LEMMA 5.3 (isomorphism lemma [EFT84]). *If $\mathcal{I}$ and $\mathcal{I}'$ are isomorphic, then for any $F \in WFF(L)$, we have*

$$\mathcal{H} \models F \Longleftrightarrow \mathcal{H}' \models F. \qquad \square$$

The following theorem is the result of application of Theorem 2.1.

THEOREM 5.1 (see [CW95, W91]). *For $\mathcal{M}$ together with the observation sequence $\mathbf{o}_1^k$, $k \geq 0$, $M\Sigma_k$ is consistent.* □

The proof is via a duplication of that of Theorem 4.2 in [CW95]. Further, we can prove the following theorem.

THEOREM 5.2 (unique model property). *For $k \geq 0$, each logical theory MTh_k^m, generated by the axiom system $M\Sigma_k$ for $\mathcal{M}$ together with the observation sequence $\mathbf{o}_1^k$, has a unique model up to isomorphism.*

Proof. The proof is by induction on k. Let

$$AXM^{obs}(\mathcal{L}_k^m) = \{Eq(U(k-1), u^p), Eq(Y(k), y^q)\},$$

where u^p and y^q shall generically denote the observations at the instant k. We begin with the base case $k = 0$: in this case, $M\Sigma_0 = \Sigma_0$ and the unique model property is given by [CW95, Thm. 4.5].

The inductive step: suppose that theorem holds for $M\Sigma_{k-1}$, $k \geq 1$.

By Lemma 5.2, any two models $\mathcal{H}_k^m = \langle \mathcal{I}_k^m, V_k^m \rangle$ and $\mathcal{H}_k'^m = \langle \mathcal{I}_k'^m, V_k'^m \rangle$ of $M\Sigma_k$ must coincide respectively on a pair of models $\mathcal{H}_{k-1}^m$ and $\mathcal{H}_{k-1}'^m$ of $M\Sigma_{k-1}$ as in Lemma 5.2(b). Further, by the induction hypothesis, there exists an isomorphism $h : \mathbf{D} \longrightarrow \mathbf{D}'$ such that

$$\mathcal{H}_{k-1}^m \overset{h}{\cong} \mathcal{H}_{k-1}'^m.$$

Since both $\mathcal{H}_k^m$ and $\mathcal{H}_k'^m$ satisfy $AXM^{obs}(\mathcal{L}_k^m)$, the following two sets of equations hold:

$$I_k^m(U(k-1)) \overset{U(k-1)=u^p}{=} I_k^m(u^p) \overset{\text{Lemma 5.2}}{=} I_{k-1}^m(u^p) \overset{h}{=} h^{-1}(I'^m_{k-1}(u^p))$$

and

$$I_k^m(Y(k)) \overset{Y(k)=y^q}{=} I_k^m(y^q) \overset{\text{Lemma 5.2}}{=} I_{k-1}^m(y^q) \overset{h}{=} h^{-1}(I_{k-1}^m(y^q)).$$

We define a bijective mapping h^e from $\mathcal{H}_k^m$ to $\mathcal{H}_k'^m$ by extending h as follows:

$$h^e(I_k^m(U(k-1))) = h(I_k^m(u^p)), \qquad h^e(I_k^m(Y(k))) = h(I_k^m(y^q)).$$

It is now straightforward to prove that h^e is a homomorphism from $\mathcal{H}_k^m$ to $\mathcal{H}_k'^m$. Since, first (suppressing subscripts and superscripts on I_k^m for simplicity),

$$\begin{aligned}
h^e(I(\Phi(x^i, U(k-1))) &= h^e(I(\Phi(x^i, u^p))), \qquad \Phi(x^i, U(k-1)) = \Phi(x^i, u^p),\\
&= h(I(\Phi(x^i, u^p))), \qquad h^e|_{\mathcal{L}_0} = h|_{\mathcal{L}_0},\\
&= I'(\Phi)(h(I(x^i)), h(I(u^p))), \qquad h \text{ is a homomorphism},\\
&= I'(\Phi)(h^e(I(x^i)), h^e(I(U(k-1)))),\\
&\quad \text{definition of } h^e,\ h^e(I(U(k-1))) = h^e(I(u^p)).
\end{aligned}$$

Similarly, second, we have $h^e(I(\eta(x^i))) = I'(\eta)(I'(x^i))$. Hence the dynamic axioms have isomorphic interpretations and, hence, so do the reachability axioms. Further, let $E_k(x^i)$ be the abbreviation of the left side of the member of $AXM^{est}(\mathcal{L}_k^m)$ associated with $CSE_k(x^i)$. Then

$$\begin{aligned} V_k^m(CSE_k(x^i)) &= V_k^m(E_k(x^i)), \quad E_k(x^i) \longleftrightarrow CSE_k(x^i), \\ &= V_{k-1}^m(E_k(x^i)), \quad \text{Lemma 5.2}, \\ &= V'^m_{k-1}(E_k(x^i)), \quad \text{inductive hypothesis}, \\ &= V'^m_k(E_k(x^i)), \quad \text{Lemma 5.2}, \\ &= V'^m_k(CSE_k(x^i)), \quad E_k(x^i) \longleftrightarrow CSE_k(x^i). \end{aligned}$$

By the inductive construction of well-formed formulas and by the recursive definitions of V'^m_k and V_k^m, it follows that $V'^m_k(F) = V_k^m(F)$ for all $F \in WFF(\mathcal{L}_k^m)$. Hence

$$\mathcal{H}_k^m \overset{h'}{\cong} \mathcal{H}'^m_k. \qquad \square$$

From the above, we have the following theorem.

THEOREM 5.3 (decidable theoremhood [CW95, W91]). *For all $k \geq 0$, the logical theory MTh_k generated by $M\Sigma_k$ for $\mathcal{M}$ together with the observed input-output pair $(\mathbf{u}^p, \mathbf{y}^q)$ is decidable.* $\square$

We observe that we do not have the nesting property for the Markovian fragment systems since there exist some members of $M\Sigma_{k-1}$ that are not theorems of $M\Sigma_k$. For example, $K(M\Sigma_{k-2}) \not\subset MTh_k$.

6. Control reasoning power of $M\Sigma_k$. In this section, we shall prove that, as long as conditional control formulas are written in $\mathcal{L}_k^m$, there is no difference between the trajectories of identical machines which use a full COCOLOG controller and those that use a Markovian fragment controller to decide which control input shall be applied at each instant of time. In the notation that we have established, we shall prove that if up to any instant k the trajectories of two copies of $\mathcal{M}$ have been identical, then for a formula $F \in WFF(\mathcal{L}_k^m)$, the following holds:

$$M\Sigma_k \vdash F \Longleftrightarrow \Sigma_k \vdash F.$$

Let the formula $D^p(\mathcal{L}_k^m)$ be the conditional formula associated with control action $\mathbf{u}^p$ at time instant k. Assume that one control system carries the axiom set Σ_k, and the other, $M\Sigma_k$. Then the implication of the equivalence above is that the same control decision will be produced by two controllers at the instant k, that is to say,

$$M\Sigma_k \vdash D^p(\mathcal{L}_k^m) \Longleftrightarrow \Sigma_k \vdash D^p(\mathcal{L}_k^m).$$

If we assume that at the instant k both systems are in the same internal state, then the application of the same input takes both systems into the same subsequent state and both emit the same output. Hence the same observation axioms will be entered into the new axiom sets of each control system, and the scenario repeats itself.

To formalize the analysis of this situation we adopt the following hypothesis.

Basic Hypothesis (BH). (1) For all $k \geq 1$, let $\Sigma(o_1^k)$ and $M\Sigma(o'^k_1)$ denote, respectively, the full (COCOLOG) and Markovian fragment axioms systems for two machines, each identical to the given machine $\mathcal{M}$, generating the sequences $\mathbf{o}_1^k$ and $\mathbf{o}'^k_1$ respectively. The two copies of $\mathcal{M}$ are assumed to be in the same state at time $k = 0$.

(2) Furthermore, at each time instant $k \geq 1$, and for all p, $1 \leq p \leq R$, the conditional control formula $D^p(\mathcal{L}_k)$, associated with u^p in the set of formulas $CCR(\mathcal{L}_k)$ attached to $\Sigma_k(o_1^k)$, is identical to $D^p(\mathcal{L}_k^m)$, the conditional control formula associated with u^p in the set of formulas $CCR(\mathcal{L}_k^m)$, attached to $M\Sigma_k(o'^k_1)$, i.e.,

$$D^p(\mathcal{L}_k) = D^p(\mathcal{L}_k^m), \qquad 1 \leq p \leq R. \qquad \square$$

Henceforce, when the context makes the meaning clear, we will write Σ_k (respectively, $M\Sigma_k$) for $\Sigma(o_1^k)$ (respectively, $M\Sigma(o'^k_1)$).

In order to establish our main result, we need the concept of a reduct, which is defined below.

DEFINITION 6.1. *Let $\mathcal{L}$ and $\mathcal{L}'$ be two languages with $S(\mathcal{L}) \subset S(\mathcal{L}')$, and let $\mathcal{H} = \langle \mathcal{I}, V \rangle$ and $\mathcal{H}' = \langle \mathcal{I}', V' \rangle$ be structures for $\mathcal{L}$ and $\mathcal{L}'$, respectively. $\mathcal{H}$ is called a* reduct *of $\mathcal{H}'$ if and only if $\mathbf{D} = \mathbf{D}'$ and V and V' agree on $\mathcal{L}$. In this case, $\mathcal{H}'$ is called an* extension *of $\mathcal{H}$, and we write $\mathcal{H} = \mathcal{H}'|_{\mathcal{L}}$.* $\square$

For a reduct as defined above, the following lemma is well known.

LEMMA 6.1. *For $F \in WFF(\mathcal{L}) \subset WFF(\mathcal{L}')$,*

$$\mathcal{H} \models F \Longleftrightarrow \mathcal{H}' \models F. \qquad \square$$

It is obvious from Lemma 6.1 that $\mathcal{H}_{k-1} = \mathcal{H}_k|_{\mathcal{L}_{k-1}}$.

We now state the main theorem.

THEOREM 6.1. *Let* BH *hold for $\langle \Sigma(o_1^k), CCR(\mathcal{L}_k) \rangle$ and $\langle M\Sigma(o'^k_1), CCR(\mathcal{L}_k^m) \rangle$ with $k \geq 0$. Then for $F \in WFF(\mathcal{L}_k^m)$, we have*

(I) $\Sigma_k \models F \Longleftrightarrow M\Sigma_k \models F$,

(II) $\Sigma_k \vdash F \Longleftrightarrow M\Sigma_k \vdash F$,

(III) $\mathbf{o}_1^k = \mathbf{o}'^k_1$. $\square$

Part (I) states that a formula in $WFF(\mathcal{L}_k^m)$ is a logical consequence of the COCOLOG axiom set if and only if it is a consequence of corresponding fragment axiom set. Part (II) indicates that any formula in $WFF(\mathcal{L}_k^m)$ is provable with respect to the full COCOLOG axiom set if and only if it is provable with respect to the corresponding Markovian fragment axiom set; in other words, $MTh_k = WFF(\mathcal{L}_k^m) \cap Th_k$. Part (III) states that the control input and output sequences generated respectively by the COCOLOG controller and the Markovian fragment controller are identical at each time instant; hence the two closed-loop systems have the same behavior. A schematic representation in terms of both logic controllers is given in Figure 3.

Proof. The proof of the theorem is by induction on k.

Base step: We observe that Theorem 6.1 holds for $k = 0, 1$, since $M\Sigma_0 = \Sigma_0$ and $M\Sigma_1 = \Sigma_1$, and hence the selected input $\mathbf{u}^p$ is the same for both systems. Next, by BH, both copies of the controlled machine are in the same initial state and hence generate the identical observed output $\mathbf{y}^q \in \mathbf{Y}$ at the instant $k = 1$.

Inductive hypothesis: For all $F \in WFF(\mathcal{L}_{k-1}^m)$, $k \geq 1$,

IH1$_{k-1}$ $\Sigma_{k-1} \models F \Longleftrightarrow M\Sigma_{k-1} \models F$,

IH2$_{k-1}$ $\Sigma_{k-1} \vdash F \Longleftrightarrow M\Sigma_{k-1} \vdash F$,

IH3$_{k-1}$ $\mathbf{o}_1^{k-1} = \mathbf{o}'^{k-1}_1$.

Inductive step: Let BH, IH1$_{k-1}$, IH2$_{k-1}$, and IH3$_{k-1}$ hold. Then for $F \in WFF(\mathcal{L}_k^m)$, the following hold:

$$\Sigma_k \models F \Longleftrightarrow M\Sigma_k \models F, \tag{14}$$

$$\Sigma_k \vdash F \Longleftrightarrow M\Sigma_k \vdash F, \tag{15}$$

$$\mathbf{o}_1^k = \mathbf{o}'^k_1. \tag{16}$$

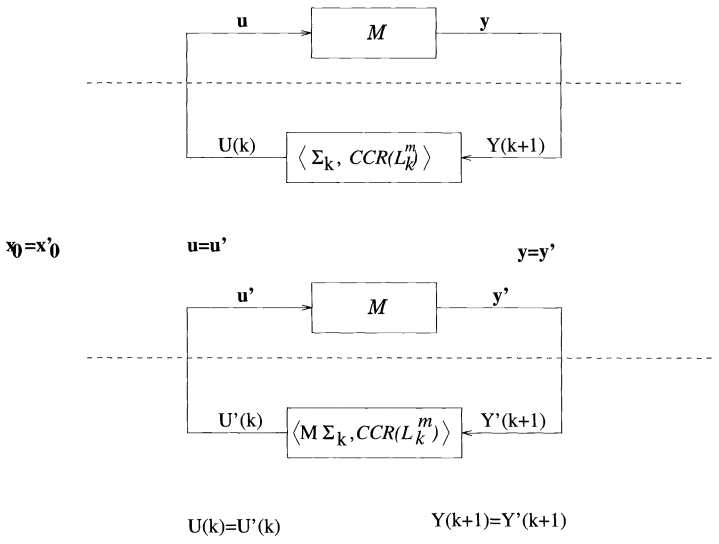


FIG. 3.

The proof of the inductive step is based upon Lemma 6.2, which states that if $\mathcal{H}^m_{k-1}$ is a reduct of $\mathcal{H}_{k-1}$, then $\mathcal{H}^m_k$ is a reduct of $\mathcal{H}_k$.

To establish the inductive step, we observe that statement (14) is an immediate consequence of Lemma 6.2. To prove (15), consider any $F \in WFF(\mathcal{L}^m_k)$. Then

$$\begin{aligned}
& \Sigma_k \vdash F \\
\iff\ & \Sigma_k \models F \quad \text{(completeness of first-order theory } \Sigma_k\text{)} \\
\iff\ & \mathcal{H}_k \models F \quad \text{(unique model property of } \Sigma_k\text{)} \\
\iff\ & \mathcal{H}^m_k \models F \quad \text{(Lemma 6.2)} \\
\iff\ & M\Sigma_k \models F \quad \text{(unique model property of } M\Sigma_k\text{)} \\
\iff\ & M\Sigma_k \vdash F \quad \text{(completeness of first-order theory } M\Sigma_k\text{)}.
\end{aligned}$$

Finally, (16) follows from (15), the identical definitions of the set of control rules $CCR(\mathcal{L}^m_{k-1})$ in each feedback control specification, and the fact that the internal state of each machine is identical at the instant $k-1$. Hence we have established Theorem 6.1 since (14), (15), and (16) of the inductive step have been shown to be the case. □

We now present the lemma.

LEMMA 6.2. *For $k \geq 0$, let $\mathcal{H}^m_k$ be the model defined in Definition* 5.1. *Then* BH, IH1_{k-1}, IH2_{k-1}, IH3_{k-1}, *and $\mathcal{H}^m_{k-1} = \mathcal{H}_{k-1}|_{\mathcal{L}^m_{k-1}}$ imply that $\mathcal{H}^m_k$ is a reduct of $\mathcal{H}_k$, i.e.,*

$$\mathcal{H}^m_k = \mathcal{H}_k|_{\mathcal{L}^m_k}.$$

Proof. We show that the model of the fragment axiom set $M\Sigma_k$ is a reduct of the model of the full-version counterpart Σ_k. First, we note that $AXM^{est}(\mathcal{L}^m_k) = AXM^{est}(\mathcal{L}_k)$. Next, by IH2_{k-1}, for any $F \in WFF(\mathcal{L}^m_{k-1} \cap \mathcal{L}^m_k)$,

$$M\Sigma_{k-1} \vdash F \iff \Sigma_{k-1} \vdash F.$$

Now, by the first part of BH, the (exclusive and exhaustive) conditional control rules attached to Σ_{k-1} and $M\Sigma_{k-1}$ are identical and are both denoted by $CCR(\mathcal{L}^m_{k-1})$. It follows from IH2_{k-1}

that the unique conditional control formula $D^p(\mathcal{L}^m_{k-1})$, $p \in \{1, \ldots, R\}$, which is provable from $M\Sigma_{k-1}$, is also the unique such formula provable from Σ_{k-1}. Hence the same control actions are applied to each system at the instant $k-1$, and by $IH3_{k-1}$, the same holds for the instants j, $1 \leq j \leq k-2$. Further, by BH, the machines controlled by $M\Sigma_{k-1}$ and Σ_{k-1} are identical, and the initial states are same. So the internal state and output sequences must be the same up to the subsequent time instant k. Consequently the same observation axiom sets $AXM^{obs}(\mathcal{L}^m_k)$ and $AXM^{obs}(\mathcal{L}_k)$ are supplied to $M\Sigma_k$ and Σ_k, i.e., $AXM^{obs}(\mathcal{L}_k) = AXM^{obs}(\mathcal{L}^m_k)$, and we have

$$\mathcal{H}_k \models Eq(U(k-1), u^p) \Longleftrightarrow \mathcal{H}^m_k \models Eq(U(k-1), u^p)$$

and

$$\mathcal{H}_k \models Eq(Y(k), y^q) \Longleftrightarrow \mathcal{H}^m_k \models Eq(Y(k), y^q).$$

Further,

$$\begin{aligned}
&\mathcal{H}_k \models CSE_{k-1}(x^i) \\
\Longleftrightarrow &\mathcal{H}_{k-1} \models CSE_{k-1}(x^i) \quad (\mathcal{H}_{k-1} \text{ is a reduct of } \mathcal{H}_k \text{ and Lemma 5.1}) \\
\Longleftrightarrow &\mathcal{H}^m_{k-1} \models CSE_{k-1}(x^i) \quad (\text{inductive hypothesis}) \\
\Longleftrightarrow &\mathcal{H}^m_k \models CSE_{k-1}(x^i) \quad (\text{Theorem 5.2}).
\end{aligned}$$

Since, in addition, $AXM^{est}(\mathcal{L}_k) = AXM^{est}(\mathcal{L}^m_k)$, the result follows from

$$\mathcal{H}^m_k \models CSE_k(x^i) \Longleftrightarrow \mathcal{H}_k \models CSE_k(x^i), \quad 1 \leq i \leq N. \qquad \square$$

As an immediate application of Theorem 6.1, we have the following corollaries.

COROLLARY 6.1. *Let* BH *hold for* $\langle \Sigma_k, CCR(\mathcal{L}_k)\rangle$ *and* $\langle M\Sigma_k, CCR(\mathcal{L}^m_k)\rangle$, $k \geq 1$. *Then*

$$\Sigma_k \vdash CSE_k(x^i) \Longleftrightarrow M\Sigma_k \vdash CSE_k(x^i).$$

Proof. Take $F = CSE_k(x^i)$, and apply Theorem 6.1. $\square$

COROLLARY 6.2. *Let* BH *hold for* $\langle \Sigma_k, CCR(\mathcal{L}_k)\rangle$ *and* $\langle M\Sigma_k, CCR(\mathcal{L}^m_k)\rangle$, $k \geq 0$. *Then for all* $k \geq 0$,

$$Eq(U(k), u^p) \in M\Sigma_{k+1} \Longleftrightarrow Eq(U(k), u^p) \in \Sigma_{k+1}.$$

Proof. Let F be $D^p(\mathcal{L}^m_k)$, and apply Theorem 6.1. $\square$

From Corollary 6.2, we conclude that $AXM^{obs}(\mathcal{L}_{k+1}) = AXM^{obs}(\mathcal{L}^m_{k+1})$, which means that MTh_k and Th_k will be incremented by the same observation axioms.

7. Analysis of MTh_k proof sequences. In this section, we show that shorter proof sequences can be obtained in $M\Sigma_k$ than in Σ_k, $k \geq 1$, for theorems that involve the state estimate predicate. These kinds of formulas are typically used to form conditional control rules. It should be noted that the analysis in this section is proof theoretic and hence does not involve semantic considerations.

Example 7.1. Suppose that Machine 1 generates the observation sequence

$$(\emptyset, \mathbf{y}^1) \quad (\mathbf{u}^1, \mathbf{y}^2) \quad (\mathbf{u}^2, \mathbf{y}^1).$$

Then $M\Sigma_3 \vdash CSE_3(x^4)$ and one of the proof sequences demonstrating $CSE_3(x^4)$ is the following:

(1) $Eq(\Phi(x^1, u^1), x^2)$ AXM^{dyn},
(2) $Eq(\Phi(x^2, u^2), x^4)$ AXM^{dyn},
(3) $Eq(\eta(x^1), y^1)$ AXM^{out},
(4) $Eq(\eta(x^2), y^2)$ AXM^{out},
(5) $Eq(\eta(x^4), y^1)$ AXM^{out},
(6) $CSE_2(x^2)$ $K(M\Sigma_2)$,
(7) $Eq(U(2), u^2)$ $AXM^{obs}(\mathcal{L}_3^m)$,
(8) $Eq(Y(3), y^1)$ $AXM^{obs}(\mathcal{L}_3^m)$,
(9) $\exists x, CSE_2(x) \wedge Eq(\Phi(x, U(2)), x^4) \wedge Eq(\eta(x^4), Y(3)) \rightarrow CSE_3(x^4)$ $AXM^{est}(\mathcal{L}_3^m)$,
(10) $CSE_2(x^2) \wedge Eq(\Phi(x^2, U(2)), x^4) \wedge Eq(\eta(x^4), Y(3))$ (2), (5), (6), (7), (8), $\wedge$–rule,
(11) $\exists x, CSE_2(x) \wedge Eq(\Phi(x, U(2)), x^4) \wedge Eq(\eta(x^4), Y(3))$ (11), $\exists$–rule,
(12) $CSE_3(x^4)$ (9), (11), MP. □

Note that the recursive chain of the Σ_3 proof in Example 2.2 is broken at (6), where $CSE_2(x^2)$ is treated as an axiom and hence needs no verification via proof. By the definition of a proof sequence with respect to Σ_k, formulas in $K(M\Sigma_{k-1})$ (see §2) cannot appear in a Σ_k-proof list $\mathcal{P}$ without being the result of deductions using previous lines in $\mathcal{P}$. This is because the formulas in $K(M\Sigma_{k-1})$ are not logical axioms, nor are they members of Σ_k.

The theorem below formally states how an $M\Sigma_k$-proof can be extended to a Σ_k-proof by extending all $K(M\Sigma_{k-1})$-lines in a $M\Sigma_k$-proof sequence.

THEOREM 7.1. *Suppose that $F \in MTh_k$, and let $\mathcal{P}$ be an $M\Sigma_k$-proof of F. Then there exists a Σ_k-proof $\mathcal{P}'$ of F such that $|\mathcal{P}'| \geq |\mathcal{P}|$.*

Proof. Let $\mathcal{P} \underline{\Delta} \{L_1, L_2, \ldots, L_n\}$ be an $M\Sigma_k$-proof sequence for F. Then the corresponding proof sequence $\mathcal{P}'$ can be constructed as follows. By definition, each of the lines L_i is either a member of $M\Sigma_k$ or a formula deduced from its predecessors in the sequence. Let L_i be the last line in $\mathcal{P}$ such that $L_i \in M\Sigma_k$. Then there are two cases to consider: first, $L_i \notin K(M\Sigma_{k-1})$. In this case,

$$L_i \in \Sigma_0 \cup AXM^{obs}(\mathcal{L}_k^m) \cup AXM^{est}(\mathcal{L}_k^m);$$

hence, by Theorem 6.1,

$$L_i \in \Sigma_0 \cup AXM^{obs}(\mathcal{L}_k) \cup AXM^{est}(\mathcal{L}_k).$$

So L_i itself is a Σ_k-line, and in this case, we leave the line unchanged.

Second, $L_i \in K(M\Sigma_{k-1})$, i.e., $L_i = ?CSE_{k-1}(x^j)$ for some $x^j \in X$; then, by the definition of $K(M\Sigma_{k-1})$,

$$M\Sigma_{k-1} \vdash L_i \in WFF(\mathcal{L}_{k-1}^m),$$

and so there exists an $M\Sigma_{k-1}$-proof sequence of L_i, say, $\mathcal{P}_i^{(1)} \underline{\Delta} \{L_1^i, \ldots, L_{q_i}^i\}$ with $L_{q_i}^i = L_i$. Now replace L_i in $\mathcal{P}$ with $\mathcal{P}_i$ to get the new proof sequence for F:

$$\{L_1, \ldots, L_{i-1}, \{\mathcal{P}_i^{(1)}\}, L_{i+1}, \ldots, L_n\}.$$

Note that L_i appears in the last line of $\mathcal{P}_i$, so if L_i is used to deduce any other formula L_j, $j > i$, it can still be invoked after $\mathcal{P}_i$ has been inserted. This procedure is repeated until all $K(M\Sigma_{k-1})$-lines have been replaced by their corresponding $M\Sigma_{k-1}$-proof sequences. Let the resulting sequence be

$$\mathcal{P}^{(1)} \underline{\Delta} \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n\},$$

where

$$\mathcal{P}_i = \begin{cases} L_i & \text{if } L_i \notin K(M\Sigma_{k-1}), \\ \text{a } M\Sigma_{k-1}\text{-proof of } L_i & \text{otherwise.} \end{cases}$$

If $\mathcal{P}_i$ is an inserted sequence and there exist $K(M\Sigma_{k-2})$-lines in this $M\Sigma_{k-1}$-proof, we apply the same procedure to each of them to get $\mathcal{P}^{(2)}$, in which there are no $K(M\Sigma_{k-1})$- nor $K(M\Sigma_{k-2})$-lines. Repeating this procedure a finite number times eventually yields $\mathcal{P}'$, in which each line is in Σ_k, and hence is such that the whole sequence $\mathcal{P}'$ is in Σ_k. Clearly $|\mathcal{P}'| \geq |\mathcal{P}|$, and this establishes the theorem. □

If the above proof itself is viewed as an expansion method for an $M\Sigma_k$-proof, then the proof in the following lemma can be viewed as a technique for shortening a Σ_k-proof.

LEMMA 7.1. *Suppose that* $\Sigma_k \vdash CSE_k(x^i)$. *Then there exists an* $M\Sigma_k$*-proof of* $CSE_k(x^i)$ *which has a fixed length with respect to* k. □

Proof. First we recall from [CW95] that $Eq(\Phi(x, U(k-1)), y)$ is the abbreviation of $Eq(\Phi(x, u^p), y) \wedge Eq(U(k-1), u^p)$ for some $u^p \in U$, and $Eq(\eta(x^i), Y(k))$ is the abbreviation of $Eq(\eta(x^i), y^q) \wedge Eq(Y(k), y^q)$ for some $y^q \in Y$.

Consider any $k > 1$. Suppose $\Sigma_k \vdash CSE_k(x^i)$. Then from the definition of $AXM^{est}(\mathcal{L}_k^m)$ we have

$$\Sigma_k \vdash \exists x, CSE_{k-1}(x) \wedge Eq(\Phi(x, U(k-1)), x^i) \wedge Eq(\eta(x^i), Y(k)).$$

Then, by the logical axioms,

$$\Sigma_k \vdash CSE_{k-1}(x^j) \wedge Eq(\Phi(x^j, U(k-1)), x^i) \wedge Eq(\eta(x^i), Y(k))$$

for some $x^j \in X$. Hence,

$$\Sigma_k \vdash CSE_{k-1}(x^j),$$

and by Corollary 6.1,

$$M\Sigma_k \vdash CSE_{k-1}(x^j).$$

Hence $CSE_{k-1}(x^j) \in K(M\Sigma_{k-1})$. Similarly we have

$$\Sigma_k \vdash Eq(\Phi(x^j, u^p), x^i) \quad \text{and} \quad \Sigma_k \vdash Eq(\eta(x^i), y^q),$$

which implies

$$Eq(\Phi(x^j, u^p), x^i) \in AXM^{dyn}(\mathcal{L}_0) \quad \text{and} \quad Eq(\eta(x^i), y^q) \in AXM^{out}(\mathcal{L}_0).$$

Now we can construct the following proof sequence for $CSE_k(x^i)$ and hence establish the lemma:

(1) $Eq(Y(k), y^q)$ $AXM^{obs}(\mathcal{L}_k^m)$,
(2) $Eq(U(k-1)), u^p)$ $AXM^{obs}(\mathcal{L}_k^m)$,
(3) $Eq(\Phi(x^j, u^p), x^i)$ $AXM^{dyn}(\mathcal{L}_0)$,
(4) $Eq(\eta(x^i), y^q))$ $AXM^{out}(\mathcal{L}_0)$,
(5) $CSE_{k-1}(x^j)$ $K(M\Sigma_{k-1})$,
(6) $Eq(\Phi(x^j, U(k-1)), x^i)$ (2), (3), AXM^{Eq}, MP,
(7) $Eq(\eta(x^i), Y(k))$ (1), (4) AXM^{Eq}, MP,
(8) $CSE_{k-1}(x^j) \wedge Eq(\Phi(x^j, U(k-1)), x^i)$ (5), (6), $\wedge_1$−rule,
(9) $CSE_{k-1}(x^j) \wedge Eq(\Phi(x^j, U(k-1)), x^i) \wedge Eq(\eta(x^i), Y(k))$ (8), (7), $\wedge_1$−rule,
(10) $\exists x, CSE_{k-1}(x) \wedge Eq(\Phi(x, U(k-1)), x^i) \wedge Eq(\eta(x^i), Y(k))$ (9), logical axiom (4), definition of $\exists$,
(11) $\exists x, CSE_{k-1}(x) \wedge Eq(\Phi(x, U(k-1)), x^i) \wedge Eq(\eta(x^i), Y(k)) \leftrightarrow CSE_k(x^i)$ $AXM^{est}(\mathcal{L}_k^m)$,

(12) $\exists x, CSE_{k-1}(x) \wedge Eq(\Phi(x, U(k-1)), x^i) \wedge Eq(\eta(x^i), Y(k)) \rightarrow CSE_k(x^i)$
(11), $\wedge_1$–rule,
(13) $CSE_k(x^i)$ (10), (12), MP. □

Remark. By use of AXM^{Eq}, we can also find a shorter proof for $CSE_k(t)$ with $t \in \text{Term}(\mathcal{L}_k^m)$ if $\Sigma_k \vdash CSE_k(t)$.

Remark. Although Lemma 7.1 can be generalized to apply to any $F \in WFF(\mathcal{L}_k^m)$, it is not our intention to give a detailed proof in this paper. In general, for a given $F \in WFF(\mathcal{L}_k^m) \cap Th_k$, which does not contain predicates CSE_k or CSE_{k-1}, the shortest Σ_k-proof sequence of such formula is also an $M\Sigma_k$-proof sequence; hence we may not be able to find a shorter proof with respect to $M\Sigma_k$. However, if F contains CSE_k or CSE_{k-1}, then the shortest $M\Sigma_k$-proof of F will be shorter than the shortest Σ_k-proof. Moreover, since the length of proof for the formula $CSE_k(x^i)$ with respect Σ_k depends upon the time index k while that with respect to $M\Sigma_k$ does not, the effect of invoking axioms in $K(M\Sigma_{k-1})$ on the proof of F will become more marked as k increases. In Appendix 2, we present two pairs of examples generated on a computer by theorem-proving software to illustrate such a phenomenon.

Example 7.2. For the seven-state machine presented in Example 2.1, suppose that the target state is x^3. The first line of a $CCR(\mathcal{L}_2)$ associated with u^1 is

$$\textbf{if } \exists x \exists l, CSE_2(x) \wedge Rbl(\Phi(x, u^1), x^3, l), \quad \textbf{then } \quad Eq(U(2), u^1).$$

Given the observation sequence $\{(\emptyset, \mathbf{y}^1), (\mathbf{u}^1, \mathbf{y}^2)\}$ of Example 2.2, we shall obtain the control action u^1 at the instant 2 by proving the antecedent clause of the **if-then** rule $CCR(\mathcal{L}_2)$. The first part of the proof of the conditional control condition is the proof of $CSE_2(x^2)$ in Th_2. For the sake of simplicity, we omit some rudimentary deductions from the two proofs below; these deductions can be carried out as both Σ_k and $M\Sigma_k$ proofs and hence are not relevant in comparisons of the length of Σ_k and $M\Sigma_k$ proofs.

(1) $Eq(Y(1), y^1)$ $AXM^{obs}(\mathcal{L}_1)$,
(2) $Eq(Y(2), y^2)$ $AXM^{obs}(\mathcal{L}_2)$,
(3) $Eq(U(1), u^1)$ $AXM^{obs}(\mathcal{L}_2)$,
(4) $Eq(\eta(x^1), Y(1)) \leftrightarrow CSE_1(x^1)$ $AXM^{est}(\mathcal{L}_1)$,
(5) $Eq(\eta(x^1), y^1)$ AXM^{out},
(6) $Eq(\eta(x^1), y^1) \rightarrow CSE_1(x^1)$ (4), $\wedge_2$–rule,
(7) $CSE_1(x^1)$ (5), (6), MP,
(8) $\exists x CSE_1(x) \wedge Eq(\Phi(x, U(1)), x^2) \wedge Eq(\eta(x^2), Y(2)) \leftrightarrow CSE_2(x^2)$
$AXM^{est}(\mathcal{L}_2)$,
(9) $\exists x CSE_1(x) \wedge Eq(\Phi(x, U(1)), x^2) \wedge Eq(\eta(x^2), Y(2)) \rightarrow CSE_2(x^2)$
(8), $\wedge_2$–rule,
(10) $CSE_1(x^1) \wedge Eq(\Phi(x^1, U(1)), x^2) \wedge Eq(\eta(x^2), Y(2))$ (6), AXM^{dyn},
AXM^{out}, $\wedge_1$–rule,
(11) $\exists x CSE_1(x) \wedge Eq(\Phi(x, U(1)), x^2) \wedge Eq(\eta(x^2), Y(2))$ (10), definition of $\exists$,
(12) $CSE_2(x^2)$ (9), (11), MP. □

The second part is the proof of $Rbl(\Phi(x^2, u^1), x^3, 1)$:

(13) $\forall x \forall y, (\exists u Eq(\Phi(x, u), y) \rightarrow Rbl(x, y, 1))$ AXM^{Rbl},
(14) $Eq(\Phi(x^1, u^2), x^3)$ AXM^{dyn},
(15) $Eq(\Phi(x^1, u^2), x^3) \rightarrow Rbl(x^1, x^3, 1)$ logic axiom (4), (13), MP,
(16) $Rbl(x^1, x^3, 1)$ (14), (15), MP,
(17) $Eq(\Phi(x^2, u^1), x^2)$ AXM^{dyn},
(18) $Eq(\Phi(x^2, u^1), x^1) \rightarrow (Rbl(x^1, x^3, 1) \rightarrow Rbl(\Phi(x^2, u^1), x^3, 1))$
logic axiom (4), AXM^{Eq},
(19) $Rbl(x^1, x^3, 1) \rightarrow Rbl(\Phi(x^2, u^1), x^3, 1)$ (17), (18), MP,
(20) $Rbl(\Phi(x^2, u^1), x^3, 1)$ (16), (19), MP.

Finally,

(21) $CSE_2(x^2) \wedge Rbl(\Phi(x^2, u^1), x^3, 1)$ (11), (20), $\wedge_1$−rule,

(22) $\exists x \exists l, CSE_2(x) \wedge Rbl(\Phi(x, u^i), x^3, l-1)$ definition of $\exists$, (21).

The extralogical control rules $CCR(L_2)$ then give $Eq(U(2), u^1)$. So $Eq(U(2), u^1) \in \Sigma_3$.

The proof of the control condition for $Eq(U(2), u^1)$ with respect to $M\Sigma_2$ can be obtained by appending lines (13) to (22) above to the initial sequence of lines shown below, which constitutes a proof of $CSE_2(x^2)$ in MTh_2. The length of the proof is then 18 steps instead of 22 as in the previous proof with respect to Th_2.

(1) $Eq(Y(2), y^2)$ $AXM^{obs}(\mathcal{L}_2)$,

(2) $Eq(U(1), u^1)$ $AXM^{obs}(\mathcal{L}_2)$,

(3) $CSE_1(x^1)$ $K(M\Sigma_1)$,

(4) $\exists x CSE_1(x) \wedge Eq(\Phi(x, U(1)), x^2) \wedge Eq(\eta(x^2), Y(2)) \rightarrow CSE_2(x^2)$ $AXM^{est}(\mathcal{L}_2)$,

(5) $\exists CSE_1(x^1) \wedge Eq(\Phi(x^1, U(1)), x^2) \wedge Eq(\eta(x^2), Y(2)) \rightarrow CSE_2(x^2)$ $AXM^{est}(\mathcal{L}_2)$,

(6) $\exists CSE_1(x^1) \wedge Eq(\Phi(x^1, U(1)), x^2) \wedge Eq(\eta(x^2), Y(2))$ (1), (2), AXM^{out}, $\wedge_1$−rule,

(7) $CSE_1(x^1) \wedge Eq(\Phi(x^1, U(1)), x^2) \wedge Eq(\eta(x^2), Y(2)) \rightarrow CSE_2(x^2)$ $\exists$ rule, (6),

(8) $CSE_2(x^2)$ (6), (7), MP. □

8. Appendix 1: A complete description of Σ_0. We list all the axioms and other components of Σ_0 except for the particular finite machine axioms, which would be given separately for any specific controlled machine $\mathcal{M}$, and the reachability axioms that appear in §2.

$AXM^{log}(\mathcal{L}_0)$:

(1) $A \rightarrow (B \rightarrow A)$.

(2) $(A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))$.

(3) $(\neg B \rightarrow \neg A) \rightarrow ((\neg B \rightarrow A) \rightarrow B)$.

(4) $\forall x A(x) \rightarrow A(t)$, $t \in \text{Term}(\mathcal{L}_0)$.

(5) $\forall x(A \rightarrow B) \rightarrow (A \rightarrow \forall x B)$, x not free in A.

$AXM^{Eq}(\mathcal{L}_0)$:

(1) $\forall x Eq(x, x)$.

(2) $\forall x \forall y Eq(x, y) \rightarrow Eq(y, x)$.

(3) $\forall x \forall y \forall z Eq(x, y) \wedge Eq(y, z) \rightarrow Eq(x, z)$.

(4) $\forall x \forall y Eq(x, y) \rightarrow Eq(\eta(x), \eta(y))$.

(5) $\forall x \forall y \forall u(Eq(x, y) \rightarrow Eq(\Phi(x, u), \Phi(y, u)))$.

(6) $\forall u \forall v \forall u(Eq(u, v) \rightarrow Eq(\Phi(x, u), \Phi(x, v)))$.

(7) $\forall x \forall y \forall z(Eq(x, y) \rightarrow Eq(+_{K(N)}(x, z), +_{K(N)}(y, z)))$.

(8) $\forall x \forall y \forall z(Eq(x, y) \rightarrow Eq(+_{K(N)}(z, x), +_{K(N)}(z, y)))$.

(9) $\forall x \forall y \forall z(Eq(x, y) \rightarrow Eq(-_{K(N)}(x, z), -_{K(N)}(y, z)))$.

(10) $\forall x \forall y \forall z(Eq(x, y) \rightarrow Eq(-_{K(N)}(z, x), -_{K(N)}(z, y)))$.

(11) $\forall x_1 \forall x_2 \forall x_3 \forall y_1 \forall y_2 \forall y_3(Eq(x_1, y_1) \wedge Eq(x_2, y_2) \wedge Eq(x_3, y_3) \rightarrow (Rbl((x_1, x_2, x_3) \rightarrow Rbl(y_1, y_2, y_3))))$.

(12) $\forall x_1 \forall x_2 \forall y_1 \forall y_2(Eq(x_1, y_1) \wedge Eq(x_2, y_2) \rightarrow (Eq(x_1, x_2) \rightarrow Eq(y_1, y_2)))$.

$AXM^{subs}(\mathcal{L}_k)$:

$\forall x \forall y(Eq(x, y) \rightarrow (CSE_k(x) \rightarrow CSE_k(y)))$.

$AXM^{arith}(\mathcal{L}_0)$: For the arithmetic functions $+_{K(N)}$, $-_{K(N)}$, specified via

$$a +_{K(N)} b = \begin{cases} a + b & \text{if } a + b \leq K(N), \\ K(N) + 1 & \text{if } a + b > K(N) \end{cases}$$

and

$$a -_{K(N)} b = \begin{cases} a - b & \text{if } a + b \geq K(N), \\ 0 & \text{if } a + b < K(N), \end{cases}$$

the axioms are

$$Eq(0 +_{K(N)} 0, 0),\ Eq(0 +_{K(N)} 1, 1), \ldots, Eq(0 +_{K(N)} (N), K(N)),$$

$$Eq(1 +_{K(N)} 1, 2), \ldots, Eq(1 +_{K(N)} K(N), K(N) + 1),$$

$$Eq(2 +_{K(N)} 2, 4), \ldots, Eq(2 +_{K(N)} K(N), K(N) + 1),$$

$$\vdots$$

$$Eq(K(N) +_{K(N)} K(N), K(N) + 1).$$

and

$$Eq(K(N) -_{K(N)} 0, K(N)), \ldots, Eq(K(N) -_{K(N)} K(N), 0),$$

$$Eq((K(N) - 1) -_{K(N)} K(N), 0), \ldots, Eq((K(N) - 1) -_{K(N)} 0, K(N) - 1),$$

$$\vdots$$

$$Eq(0 -_{K(N)} 0, 0).$$

$AXM^{size}(\mathcal{L}_0)$:

(1) For the state space:

$\neg Eq(x^1, x^2) \wedge \neg Eq(x^1, x^3) \wedge \cdots \wedge \neg Eq(x^1, x^N)$
$\wedge \neg Eq(x^2, x^3) \wedge \cdots \wedge \neg Eq(x^2, x^N)$
$\wedge \cdots$
$\wedge \neg Eq(x^{N-1}, x^N).$
$\forall x(\vee_{i=1}^{N} Eq(x, x^i))$

(2) For the input space:

$\neg Eq(u^1, u^2) \wedge \neg Eq(u^1, u^3) \wedge \cdots \wedge \neg Eq(u^1, u^R)$
$\wedge \neg Eq(u^2, u^3) \wedge \cdots \wedge \neg Eq(u^2, u^R)$
$\wedge \cdots$
$\wedge \neg Eq(u^{R-1}, u^R).$
$\forall u(\vee_{i=1}^{R} Eq(u, u^i))$

(3) For the output space:

$\neg Eq(y^1, y^2) \wedge \neg Eq(y^1, y^3) \wedge \cdots \wedge \neg Eq(y^1, y^M)$
$\wedge \neg Eq(y^2, y^3) \wedge \cdots \wedge \neg Eq(y^2, y^M)$
$\wedge \cdots$
$\wedge \neg Eq(y^{M-1}, y^M).$
$\forall y(\vee_{i=1}^{M} Eq(y, y^i))$

(4) For the integers in $\mathcal{L}_0$:

$\neg Eq(0, 1) \wedge \neg Eq(0, 2) \wedge \cdots \wedge \neg Eq(0, K(N) + 1)$
$\wedge \neg Eq(1, 2) \wedge \cdots \wedge \neg Eq(1, K(N) + 1)$
$\wedge \cdots$
$\wedge \neg Eq(K(N), K(N) + 1).$
$\forall l(\vee_{i=0}^{K(N)+1} Eq(l, i))$

Inference rules.

MP. Let A and B be well-formed formulas. Then

$$\frac{A,\ A \rightarrow B}{B}.$$

Generalization. Let A be a well-formed formula. Then

$$\frac{A}{\forall x A}.$$

Invoking the deduction theorem of first-order theory, we add the following rules:

$\bigwedge_1$-*rule.*

$$\frac{A, B}{A \wedge B}.$$

$\bigwedge_2$-*rule.*

$$\frac{A \wedge B}{A}.$$

9. Appendix 2: Two examples of ATP in Σ_k and $M\Sigma_k$. Using the seven-state machine of Example 2.1 and the observation sequences $\mathbf{o}_1^3$ and $\mathbf{o}_1^4$ displayed below, we may present a comparison of the proofs of two COCOLOG theorems in a full COCOLOG theory and its fragmentary counterpart. The first pair of proofs are for the theorem $CSE_3(x^4)$ in, respectively, Σ_3 and $M\Sigma_3$, given the observation sequence $\mathbf{o}_1^3$. The second pair of proofs are for the theorem $CSE_4(x^6)$ in, respectively, Σ_4 and $M\Sigma_4$, given the observation sequence $\mathbf{o}_1^4$. We recall that $CSE_3(x^4)$ (respectively, $CSE_4(x^6)$) means that x^4 (respectively, x^6) lies in the state estimate set at time instant $k = 3$ (respectively, $k = 4$).

The proof methodology used in the examples below is based upon the resolution principle, where the axiom set is written in clausal form. The reader may refer to [CL73] for full description of all terms and concepts from ATP used here. The proofs were generated by use of the general-purpose theorem-proving software GTP that was developed at McGill University by M. Newborn et al. [N89], in conjunction with the so-called function evaluation facility FE, which was proposed by S. Wang and P. E. Caines [WC92] and implemented by Q. X. Yu and S. Wang.

Each proof listing consists of

(1) a modified axiom set, i.e., a modified version of a full COCOLOG axiom set or a Markovian fragment axiom set. In these proof listings, phi7 stands for the transition function Φ. The FE facility of the GTP program (see [WC92]) replaces the axiomatic definition of phi7, and the predicate eta replaces the output function η. For technical reasons concerning the ATP proof procedure, we use the predicate $eta(\cdot, \cdot)$ for the specification of the output function, e.g., eta(x1,y1) replaces the formula $Eq(\eta(x^1), y^1)$. Similarly, output2(y1) (respectively, input1(u1)) replaces the formula $Eq(Y(2), y^1)$ (respectively, $Eq(U(1), u^1)$), and so on. Each axiom line begins with the capital letter A.

(2) the negated theorem $\neg CSE3(x^4)$ (or $\neg CSE4(x^6)$); such a line begins with the capital letter T.

(3) the refutation path, given by a set of lines, each of which begins with two reference indices giving the two parent clauses that generate the current clause by resolution. For example, (12a,22b) means that the current clause is generated by resolving the first literal of clause 12 and the second literal of clause 22.

(4) at the end of each listing, some performance indices.

The most important observation to be made concerning these examples is that the number of resolutions attempted during the proof with respect to $M\Sigma_3$, based upon the observation sequence $\mathbf{o}_1^3$, is only about 1.5% of that with respect to Σ_3. Further, the addition of only one more observation pair results in a dramatic increase in the number of resolutions used in the proof of $CSE_4(x^6)$ with respect to Σ_4, while the complexity of the proof of this theorem with respect to $M\Sigma_4$ is approximately equal to that of $CSE_3(x^4)$ with respect to $M\Sigma_3$.

Observation sequence $\mathbf{o}_1^3$:

$$(\emptyset, \mathbf{y}^1),\ (\mathbf{u}^1, \mathbf{y}^2),\ (\mathbf{u}^2, \mathbf{y}^1).$$

A resolution tree of Theorem $CSE3(x4)$ with Σ_3.

1: A Rbl(x,x,0)

2: A Rbl(x,z,1) ∨ ¬Eq(phi7(x,y),z)
3: A ¬Rbl(x,y,minus7(z,1)) ∨ ¬Rbl(y,u,1) ∨ Rbl(x,u,z)
4: A Eq(x,x)
5: A ¬Eq(x,y) ∨ Eq(y,x)
6: A ¬Eq(x,y) ∨ ¬Eq(y,z) ∨ Eq(x,z)
7: A eta(x1,y1) 8: A eta(x4,y1)
9: A eta(x2,y2) 10: A eta(x7,y2)
11: A eta(x5,y2) 12: A eta(x3,y3)
13: A eta(x6,y3) 14: A out1(y1)
15: A out2(y2) 16: A input1(u1)
17: A out3(y1) 18: A input2(u2)
19: A ¬eta(y,x) ∨ ¬out1(x) ∨ CSE1(y)
20: A ¬eta(y,x) ∨ ¬out1(x) ∨ ¬CSE1(y)
21: A ¬Eq(phi7(x,y),z) ∨ ¬eta(z,u) ∨ ¬out2(u) ∨ ¬input1(y) ∨ ¬CSE1(x) ∨ CSE2(z)
22: A ¬Eq(phi7(x,y),z) ∨ ¬eta(z,u) ∨ ¬out3(u) ∨ ¬input2(y) ∨ ¬CSE2(x) ∨ CSE3(z)
23: T ¬CSE3(x4)

Refutation path:
24: (23a,22f) [–] ¬Eq(phi7(x,y),x4) ∨ ¬eta(x4,z) ∨ ¬out3(z) ∨ ¬input2(y) ∨ ¬CSE2(x4)
25: (24a,4a) [–] ¬eta(x4,x) ∨ ¬out3(x) ∨ ¬input2(u2) ∨ ¬CSE2(x2)
26: (25a,8a) [–] ¬out3(y1) ∨ ¬input2(u2) ∨ ¬CSE2(x2)
27: (26a,17a) [–] ¬input2(u2) ∨ ¬CSE2(x2)
28: (27a,18a) [–] ¬CSE2(x2)
29: (28a,21f) [–] ¬Eq(phi7(x,y),x2) ∨ ¬eta(x2,z) ∨ ¬out2(z) ∨ ¬input1(y) ∨ ¬CSE1(x)
30: (29e,19c) [–] ¬Eq(phi7(x,y),x2) ∨ ¬eta(x2,z) ∨ ¬eta(x,u) ∨ ¬out1(u) ∨ ¬out2(z)
31: (30a,4a) [–] ¬eta(x2,x) ∨ ¬eta(x1,y) ∨ ¬out1(y) ∨ ¬out2(x) ∨ ¬input1(u1)
32: (31b,7a) [–] ¬eta(x2,x) ∨ ¬out1(y1) ∨ ¬out2(x) ∨ ¬input1(u1)
33: (32a,9a) [–] ¬out1(y1) ∨ ¬out2(y2) ∨ ¬input1(u1)
34: (33a,14a) [–] ¬out2(y2) ∨ ¬input1(u1)
35: (34a,15a) [–] ¬input1(u1)
36: (35a,16a) [–] []

Number of resolutions = 1474.
Depth of resolution tree = 13.

A resolution tree of Theorem $CSE3(x4)$ with $M\Sigma_3$.
1: A Rbl(x,x,0)
2: A Rbl(x,z,1) ∨ ¬Eq(phi7(x,y),z)
3: A ¬Rbl(x,y,minus7(z,1)) ∨ ¬Rbl(y,u,1) ∨ Rbl(x,u,z)
4: A Eq(x,x)
5: A ¬Eq(x,y) ∨ Eq(y,x)
6: A ¬Eq(x,y) ∨ ¬Eq(y,z) ∨ Eq(x,z)
7: A eta(x1,y1) 8: A eta(x4,y1)
9: A eta(x2,y2) 10: A eta(x7,y2)
11: A eta(x5,y2) 12: A eta(x3,y3)
13: A eta(x6,y3) 14: A out3(y1)
15: A input2(u2) 16: A CSE2(x2)
17: A CSE2(x5) 18: A ¬CSE2(x1)
19: A ¬CSE2(x3) 20: A ¬CSE2(x4)
21: A ¬CSE2(x6) 22: A ¬CSE2(x7)
23: A ¬Eq(phi7(x,y),z) ∨ ¬eta(z,u) ∨ ¬out3(u) ∨ ¬input2(y) ∨ ¬CSE2(x) ∨ CSE3(z)
24: T ¬CSE3(x4)

Refutation path:
25: (24a,23f) [–] ¬Eq(phi7(x,y),x4) ∨ ¬eta(x4,z) ∨ ¬out3(z) ∨ ¬input2(y) ∨ ¬CSE2(x)
26: (25a,4a) [–] ¬eta(x4,x) ∨ ¬out3(x) ∨ ¬input2(u2) ∨ ¬CSE2(x2)
27: (26a,8a) [–] ¬out3(y1) ∨ ¬input2(u2) ∨ ¬CSE2(x2)

28: (27a,14a) [–] ¬input2(u2) ∨ ¬CSE2(x2)
29: (28a,15a) [–] ¬CSE2(x2)
30: (29a,16a) [–] []

Number of resolutions = 21.
Depth of resolution tree = 6.
Observation sequence o_1^4:

$$(\emptyset, \mathbf{y}^1),\ (\mathbf{u}^1, \mathbf{y}^2), (\mathbf{u}^2, \mathbf{y}^1), (\mathbf{u}^2, \mathbf{y}^3).$$

A resolution tree of Theorem $CSE4(x6)$ with Σ_4.
1: A Rbl(x,x,0)
2: A Rbl(x,z,1) ∨ ¬Eq(phi7(x,y),z)
3: A ¬Rbl(x,y,minus7(z,1)) ∨ ¬Rbl(y,u,1) ∨ Rbl(x,u,z)
4: A Eq(x,x)
5: A ¬Eq(x,y) ∨ Eq(y,x)
6: A ¬Eq(x,y) ∨ ¬Eq(y,z) ∨ Eq(x,z)
7: A eta(x1,y1) 8: A eta(x4,y1)
9: A eta(x2,y2) 10: A eta(x7,y2)
11: A eta(x5,y2) 12: A eta(x3,y3)
13: A eta(x6,y3) 14: A out1(y1)
15: A out2(y2) 16: A input1(u1)
17: A out3(y1) 18: A input2(u2)
19: A out4(y3) 20: A input3(u2)
21: A ¬eta(y,x) ∨ ¬out1(x) ∨ CSE1(y)
22: A ¬eta(y,x) ∨ ¬out1(x) ∨ ¬CSE1(y)
23: A ¬Eq(phi7(x,y),z) ∨ ¬eta(z,u) ∨ ¬out2(u) ∨ ¬input1(y) ∨ ¬CSE1(x) ∨ CSE2(z)
24: A ¬Eq(phi7(x,y),z) ∨ ¬eta(z,u) ∨ ¬out3(u) ∨ ¬input2(y) ∨ ¬CSE2(x) ∨ CSE3(z)
25: A ¬Eq(phi7(x,y),z) ∨ ¬eta(z,u) ∨ ¬out4(u) ∨ ¬input3(y) ∨ ¬CSE3(x) ∨ CSE4(z)
26: T ¬CSE4(x6)

Refutation path:
27: (26a,25f) [–] ¬Eq(phi7(x,y),x6) ∨ ¬eta(x6,z) ∨ ¬out4(z) ∨ ¬input3(y) ∨ ¬CSE3(x)
28: (27a,4a) [–] ¬eta(x6,x) ∨ ¬out4(x) ∨ ¬input3(u2) ∨ ¬CSE3(x4)
29: (28a,13a) [–] ¬out4(y3) ∨ ¬input3(u2) ∨ ¬CSE3(x4)
30: (29a,19a) [–] ¬input3(u2) ∨ ¬CSE3(x4)
31: (30a,20a) [–] ¬CSE3(x4)
32: (31a,24f) [–] ¬Eq(phi7(x,y),x4) ∨ ¬eta(x4,z) ∨ ¬out3(z) ∨ ¬input2(y) ∨ ¬CSE2(x)
33: (32a,4a) [–] ¬eta(x4,x) ∨ ¬out3(x) ∨ ¬input2(u2) ∨ ¬CSE2(x2)
34: (33a,8a) [–] ¬out3(y1) ∨ ¬input2(u2) ∨ ¬CSE2(x2)
35: (34a,17a) [–] ¬input2(u2) ∨ ¬CSE2(x2)
36: (35a,18a) [–] ¬CSE2(x2)
37: (36a,23f) [–] ¬Eq(phi7(x,y),x2) ∨ ¬eta(x2,z) ∨ ¬out2(z) ∨ ¬input1(y) ∨ ¬CSE1(x)
38: (37e,21c) [–] ¬Eq(phi7(x,y),x2) ∨ ¬eta(x2,z) ∨ ¬eta(x,u) ∨ ¬out1(u) ∨ ¬out2(z)
39: (38a,4a) [–] ¬eta(x2,x) ∨ ¬eta(x1,y) ∨ ¬out1(y) ∨ ¬out2(x) ∨ ¬input1(u1)
40: (39b,7a) [–] ¬eta(x2,x) ∨ ¬out1(y1) ∨ ¬out2(x) ∨ ¬input1(u1)
41: (40a,9a) [–] ¬out1(y1) ∨ ¬out2(y2) ∨ ¬input1(u1)
42: (41a,14a) [–] ¬out2(y2) ∨ ¬input1(u1)
43: (42a,15a) [–] ¬input1(u1)
44: (43a,16a) [–] []

Number of resolutions = 114,694.
Depth of resolution tree = 18.

A resolution tree of Theorem $CSE4(x6)$ with $M\Sigma_4$.
1: A Rbl(x,x,0)
2: A Rbl(x,z,1) ∨ ¬Eq(phi7(x,y),z)

3: A ¬Rbl(x,y,minus7(z,1)) ∨ ¬Rbl(y,u,1) ∨ Rbl(x,u,z)
4: A Eq(x,x)
5: A ¬Eq(x,y) ∨ Eq(y,x)
6: A ¬Eq(x,y) ∨ ¬Eq(y,z) ∨ Eq(x,z)
7: A eta(x1,y1) 8: A eta(x4,y1)
9: A eta(x2,y2)) 10: A eta(x7,y2)
11: A eta(x5,y2) 12: A eta(x3,y3)
13: A eta(x6,y3) 14: A out4(y3)
15: A input3(u2) 16: A CSE3(x4)
17: A ¬CSE3(x1) 18: A ¬CSE3(x2)
19: A ¬CSE3(x3) 20: A ¬CSE3(x5)
21: A ¬CSE3(x6) 22: A ¬CSE3(x7)
23: A ¬Eq(phi7(x,y),z) ∨ ¬eta(z,u) ∨ ¬out4(u) ∨ ¬input3(y) ∨ ¬CSE3(x) ∨ CSE4(z)
24: T ¬CSE4(x6)

Refutation path:
25: (24a,23f) [–] ¬Eq(phi7(x,y),x6) ∨ ¬eta(x6,z) ∨ ¬out4(z) ∨ ¬input3(y) ∨ ¬CSE3(x)
26: (25a,4a) [–] ¬eta(x6,x) ∨ ¬out4(x) ∨ ¬input3(u2) ∨ ¬CSE3(x4)
27: (26a,13a) [–] ¬out4(y3) ∨ ¬input3(u2) ∨ ¬CSE3(x4)
28: (27b,15a) [–] ¬out4(y3) ∨ ¬CSE3(x4)
29: (28b,16a) [–] ¬out4(y3)
30: (29a,14a) [–] []

Number of resolutions = 18.
Depth of resolution tree = 6.

Acknowledgment. The authors gratefully acknowledge the perceptive and constructive comments of the reviewers.

REFERENCES

[C88] P. E. CAINES, *Linear Stochastic Systems*, John Wiley and Sons, New York, 1988.
[CL73] C.-L. CHANG AND R. C.-T. LEE, *Symbolic Logic and Mechanical Theorem Proving*, Academic Press, New York, 1973.
[CMW93] P. E. CAINES, T. MACKLING, AND Y. J. WEI, *Logic control via automatic theorem proving: COCOLOG fragments implemented in Blitzensturm* 5.0, in Proc. American Control Conf., San Francisco, 1993, pp. 1209–1213.
[CW91] P. E. CAINES AND S. WANG, *On a conditional observer and controller logic (COCOLOG) for finite machines and its automatic reasoning methodology*, in Recent Advances in Mathematical Theory of Systems, Control, Networks and Signal Processing, vol. II, Proceedings of MTNS-91, Kobe, Japan, June 1991, H. Kimura and S. Kodama, eds., Mita Press, Osaka, 1991, pp. 49–54.
[CW95] ———, *COCOLOG: A conditional controller and observer logic for finite machines*, SIAM J. Control Optim., 33 (1995), pp. 1687–1715.
[CWe94] P. E. CAINES AND Y. J. WEI, *Markovian fragments: Complete subtheories of COCOLOG theories*, in Discrete Event Systems, Manufacturing Systems, and Communication Networks, IMA Vol. Math. Appl. 73, P. R. Kumar and P. P. Varaiya, eds., Springer-Verlag, New York, 1994, pp. 1–40.
[EFT84] H. D. EBBINGHAUS, J. FLUM, AND W. THOMAS, *Mathematical Logic*, Undergrad. Texts Math., Springer-Verlag, New York, 1984.
[N89] M. NEWBORN, *The Great Theorem Prover*, Newborn Software, P.O. Box 429, Victoria Station, Westmount, Quebec, Canada, H3Z 2V8.
[RG87] R. GOLDBLATT, *Logics of Time and Computation*, CSLI/Stanford, Stanford, CA, 1987.
[W91] S. WANG, *Classical and Logic Based Control Theory for Finite State Machines*, Ph.D. thesis, McGill University, October 1991.
[WC92] S. WANG AND P. E. CAINES, *Automated reasoning with function evaluation for COCOLOG with examples*, Proc. 31st IEEE Conf. on Decision and Control, Tucson, AZ, 1992, pp. 3758–3763. Complete version: Research Report 1713, INRIA, Sophia-Antipolis, 1992.
[WeC92] Y. J. WEI AND P. E. CAINES, *On Markovian fragments of COCOLOG for logic control systems*, in Proc. 31st IEEE Conf. on Decision and Control, Tucson, AZ, 1992, pp. 2967–2972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1734–1766, September 1996

013

MODEL SIMPLIFICATION AND OPTIMAL CONTROL OF STOCHASTIC SINGULARLY PERTURBED SYSTEMS UNDER EXPONENTIATED QUADRATIC COST*

ZIGANG PAN† AND TAMER BAŞAR†

Abstract. We study the optimal control of a general class of stochastic singularly perturbed linear systems with perfect and noisy state measurements under positively and negatively exponentiated quadratic cost. The (expected) cost function to be minimized is actually taken as the long-term time average of the logarithm of the expected value of an exponentiated quadratic loss. We identify appropriate "slow" and "fast" subproblems, obtain their optimum solutions (compatible with the corresponding measurement structure), and subsequently study the performances they achieve on the full-order system as the singular perturbation parameter ϵ becomes sufficiently small, with the expressions given in all cases being exact to within $O(\sqrt{\epsilon})$. It is shown that the composite controller (obtained by appropriately combining the optimum slow and fast controllers) achieves a performance level close to the optimal one whenever the full-order problem has a solution. The slow controller, on the other hand, achieves (asymptotically, as $\epsilon \to 0$) only a finite performance level (but not necessarily optimal), provided that the fast subsystem is open-loop stable. If the intensity of the noise in the system dynamics decreases to zero, however, the slow controller also achieves a performance level close to the optimal one.

The paper also presents a more direct derivation (than heretofore available) of the solution to the linear exponential quadratic Gaussian (LEQG) problem under noisy state measurements, which allows for a general quadratic cost (with cross terms) in the exponent and correlation between system and measurement noises, and obtains both necessary and sufficient conditions for existence of an optimal solution. Such a general LEQG problem is encountered in the slow-fast decomposition of the full-order problem, even if the original problem does not feature correlated noises. In this general context, the paper also establishes a complete equivalence between the LEQG problem and the H^∞-optimal control problem with measurement feedback, though this equivalence does not extend to the slow and fast subproblems arrived at after time-scale separation.

Key words. linear exponential quadratic Gaussian optimal control, generalized Riccati differential equation, generalized algebraic Riccati equation, singular perturbations, H^∞-optimal control

AMS subject classifications. 93E20, 90D25, 90A46, 93C80

1. Introduction.

The problem of optimal control of stochastic linear systems under exponentiated quadratic loss (the so-called linear exponential quadratic Gaussian (LEQG) problem) has been studied extensively in the literature, with new interest aroused on the topic due to the recently established relationship with the H^∞-optimal control of similar systems (but with deterministic disturbances) under quadratic loss. Perhaps the first formulation of the LEQG problem was given by Jacobson [8], in both discrete and continuous time, and using perfect state measurements, motivated by the fact that the exponentiated quadratic cost captures risk-seeking or risk-averse behavior, not obtainable using the linear quadratic Gaussian (LQG) formulation (which is risk neutral). Indeed it was discovered in [8] that the LEQG formulation with a positive exponent is equivalent (as far as the optimal solution goes) to a deterministic zero-sum linear quadratic (LQ) differential game, which we now know [2] is equivalent to an H^∞-optimal control problem, thus completing the link. The counterparts of the results of [8] in the imperfect state measurement case for discrete and continuous time were later obtained in [21], [25], and [3], with the relationship with the H^∞-optimal control problem established in a series of subsequent publications, such as [7], [23], [24]; see also the book by Whittle [22]. Similar relationships (between exponentiated-cost stochastic control and worst-case designs) exist also for nonlinear problems, as established for some subclass of such problems in [5]; see also the recent paper [19] for connection with differential games in the

*Received by the editors November 30, 1993; accepted for publication (in revised form) June 29, 1995. This research was supported in part by U.S. Department of Energy grant DE-FG-02-88-ER-13939.

†Decision and Control Laboratory, Coordinated Science Laboratory and the Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801 (pan@markov.csl.uiuc.edu and tbasar@black.csl.uiuc.edu).

infinite horizon. Another recent paper [9] completely establishes this equivalence in the discrete-time, finite-horizon case.

Our objective in this paper is to study, under both perfect and noisy state measurements, the robustness properties of the optimal solution of the LEQG problem with respect to unmodeled fast dynamics. This study is conducted in the framework of singularly perturbed models, with a small positive parameter ϵ quantifying the extent of coupling between the slow and fast dynamics. We seek ϵ-independent controllers that provide good (in a sense to be made precise later) approximation to the optimal controller of the full-order problem in a neighborhood of $\epsilon = 0$.

As mentioned earlier, at the full-order level there is an equivalence between the positively exponentiated subclass and a class of LQ H^∞-optimal control problems with singularly perturbed dynamics, with this latter class of problems extensively studied recently from the point of view of robustness and model reduction (see [12], [14], [15], [11]). This equivalence, however, does not readily carry over to the "model-reduction" stage, and as will be seen here, the end results in the two cases are considerably different. One of the reasons for this is that (as has been studied earlier in [20]) in stochastic problems the parameter ϵ has to enter the system dynamics and the measurement equation in a certain way for the problem to be well defined as $\epsilon \to 0$. The exact problem formulation provided in §2 shows that indeed in the stochastic case a time-scale separation of the full-order system becomes much more involved. Nevertheless, we still find occasion to use some of our earlier results from [12] and [14] in the present development, to simplify some of the proofs. Furthermore, in the derivation of the optimal solution to the stochastic control problem associated with the slow subsystem, we are faced with the need to obtain a clean and complete solution to the general LEQG problem with general cost structure and correlation between system and measurement noises. This motivates us into the investigation that leads to the results of §4, which generalize the earlier results of [3].

The paper is organized as follows. In the next section (§2) we formulate the LEQG problem with perfect and noisy state measurements for singularly perturbed systems. In §3, we study the singularly perturbed stochastic control problem under perfect state measurements, where we decompose the problem into slow and fast ϵ-free subproblems, obtain optimal controllers for these subproblems, and study the optimality of the composite controller as well as that of the slow controller in terms of the attainable performance for the full-order problem. In §4, we present a clean derivation (under least-stringent conditions) of a complete solution to the general LEQG problem under noisy state measurements with general cost structure and correlation between system and measurement noises, in both finite- and infinite-horizon cases. In §5, we study the singularly perturbed LEQG problem and its decomposition under noisy state measurements, where we identify the slow and fast subproblems to the full-order problem, obtain optimal controllers for these subproblems, construct the composite controller from these controllers, and study the optimality of these suboptimal controllers in terms of the attainable performance for the full-order problem. Three numerical examples are presented in §6 to illustrate the theory, and the paper ends with the concluding remarks of §7. Details of some of the derivations, not included here, can be found in the internal report [13].

2. Problem formulation. The system under consideration, with slow and fast dynamics, is described in the "singularly perturbed" form by

$$\text{(1)} \qquad \begin{cases} dx_1 = (A_{11}x_1 + A_{12}x_2 + B_1u_t)\,dt + G_1\,dw_t, & x_1(0) = x_{10}, \\ \epsilon dx_2 = (\epsilon^\alpha A_{21}x_1 + A_{22}x_2 + B_2u_t)\,dt + \epsilon^{1/2}G_2\,dw_t, & x_2(0) = x_{20}, \end{cases}$$

$$\text{(2)} \qquad \begin{cases} dy_1 = (C_{11}x_1 + C_{12}x_2)\,dt + E_1\,dw_t, & y_1(0) = 0, \\ dy_2 = (\epsilon^{1/2}C_{21}x_1 + C_{22}x_2)\,dt + \epsilon^{1/2}E_2\,dw_t, & y_2(0) = 0, \end{cases}$$

where $x := (x_1', x_2')'$ is the n-dimensional state vector, with x_1 of dimension n_1 and x_2 of dimension $n_2 := n - n_1$; $y := (y_1', y_2')'$ is the m-dimensional measurement process, with y_1 of dimension m_1 and y_2 of dimension $m_2 := m - m_1$; $\{u_t\}$ is the p-dimensional control input, and $\{w_t\}$ is an r-dimensional vector-valued standard Wiener process with $w_0 = 0$ with probability 1, which is independent of the initial condition; ϵ is a small positive scalar, denoting the singular perturbation parameter; the underlying probability space is the triplet $(\Omega, \mathcal{F}, \boldsymbol{P})$; and the parameter α is taken to be equal to $1/2$, except in the perfect state measurements case, when it is taken to be equal to 0. The specific way the parameter ϵ enters equations (1) and (2) is crucial for the system and measurement dynamics to be well defined in the limit as $\epsilon \to 0$, as otherwise (that is, for other powers of ϵ) either the noise dominates in the limit (and hence the optimization problem loses its significance) or the stochastic terms disappear (again making the problem uninteresting)—as extensively discussed in [20] in the context of the singularly perturbed LQG problem, whose arguments equally apply here as the nature of the cost function was irrelevant in that analysis.

Associated with the system (1), we now introduce the infinite-horizon exponentiated quadratic (*risk-sensitive*) cost function

$$J_\theta(\mu) = \lim_{t_f \to \infty} \frac{2}{\theta t_f} \ln \left\{ \boldsymbol{E} \left\{ \exp \left[\frac{\theta}{2} \left(\int_0^{t_f} (x' Q x + u' u)\, dt \right) \right] \right\} \right\}. \tag{3}$$

Here ln denotes the natural logarithm, and the scalar $\theta \neq 0$ is the *risk-sensitivity* parameter, in terms of which we will parametrize the solution.

We will consider two different information structures for the controller: perfect state measurements (where the current and past values of the state are available), and the noisy (imperfect) state measurements, where the measurement equations are as given in (2) above. Thus, in the former case, the control input $u = \{u_t\}$ is generated by a closed-loop control policy μ, according to

$$u_t = \mu(t, x_{[0,t]}), \quad t \geq 0, \tag{4}$$

where $\mu \in \mathcal{M}$ is an admissible controller, satisfying the standard conditions of Lipschitz continuity in x and piecewise continuity in t. Furthermore, as indicated earlier, we take in this case $\alpha = 0$, which makes the system dynamics well defined as $\epsilon \to 0$, as shown in [20]. Our objective is to find an optimal (minimizing) controller with respect to the cost (3), i. e., a $u_t^* = \mu^*(t, x_{[0,t]})$ such that

$$J_\theta(\mu^*) = \min_{\mu \in \mathcal{M}} J_\theta(\mu) := J_\theta^*. \tag{5}$$

Here, the initial state x_0 is taken as a fixed vector in $\mathbf{R}^n$.

In the noisy state measurements case, the initial state is taken to be a Gaussian random vector with mean $\bar{x}_0$ and covariance Σ_0, where Σ_0 is assumed to be positive definite, and it depends on ϵ in a way to be specified shortly. In this case, the control input $u \in \mathcal{H}_u$ is generated by a control policy μ_I, according to[1]

$$u_t = \mu_I(t, y_{[0,t]}), \quad t \geq 0, \tag{6}$$

where $\mu_I : [0, t_f] \times \mathcal{H}_y \to \mathcal{H}_u$ is piecewise continuous in t and Lipschitz continuous in $y := (y_1' \quad y_2')' \in \mathcal{H}_y$, further satisfying the given causality condition. Let us denote the class

[1] $\mathcal{H}_u$ denotes the Hilbert space of p-dimensional square-integrable functions (the controls). Likewise, $\mathcal{H}_y$ denotes the Hilbert space of m-dimensional square-integrable functions (the measurements).

of all admissible controllers in this case by $\mathcal{M}_I$, which is defined in precise mathematical terms in §4 later. For the limiting system and measurement equations to be well defined (as $\epsilon \to 0$), we take $\alpha = 1/2$ in this case. Denoting the cost function (3) for the noisy state measurements case by $J_{I\theta}(\mu_I)$, we again seek an optimal controller with respect to $J_{I\theta}(\mu_I)$, that is, a $u_t^* = \mu_I^*(t, y_{[0,t]})$, $t \geq 0$, such that

$$J_{I\theta}(\mu_I^*) = \min_{\mu_I \in \mathcal{M}_I} J_{I\theta}(\mu_I) := J_{I\theta}^*. \tag{7}$$

To complete the formulation of this risk-sensitive stochastic control problem, we introduce two basic assumptions, which will be required to hold throughout.

A1. Σ_0 and Q in (3) are partitioned as

$$\Sigma_0 = \begin{bmatrix} \Sigma_{011} & \sqrt{\epsilon}\,\Sigma_{012} \\ \sqrt{\epsilon}\,\Sigma_{021} & \Sigma_{022} \end{bmatrix}, \quad Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix},$$

where in each case the 11-block is of dimensions $n_1 \times n_1$, and the 22-block of $n_2 \times n_2$.

A2. The matrices A_{22}, Q_{22}, G_2G_2', and $N := EE'$ are invertible, where $E' = [E_1' \; E_2']$.[2] The system noise and the measurement noise are uncorrelated, i.e., $G_1E' = 0$ and $G_2E' = 0$.

3. Model simplification under perfect state measurements. The full-order solution to the (infinite-horizon) LEQG problem with state feedback can be obtained by taking an appropriate limit of the finite-horizon solution first obtained in [8]. Toward presenting this solution, let us first introduce the following useful notation:

$$A_\epsilon := \begin{bmatrix} A_{11} & A_{12} \\ \frac{1}{\epsilon}A_{21} & \frac{1}{\epsilon}A_{22} \end{bmatrix}, \quad B_\epsilon := \begin{bmatrix} B_1 \\ \frac{1}{\epsilon}B_2 \end{bmatrix}, \quad G_\epsilon := \begin{bmatrix} G_1 \\ \frac{1}{\sqrt{\epsilon}}G_2 \end{bmatrix}.$$

Note that in terms of this notation, the n-dimensional system dynamics can be written in the compact form

$$dx = (A_\epsilon x + B_\epsilon u_t)\,dt + G_\epsilon\,dw_t, \quad x(0) = x_0.$$

Introduce the quantity

$$S_\epsilon(\theta) := B_\epsilon B_\epsilon' - \theta G_\epsilon G_\epsilon' \tag{8}$$

and consider the generalized algebraic Riccati equation (GARE)

$$A_\epsilon'\tilde{Z} + \tilde{Z}A_\epsilon - \tilde{Z}S_\epsilon\tilde{Z} + Q = 0. \tag{9}$$

Further introduce the quantity

$$\theta^*(\epsilon) := \sup\{\theta \in \boldsymbol{R} : \text{the GARE (9) admits a positive definite solution } \tilde{Z}(\epsilon)\}. \tag{10}$$

Assume that (A_ϵ, B_ϵ) is controllable and (A_ϵ, Q) is observable for every $\epsilon > 0$. Then it follows by taking an appropriate limit of the finite-horizon result of [8] (see Theorem 7 of

[2]The conditions of invertibility of Q_{22} and G_2G_2' can be further relaxed to the conditions of observability of the pairs (A_{22}, Q_{22}) and (A_{22}, G_2). Results similar to those (to be) presented in this paper can be derived under these relaxed conditions, by perturbing the matrices Q_{22} and G_2G_2' by λI, for some scalar $\lambda > 0$, applying the results of this paper, and then letting $\lambda \to 0$; the limiting quantities are all well defined, as shown in [18].

[13]) that for each $\epsilon > 0$, $\theta^*(\epsilon)$ is positive, and for $\theta < \theta^*(\epsilon)$ the LEQG problem (with perfect state measurements) admits an optimal state-feedback solution

$$u^* = \mu^*(x) = -B_\epsilon' \tilde{Z}(\epsilon)\, x, \quad t \geq 0 \tag{11}$$

with the optimal (minimum) cost being

$$J_\theta^*(\epsilon) = \mathrm{Tr}(G_\epsilon G_\epsilon' \tilde{Z}(\epsilon))\,. \tag{12}$$

Furthermore, the feedback matrix $A_\epsilon - B_\epsilon B_\epsilon' \tilde{Z}(\epsilon)$ is Hurwitz.

Thus completing presentation of the full-order solution (for all $\epsilon > 0$), we now return to the original goal of this paper, which is the derivation of the optimal solution as $\epsilon \to 0$, via model simplification. Toward the end of obtaining ϵ-free solutions, we first decompose the system into slow and fast modes as in [12].

3.1. Time-scale decomposition.

Slow subsystem. The slow subsystem is obtained by letting $\epsilon = 0$ in the system dynamics, and solving for x_2 (to be denoted $\bar{x}_2$) in terms of $x_1 =: x_s$, $u =: u_s$, and under the working assumption A2:

$$\bar{x}_2 = -A_{22}^{-1}(A_{21}x_s + B_2 u_s). \tag{13}$$

Using this in the first equation of (1), we obtain the reduced-order (slow) dynamics

$$dx_{st} = (A_0 x_{st} + B_0 u_{st})\, dt + G_1\, dw_t, \tag{14}$$

where $A_0 := A_{11} - A_{12}A_{22}^{-1}A_{21}$, $B_0 := B_1 - A_{12}A_{22}^{-1}B_2$. Use of (13) also in the cost function (3) leads to the reduced (slow) cost (with $x_1 = x_s$)

$$J_{s\theta}(\mu_s) = \lim_{t_f \to \infty} \frac{2}{\theta t_f} \ln \left\{ \boldsymbol{E} \left\{ \exp \left[\frac{\theta}{2} \left(\int_0^{t_f} (|x_s|^2_{Q_{11}} + x_s' Q_{12} \bar{x}_2 + \bar{x}_2' Q_{21} x_s + |\bar{x}_2|^2_{Q_{22}} + |u_s|^2)\, dt \right) \right] \right\} \right\}. \tag{15}$$

Note that this is another LEQG problem, which this time has a cross term in the cost between the control and the state, but a simple linear state-feedback transformation on the control brings it into the same form as the full-order LEQG problem, and hence the theory described at the beginning of this section (for the full-order problem) applies, with some obvious modifications. Hence, the *slow* LEQG problem admits an optimal solution if the GARE

$$\tilde{A}_0' \tilde{Z}_s + \tilde{Z}_s \tilde{A}_0 - \tilde{Z}_s S_0 \tilde{Z}_s + \tilde{Q} = 0 \tag{16}$$

admits a minimal positive definite solution $Z_{s\theta}$, such that the matrix $\tilde{A}_0 - S_0 Z_{s\theta}$ is Hurwitz, where the coefficient matrices above are explicit functions of the parameter θ, and are written as (see [12] for details)

$$\begin{aligned}
\tilde{A}_0(\theta) &= A_{11} - A_{12}Q_{22}^{-1}Q_{21} - (S_{12} + A_{12}Q_{22}^{-1}A_{22}')(S_{22} + A_{22}Q_{22}^{-1}A_{22}')^{-1} \\
&\quad \cdot (A_{21} - A_{22}Q_{22}^{-1}Q_{21}), \\
\tilde{Q} &= Q_{11} - Q_{12}Q_{22}^{-1}Q_{21} + (A_{21}' - Q_{12}Q_{22}^{-1}A_{22}')(S_{22} + A_{22}Q_{22}^{-1}A_{22}')^{-1} \\
&\quad \cdot (A_{21} - A_{22}Q_{22}^{-1}Q_{21}), \\
S_0(\theta) &= S_{11} + A_{12}Q_{22}^{-1}A_{12}' - (S_{12} + A_{12}Q_{22}^{-1}A_{22}')(S_{22} + A_{22}Q_{22}^{-1}A_{22}')^{-1} \\
&\quad \cdot (S_{21} + A_{22}Q_{22}^{-1}A_{12}'),
\end{aligned}$$

where S_{ij} denotes the ijth block of $S(\theta)$, of dimensions $n_i \times n_j$, $i, j = 1, 2$, where the latter is defined by

$$S(\theta) := BB' - \theta GG', \quad B := \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad G := \begin{bmatrix} G_1 \\ 0 \end{bmatrix}.$$

In view of this, let us define

$$\theta_s := \sup\{\theta \in \boldsymbol{R} : \text{the GARE (16) admits a positive definite solution}\}. \tag{17}$$

Then, the transformed LEQG problem admits an optimal solution if $\theta < \theta_s$. For $\theta < \theta_s$, let $Z_{s\theta}$ be the unique positive definite solution of (16). Then, the optimal controller (after the inverse transformation) is given by (see [13])

$$\begin{aligned} u^*_{s\theta} = \mu^*_{s\theta}(t, x_s) = (-B_1' Z_{s\theta} + B_2'(S_{22} + A_{22} Q_{22}^{-1} A_{22}')^{-1} \\ \cdot ((S_{21} + A_{22} Q_{22}^{-1} A_{12}') Z_{s\theta} - (A_{21} - A_{22} Q_{22}^{-1} Q_{21}))) x_s \end{aligned} \tag{18}$$

Fast subsystem. To arrive at the fast subproblem, let $x_f := x_2 - \bar{x}_2$, $u_f := u - u_s$, and $\tau = \frac{t'-t}{\epsilon}$, where we take t to be frozen and t' to vary on the same scale as t. We define the fast subsystem and the associated cost (as in the standard regulator problem; see [4]) by

$$\frac{d}{d\tau} x_f^t = A_{22} x_f^t + B_2 u_f^t, \quad x_f^t(0) = x_f, \tag{19a}$$

$$J^t_{f\theta}(\mu^t_f) = \frac{2}{\theta} \ln \left\{ \boldsymbol{E} \left\{ \exp \left[\frac{\theta}{2} \left(\int_0^\infty (|x_f^t|^2_{Q_{22}} + |u_f^t|^2)\, d\tau \right) \right] \right\} \right\}. \tag{19b}$$

This is a deterministic, strictly convex, optimal control problem,[3] which admits a unique optimal controller that does not depend on the parameter θ:

$$\begin{aligned} & u_f^{t*}(\tau) = \mu_f^{t*}(x_f^t(\tau)) = -B_2' Z_f x_f^t(\tau) \\ \Rightarrow \quad & \mu_f^* = \mu_f^{t*}(x_f^t(0)) = -B_2' Z_f x_f = -B_2' Z_f (x_2 - \bar{x}_2), \end{aligned} \tag{20}$$

where Z_f is the positive definite solution to the ARE[4]

$$A_{22}' Z_f + Z_f A_{22} + Q_{22} - Z_f S_{22} Z_f = 0. \tag{21}$$

Substitute (13) and (18) into (20) to obtain

$$\begin{aligned} \mu^*_{f\theta}(t, x) = -B_2' Z_f x_2 - B_2' Z_f Q_{22}^{-1} (A_{12}' Z_{s\theta} + Q_{21} - A_{22}'(S_{22} \\ + A_{22} Q_{22}^{-1} A_{22}')^{-1} ((S_{21} + A_{22} Q_{22}^{-1} A_{12}') Z_{s\theta} - (A_{21} - A_{22} Q_{22}^{-1} Q_{21}))) x_1. \end{aligned} \tag{22}$$

Also, introduce the following Lyapunov equation, when the matrix A_{22} is Hurwitz:

$$A_{22}' Z_{of} + Z_{of} A_{22} + Q_{22} = 0, \tag{23}$$

whose relevance to our problem will be seen shortly.

[3]Recall that by our standard assumption A1, $Q_{22} > 0$.

[4]This ARE admits a unique positive definite solution if the pair (A_{22}, B_2) is controllable.

3.2. Composite controller. We now introduce the composite controller

$$\mu^*_{c\theta}(t,x) = \mu^*_{s\theta}(t,x) + \mu^*_{f\theta}(t,x), \tag{24}$$

where $\mu^*_{s\theta}$ and μ^*_f were defined by (18) and (22), respectively, for $\theta < \theta_s$. After some manipulations, this composite controller can be written as

$$\mu^*_{c\theta}(t,x) = -B' \begin{bmatrix} Z_{s\theta} & 0 \\ Z_c & Z_f \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \tag{25}$$

where

$$\begin{aligned} Z_c := {} & Z_f Q_{22}^{-1}(A'_{12}Z_{s\theta} + Q_{21}) - (I + Z_f Q_{22}^{-1}A'_{22})(S_{22} + A_{22}Q_{22}^{-1}A'_{22})^{-1} \\ & \cdot((S_{21} + A_{22}Q_{22}^{-1}A'_{12})Z_{s\theta} - (A_{21} - A_{22}Q_{22}^{-1}Q_{21})) \end{aligned}$$

and

$$B := \begin{pmatrix} B'_1 & B'_2 \end{pmatrix}.$$

3.3. Performances of the suboptimal controllers. The next important set of issues is the derivation of the conditions under which the performances attained by the composite and slow controllers are finite, and study of the limiting behavior of these expected cost values as $\epsilon \to 0$. A complete answer to these questions is given in the following theorem.

THEOREM 1. *For the singularly perturbed system* (1) *with state-feedback information and the cost function* (3), *let the relevant parts of assumptions* A1–A2 *be satisfied, the pairs* (A_0, B_0) *and* (A_{22}, B_2) *be controllable, and the pair* $(A_0, Q_{11} - Q_{12}Q_{22}^{-1}Q_{21})$ *be observable. Then,*

1. $\lim_{\epsilon \to 0^+} \theta^*(\epsilon) = \theta_s$.

2. $\forall \theta < \theta_s$, $\exists \epsilon_\theta > 0$ *such that* $\forall \epsilon \in [0, \epsilon_\theta)$, *the GARE* (9) *admits a positive definite solution, and consequently, the problem has an optimal solution, and the optimal cost for the problem can be approximated by*

$$J^*_{\theta\infty}(\epsilon) = \mathrm{Tr}(G_1 G'_1 Z_{s\theta} + G_2 G'_2 Z_f) + O(\sqrt{\epsilon}). \tag{26}$$

3. $\forall \theta < \theta_s$, *if the composite controller* $\mu^*_{c\theta}$ *is applied to the system, then* $\exists \epsilon'_\theta > 0$ *such that* $\forall \epsilon \in [0, \epsilon'_\theta)$,

$$J^c_\theta := J_\theta(\mu^*_{c\theta}) = J^*_\theta(\epsilon) + O(\sqrt{\epsilon}). \tag{27}$$

4. $\forall \theta < \theta_s$, *if, in addition, the matrix* A_{22} *is Hurwitz, and the slow controller* $\mu^*_{s\theta}$ *is applied to the system, then* $\exists \tilde{\epsilon}_\theta > 0$ *such that* $\forall \epsilon \in [0, \tilde{\epsilon}_\theta)$,

$$J^s_\theta := J_\theta(\mu^*_{s\theta}) = J^*_\theta(\epsilon) + \mathrm{Tr}(G_2 G'_2 (Z_{of} - Z_f)) + O(\sqrt{\epsilon}). \tag{28}$$

Proof. By Theorem 1 in [12], result 1. follows, and there exists an $\epsilon_\theta > 0$, such that the full-order GARE (9) admits a positive definite solution, for $\epsilon \in [0, \epsilon_\theta)$, which can be approximated by

$$\tilde{Z} = \begin{bmatrix} Z_{s\theta} + O(\sqrt{\epsilon}) & \epsilon Z'_c + O(\epsilon^{3/2}) \\ \epsilon Z_c + O(\epsilon^{3/2}) & \epsilon Z_f + O(\epsilon^{3/2}) \end{bmatrix}.$$

By Corollary 2 in [11], we have that the matrix $\tilde{A}_0 - S_0 Z_{s\theta}$ is Hurwitz for $\theta < \theta_{s\infty}$, which leads to 2.

For 3. and 4., we substitute the controllers $\mu^*_{c\theta}$ and $\mu^*_{s\theta}$ into the full-order system to obtain the resulting control-free LEQG problems. We then use the detailed derivations that led to Theorem 1 of [12] to establish 3. and 4. □

3.4. A large-deviation form. Consider now a large-deviation form of the problem considered in this section, which is the case when the system noise intensity asymptotically approaches zero. To formally illustrate this situation, we consider the following setup for the problem.

State dynamics:

$$\begin{cases} dx_1 = (A_{11}x_1 + A_{12}x_2)\,dt + B_1u + \xi G_1\,dw_t, & x_1(0) = x_{10}, \\ \epsilon dx_2 = (A_{21}x_1 + A_{22}x_2 + B_2u)\,dt + \sqrt{\epsilon}\xi G_2\,dw, & x_2(0) = x_{20}, \end{cases} \tag{29}$$

Cost function:

$$J_{\theta\infty}(\mu, \xi) = \lim_{t_f\to\infty} \frac{2\xi^2}{\theta t_f} \ln\left\{\boldsymbol{E}\left\{\exp\left[\frac{\theta}{2\xi^2}\left(\int_0^{t_f}(x'Qx + u'u)\,dt\right)\right]\right\}\right\}, \tag{30}$$

where ξ is a small scalar parameter to be varied. We will study the solution as $\xi \to 0$. Note that this problem is equivalent to the one considered earlier in this section, if we introduce the substitutions

$$\theta \leftarrow \frac{\theta}{\xi^2}, \quad G_\epsilon \leftarrow \xi G_\epsilon. \tag{31}$$

Assume that (A_ϵ, B_ϵ) is controllable, and (A_ϵ, Q) is observable for every $\epsilon > 0$. Under assumptions A1–A2, we know that the optimal solution exists, for each fixed $\epsilon > 0$, if the GARE (9) admits a positive definite solution, in which case the optimal controller is given by (11) independent of the parameter ξ. We define the quantity $\theta^*(\epsilon)$ in the same way as in (10). Thus, $\forall\theta < \theta^*(\epsilon)$, the optimal cost is given by

$$J^*_{\theta\infty}(\epsilon, \xi) = \xi^2 \mathrm{Tr}(G_\epsilon G'_\epsilon \tilde{Z}(t; \epsilon)), \tag{32}$$

where $\tilde{Z}(\epsilon)$ is the minimal positive definite solution to the GARE (9). Hence, $J^*_{\theta\infty}(\epsilon, \xi) \to 0$ as $\xi \to 0$.

To obtain ϵ-free solutions to the problem, we decompose the system into slow and fast subsystems. The slow subsystem is obtained by setting $\epsilon = 0$ in the state dynamics, as well as in the cost function. This slow problem admits an optimal solution if the GARE (16) admits a positive definite solution $Z_{s\theta}$, such that the matrix $\tilde{A}_0 - S_0 Z_{s\theta}$ is Hurwitz. We define θ_s the same way as in (17). Then, for $\theta < \theta_s$, the optimal control for the slow subsystem is again given by (18).

To obtain the fast subsystem, we use the same notation as in §3.1. The fast dynamics are the same as (19a) and the cost function is the same as (19b) under the substitutions (31). Then, the optimal control is given by (20), where Z_f is the unique positive definite solution to ARE (21). Substitution of (13) and (18) into (20) yields the fast controller, which is precisely (22). We also introduce the matrix Z_{of} to be the solution of Lyapunov equation (23).

Then, we form the composite controller as in (24), which leads to (25). We now summarize this result below, as a corollary to Theorem 1.

COROLLARY 1. *For the singularly perturbed system* (29), *with state-feedback information and the cost function* (30), *let assumptions* A1–A2 *be satisfied, the pairs* (A_0, B_0) *and* (A_{22}, B_2) *be controllable, and the pair* $(A_0, Q_{11} - Q_{12}Q_{22}^{-1}Q_{21})$ *be observable. Then,*

1. $\lim_{\epsilon\to 0^+} \theta^*(\epsilon) = \theta_s$.

2. $\forall\theta < \theta_s$, $\exists\epsilon_\theta > 0$ *such that* $\forall\epsilon \in [0, \epsilon_\theta)$, *the GARE* (9) *admits a positive definite solution, and consequently, the problem has an optimal solution, and the optimal cost for the problem can be approximated by*

$$J^*_\theta(\epsilon, \xi) = \xi^2(\mathrm{Tr}(G_1G'_1Z_{s\theta} + G_2G'_2Z_f) + O(\sqrt{\epsilon})). \tag{33}$$

3. $\forall\theta < \theta_s$, *if the composite controller* $\mu^*_{c\theta}$ *is applied to the system, then* $\exists\,\epsilon'_\theta > 0$ *such that* $\forall\epsilon \in [0, \epsilon'_\theta)$,

$$J^c_\theta(\xi) := J_\theta(\mu^*_{c\theta}, \xi) = J^*_\theta(\epsilon, \xi) + O(\xi^2\sqrt{\epsilon}). \tag{34}$$

4. $\forall\theta < \theta_s$, *if, in addition, the matrix* A_{22} *is Hurwitz, and the slow controller* $\mu^*_{s\theta}$ *is applied to the system, then* $\exists\tilde{\epsilon}_\theta > 0$ *such that* $\forall\epsilon \in [0, \tilde{\epsilon}_\theta)$,

$$J^s_\theta(\xi) := J_\theta(\mu^*_{s\theta}, \xi) = J^*_\theta(\epsilon, \xi) + O(\xi^2). \tag{35}$$

4. Imperfect state measurements: Full-order solution. In order to obtain the counterparts of the results of §3 in the noisy state measurements case, we will first need the solution to the full-order case, which is derived in this section. The derivation proceeds in two steps: first the solution to the finite-horizon LEQG problem is obtained, and then an appropriate limit of that solution is taken.

The model we adopt in this section involves, by necessity, a more general version of the problem formulated in §2 (for fixed $\epsilon > 0$), where the system and measurement noises are correlated, and the exponent of the loss function contains an additional cross term between the state and the control. Such a generalized model will be encountered in the next section, when we study the slow subproblem. Hence, the solution derived in this section will serve two purposes: it provides a solution to the full-order problem and also a solution to the slow subproblem which is the counterpart of the one of §3.1 in the noisy measurement case.

Accordingly, we now consider the following system and measurement dynamics:

$$\begin{cases} dx = (Ax + Bu_t)\,dt + G\,dw_t, & x(0) = x_0, \\ dy = Cx\,dt + E\,dw_t, & y(0) = 0, \end{cases} \tag{36}$$

where the correlation between the system and measurement noises is given by $L := GE' \neq 0$, and $x_0 \sim N(\bar{x}_0, \Sigma_0)$, $\Sigma_0 > 0$, $N := EE' > 0$.

4.1. The finite-horizon problem. Along with the system described by (36), consider the cost function

$$J_{I\theta}(\mu_I) = \frac{2}{\theta}\ln\left\{E\left\{\exp\left[\frac{\theta}{2}\left(x'_{t_f}Q_f x_{t_f} + \int_0^{t_f}(x'Qx + 2x'Pu + u'Ru)\,dt\right)\right]\right\}\right\}, \tag{37}$$

where $R > 0$, $Q_f \geq 0$, and $Q - PR^{-1}P \geq 0$.[5]

Introduce the notation

$$\begin{aligned} \bar{A} &:= A - BR^{-1}P', \quad \bar{Q} := Q - PR^{-1}P', \quad \bar{S} := BR^{-1}B' - \theta GG', \\ \tilde{A} &:= A - LN^{-1}C, \quad \tilde{M} := GG' - LN^{-1}L', \quad \tilde{R} := C'N^{-1}C - \theta Q, \end{aligned}$$

and in terms of this the backward generalized (game) Riccati differential equation (GRDE)

$$\dot{\tilde{Z}} + \bar{A}'\tilde{Z} + \tilde{Z}\bar{A} - \tilde{Z}\bar{S}\tilde{Z} + \bar{Q} = 0, \quad \tilde{Z}(t_f) = Q_f\,, \tag{38}$$

and the forward GRDE

$$\dot{\tilde{\Sigma}} = \tilde{\Sigma}\tilde{A}' + \tilde{A}\tilde{\Sigma} - \tilde{\Sigma}\tilde{R}\tilde{\Sigma} + \tilde{M}, \quad \tilde{\Sigma}(0) = \Sigma_0. \tag{39}$$

[5]The analysis and results of this subsection are valid even if the system and cost matrices are time varying, satisfying some natural smoothness conditions, such as continuous differentiability in t; see [13].

Define the quantity

$$\theta_I^* := \sup\{\theta \in \boldsymbol{R} : \text{the GRDEs (38) and (39) admit nonnegative definite solutions } \tilde{Z} \text{ and } \tilde{\Sigma}, \text{ respectively, on } [0, t_f], \text{ and the matrix } I - \theta\tilde{\Sigma}\tilde{Z} \text{ has only positive eigenvalues, for each } t \in [0, t_f]\}, \tag{40}$$

and for $\theta < \theta_I^*$, introduce the filter

$$d\check{x} = (A + \theta\tilde{\Sigma}Q)\check{x}\,dt + (B + \theta\tilde{\Sigma}P)u\,dt + (\tilde{\Sigma}C' + L)N^{-1}(dy - C\check{x}\,dt) \tag{41}$$

with initial state $\check{x}(0) = \bar{x}_0$. Letting $\hat{x} := (I - \theta\tilde{\Sigma}\tilde{Z})^{-1}\check{x}$, it can be shown (see [13]) that $\hat{x}$ is generated by the following dynamics:

$$\begin{aligned} d\hat{x} = {} & (\bar{A} - \bar{S}\tilde{Z})\hat{x}\,dt + (I - \theta\tilde{\Sigma}\tilde{Z})^{-1}(B + \theta\tilde{\Sigma}P)\tilde{u}\,dt + (I - \theta\tilde{\Sigma}\tilde{Z})^{-1} \\ & \cdot (\tilde{\Sigma}C' + L)N^{-1}(dy - (C + \theta L'\tilde{Z})\hat{x}\,dt), \end{aligned} \tag{42}$$

where $\tilde{u} := u + R^{-1}(B'\tilde{Z} + P')\hat{x}$.

Let $\varepsilon := x - \check{x}$ and $e := x - \hat{x}$. We will now restrict our attention to the class of controllers such that the following process ζ defines a martingale on $[0, t_f]$:

$$\begin{aligned} \zeta = \exp\Bigg\{ & \int_0^t (\varepsilon'\tilde{\Sigma}^{-1}G - e'\tilde{\Sigma}^{-1}(\tilde{\Sigma}C' + L)N^{-1}E)dw_t \\ & - \frac{1}{2}\int_0^t |G'\tilde{\Sigma}^{-1}\varepsilon - E'N^{-1}(C\tilde{\Sigma} + L')\tilde{\Sigma}^{-1}e|^2\,dt \Bigg\}. \end{aligned} \tag{43}$$

Hence, we define the set of admissible controllers $\mathcal{M}_I$ to be all mappings $\mu_I : [0, t_f] \times \mathcal{H}_y \to \mathcal{H}_u$ that are piecewise continuous in t and Lipschitz continuous in y, and further satisfying the given causality condition such that ζ is a martingale on $[0, t_f]$.

The above condition will be needed in the application of the Girsanov theorem [6] for a change of probability measures. A sufficient condition for this condition to be satisfied is the existence of positive constants δ and κ such that

$$\boldsymbol{E}\{\exp\{\delta|G'\tilde{\Sigma}^{-1}\varepsilon - E'N^{-1}(C\tilde{\Sigma} + L')\tilde{\Sigma}^{-1}e|^2\}\} \le \kappa \quad \forall t \in [0, t_f].$$

It should be obvious that any linear control law renders this condition valid, and hence the class of admissible controllers is bigger than the class of all linear controllers. We refer the reader to the recent book [6] for a thorough coverage of this topic.

We note at this point that this definition of the admissible controllers involves only one martingale condition, whereas the one of [3] involved two such (separate) conditions. We can do away with the second condition because of fact that in the proof to follow, we pursue a line of reasoning motivated by the "completions of squares" proof for the H^∞-optimal control problem, which avoids recasting of the problem in a new probability space, where the processes $\{w_t\}$ and $\{y_t\}$ are independent Wiener processes. As a result of this alternative approach, the existence of an optimal controller is not restricted by the additional condition of existence of a solution to GRDE (3.5) of [3] which clearly is unnecessary. Hence, the ensuing derivation is much simpler than that of [3], and it is more complete in the sense that it leads to both necessary and sufficient conditions for the existence of an optimal controller.

Now, we prove the following result.

THEOREM 2. *Consider the general finite-horizon LEQG problem described by* (36) *and* (37). *Assume that* $\Sigma_0 > 0$, $R > 0$, $Q - PR^{-1}P' \ge 0$, *and* $N > 0$. *Then, for each* $\theta < \theta_I^*$, *the optimal controller is given by*

$$u_{It}^* = \mu_I^*(t, y_{[0,t]}) = -R^{-1}(B'\tilde{Z} + P')(I - \theta\tilde{\Sigma}\tilde{Z})^{-1}\check{x} \equiv -R^{-1}(B'\tilde{Z} + P')\hat{x}, \tag{44}$$

where $\check{x}$ is generated by the filter (41), *or equivalently, $\hat{x}$ is generated by the filter* (42). *The optimal cost can be written as*

$$J_{I\theta}^* = \inf_{\mu_I \in \mathcal{M}_I} J_{I\theta}(\mu_I) = \bar{x}_0' \tilde{Z}(0)(I - \theta \Sigma_0 \tilde{Z}(0))^{-1} \bar{x}_0 - \frac{1}{\theta} \ln\{\det(I - \theta \tilde{\Sigma}(t_f) Q_f)\} + \int_0^{t_f} \operatorname{Tr}(\tilde{\Sigma} Q + (\tilde{\Sigma} C' + L) N^{-1} (C \tilde{\Sigma} + L') \tilde{Z} (I - \theta \tilde{\Sigma} \tilde{Z})^{-1})\, dt. \tag{45}$$

Furthermore, the above controller is also conditionally optimal.[6]

Proof. The differential equation for ε is easily obtained to be

$$d\varepsilon = (\tilde{A} - \tilde{\Sigma} C' N^{-1} C)\varepsilon\, dt - \theta \tilde{\Sigma} Q \check{x}\, dt - \theta \tilde{\Sigma} P u\, dt + (G - (\tilde{\Sigma} C' + L) N^{-1} E)\, dw_t.$$

Let $\tilde{\Psi} := \tilde{Z}(I - \theta \tilde{\Sigma} \tilde{Z})^{-1}$, and define

$$\Upsilon(t, \varepsilon, \check{x}) := |\varepsilon|^2_{\frac{1}{\theta}\tilde{\Sigma}^{-1}} + |\check{x}|^2_{\tilde{\Psi}} =: \Upsilon_1(t, \varepsilon) + \Upsilon_2(t, \check{x}).$$

Note that the matrix $\tilde{\Psi}$ satisfies the following GRDE:

$$\dot{\tilde{\Psi}} + \tilde{\Psi}(A + \theta \tilde{\Sigma} Q) + (A + \theta \tilde{\Sigma} Q)' \tilde{\Psi} + Q + \theta \tilde{\Psi} (\tilde{\Sigma} C' + L) N^{-1} (C \tilde{\Sigma} + L') \tilde{\Psi} - ((I - \theta \tilde{Z} \tilde{\Sigma})^{-1} P + \tilde{\Psi} B) R^{-1} (P' (I - \theta \tilde{\Sigma} \tilde{Z})^{-1} + B' \tilde{\Psi}) = 0. \tag{46}$$

This, with some additional lengthy but straightforward algebraic manipulations, leads to the following expression of the differential for Υ:

$$\begin{aligned} d\Upsilon = {} & \operatorname{Tr}(E' N^{-1} (C \tilde{\Sigma} + L') \tilde{\Psi} (\tilde{\Sigma} C' + L) N^{-1} E + (G' - E' N^{-1} (C \tilde{\Sigma} + L')) \frac{1}{\theta} \tilde{\Sigma}^{-1} (G \\ & - (\tilde{\Sigma} C' + L) N^{-1} E))\, dt - (x' Q x + 2 x' P u + u' R u)\, dt + |\tilde{u}|^2_R\, dt + \frac{2}{\theta} (\varepsilon' \tilde{\Sigma}^{-1} G \\ & - e' \tilde{\Sigma}^{-1} (\tilde{\Sigma} C' + L) N^{-1} E) dw_t - \frac{1}{\theta} |G' \tilde{\Sigma}^{-1} \varepsilon - E' N^{-1} (C \tilde{\Sigma} + L') \tilde{\Sigma}^{-1} e|^2\, dt. \end{aligned}$$

Adding the identically zero quantity

$$(2/\theta) \int_0^{t_f} d\Upsilon - (2/\theta)(\Upsilon(t_f, \varepsilon(t_f), \check{x}(t_f)) + \Upsilon(0, \varepsilon(0), \bar{x}_0))$$

to the exponent of $J_{I\theta}$ yields, after some rearrangement,

$$J_{I\theta} = \int_0^{t_f} \operatorname{Tr}(\tilde{\Psi}(\tilde{\Sigma} C' + L) N^{-1} (C \tilde{\Sigma} + L') + \frac{1}{\theta} \tilde{\Sigma}^{-1} (\tilde{M} + \tilde{\Sigma} C' N^{-1} C \tilde{\Sigma}))\, dt + \frac{2}{\theta} \ln \left\{ \boldsymbol{E} \left\{ \exp \left\{ \frac{1}{2} |\varepsilon(0)|^2_{\Sigma_0^{-1}} - \frac{1}{2} |e(t_f)|^2_{\tilde{\Sigma}(t_f)^{-1} - \theta Q_f} + \frac{\theta}{2} \int_0^{t_f} |\tilde{u}|^2_R\, dt \right\} \zeta(t_f) \right\} \right\} + |\bar{x}_0|^2_{\tilde{\Psi}}. \tag{47}$$

Introduce the change of probability [6]

$$\frac{d\tilde{\boldsymbol{P}}}{d\boldsymbol{P}} = \zeta(t_f). \tag{48}$$

The measure $\tilde{\boldsymbol{P}}$ is a valid probability measure for all $\mu_I \in \mathcal{M}_I$, since ζ is a martingale on $[0, t_f]$ by our construction of $\mathcal{M}_I$.

[6]For a precise definition, see [3]. This property is also referred to as *strong time consistency* [1].

Under the new probability measure $\tilde{P}$, the process v_t, defined by

$$v_t := w_t - \int_0^{t_f} (G'\tilde{\Sigma}^{-1}\varepsilon - E'N^{-1}(C\tilde{\Sigma} + L')\tilde{\Sigma}^{-1}e)\,dt,$$

is a standard Wiener process starting at 0, and it is independent of x_0.

It is straightforward to derive the following expression for the stochastic differential equation satisfied by y, under the new measure $\tilde{P}$:

$$dy = (C + \theta L'\tilde{Z})\hat{x}\,dt + E\,dv_t.$$

Hence, we conclude the following equivalence of sigma-fields:

$$Y_0^t := \sigma\{y_s : \ 0 \le s \le t\} = \sigma\{Ev_s : \ 0 \le s \le t\},$$

and that Y_0^t is independent of x_0, for each $t \in [0, t_f]$.

Let the expectation with respect to the probability measure $\tilde{P}$ be denoted by $\tilde{E}$. Then,

$$\bar{J}_{I\theta} = \frac{2}{\theta} \ln\left\{\tilde{E}\left\{\exp\left\{\frac{\theta}{2}\int_0^{t_f} |\tilde{u}|_R^2\,dt\right\}\tilde{E}\left\{\exp\left\{\frac{1}{2}|\varepsilon(0)|^2_{\Sigma_0^{-1}} - \frac{1}{2}|e(t_f)|^2_{\tilde{\Sigma}(t_f)^{-1}-\theta Q_f}\right\}|Y_0^{t_f}\right\}\right\}\right\}.$$

We will first obtain an expression for the quantity

$$J_b := \tilde{E}\left\{\exp\left\{\frac{1}{2}|\varepsilon(0)|^2_{\Sigma_0^{-1}} - \frac{1}{2}|e(t_f)|^2_{\tilde{\Sigma}(t_f)^{-1}-\theta Q_f}\right\}|Y_0^{t_f}\right\}.$$

Toward that end, we first derive a differential equation for e in terms of v_t:

$$de = (\tilde{A} + \tilde{M}\tilde{\Sigma}^{-1})e\,dt + (B - (I - \theta\tilde{\Sigma}\tilde{Z})^{-1}(B + \theta\tilde{\Sigma}P))\tilde{u}\,dt$$
$$+ (L - (I - \theta\tilde{\Sigma}\tilde{Z})^{-1}\cdot(\tilde{\Sigma}C' + L))N^{-1}E\,dv_t + (G - LN^{-1}E)\,dv_t.$$

Note that the processes $\{Ev_t\}_{0\le t\le t_f}$ and $\{(G - LN^{-1}E)v_t\}_{0\le t\le t_f}$ are independent, since $(G - LN^{-1}E)E' = 0$. Then, we can decompose e into $e = \hat{e} + \tilde{e}$, where

$$d\hat{e} = (\tilde{A} + \tilde{M}\tilde{\Sigma}^{-1})\hat{e}\,dt + (B - (I - \theta\tilde{\Sigma}\tilde{Z})^{-1}(B + \theta\tilde{\Sigma}P))\tilde{u}\,dt + (L - (I - \theta\tilde{\Sigma}\tilde{Z})^{-1}$$
$$\cdot(\tilde{\Sigma}C' + L))N^{-1}E\,dv_t,$$
$$d\tilde{e} = (\tilde{A} + \tilde{M}\tilde{\Sigma}^{-1})\tilde{e}\,dt + (G - LN^{-1}E)\,dv_t, \quad \tilde{e}(0) = \varepsilon(0).$$

The process $\hat{e}$ belongs to $Y_0^{t_f}$, and the process $\tilde{e}$ is independent of $Y_0^{t_f}$. Therefore, the process $\hat{e}$ is the conditional expectation of the process e given $Y_0^{t_f}$. The conditional distribution of the vector $[\tilde{e}(0)'\ e(t_f)']'$, given the measurement sigma-field $Y_0^{t_f}$, is Gaussian with mean and covariance

$$\begin{bmatrix} 0 \\ \hat{e}(t_f) \end{bmatrix}, \quad \begin{bmatrix} \Sigma_0 & D(t_f)' \\ D(t_f) & \Phi(t_f) \end{bmatrix} =: \Lambda, \quad \text{respectively,}$$

where D satisfies

$$\dot{D} = (\tilde{A} + \tilde{M}\tilde{\Sigma}^{-1})D, \quad D(0) = \Sigma_0,$$

and Φ is the solution to the following Lyapunov differential equation:

$$\dot{\Phi} = (\tilde{A} + \tilde{M}\tilde{\Sigma}^{-1})\Phi + \Phi(\tilde{A} + \tilde{M}\tilde{\Sigma}^{-1})' + \tilde{M}, \quad \Phi(0) = \Sigma_0.$$

Hence, J_b can be evaluated as

$$J_b = \int_{R^{2n}} \frac{1}{(2\pi)^n(\det(\Lambda))^{1/2}} \exp\left\{\frac{1}{2}|\tilde{e}(0)|^2_{\Sigma_0^{-1}} - \frac{1}{2}|\tilde{e}(t_f)+\hat{e}(t_f)|^2_{\tilde{\Sigma}(t_f)^{-1}-\theta Q_f}\right.$$

$$\left. - \frac{1}{2}\left|\left[\begin{array}{c}\tilde{e}(0)\\ \tilde{e}(t_f)\end{array}\right]\right|^2_{\Lambda^{-1}}\right\} d\tilde{e}(0)\,d\tilde{e}(t_f) = \int_{R^{2n}} \frac{1}{(2\pi)^n(\det(\Lambda))^{1/2}} \exp\left\{-\frac{1}{2}\left|\left[\begin{array}{c}\tilde{e}(0)\\ \tilde{e}(t_f)\end{array}\right]\right.\right.$$

$$\left.\left. + \tilde{\Lambda}\left[\begin{array}{c}0\\ (\tilde{\Sigma}^{-1}-\theta Q_f)\hat{e}(t_f)\end{array}\right]\right|^2_{\tilde{\Lambda}^{-1}}\right\} d\tilde{e}(0)\,d\tilde{e}(t_f) = \sqrt{\frac{\det(\tilde{\Lambda})}{\det(\Lambda)}},$$

where

$$\tilde{\Lambda} = \left[\begin{array}{cc}\Sigma_0^{-1}D'\tilde{\Phi}^{-1}D\Sigma_0^{-1} & -\Sigma_0^{-1}D'\tilde{\Phi}^{-1}\\ -\tilde{\Phi}^{-1}D\Sigma_0^{-1} & \tilde{\Phi}^{-1}+\tilde{\Sigma}^{-1}-\theta Q_f\end{array}\right]^{-1}\Bigg|_{t=t_f},$$

$$\tilde{\Phi} = \Phi - D\Sigma_0^{-1}D'.$$

Thus, the cost function can be written as follows:

$$J_{I\theta} = \int_0^{t_f} \mathrm{Tr}(\tilde{\Psi}(\tilde{\Sigma}C'+L)N^{-1}(C\tilde{\Sigma}+L') + \frac{1}{\theta}\tilde{\Sigma}^{-1}(\tilde{M}+\tilde{\Sigma}C'N^{-1}C\tilde{\Sigma}))\,dt$$

$$+ \frac{1}{\theta}\ln(\det(\tilde{\Lambda})) - \frac{1}{\theta}\ln(\det(\Lambda)) + \frac{2}{\theta}\ln\left\{E\left\{\exp\left\{\frac{\theta}{2}\int_0^{t_f}|\tilde{u}|^2_R\,dt\right\}\right\}\right\} + |\bar{x}_0|^2_{\tilde{\Psi}}$$

$$J_{I\theta} \geq \int_0^{t_f} \mathrm{Tr}(\tilde{\Psi}(\tilde{\Sigma}C'+L)N^{-1}(C\tilde{\Sigma}+L') + \frac{1}{\theta}\tilde{\Sigma}^{-1}(\tilde{M}+\tilde{\Sigma}C'N^{-1}C\tilde{\Sigma}))\,dt \tag{49}$$

$$+ \frac{2}{\theta}\ln(J_b) + |\bar{x}_0|^2_{\tilde{\Psi}}.$$

The controller (44) achieves the lower bound above, and hence is optimal. It is easy to see that the controller (44) is also conditionally optimal.

Some straightforward algebraic manipulations lead to the conclusion that the lower bound in (49) is indeed the same as (45).

This completes the proof. □

Remark 1. We observe that the optimal controller obtained for the LEQG problem is precisely the central controller for the H^∞-optimal control problem [2]. The preceding theorem also subsumes the main result of [3] as a special case, and obtains it under less restrictive conditions.

Remark 2. Theorem 2 also holds when $\Sigma_0 \geq 0$, if we restrict the set of admissible controllers to be the set of linear controllers. This generalization can be proved via a standard perturbation analysis, by first replacing Σ_0 by $\Sigma_0 + \rho I$, $\rho > 0$, and then letting $\rho \downarrow 0$.

4.2. The infinite-horizon problem. We now study the infinite-horizon case, where the cost function is taken to be the time-average of (37):

$$J_{I\theta}(\mu_I) = \lim_{t_f\to\infty}\frac{2}{\theta t_f}\ln\left\{E\left\{\exp\left[\frac{\theta}{2}\left(\int_0^{t_f}(x'Qx+2x'Pu+u'Ru)\,dt\right)\right]\right\}\right\}. \tag{50}$$

Introduce two GAREs,

$$\bar{A}'\tilde{Z} + \tilde{Z}\bar{A} - \tilde{Z}\bar{S}\tilde{Z} + \bar{Q} = 0 \tag{51}$$

and

$$\tilde{\Sigma}\tilde{A}' + \tilde{A}\tilde{\Sigma} - \tilde{\Sigma}\tilde{R}\tilde{\Sigma} + \tilde{M} = 0. \tag{52}$$

Define the quantity

$$\theta_I^* := \sup\{\theta \in \boldsymbol{R} : \text{the GAREs (51) and (52) admit minimal positive definite solutions } \tilde{Z} \text{ and } \tilde{\Sigma}, \text{ respectively, and the matrix } I - \theta\tilde{\Sigma}\tilde{Z} \text{ has only positive eigenvalues}\}. \tag{53}$$

For $\theta < \theta_I^*$, we introduce the filter

$$d\check{x} = (A + \theta\tilde{\Sigma}Q)\check{x}\,dt + (B + \theta\tilde{\Sigma}P)u\,dt + (\tilde{\Sigma}C' + L)N^{-1}(dy - C\check{x}\,dt) \tag{54}$$

with the initial state $\check{x}(0) = \bar{x}_0$. Let $\hat{x} := (I - \theta\tilde{\Sigma}\tilde{Z})^{-1}\check{x}$; then it is not difficult to show, as in the finite-horizon case, that $\hat{x}$ is generated by the following differential equation:

$$d\hat{x} = (\bar{A} - \bar{S}\tilde{Z})\hat{x}\,dt + (I - \theta\tilde{\Sigma}\tilde{Z})^{-1}(B + \theta\tilde{\Sigma}P)\tilde{u}\,dt + (I - \theta\tilde{\Sigma}\tilde{Z})^{-1} \cdot(\tilde{\Sigma}C' + L)N^{-1}(dy - (C + \theta L'\tilde{Z})\hat{x}\,dt), \tag{55}$$

where $\tilde{u} = u + R^{-1}(B'\tilde{Z} + P')\hat{x}$.

Suppose $\Sigma_0 > 0$, but $\Sigma_0 \leq \tilde{\Sigma}$; then, the solution of GRDE (39) converges to $\tilde{\Sigma}$ exponentially as $t \to \infty$.

We will define the set of admissible controllers, $\mathcal{M}_I$, to be all mappings $\mu_I : [0, \infty) \times \mathcal{H}_y \to \mathcal{H}_u$ that are admissible for every finite-horizon problem with initial time 0 and final time $t_f > 0$, for all $t_f \in \boldsymbol{R}^+$. Then, we have the following counterpart of Theorem 2.

THEOREM 3. *Consider the general LEQG problem described by* (36) *and* (50). *Let* $\Sigma_0 > 0$, $R > 0$, $Q - PR^{-1}P' \geq 0$, *and* $N > 0$, *and assume that the pairs* (A, B) *and* (A, G) *are controllable, and the pairs* (A, C) *and* (A, Q) *are observable. For each* $\theta < \theta_I^*$, *if* $\Sigma_0 \leq \tilde{\Sigma}$, *then the optimal controller is given by*

$$u_{It}^* = \mu_I^*(t, y_{[0,t]}) \equiv -R^{-1}(B'\tilde{Z} + P')(I - \theta\tilde{\Sigma}\tilde{Z})^{-1}\check{x} = -R^{-1}(B'\tilde{Z} + P')\hat{x}, \tag{56}$$

where $\check{x}$ *is generated by the filter* (54), *or equivalently,* $\hat{x}$ *is generated by filter* (55). *The optimal cost can be written as*

$$J_{I\theta}^* = \inf_{\mu_I \in \mathcal{M}_I} J_{I\theta}(\mu_I) = Tr(\tilde{\Sigma}Q + (\tilde{\Sigma}C' + L)N^{-1}(C\tilde{\Sigma} + L')\tilde{Z}(I - \theta\tilde{\Sigma}\tilde{Z})^{-1}). \tag{57}$$

Proof. By Theorem 2, we have the following relationships, for any admissible controller:

$$\frac{2}{\theta t_f}\ln\left\{\boldsymbol{E}\left\{\exp\left[\frac{\theta}{2}\left(\int_0^{t_f}(x'Qx + u'u)\,dt\right)\right]\right\}\right\}$$

$$\geq \frac{1}{t_f}(\bar{x}_0'\tilde{Z}^{t_f}(0)(I - \theta\Sigma_0\tilde{Z}^{t_f}(0))^{-1}\bar{x}_0 - \frac{1}{\theta}\ln\{\det(I - \theta\tilde{\Sigma}^{t_f}(t_f)Q_f)\}$$

$$+ \int_0^{t_f}\mathrm{Tr}(\tilde{\Sigma}^{t_f}Q + (\tilde{\Sigma}^{t_f}C' + L)N^{-1}(C\tilde{\Sigma}^{t_f} + L')\tilde{Z}^{t_f}(I - \theta\tilde{\Sigma}^{t_f}\tilde{Z}^{t_f})^{-1})\,dt),$$

where $\tilde{Z}^{t_f}$ and $\tilde{\Sigma}^{t_f}$ are the solutions to GRDEs (38) and (39), respectively, on the time interval $[0, t_f]$, in the former case with $Q_f = 0$. Hence,

$$J_{I\theta}(\mu_I) \geq \mathrm{Tr}(\tilde{\Sigma}Q + (\tilde{\Sigma}C' + L)N^{-1}(C\tilde{\Sigma} + L')\tilde{Z}(I - \theta\tilde{\Sigma}\tilde{Z})^{-1}) = J_{I\theta}^*,$$

since $\tilde{Z}^{t_f}$ converges to $\tilde{Z}$ as $t_f \to \infty$ and $\tilde{\Sigma}_f^t$ converges to $\tilde{\Sigma}$ as $t \to \infty$ exponentially.

To prove the theorem, it is sufficient to show that the controller defined by (56) and (54), or equivalently, the one given by (56) and (55), achieves a performance level that is equal to $J^*_{I\theta}$ given by (57).

Let

$$J^{t_f}_{I\theta}(\mu_I) := \frac{2}{\theta}\ln\left\{E\left\{\exp\left[\frac{\theta}{2}\left(\int_0^{t_f}(x'Qx+u'u)\,dt\right)\right]\right\}\right\},$$

$$\bar{J}^{t_f}_{I\theta}(\mu_I) := \frac{2}{\theta}\ln\left\{E\left\{\exp\left[\frac{\theta}{2}\left(\int_0^{t_f}(x'Qx+u'u)\,dt + x(t_f)'\tilde{Z}x(t_f)\right)\right]\right\}\right\}.$$

Then, clearly $\bar{J}^{t_f}_{I\theta}(\mu_I) \geq J^{t_f}_{I\theta}(\mu_I)$, and

$$J_{I\theta}(\mu_I) = \lim_{t_f\to\infty}\frac{1}{t_f}J^{t_f}_{I\theta}(\mu_I).$$

Hence, we have

$$J_{I\theta}(\mu_I) \leq \lim_{t_f\to\infty}\frac{1}{t_f}\bar{J}^{t_f}_{I\theta}(\mu_I).$$

By the proof of Theorem 2, and in particular by (47), we have the identity

$$\bar{J}^{t_f}_{I\theta}(\mu_I) = |\bar{x}_0|^2_{\tilde{\Psi}} + \int_0^{t_f}\mathrm{Tr}(\tilde{\Psi}(\tilde{\Sigma}C'+L)N^{-1}(C\tilde{\Sigma}+L') + \frac{1}{\theta}\tilde{\Sigma}^{-1}(\tilde{M}+\tilde{\Sigma}C'$$
$$\cdot N^{-1}C\tilde{\Sigma}))\,dt + \frac{2}{\theta}\ln\left\{E\left\{\exp\left\{\frac{1}{2}|\varepsilon(0)|^2_{\tilde{\Sigma}^{-1}} - \frac{1}{2}|e(t_f)|^2_{\tilde{\Sigma}^{-1}-\theta\tilde{Z}}\right.\right.\right.$$
$$\left.\left.\left. + \frac{\theta}{2}\int_0^{t_f}|\tilde{u}|^2_R\,dt\right\}\zeta(t_f)\right\}\right\},$$

where $\varepsilon := x-\check{x}$, $e := x-\hat{x}$, $\tilde{u} := u + R^{-1}(B'\tilde{Z}+P')\hat{x}$, $\tilde{\Psi} := \tilde{Z}(I-\theta\tilde{\Sigma}\tilde{Z})^{-1}$, and ζ is defined by

$$\zeta = \exp\left\{\int_0^t(\varepsilon'\tilde{\Sigma}^{-1}G - e'\tilde{\Sigma}^{-1}(\tilde{\Sigma}C'+L)N^{-1}E)dw_t\right.$$
$$\left. - \frac{1}{2}\int_0^t|G'\tilde{\Sigma}^{-1}\varepsilon - E'N^{-1}(C\tilde{\Sigma}+L')\tilde{\Sigma}^{-1}e|^2\,dt\right\}.$$

It is clear that, under the controller μ^*_I, as defined by (56) and (54) (as well as (55)), the process ζ is a martingale on $[0, t_f]$. Introduce a change of probability

$$\frac{d\tilde{P}}{dP} = \zeta(t_f).$$

Then, under the new measure,

$$\bar{J}^{t_f}_{I\theta}(\mu^*_I) = |\bar{x}_0|^2_{\tilde{\Psi}} + \int_0^{t_f}\mathrm{Tr}(\tilde{\Psi}(\tilde{\Sigma}C'+L)N^{-1}(C\tilde{\Sigma}+L') + \frac{1}{\theta}\tilde{\Sigma}^{-1}(\tilde{M}+\tilde{\Sigma}C'$$
$$\cdot N^{-1}C\tilde{\Sigma}))\,dt + \frac{2}{\theta}\ln\left\{\tilde{E}\left\{\exp\left\{\frac{1}{2}|\varepsilon(0)|^2_{\tilde{\Sigma}^{-1}} - \frac{1}{2}|e(t_f)|^2_{\tilde{\Sigma}^{-1}-\theta\tilde{Z}}\right\}\right\}\right\}.$$

This leads to the inequality

$$J_{I\theta}(\mu_I^*) \le \lim_{t_f\to\infty} \frac{1}{t_f} \bar{J}_{I\theta}^{t_f}(\mu_I^*) \le J_{I\theta}^*$$

from which the theorem follows. □

Remark 3. The theorem can be generalized to the case when $\Sigma_0 \ge 0$ if we restrict the set of admissible controllers to linear controllers, in view of Remark 2 and the independence of the solution from the initial condition.

Before concluding this subsection, we write the full-order solution (optimal controller) to the original problem formulated in §2 as a special case of the one given in Theorem 3 above.

First we introduce some additional notation:

$$C_\epsilon := \begin{bmatrix} C_{11} & C_{12} \\ \epsilon^{1/2}C_{21} & C_{22} \end{bmatrix}, \quad E_\epsilon := \begin{bmatrix} E_1 \\ \epsilon^{1/2}E_2 \end{bmatrix}, \quad N_\epsilon := E_\epsilon E_\epsilon',$$

$$R_\epsilon(\theta) := C_\epsilon' N_\epsilon^{-1} C_\epsilon - \theta Q =: \begin{bmatrix} R_{\epsilon 11} & \frac{1}{\sqrt{\epsilon}}R_{\epsilon 12} \\ \frac{1}{\sqrt{\epsilon}}R_{\epsilon 21} & \frac{1}{\epsilon}R_{\epsilon 22} \end{bmatrix}, \quad R_{12}(\theta) = R_{21}(\theta)' := C_1' N^{-1} C_2,$$

$$R_{22}(\theta) := C_2 N^{-1} C_2, \quad R_{11}(\theta) := C_1' N^{-1} C_1 - \theta Q_{11}, \quad R(\theta) := \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}.$$

Let us assume that the pairs (A_ϵ, B_ϵ) and (A_ϵ, G_ϵ) are controllable, and the pairs (A_ϵ, Q) and (A_ϵ, C_ϵ) are observable for every $\epsilon > 0$. Then the controller is given by

$$u_{I\epsilon}^*(t) = \mu_{I\epsilon}^*(t, y_{[0,t]}) = -B_\epsilon' \tilde{Z}(\epsilon)\hat{x}_t, \tag{58a}$$

$$d\hat{x}_t = (A_\epsilon - S_\epsilon \tilde{Z}(\epsilon))\hat{x}_t\, dt + (I - \theta\tilde{\Sigma}(\epsilon)\tilde{Z}(\epsilon))^{-1}\tilde{\Sigma}(\epsilon)C_\epsilon' N_\epsilon^{-1}(dy_t - C_\epsilon \hat{x}_t\, dt), \tag{58b}$$

where $\hat{x}(0) = (I - \theta\tilde{\Sigma}(\epsilon)\tilde{Z}(\epsilon))^{-1}\bar{x}_0$.[7] Here $\tilde{Z}(\epsilon)$ is the minimal positive definite solution to the GARE (9), and $\tilde{\Sigma}(\epsilon)$ is the minimal positive definite solution to

$$A_\epsilon \tilde{\Sigma} + \tilde{\Sigma} A_\epsilon' - \tilde{\Sigma} R_\epsilon \tilde{\Sigma} + G_\epsilon G_\epsilon' = 0, \tag{59}$$

with

$$R_\epsilon := C_\epsilon' N_\epsilon^{-1} C_\epsilon - \theta Q. \tag{60}$$

This is a valid solution for all $\theta < \theta_I^*(\epsilon)$, provided that $\Sigma_0 \le \tilde{\Sigma}(\epsilon)$, where

$$\begin{aligned} \theta_I^*(\epsilon) := \sup\{\theta \in \boldsymbol{R} :\ & \text{the GAREs (9) and (59) admit minimal} \\ & \text{positive definite solutions, and the matrix} \\ & I - \theta\tilde{\Sigma}(\epsilon)\tilde{Z}(\epsilon) \text{ has only positive eigenvalues}\}. \end{aligned} \tag{61}$$

For all $\theta < \theta_I^*(\epsilon)$ the optimal cost is

$$J_{I\theta}^*(\epsilon) = \mathrm{Tr}(\tilde{\Sigma}(\epsilon)Q + \tilde{\Sigma}(\epsilon)C_\epsilon' N_\epsilon^{-1} C_\epsilon \tilde{\Sigma}(\epsilon)\tilde{Z}(\epsilon)(I - \theta\tilde{\Sigma}(\epsilon)\tilde{Z}(\epsilon))^{-1}). \tag{62}$$

Remark 4. The condition $\Sigma_0 \le \tilde{\Sigma}(\epsilon)$ is essential for the existence of an optimal time-invariant control policy. Accordingly, we will assume henceforth that Σ_0 is less than or equal to (in matrix sense) the solution to GARE (59) at $\theta = 0$ for sufficiently small $\epsilon > 0$.

[7] It follows from a result of [2] that $\hat{r}_t$ of (4.2) in [3] is precisely $(I - \theta\tilde{\Sigma}\tilde{Z})^{-1}\hat{x}_t$, where $\hat{x}_t$ is generated by (58b).

5. Imperfect state measurements: Model simplification. We now turn to model simplification for the noisy state measurements case, with (36) replaced by (1)–(2), with $\alpha = 1/2$, and cost taken as given by (3) (instead of (50)). As in the perfect state measurements case discussed in §3, we first decompose the system into slow and fast subsystems.

5.1. Time-scale decomposition.

Slow subsystem. The slow subsystem is obtained by letting $dx_2 = 0$ and solving for $x_2\,dt$ (to be denoted $\bar{x}_2\,dt$) in terms of x_1, u, dw_t, and under the working assumption A2:

$$\bar{x}_2\,dt = -A_{22}^{-1}(\sqrt{\epsilon}A_{21}x_1\,dt + B_2u\,dt + \sqrt{\epsilon}G_2\,dw_t). \tag{63}$$

Using this in (1)–(2) and denoting $y_s := y$, we obtain

$$dx_1 = ((A_{11} + O(\sqrt{\epsilon}))x_1 + (B_1 - A_{12}A_{22}^{-1}B_2)u)\,dt + (G_1 + O(\sqrt{\epsilon}))\,dw_t, \quad x_1(0) = x_{10}, \tag{64a}$$

$$dy_{s1} = ((C_{11} + O(\sqrt{\epsilon}))x_s - C_{12}A_{22}^{-1}B_2u)\,dt + (E_1 + \sqrt{\epsilon}C_{12}A_{22}^{-1}G_2)\,dw_t, \quad y_1(0) = 0, \tag{64b}$$

$$dy_{s2} = (\sqrt{\epsilon}(C_{21} - C_{22}A_{22}^{-1}A_{21})x_s - C_{22}A_{22}^{-1}B_2u)\,dt + \sqrt{\epsilon}(E_2 - C_{22}A_{22}^{-1}G_2)\,dw_t, \quad y_2(0) = 0. \tag{64c}$$

For each $\epsilon > 0$, measurements (64b) and (64c) are equivalent to the vector measurement $\tilde{y}$, defined by

$$\tilde{y}_s := \begin{bmatrix} y_{s1} \\ \frac{1}{\sqrt{\epsilon}}\,y_{s2} \end{bmatrix} + \begin{bmatrix} C_{12} \\ \frac{1}{\sqrt{\epsilon}}\,C_{22} \end{bmatrix} A_{22}^{-1}B_2u. \tag{65}$$

Now set $\epsilon = 0$ in (64a) and (65) to arrive at the slow subproblem in terms of $x_s := x_1$, w_t, u, and $\tilde{y}_s$, with $\bar{x}_2 = -A_{22}^{-1}B_2u$, reduced (slow) dynamics

$$dx_s = (A_{11}x_s + B_0u)\,dt + G_1\,dw_t, \quad x_s(0) = x_{10}, \tag{66}$$

slow measurements

$$d\tilde{y}_s = C_0x_s\,dt + E^{\square}\,dw_t, \quad \tilde{y}_s(0) = 0, \tag{67}$$

and reduced (slow) cost function

$$J_{Is\theta}(\mu_s) = \lim_{t_f\to\infty} \frac{2}{\theta t_f} \ln\left\{E\left\{\exp\left[\frac{\theta}{2}\int_0^{t_f} (|x_s|^2_{Q_{11}} - 2x_s'Q_{12}A_{22}^{-1}B_2u + |u|^2_{I+B_2'A_{22}^{-1\prime}Q_{22}A_{22}^{-1}B_2})\,dt\right]\right\}\right\}, \tag{68}$$

where

$$B_0 := B_1 - A_{12}A_{22}^{-1}B_2, \quad C_0 := C_1 - C_2A_{22}^{-1}A_{21}, \quad E^{\square} := E - C_2A_{22}^{-1}G_2,$$

$$C_1 := \begin{bmatrix} C_{11}' & C_{21}' \end{bmatrix}', \quad C_2 := \begin{bmatrix} 0 & C_{22}' \end{bmatrix}'.$$

Note that even though we started with independent system and measurement noises, the slow LEQG involves correlated noises and a coupling term between the state and the control in the

exponent of the cost. But, thanks to the general solution presented in the previous section for this class of LEQG problems, we can now solve this slow LEQG problem. By Theorem 3, it admits an optimal solution if

(i) the following GARE admits a (minimal) positive definite solution, $Z_{s\theta}$:

$$\bar{A}_s' Z_{s\theta} + Z_{s\theta}\bar{A}_s - Z_{s\theta}\bar{S}_s Z_{s\theta} + \bar{Q}_s = 0, \tag{69}$$

where

$$\begin{aligned}
\bar{A}_s &:= A_{11} + B_0(I + B_2' A_{22}^{-1\prime} Q_{22} A_{22}^{-1} B_2)^{-1} B_2' A_{22}^{-1\prime} Q_{21},\\
\bar{S}_s(\theta) &:= B_0(I + B_2' A_{22}^{-1\prime} Q_{22} A_{22}^{-1} B_2)^{-1} B_0' - \theta G_1 G_1',\\
\bar{Q}_s &:= Q_{11} - Q_{12} A_{22}^{-1} B_2 (I + B_2' A_{22}^{-1\prime} Q_{22} A_{22}^{-1} B_2)^{-1} B_2' A_{22}^{-1\prime} Q_{21};
\end{aligned}$$

(ii) the following GARE admits a (minimal) positive definite solution, $\Sigma_{s\theta}$:

$$\tilde{A}_s \Sigma_{s\theta} + \Sigma_{s\theta}\tilde{A}_s' - \Sigma_{s\theta}\tilde{R}_s \Sigma_{s\theta} + \tilde{M}_s = 0, \tag{70}$$

where

$$\begin{aligned}
\tilde{A}_s &:= A_{11} - G_1 E^{\square\prime} N^{\square -1} C_0, \quad \tilde{R}_s := C_0' N^{\square -1} C_0 - \theta Q_{11},\\
\tilde{M}_s &:= G_1 G_1' - G_1 E^{\square\prime} N^{\square -1} E^{\square} G_1', \quad N^{\square} := E^{\square} E^{\square\prime};
\end{aligned}$$

and

(iii) the solutions to (69) and (70) satisfy the spectral radius condition

$$I - \theta \Sigma_{s\theta} Z_{s\theta} \quad \text{has only positive eigenvalues.} \tag{71}$$

Hence, let us introduce the quantity

$$\begin{aligned}
\theta_{Is} := \sup\{\theta \in \boldsymbol{R} : &\text{ the GAREs (69) and (70) admit minimal positive}\\
&\text{definite solutions, and further satisfy (71)}\}.
\end{aligned} \tag{72}$$

For $\theta < \theta_{Is}$, the slow LEQG problem admits an optimal controller, given by

$$\tilde{u}_{Is}^* = \tilde{\mu}_{Is}^*(t, \tilde{y}_{s[0,t]}) = -(I + B_2' A_{22}^{-1\prime} Q_{22} A_{22}^{-1} B_2)^{-1}(B_0' Z_{s\theta} - B_2' A_{22}^{-1\prime} Q_{21})\hat{x}_s, \tag{73a}$$

$$\begin{aligned}
d\hat{x}_s = {} & (A_{11} + \theta G_1 G_1' Z_{s\theta})\hat{x}_s\, dt + B_0 \tilde{u}_{Is}^*\, dt + (I - \theta \Sigma_{s\theta} Z_{s\theta})^{-1}(\Sigma_{s\theta} C_0'\\
& + G_1 E^{\square\prime}) N^{\square -1}(d\tilde{y}_s - C_0 \hat{x}_s\, dt - E^{\square}\theta G_1' Z_{s\theta}\hat{x}_s\, dt), \quad \hat{x}_s(0) = (I - \theta\Sigma_{s\theta} Z_{s\theta})^{-1}\bar{x}_{10}.
\end{aligned} \tag{73b}$$

Fast subsystem. To obtain the fast subsystem, we let $x_f := x_2 - \bar{x}_2 = x_2 + A_{22}^{-1} B_2 u_{Is}^*$, $u_f := u - u_s$, $y_f := y - y_s$, and $\tau = \frac{t' - t}{\epsilon}$, where we take t to be frozen, and t' to vary on the same time scale as t. In terms of the equivalent measurements

$$\tilde{y}_f := \begin{bmatrix} \frac{1}{\sqrt{\epsilon}}\, y_{f1}' & \frac{1}{\epsilon}\, y_{f2}' \end{bmatrix}, \tag{74}$$

we define the fast subsystem and the associated cost, respectively, by

$$dx_f^t = (A_{22} x_f^t + B_2 u_f^t)\, d\tau + G_2 dw_\tau, \quad x_f^t(0) = x_f, \tag{75a}$$

$$d\tilde{y}_f^t = C_2 x_f^t\, d\tau + E dw_\tau, \quad y_f^t(0) = 0, \tag{75b}$$

$$J_{If\theta}^t(\mu_f^t) = \boldsymbol{E}\left(\int_0^\infty (|x_f^t|_{Q_{22}}^2 + |u_f^t|^2)\, d\tau\right). \tag{75c}$$

This is a risk-neutral LQG problem, which does not depend on the parameter θ. It admits an optimal controller

$$u_{If}^{t*}(\tau) = \mu_{If}^{t*}(\tau, \tilde{y}_{f(-\infty,\tau]}^{t}) = -B_2' Z_f \hat{x}_f^t(\tau),$$

$$d\hat{x}_f^t = (A_{22} - S_{22}Z_f)\hat{x}_f^t\, d\tau + \Sigma_f C_2' N^{-1}(d\tilde{y}_f^t - C_2\hat{x}_f^t\, d\tau),$$

where Z_f and Σ_f are the unique nonnegative definite solutions to the AREs[8]

$$A_{22}'Z_f + Z_f A_{22} + Q_{22} - Z_f S_{22} Z_f = 0 \tag{76}$$

and

$$A_{22}\Sigma_f + \Sigma_f A_{22}' + G_2 G_2' - \Sigma_f R_{22}\Sigma_f = 0\,, \tag{77}$$

respectively. Transforming the control policy μ_{If}^{t*} back to the t time scale, we obtain the fast controller

$$u_{If}^{*} = \mu_{If}^{*} = \mu_f^{t*}(x_f^t(0)) = -B_2' Z_f \hat{x}_f, \tag{78a}$$

$$\epsilon d\hat{x}_f = (A_{22} - S_{22}Z_f)x_f\, dt + \Sigma_f C_2' N^{-1}\left(\begin{bmatrix} \sqrt{\epsilon}\, y_{f1} \\ y_{f2} \end{bmatrix} - C_2\hat{x}_f\, dt\right), \tag{78b}$$

$$\hat{x}_f(0) = x_f(0).$$

Also, we introduce two Lyapunov equations, for the case when A_{22} is Hurwitz:

$$A_{22}'Z_{of} + Z_{of}A_{22} + Q_{22} = 0, \tag{79}$$

$$A_{22}\Sigma_{of} + \Sigma_{of}A_{22}' + G_2G_2' = 0, \tag{80}$$

which, as we shall see shortly, are relevant to the problem under consideration.

5.2. Performance of suboptimal controllers. We now address the performance evaluation under the slow controller and the composite controller (to be defined), when applied to the full-order system, and the resulting degree of suboptimality with respect to the optimal performance of the full-order system. Toward this end, we first simplify the slow controller (73) using some straightforward algebraic manipulations of the type that can be found in [14] (derived in the context of the corresponding H^∞-optimal control problem):

$$\tilde{u}_{Is}^{*} = \tilde{\mu}_{Is}^{*}(t, \tilde{y}_{s[0,t]}) = -(B_1'Z_{s\theta} + B_2'V)\hat{x}_s, \tag{81a}$$

$$d\hat{x}_s = (\bar{A}_s - \bar{S}_s Z_{s\theta})\hat{x}_s\, dt + (I - \theta\Sigma_{s\theta}Z_{s\theta})^{-1}(\Sigma_{s\theta}C_1' + Y'C_2')N^{-1}(d\tilde{y}_s - (C_1 - C_2A_{22}^{-1}(A_{21} + \theta G_2G_1'Z_{s\theta}))\hat{x}_s\, dt), \tag{81b}$$

$$\hat{x}_s(0) = (I - \theta\Sigma_{s\theta}Z_{s\theta})^{-1}\bar{x}_{10},$$

where

$$Y := Y_1\Sigma_{s\theta} + Y_2, \quad V := V_1 Z_{s\theta} + V_2,$$

$$Y_1 := -(R_{22} + A_{22}'(G_2G_2)^{-1}A_{22})^{-1}(R_{21} + A_{22}'(G_2G_2)^{-1}A_{21}),$$

$$Y_2 := -(R_{22} + A_{22}'(G_2G_2)^{-1}A_{22})^{-1}A_{22}'(G_2G_2')^{-1}G_2G_1,$$

$$V_1 := -(S_{22} + A_{22}Q_{22}^{-1}A_{22}')^{-1}(S_{21} + A_{22}Q_{22}^{-1}A_{12}'),$$

$$V_2 := -(S_{22} + A_{22}Q_{22}^{-1}A_{22}')^{-1}A_{22}Q_{22}^{-1}Q_{21}.$$

[8] As is well known, these AREs admit positive definite solutions if (A_{22}, B_2) is controllable and (A_{22}, C_2) is observable.

Using (65) in (81b), we obtain the following alternative form for the slow controller:

$$u^*_{Is} = \mu^*_{Is}(t, y_{[0,t]}) = -(B_1' Z_{s\theta} + B_2' V)\hat{x}^s_s, \tag{82a}$$

$$\begin{aligned} d\hat{x}^s_s = {} & (\bar{A}_s - \bar{S}_s Z_{s\theta})\hat{x}^s_s\, dt + (I - \theta\Sigma_{s\theta} Z_{s\theta})^{-1}(\Sigma_{s\theta} C_1' + Y' C_2')N^{-1} \\ & \cdot \left(\begin{bmatrix} dy_1' & \frac{1}{\sqrt{\epsilon}} dy_2' \end{bmatrix}' - (C_{2\epsilon} A_{22}^{-1} B_2 (B_1' Z_{s\theta} + B_2' V) + C_1 - C_2 A_{22}^{-1}(A_{21} \right. \\ & \left. + \theta G_2 G_1' Z_{s\theta}))\hat{x}^s_s\, dt \right), \quad \hat{x}^s_s(0) = (I - \theta\Sigma_{s\theta} Z_{s\theta})^{-1}\bar{x}_{10}. \end{aligned} \tag{82b}$$

Before deriving an expression for the composite controller, let us introduce the notation

$$\begin{aligned} X &:= X_1 \Sigma_{s\theta} + X_2, \quad U := U_1 Z_{s\theta} + U_2, \\ X_1 &:= (G_2 G_2)^{-1} A_{21} - (G_2 G_2)^{-1} A_{22}(R_{22} + A_{22}'(G_2 G_2)^{-1} A_{22})^{-1} \\ &\qquad (R_{21} + A_{22}'(G_2 G_2)^{-1} A_{21}), \\ X_2 &:= (G_2 G_2)^{-1} G_2 G_1' - (G_2 G_2)^{-1} A_{22}(R_{22} + A_{22}'(G_2 G_2)^{-1} A_{22})^{-1} \\ &\qquad A_{22}'(G_2 G_2')^{-1} G_2 G_1, \\ U_1 &:= Q_{22}^{-1} A_{12}' - Q_{22}^{-1} A_{22}'(S_{22} + A_{22} Q_{22}^{-1} A_{22}')^{-1}(S_{21} + A_{22} Q_{22}^{-1} A_{12}'), \\ U_2 &:= Q_{22}^{-1} Q_{21} - Q_{22}^{-1} A_{22}'(S_{22} + A_{22} Q_{22}^{-1} A_{22}')^{-1} A_{22} Q_{22}^{-1} Q_{21}. \end{aligned}$$

Then, combining (78a) and (82a), we arrive at the composite controller, expressed as

$$\begin{aligned} u^*_{Ic} &= \mu^*_{Ic}(t, y_{[0,t]}) := \mu^*_{Is}(t, y_{[0,t]}) + \mu^*_{If}(t, y_{[0,t]}) \\ &= -(B_1' Z_{s\theta} + B_2' V)\hat{x}^c_s - B_2' Z_f \hat{x}^c_f, \end{aligned} \tag{83}$$

where a differential equation representation for $\hat{x}^c_s$ can be obtained by using (65) in (81b):

$$\begin{aligned} d\hat{x}^c_s = {} & (\bar{A}_s - \bar{S}_s Z_{s\theta})\hat{x}^c_s\, dt + (I - \theta\Sigma_{s\theta} Z_{s\theta})^{-1}(\Sigma_{s\theta} C_1' + Y' C_2')N^{-1} \\ & \cdot \left(\begin{bmatrix} dy_1' & \frac{1}{\sqrt{\epsilon}} dy_2' \end{bmatrix}' - (C_{2\epsilon} A_{22}^{-1} B_2 (B_1' Z_{s\theta} + B_2' V) + C_1 - C_2 A_{22}^{-1}(A_{21} \right. \\ & \left. + \theta G_2 G_1' Z_{s\theta}))\hat{x}^c_s\, dt - C_{2\epsilon} A_{22}^{-1} B_2 B_2' Z_f \hat{x}^c_f \right), \quad \hat{x}^c_s(0) = (I - \theta\Sigma_{s\theta} Z_{s\theta})^{-1}\bar{x}_{10}. \end{aligned} \tag{84}$$

To obtain the differential equation governing $\hat{x}^c_f$, we let

$$\begin{aligned} \begin{bmatrix} \sqrt{\epsilon} dy_{f1} \\ dy_{f2} \end{bmatrix} = {} & \begin{bmatrix} \sqrt{\epsilon} dy_1 \\ dy_2 \end{bmatrix} - \sqrt{\epsilon}(C_{2\epsilon} A_{22}^{-1} B_2 (B_1' Z_{s\theta} + B_2' V) + C_1 \\ & - C_2 A_{22}^{-1}(A_{21} + \theta G_2 G_1' Z_{s\theta}))\hat{x}^s_s\, dt \end{aligned}$$

and

$$\bar{x}_2 = -A_{22}^{-1} B_2 u^*_{Is} = -U\hat{x}^c_s$$

and substitute the above into (78b), to arrive at

$$\begin{aligned} \epsilon d\hat{x}^c_f = {} & (A_{22} - S_{22} Z_f)\hat{x}^c_f\, dt + \Sigma_f C_2' N^{-1}(\begin{bmatrix} \sqrt{\epsilon} dy_1' & dy_2' \end{bmatrix} - \sqrt{\epsilon}(C_{2\epsilon} A_{22}^{-1} B_2 \\ & \cdot (B_1' Z_{s\theta} + B_2' V) + C_1 - C_2 A_{22}^{-1}(A_{21} + \theta G_2 G_1' Z_{s\theta}))\hat{x}^c_s\, dt - C_2 \hat{x}^c_f\, dt); \\ & \hat{x}^c_f(0) = \bar{x}_{20} + U\hat{x}^c_s(0). \end{aligned} \tag{85}$$

The main results of this section are now given in Theorem 4 below, which provides expressions for the performances of the full-order system under full-order, slow, and composite controllers and establishes their asymptotic optimality (as $\epsilon \to 0$). We note that the condition of Remark 4 is guaranteed in this case by the conditions that $\Sigma_{011} < \Sigma_{s\theta}$ at $\theta = 0$, and $\Sigma_{022} < \Sigma_f$.

THEOREM 4. *For the singularly perturbed system* (1)–(2) *with* $\alpha = 1/2$, *and under cost function* (3):

1. *For each* $\epsilon > 0$, *if* (A_ϵ, B_ϵ) *and* (A_ϵ, G_ϵ) *are controllable,* (A_ϵ, C_ϵ) *and* (A_ϵ, Q) *are observable, and* N_ϵ *is invertible, then* $\forall \theta < \theta^*_{I\infty}(\epsilon)$, *and the optimal cost for the full-order LEQG problem can be written as*

$$J^*_{I\theta}(\epsilon) = \mathrm{Tr}(\tilde{\Sigma} Q + (\tilde{\Pi}^{-1} + \theta(\tilde{Z} - \theta \tilde{Z}\tilde{\Sigma}\tilde{Z}))^{-1}((I - \theta\tilde{Z}\tilde{\Sigma})Q(I - \theta\tilde{\Sigma}\tilde{Z}) + \tilde{Z}B_\epsilon B'_\epsilon \tilde{Z})), \tag{86}$$

where the matrix $\tilde{\Pi}$ *is the unique positive definite solution to the following Lyapunov equation:*

$$(A_\epsilon - S_\epsilon \tilde{Z})\tilde{\Pi} + \tilde{\Pi}(A'_\epsilon - \tilde{Z}S_\epsilon) + (I - \theta\tilde{\Sigma}\tilde{Z})^{-1}\tilde{\Sigma}C'_\epsilon N_\epsilon^{-1} C_\epsilon (I - \theta\tilde{Z}\tilde{\Sigma})^{-1} = 0. \tag{87}$$

2. *Let assumptions* A1 *and* A2 *be satisfied, the pairs* (A_{11}, B_0), $(A_{11}, G_1G'_1 - G_1G'_2(G_2G'_2)^{-1}G_2G'_1)$ *and* (A_{22}, B_2) *be controllable, and the pairs* (A_{11}, C_0), $(A_{11}, Q_{11} - Q_{12}Q_{22}^{-1}Q_{21})$ *and* (A_{22}, C_2) *be observable. Let* $\theta^*_I(\epsilon)$ *be as defined by* (61). *Then,*

(i) $\lim_{\epsilon \to 0^+} \theta^*_I(\epsilon) = \theta_{Is}$.

(ii) $\forall \theta < \theta_{Is}$, *there exists* $\epsilon_\theta > 0$, *such that* $\forall 0 < \epsilon \le \epsilon_\theta$, *the GAREs* (9) *and* (59) *admit minimal positive definite solutions, which can be approximated by*

$$\tilde{Z}(\epsilon) = \begin{bmatrix} Z_{s\theta} + O(\sqrt{\epsilon}) & \epsilon(Z_f U + V)' + O(\epsilon^{3/2}) \\ \epsilon(Z_f U + V) + O(\epsilon^{3/2}) & \epsilon Z_f + O(\epsilon^{3/2}) \end{bmatrix} \tag{88}$$

and

$$\tilde{\Sigma}(\epsilon) = \begin{bmatrix} \Sigma_{s\theta} + O(\sqrt{\epsilon}) & \sqrt{\epsilon}(\Sigma_f X + Y)' + O(\epsilon) \\ \sqrt{\epsilon}(\Sigma_f X + Y) + O(\epsilon) & \Sigma_f + O(\sqrt{\epsilon}) \end{bmatrix}. \tag{89}$$

Furthermore, $I - \theta\tilde{\Sigma}(\epsilon)\tilde{Z}(\epsilon)$ *has only positive eigenvalues.*

(iii) $\forall \theta < \theta_{Is}$, *there exists* $\tilde{\epsilon}_\theta \in (0, \epsilon_\theta]$, *such that* $\forall 0 < \epsilon \le \tilde{\epsilon}_\theta$, *the Lyapunov equation* (87) *admits a positive definite solution, which can be approximated by*

$$\tilde{\Pi}(\epsilon) = \begin{bmatrix} I & 0 \\ -U & I \end{bmatrix} \begin{bmatrix} \Pi_{s\theta} + O(\sqrt{\epsilon}) & O(\sqrt{\epsilon}) \\ O(\sqrt{\epsilon}) & \Pi_f + O(\sqrt{\epsilon}) \end{bmatrix} \begin{bmatrix} I & -U' \\ 0 & I \end{bmatrix}, \tag{90}$$

where $\Pi_{s\theta}$ *is the unique positive definite solution to the Lyapunov equation*

$$(\bar{A}_s - \bar{S}_s Z_{s\theta})\Pi_{s\theta} + \Pi_{s\theta}(\bar{A}'_s - Z_{s\theta}\bar{S}_s) + (I - \theta\Sigma_{s\theta}Z_{s\theta})^{-1}((\Sigma_{s\theta}C'_1 + Y'C'_2) \cdot N^{-1}(C_1\Sigma_{s\theta} + C_2 Y) + X'G_2G'_2X)(I - \theta Z_{s\theta}\Sigma_{s\theta})^{-1} = 0 \tag{91}$$

and Π_f *is the unique positive definite solution to the Lyapunov equation*

$$(A_{22} - B_2B'_2Z_f)\Pi_f + \Pi_f(A'_{22} - Z_fB_2B'_2) + \Sigma_f C'_2 N^{-1} C_2 \Sigma_f = 0. \tag{92}$$

(iv) $\forall \theta < \theta_{Is}$, $\forall 0 < \epsilon \le \tilde{\epsilon}_\theta$, *the optimal cost for the full-order LEQG problem can be approximated by*

$$\begin{aligned} J^*_{I\theta}(\epsilon) = \mathrm{Tr}(&\Sigma_{s\theta}Q_{11} + (\Pi_{s\theta}^{-1} + \theta(Z_{s\theta} - \theta Z_{s\theta}\Sigma_{s\theta}Z_{s\theta}))^{-1}((Z_{s\theta}B_1 + V'B_2) \\ &\cdot (B'_1Z_{s\theta} + B'_2V) + (I - \theta Z_{s\theta}\Sigma_{s\theta})Q_{11}(I - \theta\Sigma_{s\theta}Z_{s\theta}) \\ &- (I - \theta Z_{s\theta}\Sigma_{s\theta})Q_{12}U - U'Q_{21}(I - \theta\Sigma_{s\theta}Z_{s\theta}) + U'Q_{22}U) \\ &+ \Sigma_f Q_{22} + \Pi_f(Q_{22} + Z_fB_2B'_2Z_f)) + O(\sqrt{\epsilon}). \end{aligned} \tag{93}$$

(v) $\forall\theta < \theta_{Is}$, *if the composite controller* $\mu^*_{Ic\theta}$ *is applied to the system, then* $\exists\epsilon'_\theta > 0$ *such that* $\forall\epsilon \in [0, \epsilon'_\theta)$,

$$J^c_{I\theta} := J_{I\theta}(\mu^*_{Ic\theta}) = J^*_{I\theta}(\epsilon) + O(\sqrt{\epsilon}). \tag{94}$$

(vi) $\forall\theta < \theta_{Is}$, *if, in addition, the matrix* A_{22} *is Hurwitz, and the slow controller* $\mu^*_{Is\theta}$ *is applied to the system, then* $\exists\hat{\epsilon}_\theta > 0$ *such that* $\forall\epsilon \in [0, \hat{\epsilon}_\theta)$,

$$\begin{aligned} J^s_{I\theta} := J_{I\theta}(\mu^*_{Is\theta}) = J^*_{I\theta}(\epsilon) + \operatorname{Tr}(&\Sigma_{of} Q_{22} - \Sigma_f Q_{22} \\ &- \Pi_f(Q_{22} + Z_f B_2 B_2' Z_f)) + O(\sqrt{\epsilon}). \end{aligned} \tag{95}$$

Proof. We first substitute the optimal controller (58a)–(58b) into the full-order system, for any $\theta < \theta^*_{Is}(\epsilon)$, to obtain the following control-free LEQG problem in terms of $x^e := [x'\ \hat{x}']'$ and w:

$$\begin{aligned} dx^e &= \begin{bmatrix} A_\epsilon & -B_\epsilon B_\epsilon' \tilde{Z} \\ (I - \theta\tilde{\Sigma}\tilde{Z})^{-1}\tilde{\Sigma}C_\epsilon' N_\epsilon^{-1} C_\epsilon & A_\epsilon - S_\epsilon\tilde{Z} - (I - \theta\tilde{\Sigma}\tilde{Z})^{-1}\tilde{\Sigma}C_\epsilon' N_\epsilon^{-1} C_\epsilon \end{bmatrix} \\ &\quad \cdot x^e\, dt + \begin{bmatrix} G_\epsilon \\ (I - \theta\tilde{\Sigma}\tilde{Z})^{-1}\tilde{\Sigma}C_\epsilon' N_\epsilon^{-1} E_\epsilon \end{bmatrix} dw_t \\ &:= F^e_\epsilon x^e\, dt + G^e_\epsilon\, dw_t, \\ J^*_{I\theta}(\epsilon) &= \lim_{t_f\to\infty} \frac{2}{\theta t_f} \ln\left\{E\left\{\exp\left[\frac{\theta}{2}\int_0^{t_f} x^{e\prime} H^e x^e\, dt\right]\right\}\right\}, \end{aligned}$$

where

$$H^e := \begin{bmatrix} Q & 0 \\ 0 & \tilde{Z}B_\epsilon B_\epsilon'\tilde{Z} \end{bmatrix}$$

and $x^e(0)$ is a Gaussian random variable with mean and variance

$$\boldsymbol{E}x^e(0) = \begin{bmatrix} \bar{x}_0 \\ (I - \theta\tilde{\Sigma}\tilde{Z})^{-1}\bar{x}_0 \end{bmatrix} =: \bar{x}^e_0, \quad \operatorname{Var}(x^e(0)) = \begin{bmatrix} \Sigma_0 & 0 \\ 0 & 0 \end{bmatrix} =: \Theta^e_0.$$

To compute $J^*_{I\theta}(\epsilon)$ explicitly, we associate with the above system a fictitious measurement

$$dy^e = dv_t, \qquad y^e(0) = 0, \tag{96}$$

where v_t, or y^e, is a standard Wiener process independent of the initial condition and w_t.

Then, the two GAREs associated with this problem are (from §4.2)

$$F^{e\prime}_\epsilon\tilde{\Xi} + \tilde{\Xi}F^e_\epsilon + \tilde{\Xi}\theta G^e_\epsilon G^{e\prime}_\epsilon\tilde{\Xi} + H^e = 0 \tag{97}$$

and

$$F^e_\epsilon\tilde{\Theta} + \tilde{\Theta}F^{e\prime}_\epsilon + \tilde{\Theta}\theta H^e\tilde{\Theta} + G^e_\epsilon G^{e\prime}_\epsilon = 0. \tag{98}$$

It is shown in [13] that the minimal solutions to these GAREs are

$$\begin{aligned} \tilde{\Xi} &= \begin{bmatrix} \tilde{Z}^{-1} & \tilde{Z}^{-1} \\ \tilde{Z}^{-1} & (\tilde{Z} - \theta\tilde{Z}\tilde{\Sigma}\tilde{Z})^{-1} + \tilde{\Delta}^{-1} \end{bmatrix}^{-1}, \\ \tilde{\Theta} &= \begin{bmatrix} \tilde{\Sigma}^{-1} & -\tilde{\Sigma}^{-1} + \theta\tilde{Z} \\ -\tilde{\Sigma}^{-1} + \theta\tilde{Z} & \tilde{\Sigma}^{-1} - \theta\tilde{Z} + \tilde{\Pi}^{-1} \end{bmatrix}^{-1}, \end{aligned}$$

where $\tilde{\Delta}$ is the unique positive definite solution to the Lyapunov equation

$$(I-\theta\tilde{Z}\tilde{\Sigma})(A_\epsilon'-R_\epsilon\tilde{\Sigma})(I-\theta\tilde{Z}\tilde{\Sigma})^{-1}\tilde{\Delta}+\tilde{\Delta}(I-\theta\tilde{\Sigma}\tilde{Z})^{-1}(A_\epsilon-\tilde{\Sigma}R_\epsilon)(I-\theta\tilde{\Sigma}\tilde{Z})+\tilde{Z}B_\epsilon B_\epsilon'\tilde{Z}=0.$$

Furthermore, the matrices $F_\epsilon^e+\theta G_\epsilon^e G_\epsilon^{e\prime}\tilde{\Xi}$ and $F_\epsilon^{e\prime}+\theta H^e\tilde{\Theta}$ are Hurwitz.

We note that $\frac{1}{\theta}\tilde{\Theta}^{-1}$ is the maximal solution to the GARE (97), since $F_\epsilon^e+\theta G_\epsilon^e G_\epsilon^{e\prime}\tilde{\Theta}$ is an antistable matrix. Thus, by Theorem 5 of [26], we have $\frac{1}{\theta}\tilde{\Theta}^{-1}>\tilde{\Xi}$. It is easily seen that $\tilde{\Theta}>\Theta_0^e$. Hence, we obtain

$$J_{I\theta}^*=\mathrm{Tr}(\tilde{\Theta}H^e)$$

in view of Theorem 3. Using a matrix inversion identity, we obtain (86), which is an alternative expression to (62). Hence, part (1) is proven.

The GARE (9) also arises in the singularly perturbed H^∞-optimal control problem, whose approximate solution has been studied extensively in [12]. Here, due to the factor $\sqrt{\epsilon}$ multiplying A_{21} and G_2, these two matrices do not enter the slow GARE (69), nor do they enter the zeroth order approximation of the solution. Thus, GARE (9) admits a minimal positive definite solution, which is approximated by (88), for sufficiently small ϵ if $\theta<\theta_{Is}$.

To study the behavior of the solution of GARE (59), we first partition $\tilde{\Sigma}$ as follows (in a way consistent with the given partitioning on Σ_0):

$$\tilde{\Sigma}:=\begin{bmatrix}\tilde{\Sigma}_{11} & \sqrt{\epsilon}\,\tilde{\Sigma}_{12}\\ \sqrt{\epsilon}\,\tilde{\Sigma}_{21} & \tilde{\Sigma}_{22}\end{bmatrix}, \tag{99}$$

where $\tilde{\Sigma}_{12}=\tilde{\Sigma}_{21}'$. Then, by substituting this structure back into GARE (59), we obtain the following coupled matrix Riccati equations for the matrices $\tilde{\Sigma}_{11}$, $\tilde{\Sigma}_{12}$, and $\tilde{\Sigma}_{22}$:

$$\begin{aligned}(100a)\quad & A_{11}\tilde{\Sigma}_{11}+\sqrt{\epsilon}A_{12}\tilde{\Sigma}_{12}'+\tilde{\Sigma}_{11}A_{11}'+\sqrt{\epsilon}\tilde{\Sigma}_{12}A_{12}'+G_1G_1'-\tilde{\Sigma}_{11}R_{\epsilon 11}\tilde{\Sigma}_{11}\\ & -\tilde{\Sigma}_{12}R_{\epsilon 21}\tilde{\Sigma}_{11}-\tilde{\Sigma}_{11}R_{\epsilon 12}\tilde{\Sigma}_{12}'-\tilde{\Sigma}_{12}R_{\epsilon 22}\tilde{\Sigma}_{12}'=0,\\ (100b)\quad & \epsilon A_{11}\tilde{\Sigma}_{12}+\sqrt{\epsilon}A_{12}\tilde{\Sigma}_{22}+\tilde{\Sigma}_{11}A_{21}'+\tilde{\Sigma}_{12}A_{22}'+G_1G_2'-\epsilon\tilde{\Sigma}_{11}R_{\epsilon 11}\tilde{\Sigma}_{12}\\ & -\epsilon\tilde{\Sigma}_{12}R_{\epsilon 21}\tilde{\Sigma}_{12}-\tilde{\Sigma}_{11}R_{\epsilon 12}\tilde{\Sigma}_{22}-\tilde{\Sigma}_{12}R_{\epsilon 22}\tilde{\Sigma}_{22}=0,\\ (100c)\quad & \epsilon A_{21}\tilde{\Sigma}_{12}+A_{22}\tilde{\Sigma}_{22}+\epsilon\tilde{\Sigma}_{12}'A_{21}'+\tilde{\Sigma}_{22}A_{22}'+G_2G_2'-\epsilon^2\tilde{\Sigma}_{12}'R_{\epsilon 11}\tilde{\Sigma}_{12}\\ & -\epsilon\tilde{\Sigma}_{22}R_{\epsilon 21}\tilde{\Sigma}_{12}-\epsilon\tilde{\Sigma}_{12}'R_{\epsilon 12}\tilde{\Sigma}_{22}-\tilde{\Sigma}_{22}R_{\epsilon 22}\tilde{\Sigma}_{22}=0.\end{aligned}$$

The above set of equations are the same as (2.26)–(2.28) of [14] for $\epsilon\to 0$ (except for certain obvious modifications), which permits us to apply the results of [14] to the present case. Hence, for $\theta<\theta_{Is}$, (100) admit solutions for sufficiently small ϵ, which can be approximated by $\tilde{\Sigma}_{11}=\Sigma_{s\theta}+O(\sqrt{\epsilon})$, $\tilde{\Sigma}_{12}=X'\Sigma_f+Y'+O(\sqrt{\epsilon})$, and $\tilde{\Sigma}_{22}=\Sigma_f+O(\sqrt{\epsilon})$. Thus, the solution to (59) can be approximated by (89), for sufficiently small ϵ and for $\theta<\theta_{Is}$.

Furthermore, for $\theta<\theta_{Is}$ and sufficiently small ϵ, the matrix $I-\theta\tilde{\Sigma}(\epsilon)\tilde{Z}(\epsilon)$ can be approximated by

$$\begin{bmatrix}I-\theta\Sigma_{s\theta}Z_{s\theta}+O(\sqrt{\epsilon}) & O(\epsilon)\\ O(\sqrt{\epsilon}) & I+O(\sqrt{\epsilon})\end{bmatrix}.$$

Hence, it can have only positive eigenvalues. Thus, part 2 (ii) is proved.

Fix any $\theta > \theta_{Is}$; then, either one of the GAREs (69) and (70) does not admit any positive definite solution, or the matrix $I - \theta\Sigma_{s\theta}Z_{s\theta}$ has at least one negative eigenvalue. The former implies, in view of a result of [12], that one of the GAREs (9) or (59) does not admit any positive definite solution for sufficiently small ϵ, which further implies that $\theta > \theta_I^*(\epsilon)$. The latter, on the other hand, leads to the conclusion that the matrix $I - \theta\tilde{\Sigma}\tilde{Z}$ has at least one negative eigenvalue for sufficiently small ϵ, which again implies that $\theta > \theta_I^*(\epsilon)$. Hence, $\theta > \theta_I^*(\epsilon)$ for sufficiently small ϵ, $\forall\theta > \theta_{Is}$. Thus, part 2 (i) is also proved.

Let $T = \begin{bmatrix} I & 0 \\ U & I \end{bmatrix}$ and $\Pi = T\tilde{\Pi}T'$. Then, premultiplying (91) by T and postmultiplying it by T' yield the following Lyapunov equation for Π:

$$T(A_\epsilon - S_\epsilon\tilde{Z})T^{-1}\Pi + \Pi T^{-1\prime}(A_\epsilon' - \tilde{Z}S_\epsilon)T' + T(I - \theta\tilde{\Sigma}\tilde{Z})^{-1}\tilde{\Sigma}C_\epsilon' N_\epsilon^{-1}C_\epsilon \cdot (I - \theta\tilde{Z}\tilde{\Sigma})^{-1}T' = 0. \tag{101}$$

Note the following approximations for $\epsilon \in (0, \epsilon_\theta]$, which are easily obtained in view of the given approximations for $\tilde{Z}$ and $\tilde{\Sigma}$:

$$T(A_\epsilon - S_\epsilon\tilde{Z})T^{-1} = \begin{bmatrix} \bar{A}_s - \bar{S}_s Z_{s\theta} & O(1) \\ O(\frac{1}{\sqrt{\epsilon}}) & \frac{1}{\epsilon}(A_{22} - B_2B_2'Z_f) + O(\frac{1}{\sqrt{\epsilon}}) \end{bmatrix},$$

$$T(I - \theta\tilde{\Sigma}\tilde{Z})^{-1}\tilde{\Sigma}C_\epsilon' N_\epsilon^{-1}C_\epsilon(I - \theta\tilde{Z}\tilde{\Sigma})^{-1}T' = \begin{bmatrix} L_s + O(\sqrt{\epsilon}) & O(\frac{1}{\sqrt{\epsilon}}) \\ O(\frac{1}{\sqrt{\epsilon}}) & \frac{1}{\epsilon}\Sigma_f C_2' N^{-1} C_2 \Sigma_f + O(\frac{1}{\sqrt{\epsilon}}) \end{bmatrix},$$

$$L_s = (I - \theta\Sigma_{s\theta}Z_{s\theta})^{-1}((\Sigma_{s\theta}C_1' + Y'C_2')N^{-1}(C_1\Sigma_{s\theta} + C_2Y) + X'G_2G_2'X)(I - \theta Z_{s\theta}\Sigma_{s\theta})^{-1}.$$

Suppose that Π takes the form

$$\begin{bmatrix} \Pi_{11} & \sqrt{\epsilon}\Pi_{12} \\ \sqrt{\epsilon}\Pi_{21} & \Pi_{22} \end{bmatrix},$$

where $\Pi_{21} = \Pi_{12}'$, and substitute it into the Lyapunov equation (101) to arrive at the following equations for Π_{11}, Π_{12}, and Π_{22}:

$$(\bar{A}_s - \bar{S}_s Z_{s\theta})\Pi_{11} + \Pi_{11}(\bar{A}_s - \bar{S}_s Z_{s\theta})' + L_s + O(\sqrt{\epsilon}) = 0, \tag{102a}$$

$$\Pi_{11}O(1) + \Pi_{12}(A_{22} - B_2B_2'Z_f) + O(1) + O(\sqrt{\epsilon}) = 0, \tag{102b}$$

$$(A_{22} - B_2B_2'Z_f)\Pi_{22} + \Pi_{22}(A_{22} - B_2B_2'Z_f)' + \Sigma_f C_2' N^{-1} C_2 \Sigma_f + O(\sqrt{\epsilon}) = 0. \tag{102c}$$

Then, it follows that $\Pi_{11} = \Pi_{s\theta}$, $\Pi_{22} = \Pi_f$, and some Π_{12} (which exists) solve equations (102a)–(102c) at $\epsilon = 0$. By an application of the implicit function theorem [10] as in the proof of Theorem 1 of [12], the solutions to (102a)–(102c) are approximated by $\Pi_{11} = \Pi_{s\theta} + O(\sqrt{\epsilon})$, $\Pi_{22} = \Pi_f + O(\sqrt{\epsilon})$, and $\Pi_{12} = O(1)$, for sufficiently small ϵ. This then completes the proof of part 2 (iii).

A mere substitution of (88), (89), and (90) into (86) yields the desired result (93) (detailed algebraic manipulations can be found in [13]), which proves part 2 (iv).

Now substitute the composite controller μ_{Ic}^* into the full-order system to arrive at an infinite-horizon, control-free LEQG problem. Let

$$x_f^c := x_2 + U\hat{x}_s^c,$$
$$\tilde{x}^{c\prime} := [x_1', \hat{x}_s^{c\prime} + \sqrt{\epsilon}x_f^{c\prime}X(I - \theta Z_{s\theta}\Sigma_{s\theta})^{-1}, x_f^{c\prime}, \hat{x}_f^{c\prime}].$$

In terms of the state variable $\tilde{x}^c$, this LEQG problem can be written as (see [13] for details)

$$d\tilde{x}^c = \begin{bmatrix} F_{11}^c + O(\sqrt{\epsilon}) & O(1) \\ O(\frac{1}{\sqrt{\epsilon}}) & \frac{1}{\epsilon}F_{22}^c + O(\frac{1}{\sqrt{\epsilon}}) \end{bmatrix}\tilde{x}^c\,dt + \begin{bmatrix} G_1^c + O(\sqrt{\epsilon}) \\ \frac{1}{\sqrt{\epsilon}}G_2^c + O(1) \end{bmatrix} dw_t,$$
$$:= F_\epsilon^c\tilde{x}^c\,dt + G_\epsilon^c\,dw_t,$$
$$J_{I\theta}^c := \lim_{t_f\to\infty}\frac{2}{\theta t_f}\ln\left\{E\left\{\exp\left[\frac{\theta}{2}\int_0^{t_f}\tilde{x}^{c\prime}(H^c + O(\sqrt{\epsilon}))\tilde{x}^c\,dt\right]\right\}\right\},$$

where

$$F_{11}^c = \begin{bmatrix} A_{11} \\ (I - \theta\Sigma_{s\theta}Z_{s\theta})^{-1}((\Sigma_{s\theta}C_1' + Y'C_2')N^{-1}C_1 + X'A_{21}) \\ -B_1B_1'Z_{s\theta} - B_1B_2'V - A_{12}U \\ F_{cs} \end{bmatrix}, \tag{103a}$$

$$F_{cs} = \bar{A}_s - \bar{S}_sZ_{s\theta} - (I - \theta\Sigma_{s\theta}Z_{s\theta})^{-1}((\Sigma_{s\theta}C_1' + Y'C_2')N^{-1}C_1 + X'A_{21} + \theta X'G_2G_1'Z_{s\theta}), \tag{103b}$$

$$F_{22}^c = \begin{bmatrix} A_{22} & -B_2B_2'Z_f \\ \Sigma_fC_2'N^{-1}C_2 & A_{22} - B_2B_2'Z_f - \Sigma_fC_2'N^{-1}C_2 \end{bmatrix}, \tag{103c}$$

$$G_1^c = \begin{bmatrix} G_1 \\ (I - \theta\Sigma_{s\theta}Z_{s\theta})^{-1}(X'G_2 + (\Sigma_{s\theta}C_1' + Y'C_2')N^{-1}E) \end{bmatrix}, \tag{103d}$$

$$G_2^c = \begin{bmatrix} G_2 \\ \Sigma_fC_2'N^{-1}E \end{bmatrix}, \tag{103e}$$

$$H^c = \begin{bmatrix} H_{11}^c & O(1) \\ O(1) & H_{22}^c \end{bmatrix}, \tag{103f}$$

$$H_{11}^c = \begin{bmatrix} Q_{11} & -Q_{12}U \\ -U'Q_{21} & (Z_{s\theta}B_1 + V'B_2)(B_1Z_{s\theta} + B_2'V) + U'Q_{22}U \end{bmatrix}, \tag{103g}$$

$$H_{22}^c = \begin{bmatrix} Q_{22} & 0 \\ 0 & Z_fB_2B_2'Z_f \end{bmatrix}. \tag{103h}$$

The initial state $\tilde{x}^c(0)$ is a Gaussian random vector with mean $\bar{x}_0^c$ and covariance Σ_0^c, given by

$$\bar{x}_0^c := \begin{bmatrix} \bar{x}_{10} \\ (I - \theta\Sigma_{s\theta}Z_{s\theta})^{-1}\bar{x}_{10} + O(\sqrt{\epsilon}) \\ \bar{x}_{20} + U(I - \theta\Sigma_{s\theta}Z_{s\theta})^{-1}\bar{x}_{10} \\ \bar{x}_{20} + U(I - \theta\Sigma_{s\theta}Z_{s\theta})^{-1}\bar{x}_{10} \end{bmatrix},$$

$$\Sigma_0^c := \begin{bmatrix} \begin{bmatrix} \Sigma_{011} & 0 \\ 0 & 0 \end{bmatrix} + O(\epsilon) & O(\sqrt{\epsilon}) \\ O(\sqrt{\epsilon}) & \begin{bmatrix} \Sigma_{022} & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix}.$$

To evaluate the cost $J_{I\theta\infty}^c$, we associate a fictitious measurement (96) with this LEQG problem (as in the proof of part 1), where v_t, or y^e, is a standard Wiener process independent of the initial state and w_t.

Then, the two GAREs associated with this problem are

$$F_\epsilon^{c\prime}\tilde{\Xi}^c + \tilde{\Xi}^c F_\epsilon^c + \tilde{\Xi}^c \theta G_\epsilon^c G_\epsilon^{c\prime} \tilde{\Xi}^c + H^c = 0$$

and

$$F_\epsilon^c \tilde{\Theta}^c + \tilde{\Theta}^c F_\epsilon^{c\prime} + \tilde{\Theta}^c \theta H^c \tilde{\Theta}^c + G_\epsilon^c G_\epsilon^{c\prime} = 0.$$

It is shown in [13] that the minimal solutions to these GAREs exist for $\theta < \theta_{Is}$ and sufficiently small ϵ, and can be approximated by

$$\tilde{\Xi}^c = \begin{bmatrix} \tilde{\Xi}_{11}^c + O(\sqrt{\epsilon}) & O(\epsilon) \\ O(\epsilon) & \epsilon\tilde{\Xi}_{22}^c + O(\epsilon^{3/2}) \end{bmatrix}, \tag{104a}$$

$$\tilde{\Xi}_{11}^c = \begin{bmatrix} Z_{s\theta}^{-1} & Z_{s\theta}^{-1} \\ Z_{s\theta}^{-1} & (Z_{s\theta} - \theta Z_{s\theta}\Sigma_{s\theta}Z_{s\theta})^{-1} + \Delta_{s\theta}^{-1} \end{bmatrix}^{-1}, \tag{104b}$$

$$\tilde{\Xi}_{22}^c = \begin{bmatrix} Z_f + \Delta_f & -\Delta_f \\ -\Delta_f & \Delta_f \end{bmatrix}, \tag{104c}$$

$$\Delta_{s\theta}(I - \theta\Sigma_{s\theta}Z_{s\theta})^{-1}(\tilde{A}_s - \Sigma_{s\theta}\tilde{R}_s)(I - \theta\Sigma_{s\theta}Z_{s\theta}) + (I - \theta Z_{s\theta}\Sigma_{s\theta})(\tilde{A}_s' \tag{104d}$$

$$- \tilde{R}_s\Sigma_{s\theta})(I - \theta Z_{s\theta}\Sigma_{s\theta})^{-1}\Delta_{s\theta} + (Z_{s\theta}B_1 + V'B_2)(B_1 Z_{s\theta} + B_2'V) + U'Q_{22}U = 0,$$

$$\Delta_f(A_{22} - \Sigma_f R_{22}) + \Delta_f(A_{22}' - R_{22}\Sigma_f) + Z_f B_2 B_2' Z_f = 0, \tag{104e}$$

$$\tilde{\Theta}^c = \begin{bmatrix} \tilde{\Theta}_{11}^c + O(\sqrt{\epsilon}) & O(\sqrt{\epsilon}) \\ O(\sqrt{\epsilon}) & \tilde{\Theta}_{22}^c + O(\sqrt{\epsilon}) \end{bmatrix}, \tag{104f}$$

$$\tilde{\Theta}_{11}^c = \begin{bmatrix} \Sigma_{s\theta}^{-1} & \theta Z_{s\theta} - \Sigma_{s\theta}^{-1} \\ \theta Z_{s\theta} - \Sigma_{s\theta}^{-1} & \Sigma_{s\theta}^{-1} - \theta Z_{s\theta} + \Pi_{s\theta}^{-1} \end{bmatrix}^{-1}, \tag{104g}$$

$$\tilde{\Theta}_{22}^c = \begin{bmatrix} \Sigma_f + \Pi & \Pi_f \\ \Pi_f & \Pi_f \end{bmatrix}. \tag{104h}$$

Furthermore, the matrices $F_\epsilon^c + \theta G_\epsilon^c G_\epsilon^{c\prime}\tilde{\Xi}^c$ and $F_\epsilon^{c\prime} + \theta H^c\tilde{\Theta}^c$ are Hurwitz. By Theorem 5 of [26], the matrix $I - \theta\tilde{\Theta}^c\tilde{\Xi}^c$ has only positive eigenvalues.

Obviously $\tilde{\Theta}^c > \Theta_0^c$. Hence, by Theorem 3, $J_{I\theta}^c = \mathrm{Tr}(\tilde{\Theta}^c H^c)$. Some straightforward algebraic manipulations lead to part 2 (v).

Now substitute the slow controller μ_{Is}^* into the full-order system to arrive at another control-free LEQG problem. Let

$$x_f^s := x_2 + U\hat{x}_s^s,$$

$$\tilde{x}^{s\prime} := [x_1',\ \hat{x}_s^{s\prime} + \sqrt{\epsilon}x_f^{s\prime}X(I - \theta Z_{s\theta}\Sigma_{s\theta})^{-1},\ x_f^{s\prime}].$$

In terms of the state variable $\tilde{x}^s$, this LEQG problem can be written as (see again [13] for details)

$$d\tilde{x}^s = \begin{bmatrix} F_{11}^c + O(\sqrt{\epsilon}) & O(1) \\ O(\frac{1}{\sqrt{\epsilon}}) + & \frac{1}{\epsilon}A_{22} + O(\frac{1}{\sqrt{\epsilon}}) \end{bmatrix} \tilde{x}^s\, dt + \begin{bmatrix} G_1^c + O(\sqrt{\epsilon}) \\ \frac{1}{\sqrt{\epsilon}}G_2 + O(1) \end{bmatrix} dw_t$$

$$:= F_\epsilon^s \tilde{x}^s\, dt + G_\epsilon^s\, dw_t,$$

$$J_{I\theta}^s := \lim_{t_f \to \infty} \frac{2}{\theta t_f} \ln \left\{ E \left\{ \exp \left[\frac{\theta}{2} \int_0^{t_f} \tilde{x}^{c\prime}(H^s + O(\sqrt{\epsilon}))\tilde{x}^c\, dt \right] \right\} \right\},$$

where

$$H^s = \begin{bmatrix} H_{11}^c & O(1) \\ O(1) & Q_{22} \end{bmatrix}$$

and F_{11}^c, G_1^c, and H_{11}^c are as defined in (103a), (103d), and (103g).

The initial state $\tilde{x}^s(0)$ is a Gaussian random vector with mean $\bar{x}_0^s$ and covariance Σ_0^s, given by

$$\bar{x}_0^s := \begin{bmatrix} \bar{x}_{10} \\ (I - \theta\Sigma_\theta Z_{s\theta})^{-1}\bar{x}_{10} + O(\sqrt{\epsilon}) \\ \bar{x}_{20} + U(I - \theta\Sigma_\theta Z_{s\theta})^{-1}\bar{x}_{10} \end{bmatrix},$$

$$\Sigma_0^s := \begin{bmatrix} \begin{bmatrix} \Sigma_{011} & 0 \\ 0 & 0 \end{bmatrix} + O(\sqrt{\epsilon}) & O(\sqrt{\epsilon}) \\ O(\sqrt{\epsilon}) & \Sigma_{022} + O(\sqrt{\epsilon}) \end{bmatrix}.$$

To evaluate the cost $J_{I\theta\infty}^s$, we associate (as in the composite controller case) a fictitious measurement (96) with this LEQG problem, where v_t, or y^e, is a standard Wiener process independent of the initial state and w_t.

Then, the two GAREs associated with this problem are

$$F_\epsilon^{s\prime}\tilde{\Xi}^s + \tilde{\Xi}^s F_\epsilon^s + \tilde{\Xi}^s\theta G_\epsilon^s G_\epsilon^{s\prime}\tilde{\Xi}^s + H^s = 0$$

and

$$F_\epsilon^s\tilde{\Theta}^s + \tilde{\Theta}^s F_\epsilon^{s\prime} + \tilde{\Theta}^s\theta H^s\tilde{\Theta}^s + G_\epsilon^s G_\epsilon^{s\prime} = 0.$$

It is shown in [13] that, if A_{22} is Hurwitz, the minimal solutions to these GAREs exist for $\theta < \theta_{Is}$ and sufficiently small ϵ, and can be approximated by

$$\tilde{\Xi}^s = \begin{bmatrix} \tilde{\Xi}_{11}^c + O(\sqrt{\epsilon}) & O(\epsilon) \\ O(\epsilon) & \epsilon Z_{of} + O(\epsilon^{3/2}) \end{bmatrix},$$

$$\tilde{\Theta}^s = \begin{bmatrix} \tilde{\Theta}_{11}^c + O(\sqrt{\epsilon}) & O(\sqrt{\epsilon}) \\ O(\sqrt{\epsilon}) & \Sigma_{of} + O(\sqrt{\epsilon}) \end{bmatrix},$$

where $\tilde{\Xi}_{11}^c$ and $\tilde{\Theta}_{11}^c$ are as defined in (104b) and (104g). Furthermore, the matrices $F_\epsilon^s + \theta G_\epsilon^s G_\epsilon^{s\prime}\tilde{\Xi}^s$ and $F_\epsilon^{s\prime} + \theta H^s\tilde{\Theta}^s$ are Hurwitz. By Theorem 5 of [26], the matrix $I - \theta\tilde{\Theta}^s\tilde{\Xi}^s$ has only positive eigenvalues.

Hence, by Theorem 3, $J_{I\theta}^s = \mathrm{Tr}(\tilde{\Theta}^s H^s)$. Some straightforward algebraic manipulations lead to (95). This completes the proof of the theorem. □

5.3. A large-deviation form. As the counterpart of the analysis of §3.4, we again consider a large-deviation form of the problem. The system under consideration is described by

$$\text{(105)}\qquad \begin{cases} dx_1 = (A_{11}x_1 + A_{12}x_2 + B_1u_t)\,dt + \xi G_1\,dw_t; \quad x_1(0) = x_{10}, \\ \epsilon dx_2 = (\epsilon^{1/2}A_{21}x_1 + A_{22}x_2 + B_2u_t)\,dt + \epsilon^{1/2}\xi G_2\,dw_t; \quad x_2(0) = x_{20}, \\ dy_1 = (C_{11}x_1 + C_{12}x_2)\,dt + \xi E_1\,dv_t; \quad y_1(0) = 0, \\ dy_2 = (\epsilon^{1/2}C_{21}x_1 + C_{22}x_2)\,dt + \epsilon^{1/2}\xi E_2\,dv_t; \quad y_2(0) = 0, \end{cases}$$

and the cost function is given as

$$\text{(106)}\qquad J_{I\theta}(\mu_I, \xi) = \lim_{t_f\to\infty} \frac{2\xi^2}{\theta t_f} \ln\left\{E\left\{\exp\left[\frac{\theta}{2\xi^2}\int_0^{t_f}(x'Qx + u'u)\,dt\right]\right\}\right\},$$

where the initial state x_0 is a Gaussian random variable with mean $\bar{x}_0$ and variance $\xi^2\Sigma_0$, and ξ is a small scalar parameter to be varied. We will again study the solution as the parameter $\xi \to 0$. This problem is equivalent to the one considered in §4, if we introduce the following substitutions:

$$\text{(107)}\qquad \theta \leftarrow \frac{\theta}{\xi^2}; \quad G_\epsilon \leftarrow \xi G_\epsilon; \quad E_\epsilon \leftarrow \xi E_\epsilon; \quad \Sigma_0 \leftarrow \xi^2\Sigma_0.$$

Define the quantity $\theta_I^*(\epsilon)$ exactly as in (61). Then, for any $\theta < \theta_I^*(\epsilon)$, the problem admits an optimal controller, given by (58a) and (58b).

The optimal solution to the full-order problem again depends on the value of ϵ explicitly. To obtain ϵ-free solutions, we decompose the system into slow and fast subsystems. The slow subsystem can be obtained in the same way as before. Under the substitution law (107), the slow LEQG problem is again described by (66), (67), and (68). This leads to a definition of θ_{Is} exactly as in (72). For any $\theta < \theta_{Is}$, the optimal controller for the slow subproblem is as in (73). The fast subsystem is again (75), under the substitution law (107). Thus, the fast controller is exactly the same as (78a)–(78b). Hence, we can form the slow and composite controllers μ_{Is}^* and μ_{Ic}^* as before. The slow controller μ_{Is}^* is as in (82a) and (82b), and the composite controller μ_{Ic}^* is as in (83), (84), and (85), of course now parametrized also by ξ.

Then, all this leads to the following corollary to Theorem 4.

COROLLARY 2. *For the ξ-parametrized singularly perturbed system* (105) *under the cost function* (106):

1. *For each $\epsilon > 0$, if the pairs (A_ϵ, B_ϵ) and (A_ϵ, G_ϵ) are controllable, the pairs (A_ϵ, C_ϵ) and (A_ϵ, Q) are observable, and the matrix N_ϵ is invertible, then, $\forall\theta < \theta_I^*(\epsilon)$, the optimal cost for the full-order LEQG problem can be written as*

$$\text{(108)}\qquad J_{I\theta}^*(\epsilon; \xi) = O(\xi^2).$$

2. *Let assumption* A2 *hold, the pairs (A_{11}, B_0), $(A_{11}, G_1G_1' - G_1G_2'(G_2G_2')^{-1}G_2G_1')$, and (A_{22}, B_2) be controllable, and the pairs (A_{11}, C_0), $(A_{11}, Q_{11} - Q_{12}Q_{22}^{-1}Q_{21})$, and (A_{22}, C_2) be observable. Then,*

(i) $\lim_{\epsilon\to 0^+}\theta_I^*(\epsilon) = \theta_{Is}$.

(ii) *$\forall\theta < \theta_{Is}$, if the composite controller $\mu_{Ic\theta}^*$ is applied to the system, then $\exists\epsilon_\theta' > 0$ such that $\forall\epsilon \in [0, \epsilon_\theta')$,*

$$\text{(109)}\qquad J_{I\theta}^c := J_{I\theta}(\mu_{Ic\theta}^*) = J_{I\theta}^*(\epsilon; \xi) + O(\xi^2\sqrt{\epsilon}).$$

(iii) *$\forall\theta < \theta_{Is}$, if, in addition, the matrix A_{22} is Hurwitz, and the slow controller $\mu_{Is\theta}^*$ is applied to the system, then $\exists\hat{\epsilon}_\theta > 0$ such that $\forall\epsilon \in [0, \hat{\epsilon}_\theta)$,*

$$\text{(110)}\qquad J_{I\theta}^s := J_{I\theta}(\mu_{Is\theta}^*) = O(\xi^2).$$

6. Examples. We present here three sets of numerical results, one for perfect state measurements and two for noisy state measurements. As stressed earlier, the quantities θ_s and θ_{Is} play important roles in the computation of approximate values for $\theta^*(\epsilon)$ and $\theta_I^*(\epsilon)$.

Example 1. Consider the system and cost function

$$\text{(111a)} \quad \begin{bmatrix} dx_1 \\ \epsilon dx_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} dt + \begin{bmatrix} 1 \\ 2 \end{bmatrix} u\, dt + \begin{bmatrix} 1 \\ \sqrt{\epsilon} \end{bmatrix} dw_t,$$

$$\text{(111b)} \quad J_\theta = \lim_{t_f \to \infty} \frac{2}{\theta t_f} \ln \left\{ \boldsymbol{E} \left\{ \exp \left[\frac{\theta}{2} \left(\int_0^{t_f} (2x_1^2 + 2x_1 x_2 + x_2^2 + u'u)\, dt \right) \right] \right\} \right\}.$$

By using a particular search algorithm, we can compute the quantity

$$\theta_s = 1.8892.$$

We next compute the maximum allowable θ-levels for the full-order system (111) for different fixed values of ϵ.

$\theta^*(\epsilon)$	1.5616	1.9842	1.9274	1.9019	1.8930	1.8903
ϵ	1	0.1	0.01	0.001	0.0001	0.00001

Note that as $\epsilon \to 0$, $\theta^*(\epsilon) \to \theta_s$.

Now, we choose $\theta = 1.6 < \theta_s$ and design the suboptimal controllers for the system based on this value of θ:

$$\mu_s^*(x_1) = -1.7633x_1, \quad \mu_c^*(x_1, x_2) = -3.3249x_1 - 0.61803x_2.$$

Then, we apply these controllers, μ_s^* and μ_c^*, to system (111) and obtain the corresponding performance levels J_s and J_c, respectively. These values are tabulated below along with the optimal cost levels, for different values of ϵ:

ϵ	1	0.1	0.01	0.001	0.0001	0.00001
$J^*(\epsilon)$	∞	3.0434	2.1560	1.9818	1.9350	1.9209
$J_c(\epsilon)$	∞	∞	2.1638	1.9828	1.9351	1.9209
$J_s(\epsilon)$	∞	∞	∞	2.2150	2.1366	2.1151

We also compute the cost level at $\epsilon = 0$,

$$J^*(0) = J_c(0) = 1.9146, \quad J_s(0) = 2.1056,$$

and observe that the composite controller asymptotically achieves the optimal performance level as $\epsilon \to 0$; however, the slow controller achieves only a suboptimal, but finite performance level asymptotically—all this being consistent with the result of Theorem 1. Also, the composite controller appears to be more robust than the slow one with respect to changes in the value of ϵ.

Example 2. Consider the setup

$$\text{(112a)} \quad \begin{bmatrix} dx_1 \\ \epsilon dx_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \sqrt{\epsilon} & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} dt + \begin{bmatrix} 1 \\ 2 \end{bmatrix} u\, dt + \begin{bmatrix} 1 \\ \sqrt{\epsilon} \end{bmatrix} dw_t,$$

$$\text{(112b)}\quad \begin{bmatrix} dy_1 \\ dy_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \sqrt{\epsilon} & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} dt + \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{\epsilon} \end{bmatrix} dv_t,$$

$$\text{(112c)}\quad J_{I\theta} = \lim_{t_f \to \infty} \frac{2}{\theta t_f} \ln \left\{ \boldsymbol{E} \left\{ \exp \left[\frac{\theta}{2} \int_0^{t_f} (2x_1^2 + 2x_1 x_2 + x_2^2 + u'u)\, dt \right] \right\} \right\}.$$

The maximum allowable θ-level for the slow subsystem is

$$\theta_{Is} = 1.4212.$$

We can also compute the maximum allowable θ-level, $\theta_I^*(\epsilon)$, for this system for different values of ϵ:

$\theta_I^*(\epsilon)$	1.4133	1.4189	1.4204	1.4209	1.4211
ϵ	0.001	10^{-4}	10^{-5}	10^{-6}	10^{-7}

Note again that as $\epsilon \to 0$, $\theta_I^*(\epsilon) \to \theta_{Is}$.

Now, we choose $\theta = 1 < \theta_{Is}$, and design the slow and composite controllers under the corresponding cost functions

$$\mu_{Is}^* = -\hat{x}_s^s, \quad \mu_{Ic}^* = -\hat{x}_s^c - 0.61803\hat{x}_f^c,$$

where

$$d\hat{x}_s^s = -5.6833\hat{x}_s^s\, dt + \begin{bmatrix} 0.44721 & 0.84721/\sqrt{\epsilon} \end{bmatrix} \begin{bmatrix} dy_1 + 2\hat{x}_s^s\, dt \\ dy_2 + 4\hat{x}_s^s\, dt \end{bmatrix},$$

$$\begin{bmatrix} d\hat{x}_s^c \\ \epsilon d\hat{x}_f^c \end{bmatrix} = \begin{bmatrix} -5.6833 & 2.0944/\sqrt{\epsilon} \\ -3.0902\sqrt{\epsilon} & -3.4721 \end{bmatrix} \begin{bmatrix} \hat{x}_s^c \\ \hat{x}_f^c \end{bmatrix} dt + \begin{bmatrix} 0.44721 & 0.84721/\sqrt{\epsilon} \\ 0 & 0.61803 \end{bmatrix} \begin{bmatrix} dy_1 + 2\hat{x}_s^c\, dt \\ dy_2 + 4\hat{x}_s^c\, dt \end{bmatrix}.$$

Then, we apply μ_{Is}^* and μ_{Ic}^* to system (112) and obtain the corresponding performance levels J_{Is} and J_{Ic}. They are tabulated below for different values of ϵ, along with the corresponding optimal cost levels.

ϵ	0.001	10^{-4}	10^{-5}	10^{-6}	10^{-7}
$J_I^*(\epsilon)$	3.9236	3.7409	3.6873	3.6707	3.6655
$J_{Ic}(\epsilon)$	∞	∞	3.6921	3.6710	3.6655
$J_{Is}(\epsilon)$	∞	3.9147	3.7691	3.7361	3.7390

We can also compute the optimal cost level at $\epsilon = 0$,

$$J_I^*(0) = J_{Ic}(0) = 3.6631, \quad J_{Is}(0) = 3.7361.$$

We see that the composite controller asymptotically achieves the optimal performance level as $\epsilon \to 0$. The slow controller, on the other hand, achieves a suboptimal but finite performance level asymptotically, which is again consistent with the statement of Theorem 4. We also note that in this case the composite controller is more sensitive than the slow one to changes in the value of ϵ. A possible explanation for this behavior is the following: since the quantity $J_{Is}(0)$ is very close to $J_{Ic}(0)$, this means there is little for the fast controller to do to improve the performance of the overall system. Furthermore, since the fast controller is an LQG design for the fast subsystem, the closed-loop fast subsystem under such a controller may not exhibit better H^∞ performance than the open-loop fast subsystem.

Example 3. Consider the following system and cost function:

$$\begin{bmatrix} dx_1 \\ \epsilon dx_2 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ \sqrt{\epsilon} & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} dt + \begin{bmatrix} 2 \\ 1 \end{bmatrix} u\, dt + \begin{bmatrix} 1 \\ 2\sqrt{\epsilon} \end{bmatrix} dw_t, \tag{113a}$$

$$\begin{bmatrix} dy_1 \\ dy_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2\sqrt{\epsilon} & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} dt + \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{\epsilon} \end{bmatrix} dv_t, \tag{113b}$$

$$J_{I\theta} = \lim_{t_f \to \infty} \frac{2}{\theta t_f} \ln \left\{ E \left\{ \exp \left[\frac{\theta}{2} \int_0^{t_f} (3x_1^2 + 2x_1x_2 + 2x_2^2 + u'u)\, dt \right] \right\} \right\}. \tag{113c}$$

The maximum allowable θ-level for the slow subsystem is

$$\theta_{Is} = 6.3756.$$

We can also compute the maximum allowable θ-level, $\theta_I^*(\epsilon)$, for this system, for different values of ϵ:

$\theta_I^*(\epsilon)$	6.8401	6.8836	6.5390	6.4289	6.3923
ϵ	0.01	0.0015	10^{-4}	10^{-5}	10^{-6}

Note again that as $\epsilon \to 0$, $\theta_I^*(\epsilon) \to \theta_{Is}$.

Now, we choose $\theta = 5 < \theta_{Is}$, and design the slow and composite controllers under the corresponding cost functions

$$\mu_{Is}^* = -0.87591\hat{x}_s^s, \quad \mu_{Ic}^* = -0.87591\hat{x}_s^c - 0.73205\hat{x}_f^c,$$

where

$$d\hat{x}_s^s = -5.5564\hat{x}_s^s\, dt + \begin{bmatrix} 0.010525 & 0.24449/\sqrt{\epsilon} \end{bmatrix} \begin{bmatrix} dy_1 + 0.87591\hat{x}_s^s\, dt, \\ dy_2 + 1.7518\hat{x}_s^s\, dt \end{bmatrix},$$

$$\begin{bmatrix} d\hat{x}_s^c \\ \epsilon d\hat{x}_f^c \end{bmatrix} = \begin{bmatrix} -5.5564 & 0.35796/\sqrt{\epsilon} \\ -23.191\sqrt{\epsilon} & -4.8552 \end{bmatrix} \begin{bmatrix} \hat{x}_s^c \\ \hat{x}_f^c \end{bmatrix} dt + \begin{bmatrix} 0.010525 & 0.24449/\sqrt{\epsilon} \\ 0 & 1.5616 \end{bmatrix} \begin{bmatrix} dy_1 + 0.87591\hat{x}_s^c\, dt \\ dy_2 + 1.7518\hat{x}_s^c\, dt \end{bmatrix}.$$

Then, we apply μ_{Is}^* and μ_{Ic}^* to system (113) and obtain the corresponding performance levels J_{Is} and J_{Ic}. They are tabulated below for different values of ϵ, along with the corresponding optimal cost levels.

ϵ	0.01	0.0015	10^{-4}	10^{-5}	10^{-6}
$J_I^*(\epsilon)$	4.6326	4.0662	3.9404	3.9210	3.9157
$J_{Ic}(\epsilon)$	∞	4.1528	3.9408	3.9210	3.9157
$J_{Is}(\epsilon)$	∞	∞	4.6194	4.5808	4.5710

We can also compute the cost level at $\epsilon = 0$,

$$J_I^*(0) = J_{Ic}(0) = 3.9134, \quad J_{Is}(0) = 4.5668.$$

Again the composite controller asymptotically achieves the optimal performance level as $\epsilon \to 0$, and the slow controller achieves a suboptimal but finite performance level asymptotically. In contradistinction with what was observed in Example 2, however, here the composite controller is less sensitive to changes in the value of ϵ than the slow one.

7. Conclusion. In this paper, we have presented a model reduction technique for the LEQG problem for linear singularly perturbed systems under perfect and imperfect state measurements. We have developed a time-scale decomposition procedure that breaks the full-order problem into appropriate slow and fast lower-order subproblems, the optimal solutions of which yield slow and fast controllers. When combined in an appropriate way, these lead to a composite controller under which the optimal performance level for the full-order system is achieved asymptotically as the singular perturbation parameter $\epsilon \to 0$. It has also been shown that when the fast subsystem is open-loop stable, the slow controller can achieve asymptotically some finite (but not optimal) performance level whenever the full-order problem admits a solution. In this case there is a clear positive gap between the asymptotic performance level a slow controller can achieve and the asymptotic performance level achieved by a full-order optimal controller, and the paper has provided a clear characterization of this performance loss. This indicates that there is a definite tradeoff between controller simplicity (due to model reduction) and loss of performance. In a large-deviations context, i.e., when the intensity of the noise in the system dynamics goes to zero, however, the slow controller can achieve asymptotically the performance level for the full-order system, provided that the fast subsystem is open-loop stable. Counterparts of these results in the finite-horizon case exist, and can be found in the internal report [13].

To obtain the optimal solution to the slow subsystem arrived at as a result of model reduction, it has turned out that one needs to develop a theory for the general LEQG problem with general cost structure (with cross terms between the state and control) and correlation between system and measurement noises. Since the LEQG problem has not been solved in the literature in such a full generality, we have also provided in the paper (§4) a clean and complete solution to this problem in both finite and infinite horizons via a different line from that of [3], which had addressed the finite-horizon case only, and under some restrictive assumptions. The solution obtained in §4 is precisely the central solution of the corresponding H^∞-optimal control problem [2], and the line of proof there would be useful even for the standard LEQG problem (i.e., without the cross term and with uncorrelated system and measurement noises) since it requires the least restrictive assumptions (leading to both necessary and sufficient conditions).

One possible nontrivial extension of the results of this paper would be the derivation of higher-order correction terms. The composite controller constructed in the paper achieves a performance level that is $O(\sqrt{\epsilon})$ close to the optimal one. This, however, may not be sufficient in some applications. Hence, high-order correction terms for the composite controller are of some interest. Another extension would be to the multiple time-scale problems, so as to obtain the counterparts of [16], which deals with the H^∞-optimal control problem. One other challenging extension would be to the nonlinear case, under both regular [17] and singular [18] perturbations.

REFERENCES

[1] T. BAŞAR, *Time consistency and robustness of equilibria in noncooperative dynamic games*, in Dynamic Policy Games in Economics, F. Van der Ploeg and A. de Zeeuw, eds., North-Holland, Amsterdam, 1989, pp. 9–54.

[2] T. BAŞAR AND P. BERNHARD, *H^∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, 2nd ed., Birkhäuser Boston, Cambridge, MA, 1995.

[3] A. BENSOUSSAN AND J. H. VAN SCHUPPEN, *Optimal control of partially observable stochastic systems with an exponential-of-integral performance index*, SIAM J. Control Optim., 23 (1985), pp. 599–613.

[4] J. H. CHOW AND P. V. KOKOTOVIĆ, *A decomposition of near-optimum regulators for systems with slow and fast modes*, IEEE Trans. Automat. Control, 21 (1976), pp. 701–705.

[5] W. H. FLEMING AND W. M. MCENEANEY, *Risk Sensitive Control and Differential Games*, Lecture Notes in Control and Info. Sci., Springer-Verlag, New York, 1992, pp. 185–197.

[6] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, vol. 25, Springer-Verlag, Berlin, New York, 1993.

[7] K. GLOVER AND J. C. DOYLE, *State-space formulae for all stabilizing controllers that satisfy an H^∞-norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[8] D. H. JACOBSON, *Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games*, IEEE Trans. Automat. Control, 18 (1973), pp. 124–131.

[9] M. R. JAMES, J. BARAS, AND R. J. ELLIOTT, *Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 780–792.

[10] L. A. LIUSTERNIK AND V. J. SOBOLEV, *Elements of Functional Analysis*, Ungar, New York, 1965.

[11] Z. PAN AND T. BAŞAR, *A Tight Bound for the H^∞ Performance of Singularly Perturbed Systems*, CSL Report, University of Illinois, Urbana, IL, March 1992.

[12] ———, *H^∞-optimal control for singularly perturbed systems. Part* I: *Perfect state measurements*, Automatica J. IFAC, 29 (1993), pp. 401–423.

[13] ———, *Model Simplification and Optimal Control of Stochastic Singularly Perturbed Systems Under Exponentiated Quadratic Cost*, Internal Report UILU-ENG-93-2249/DC-157, Decision and Control Lab., CSL, University of Illinois at Urbana–Champaign, November 1993.

[14] ———, *H^∞-optimal control for singularly perturbed systems. Part* II: *Imperfect state measurements*, IEEE Trans. Automat. Control, 39 (1994), pp. 280–299.

[15] ———, *H^∞-optimal control of singularly perturbed systems with sampled-state measurements*, in Advances in Dynamic Games and Applications, T. Başar and A. Haurie, eds., Birkhäuser Boston, Cambridge, MA, 1994, pp. 23–55.

[16] ———, *Multi-time scale zero-sum differential games with perfect state measurements*, Dynamics Control, 5 (1995), pp. 7–30.

[17] ———, *Robustness of minimax controllers to nonlinear perturbations*, J. Optim. Theory Appl., 87 (1995), pp. 631–678.

[18] ———, *Time-scale separation and robust controller design for uncertain nonlinear singularly perturbed systems under perfect state measurements*, Int. J. Robust Nonlinear Control, (1996), to appear.

[19] T. RUNOLFSSON, *The equivalence between infinite-horizon optimal control of stochastic systems with exponential-of-integral performance index and stochastic differential games*, IEEE Trans. Automat. Control, 39 (1994), pp. 1551–1563.

[20] V. R. SAKSENA AND T. BAŞAR, *Multimodeling, singular perturbations and stochastic decision problems*, in Advances in Control and Dynamic Systems, vol. XXII, C. T. Leondes, ed., Academic Press, New York, San Diego, CA, 1986, pp. 1–58.

[21] P. WHITTLE, *Risk-sensitive linear-quadratic-Gaussian control*, Adv. in Appl. Probab., 13 (1981), pp. 764–777.

[22] ———, *Risk-Sensitive Optimal Control*, John Wiley, Chichester, New York, 1990.

[23] ———, *Likelihood and cost as path integrals*, J. Roy. Statist. Soc. Ser. B, 53 (1991), pp. 505–538.

[24] ———, *A risk-sensitive maximum principle: The case of imperfect state observation*, IEEE Trans. Automat. Control, 36 (1991), pp. 793–801.

[25] P. WHITTLE AND J. KUHN, *A Hamiltonian formulation of risk-sensitive linear-quadratic-Gaussian control*, Internat. J. Control, 43 (1986), pp. 1–12.

[26] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1767–1780, September 1996

014

ANALYSIS AND OPTIMIZATION OF FEEDBACK CONTROL ALGORITHMS FOR DATA TRANSFERS IN HIGH-SPEED NETWORKS*

RAUF IZMAILOV†

Abstract. Two linear feedback control algorithms for handling and preventing congestion in high-speed networks are proposed and analyzed. The fluid approximation model is described with a continuous-time system of delay-differential equations. The algorithms are asymptotically stable, and the transient processes are nonoscillatory. The control parameters are locally optimal (optimality is based on the asymptotic rate of convergence). The results of numerical experiments suggest that these parameters are globally optimal as well. The dependence of the quality of service on the duration of the control intervals is analyzed, and the performance of algorithms in a nonstationary environment is addressed.

Key words. delay-differential equations, stability, feedback, transient process

AMS subject classifications. 93C30, 93D20

1. Introduction.

Asynchronous transfer mode (ATM) transport technology [7] is generally considered as a basis for future integrated telecommunications service. Since there would be an inevitable interaction and interference among users in the communication network, an increasing amount of research has been devoted to different control issues (see [1], [18], [19], [22], [23], [26] and their references). One of the basic problems arising here is the presence of propagation delays which pose a challenge for stability, since speed of data transmission in modern high-speed networks keeps increasing.

In most of the proposed algorithms and models (see [4], [5], [6], [11], [17], [25], and [27] and their references) control decisions are based on a single number (the deviation of the state of the system from the target value) or a single bit (the sign of such deviation). Analysis and numerical simulations [4], [10], [12], [9], [24] demonstrated that the stability of such algorithms in the presence of propagation delays has the form of bounded oscillations (occurring even in the deterministic setting).

The single number limitation on the number of control parameters appears to be the key obstacle to the elimination of these undesirable oscillations. As proved in [10], a large class of feedback algorithms based on a single number always has an unstable equilibrium. Thus it seems natural to address the question of what additional parameters should be considered and how to translate them into control algorithms. However, this question has begun to be addressed only recently [3], [8], [15], [16].

We start this paper with a description of a simple fluid flow model describing a single ATM connection. For this model we consider two feedback control algorithms proposed in the previous report [15]: the first algorithm uses two control parameters, whereas the second one uses three control parameters. The control decision is based on the system states separated with constant control time intervals. The closed-loop system is described by differential-difference equations (see [2] and [20] and their references). We formulate results about stability and local optimality of the algorithms (all proofs are in the appendix). In the next section we analyze the performance of the algorithms: global optimality, asymptotic rates of convergence, optimal duration of the control intervals, transient regimes, and performance in nonstationary conditions. In particular, we demonstrate that faster asymptotic convergence (which could be obtained by decreasing the duration of the control intervals) may lead to worse transient processes and, in the nonstationary case, to worse frequency response. In the

*Received by the editors November 4, 1994; accepted for publication (in revised form) June 29, 1995.

†NEC USA, Inc., C&C Research Laboratories, 4 Independence Way, Princeton, NJ 08540 (rauf@ccrl.nj.nec.com).

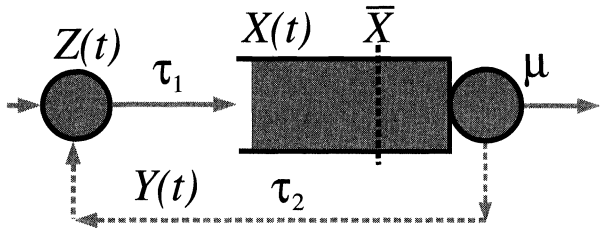


FIG. 1. *Single connection.*

conclusion we present a brief summary of results, recommendations on a proper choice of algorithm, and an outline of future directions of research.

2. Basic model. Consider a single connection (Figure 1) between a source controlled by an access regulator and a distant node served with a constant transmission capacity μ. The traffic source is monitored and regulated by the access regulator, and the distant bottleneck node sends back the information on its congestion status, defined as the difference between the current buffer contents and the target value (a fixed threshold) $\overline{X}$.

In order to describe large data transfers and isolate the issue of control mechanism from other considerations, assume (as in [10]) that there is an infinite amount of traffic to be sent to the remote node. In order to capture the small size of ATM packets as well as high rates of the network, we approximate the traffic by fluid flows. This assumption is not restrictive, and the basic results of our analysis could be extended to the discrete version of the model. The access regulator controls the current input rate $Z(t)$, basing its decisions on the buffer contents $X(t)$ of the distant node, which is continually monitored by the source. A target value $\overline{X}$ of the remote buffer contents is fixed: if $X(t) > \overline{X}$, the node is considered congested. The propagation delays from the source to the bottleneck and back are τ_1 and τ_2, which add up to the round-trip delay $\tau = \tau_1 + \tau_2$. The control objective is to adapt $Z(t)$ to μ dynamically while keeping $X(t)$ at an acceptable level.

In the next section we present two linear feedback algorithms. Each algorithm controls the source rate $Z(t)$ and varies it in proportion (determined by two or three gain parameters) to the differences between the buffer contents $X(t)$ and the target value $\overline{X}$.

3. Control algorithms. The first algorithm takes into account the deviations of $X(t)$ from the target value $\overline{X}$ during two consecutive time slots, separated by the control time interval r. (In [15] the control interval r was equal to the round-trip delay τ.) These deviations are weighted with linear gain parameters a and b, so in a neighborhood of the threshold $\overline{X}$ the system evolution is described by

$$\begin{cases} X'(t) &= Z(t-\tau_1) - \mu, \\ Z'(t) &= -a(X(t-\tau_2) - \overline{X}) - b(X(t-\tau_2 - r) - \overline{X}). \end{cases}$$

We take the derivative of the first equation here and substitute $Y(t) = X(t) - \overline{X}$ to obtain the delay-differential equation in the normalized time scale $T = t/\tau$ (where $A = a\tau^2$, $B = b\tau^2$, $R = r/\tau$):

$$Y''(T) + AY(T-1) + BY(T-1-R) = 0.$$

Its characteristic equation is

$$f(z) = z^2 e^{(R+1)z} + Ae^{Rz} + B = 0,$$

which has an infinite number of roots λ_i. The location of these roots on the complex plane determines [2, Thm. 6.7] the asymptotic behavior of $Y(T)$. In particular, the degree of stability

$\lambda = \sup_i\{\Re\lambda_i\}$ guarantees the asymptotic convergence of $Y(T)$ with the exponential rate λ: $|Y(T)| \le Ke^{-\lambda T}$ for some K [20, Chap. 3, Thm. 2.1]. The location of these roots on the complex plane determines [2] the behavior of $Y(t)$ around the equilibrium point 0.

THEOREM 1. *Denote*

$$\begin{cases} V &= \dfrac{\sqrt{R^2+2R+2}-R-2}{1+R}, \\ A^* &= -2\exp(V)\dfrac{1+R-\sqrt{2+2R+R^2}}{R}, \\ B^* &= 2\exp(V(1+R))\dfrac{1-1\sqrt{2+2R+R^2}}{R(1+R)^2}. \end{cases}$$

Then V is the degree of stability of $f(z)$, and any small deviation of A and B from A^* and B^* decreases the degree of stability of $f(z)$.

It may seem surprising that $B^*(R) < 0$ since it increases the rate $Z(t)$ when the buffer contents $X(t-\tau_2-r)$ exceeds $\overline{X}$ and vice versa. One may view this effect as a counterbalance dampening the oscillations generated by the regular feedback with the positive coefficient $A^*(R)$.

The second algorithm takes into account the deviations of $X(t)$ from the target value $\overline{X}$ during two consecutive time slots, separated by the time interval $r/2$. These deviations are weighted with linear gain parameters a, b, and c, so in a neighborhood of the threshold $\overline{X}$ the system evolution is described by

$$\begin{cases} X'(t) &= Z(t-\tau_1)-\mu, \\ Z'(t) &= -a(X(t-\tau_2)-\overline{X})-b(X(t-\tau_2-r/2)-\overline{X})-c(X(t-\tau_2-r)-\overline{X}). \end{cases}$$

We take the derivative of the first equation here and substitute $Y(t) = X(t) - \overline{X}$ to obtain the delay-differential equation in the normalized time scale $T = t/\tau$ (where $A = a\tau^2$, $B = b\tau^2$, $C = c\tau^2$, $R = r/\tau$):

$$Y''(T) + AY(T-1) + BY(T-1-R/2) + CY(T-1-R) = 0.$$

Its characteristic equation is

$$f(z) = z^2e^{(R+1)z} + Ae^{Rz} + Be^{\frac{Rz}{2}} + C = 0,$$

which has an infinite number of roots (eigenvalues). The location of these roots on the complex plane determines [2] the behavior of $Y(t)$ around the equilibrium point 0.

THEOREM 2. *Denote*

$$\begin{cases} W = \dfrac{-6-6R-R^2+\sqrt{12+24R+18R^2+6R^3+R^4}}{2+3R+R^2}, \\ A^* = \dfrac{2e^W\left(-2-3R-R^2+\sqrt{12+24R+18R^2+6R^3+R^4}\right)}{R^2}, \\ B^* = \dfrac{16e^{\frac{W(2+R)}{2}}\left(2+2R-\sqrt{12+24R+18R^2+6R^3+R^4}\right)}{(2+R)^2R^2}, \\ C^* = \dfrac{2e^{W(1+R)}\left(-2-R+\sqrt{12+24R+18R^2+6R^3+R^4}\right)}{R^2(1+R)^2}. \end{cases}$$

Then W is the degree of stability of $f(z)$, and any small deviation of A, B, and C from A^*, B^*, and C^* decreases the degree of stability of $f(z)$.

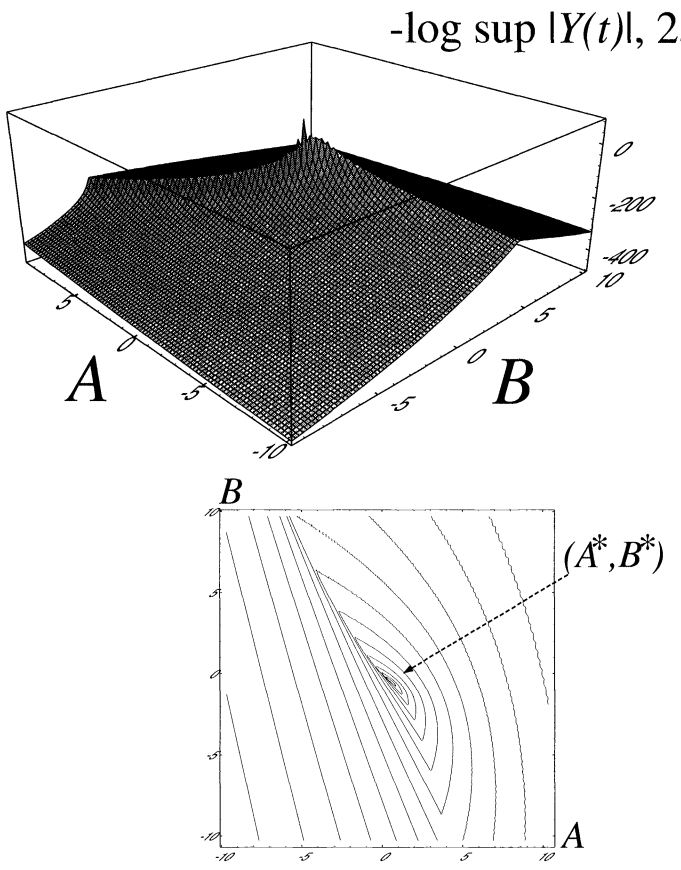


FIG. 2. *Numerical optimization.*

4. Performance analysis. Since the principal roots V and W for both algorithms are negative and real, the choice of damping parameters $A = A^*$, $B = B^*$, and $C = C^*$ suggested in the theorems of the previous section guarantees the exponential and nonoscillatory convergence of algorithms.

The structure of the proof also demonstrates the robustness of algorithms to small uncertainties in the knowledge of the round-trip delay. In other words, if the algorithms are constructed on the assumption that the round-trip delay is τ that whereas the actual round-trip delay is $\tau^* \neq \tau$, then the first algorithm has the degree of stability V^* (the second algorithm the has degree of stability W^*), where $V^* \to V$ as $\tau^* \to \tau$ (for the second algorithm, $W^* \to W$ as $\tau^* \to \tau$). The exponential stability of the control algorithms means the discrete versions of the algorithms are also stable.

The theorems also state the control coefficients are locally optimal. Extensive numerical calculations have been carried out, and the results obtained so far suggest the gain parameters described in Theorems 1 and 2 are globally optimal as well. In particular, to analyze the global optimality of $A = A^*$ and $B = B^*$ in Theorem 1, the number $S = \sup_{250<T<300} |Y(T)|$ has been calculated for each pair A and B in the square lattices $A^* - 6/R \leq A \leq A^* + 6/R$, $B^* - 6/R \leq B \leq B^* + 6/R$ of 300×300 points with the initial condition $Y(T) = T$ for $0 \leq T \leq 1$. The data were used to plot $d = -\log S$ as a function of A and B for $R = 1,\ 0.1,\ 0.01$. For all R considered, the function d has the single maximum (a typical picture for $R = 1$ is shown in Figure 2) corresponding to the single optimal point $A = A^*$, $B = B^*$. According

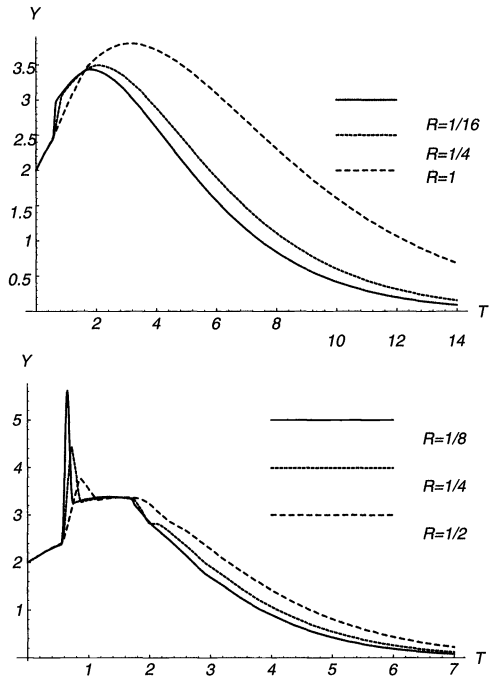


FIG. 3. *Transient processes for the first algorithm (upper graph) and for the second algorithm (lower graph) under impulse initial disturbance.*

to the definition of d, the pair (A^*, B^*) minimizes the sum S, which translates into the fastest convergence of $Y(T)$ to zero. Similar results were obtained for other target values S with different initial conditions and different time boundaries T. All this suggests that the pair (A^*, B^*) is indeed the global optimum for all R. Similar numerical calculations suggest that the triple (A^*, B^*, C^*) is the global optimum for any R.

The formulas of theorems show that the shorter the control time interval R, the better the asymptotic performance (degree of stability). If $R \to 0$, the asymptotic rate of convergence V of the first algorithm tends to $\sqrt{2} - 2 = -0.5857\ldots$ and the asymptotic rate of convergence W of the second algorithm tends to $\sqrt{3} - 3 = -1.26795\ldots$. Small control time intervals R require large gain parameters. As the control interval R decreases, the absolute values of gain parameters increase as R^{-1} for the first algorithm and as R^{-2} for the second one. The different behavior of gain parameters for the control algorithms leads to different transient performance. To illustrate it, consider the transient reaction of two algorithms on a unit jump (delta-function) disturbance. As shown in Figure 3, the first algorithm "absorbs" the jump for all values of control time intervals R uniformly, whereas the second algorithm exhibits sharp deterioration of the transient behavior (short and large peaks) for small R.

Suppose now the server rate μ is not constant: $\mu = \mu(t)$. Then [2] the function $Y(T)$ tends (as $T \to \infty$) to the stabilized solution

$$y(T) = \int_0^T h(T - z)\mu(z)\, dz, \tag{1}$$

where

$$h(s) = \frac{s}{s^2 + Ae^{-s} + Be^{-s(1+R)}}.$$

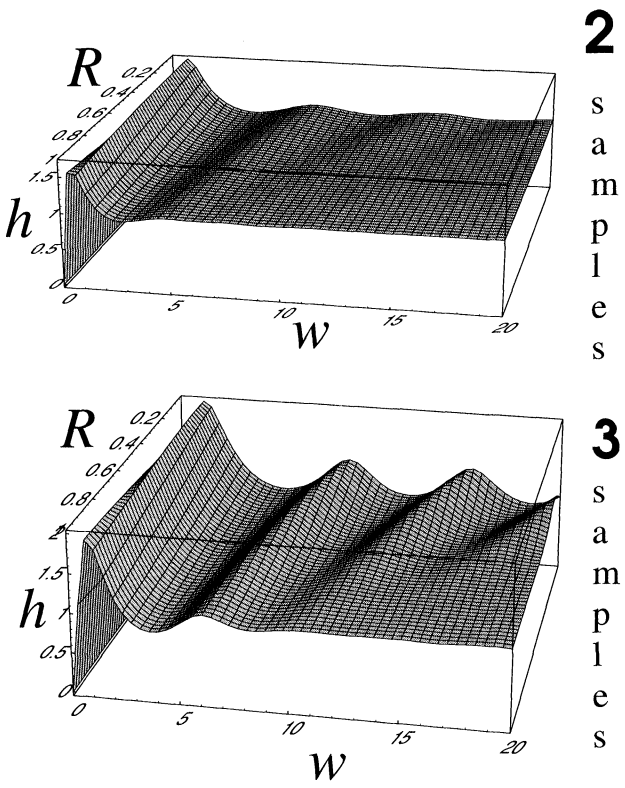


FIG. 4. *Frequency responses.*

Similarly, the stabilized reaction for the second algorithm has the form (1) with the function

$$h(s) = \frac{s}{s^2 + Ae^{-s} + Be^{\frac{-s(1+R)}{2}} + Ce^{-s(1+R)}}.$$

These integral representations give an opportunity to analyze the stabilized reaction of the proposed algorithms on variable (deterministic or random) server rate $\mu(t)$.

Since the server rate $\mu(T)$ is bounded and the model is linear, it is helpful to analyze the reaction on the harmonically changing server rate. If $\mu(T) = e^{iwT}$, then the reaction $Y(T)$ tends to $h(iw)e^{iwT}$. The absolute value $|h(iw)|$ describes the ratio of the oscillations of the buffer occupancy versus the server rate oscillations. Figure 4 displays $|h(iw)|$ for $0 \leq w \leq 20$ and $0 < R < 1$ for both algorithms.

All these observations indicate that the first algorithm has certain advantages over the second one. Although in the case of the fixed server rate, three control parameters lead to a better asymptotic convergence (the more control information is available, the better asymptotic properties could be achieved), the two-parameter scheme gives more robust transient performance as well as better performance in a nonstationary environment: since the second algorithm has large gain coefficients of order R^{-2}, it underperforms the first algorithm with smaller gain coefficients of order R^{-1} when the conditions change (transients and/or changing server rate). Similar effects (where large gain parameters, while being beneficial for the asymptotic behavior, lead to unsatisfactory transient regimes) have been observed in other control systems as well (see, for example, [14], [28], and their references).

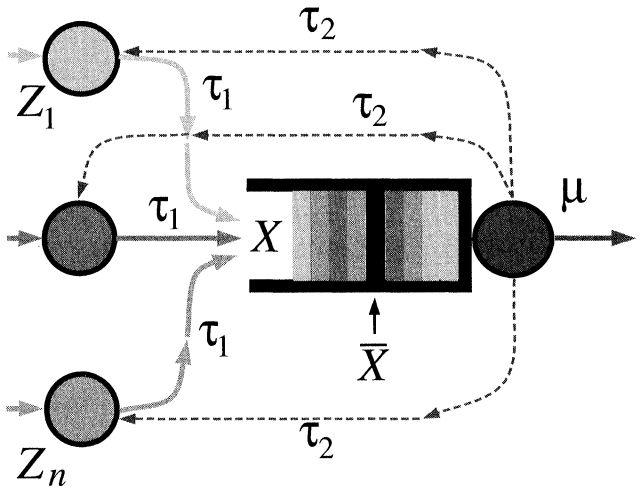


FIG. 5. *Dynamic access node.*

We conclude our analysis by noting that both control algorithms are applicable to the case where multiple connections share the bottleneck node (Figure 5) and the propagation delays are the same for each connection. For example, the first algorithm could be modified to

$$
(2)\qquad \begin{cases} X'(t) = \sum_i Z(t-\tau_1) - \mu, \\ Z_1'(t) = -(a/n)(X(t-\tau_2) - \overline{X}) - (b/n)(X(t-\tau_2-r) - \overline{X}), \\ \vdots \\ Z_n'(t) = -(a/n)(X(t-\tau_2) - \overline{X}) - (b/n)(X(t-\tau_2-r) - \overline{X}). \end{cases}
$$

Denoting $Z(t) = \sum_i Z_i(t)$, we can describe the behavior of algorithm (2) by the same formulas as in the single connection case. The stability properties of the new algorithm (2) are identical to those obtained for single connection case. The common asymptotic rate of convergence for all connections is determined by the principal root of the characteristic equation associated with the global system. The same generalization is applicable to the second algorithm.

5. Conclusion. We analyzed two linear feedback control algorithms and their convergence and optimality. The problems of the optimal choice of control intervals were addressed. We also analyzed the robustness of the algorithms for nonstationary server rate and performance of transient regimes. The results obtained suggest that the two-control parameter scheme is probably better than the three-parameter one: although it has slightly worse asymptotic properties, it has much better transient response as well as a better rate of convergence to changing server rate. It would be important to continue the analysis of these issues. In particular, the next challenging question to be addressed is to extend these algorithms to control of traffic mixture with different propagation delays.

6. Appendix.

6.1. Proof of Theorem 1. Since $f(V) = f'(V) = f''(V) = 0$ for $A = A^*$ and $B = B^*$, then V is a triple root of $f(z)$. To prove that all other roots have lesser real parts than V, the following three steps are accomplished. First we prove that there are exactly three roots in the domain $Q = \{z : |\Im z| \le 2\pi/(1+R)\}$. Next we prove that there are no roots of $f(z)$ in the domain $P = \{z : \Re z \ge -0.6,\ |z| \ge 2\pi/(1+R)\}$. Since the triple root V belongs to Q, that

would mean the absence of other roots of $f(z)$ in Q. The final step is to combine two previous steps and to conclude that there are no zeros (but V) in the domain $R = \{z : |\Re z| \geq -0.6\}$.

The first step uses theory of rotation of planar vector fields [21]. Consider the rotation of the vector field $(\Re f(z), \Im f(z))$ on the boundaries of rectangular domains

$$Q_N = \{z : |\Im z| \leq 2\pi/(1+R),\ |\Re z| \leq N\}, \qquad N \to \infty.$$

As $N \to \infty$, $e^{(1+R)z}z^2$ becomes the dominant member in $f(z)$ on the right vertical boundary of Q_N and, as such, determines the rotation of $(\Re f(z), \Im f(z))$ on the segment $(x = N,\ y \in [-2\pi/(1+R), 2\pi/(1+R)])$. The rotation of $e^{(1+R)z}z^2$ on this segment tends (as $N \to \infty$) to the rotation of $e^{(1+R)z}$, which is 4π. On the left vertical boundary, $-B$ is the dominant member, so the rotation tends to zero as $N \to \infty$. On the upper boundary (where $y \equiv 2\pi/(1+R)$) the function $g(x) = f(x+iy)$ is equal to

$$\left(e^{(1+R)x}\left(x^2 - \frac{4\pi^2}{(1+R)^2}\right) + Ae^{Rx}\cos\left(\frac{2\pi R}{1+R}\right) + B\right)$$
$$+ i\left(\frac{4\pi x}{1+R}e^{(1+R)x} + A\sin\left(\frac{2\pi R}{1+R}\right)e^{Rx}\right).$$

The rotation on the upper boundary of Q_N tends to π as $N \to \infty$: in order to prove it, it is sufficient to show that $\Re g(x) < 0$ for $x \leq 0$ (since $\Im g(x) > 0$ for $x \geq 0$ and the quarter $\{\Re g(x), \Im g(x)\}$ is not visited by the point $g(x)$, which would imply that the rotation is defined correctly $(g(x) \neq 0)$ and is equal to π).

To prove that $\Re g(x) < 0$ for $x \leq 0$, we consider two cases: case α, $1/3 \leq R \leq 1$; case β, $0 \leq R < 1/3$.

Consider case α. If $-2\pi/(1+R) \leq x \leq 0$, then all three components of $\Re g(x)$ are negative. Let $-\infty \leq x \leq -2\pi/(1+R)$. Then $\exp((1+R)x)(x^2 - 4\pi^2/(1+R)^2) < \exp((1+R)x)x^2$, and the latter function has its maximum at $x = -2/(1+R)$ at the considered interval. The first member of $\Re g(x)$ is less than $e^{-2\pi}4\pi^2/(1+R)^2 < 0.042$ (since $R \geq 1/3$). The second member of $\Re g(x)$ is negative, and the third member is less than -0.28 (since $R \leq 1$). Hence $\Re g(x)$ is negative.

Consider case β and its three subcases: case (i), $-2\pi/(1+R) < x < -\pi/(1+R)$; case (ii), $-\infty < x \leq -2\pi/(1+R)$; case (iii), $-\pi/R \leq x \leq 0$.

In case (i) the first member in $\Re g(x)$ is negative, whereas the sum of other two members is less than $A\exp(-\pi R/(1+R))\cos(2\pi R/(1+R)) + B$, which is less than -0.98 on $1/3 \leq R \leq 1$. In case (ii) the first member in $\Re g(x)$ is less than 0.041, whereas the sum of other two members is less than $A\exp(-2\pi R/(1+R))\cos(2\pi R/(1+R)) + B$, which is less than -0.919 on $1/3 \leq R \leq 1$. In case (iii) the sum of the second and third members is less than $A\cos(2\pi R/(1+R)) + B$, which is less than -0.919 on $1/3 \leq R \leq 1$, whereas the first member is less than $e^{-\pi}(3\pi^2/(1+R)^2) < -0.72$.

The same analysis is applicable for the lower boundary ($y \equiv \pi$): the rotation there tends to π.

Therefore, the rotation of $(\Re f(z), \Im f(z))$ on the boundaries of Q_N is 6π for sufficiently large N. Hence $f(z)$ has three zeros (counted with their multiplicity) inside Q.

The second step is proven as follows. For $1/3 \leq R \leq 1$ and $x \geq -0.6$, $y \geq 2\pi/(1+R)$ we have

$$|e^{(1+R)z}z^2| = |e^{Rz}||e^z||z^2| \geq 5.416|e^{Rz}|, \quad |Ae^{Rz}| \leq 1.213|e^{Rz}|.$$

Hence $|e^{(1+R)z}z^2 + Ae^{Rz}| \geq (5.41 - 1.22)|e^{Rz}| \geq 4.19|e^{Rz}| > 3 > |B|$. Assume further that $0 < R < 1/3$. Since $|AR| < 1/2$ and $|BR| < 1/2$ for $0 < R \leq 1/3$, then

$$|Ae^{Rz} + B| \leq |Ae^{Rz}| + B \leq |e^{Rz}|\frac{0.7}{R} + \frac{0.7}{R} < |e^{Rz}|\frac{0.7}{R} + |e^{Rz}|\frac{0.7}{R}e^{0.6} \leq 2\frac{|e^{Rz}|}{R}$$

on the domain $\{x \geq -0.6,\ y \geq \pi/R\}$. (The last relations follow from the inequality $|e^{Rz}| \geq e^{-0.6}$.) On the other hand, $|e^{(R+1)z}z^2| = |e^{Rz}||z^2||e^z| \geq e^{-0.6}|e^{Rz}|\pi^2/R^2 \geq |e^{Rz}|5.5/R^2$, which exceeds the previous expression for $R \leq 1$. This means there are no roots in the domain $\{x \geq -0.6,\ y \geq \pi/R\}$.

Consider now the domain $D = \{N \geq x \geq -0.6,\ 2\pi/(1+R) \leq y \geq \pi/R\}$. It is sufficient to prove that the rotation of f on the boundaries is the same as the rotation of the function $e^{(R+1)z}z^2$ (which has no zeros inside D).

Consider all four boundaries of D separately. On the upper boundary $|e^{(R+1)z}z^2| > |Ae^{Rz} + B|$ since the right part is more than $|e^{Rz}|0.44\pi^2/R^2$, whereas the left part is less than $|e^{Rz}|$. On the right boundary $|e^{(R+1)z}z^2| \gg |Ae^{Rz} + B|$ for sufficiently large N. On the left boundary $|e^{(R+1)z}z^2| \geq |y|^2$, which is larger than

$$\begin{aligned} |Ae^{Rz} + B| &\leq 1.5\max\{B + A\cos Ry,\ A\sin Ry\} \\ &\leq 1.5\max\{0.3 + 0.2Ry^2,\ 0.5y\} \leq 0.45 + 0.3y^2 \end{aligned}$$

for $|y| \geq 2\pi/(1+R)$ and $R < 1/3$. (Since $(B+A)R \to 0$ as $R \to 0$, then $h(R, y) = B + Ae^{Rx}\cos(Ry)$ is less than $0.242 + 0.2y^2$ (for $R \leq 1/3$ and $y < \pi/R$).) Finally, on the lower boundary for $x = -0.6$ the same analysis as for the left boundary shows that $|e^{(R+1)z}z^2| > |Ae^{Rz} + B|$; for $x = N$ this follows from the analysis of the right boundary. Since (see the first step of the proof) both $(\Re f, \Im f)$ and $(\Re g, \Im g)$, where

$$g(x) = \left(e^{(1+R)x}\left(x^2 - \frac{4\pi^2}{(1+R)^2}\right)\right) + i\left(\frac{4\pi x}{1+R}e^{(1+R)x}\right),$$

do not belong to the quarter $\{\Re Z > 0, \Im Z < 0\}$, a linear homotopy exists between f and g. Therefore, the final step is completed, and V is the degree of stability of $f(z)$.

To prove the local optimality of A^* and B^*, denote $A = A^* + \alpha$, $B = B^* + \beta$, and consider the Taylor series $f(z) = \sum_i f_i(z - V)^i$:

$$\begin{aligned} f_0 &= \beta + \alpha e^{RV},\ f_1 = \alpha R e^{RV},\ f_2 = \frac{\alpha R^2 e^{RV}}{2}, \\ f_3 &= \frac{\alpha R^3 e^{RV}}{6} + \frac{\sqrt{2+2R+R^2}e^{(R+1)V}}{3}, \\ f_4 &= \frac{\alpha R^4 e^{RV}}{24} + \frac{(1+R+(2+3R)\sqrt{2+2R+R^2})e^{(R+1)V}}{12}. \end{aligned}$$

Then the function $g(s) + \delta(s)s^4 = (g_0 + \delta(s))s^4 + g_1 s^3 + g_2 s^2 + g_3 s + g_4$ has the same roots as $f(z - V)$, where $g_i = f_{4-i}$ $(i = 0, \ldots, 4)$ and $\delta(s) = o(1)$.

Let $\varepsilon < 0.001$. If $\alpha = \beta = 0$, then $\max_U |\delta(s)| < \varepsilon/2$, $\min_{\partial U} |g(s)/s^4| > 2\varepsilon$, $\min_U(g_0 + \delta(s)) > g_0/2$ for sufficiently small ε_0, where ∂U is the boundary of $U = \{z : |z| \leq \varepsilon_0\}$. Then for sufficiently small $\varepsilon_1 < \varepsilon$ (i) $\max_U |\delta(s)| = \varepsilon_2 < \varepsilon$, (ii) $\min_{\partial U} |g(s)/s^4| > \varepsilon$, (iii) $\min_U(g_0 + \delta(s)) > 0$ for any $|\alpha|, |\beta| < \varepsilon_1$. Parts (i) and (ii) imply that $g(s)$ and $g(s) + \delta(s)s^4$ have the same rotation on δU and the same number of roots inside U for any $|\alpha|, |\beta| < \varepsilon_1$. Part (ii) implies that any polynomial $G_\delta(s) = g(s) + \delta s^4$ (where $|\delta| < \varepsilon_2$) has three roots in U and a negative root outside of U for any $|\alpha|, |\beta| < \varepsilon_1$. Part (iii) implies $G_\delta(s)$ has no roots in the right half-plane (RHP) iff all its Hurwitz determinants $\Delta_1 = g_1$, $\Delta_2 = g_1 g_2 - (g_0 + \delta)g_3$, $\Delta_3 = g_1 g_2 g_3 - (g_0 + \delta)g_3^2 - g_4 g_1^2$, $\Delta_4 = g_4\Delta_3$ are positive (Routh–Hurwitz criterion, [13, Chap. XV, Thm. 4]).

If $g_4 = 0$, then one of the roots of $G_\delta(s)$ is zero and one of the Hurwitz determinants $\Delta_2 = g_1 g_2 - (g_0 + \delta)g_3$, $\Delta_3 = \Delta_2 g_3$ of the polynomial $G_\delta(s)/s$ is negative for $|\delta|, |\alpha|, |\beta| < \varepsilon$,

which means that $G_\delta(s)$ has a root in the RHP. If $g_4 \neq 0$, then $G_\delta(s)$ has a root in the RHP both for $g_4 < 0$ (since $\Delta_4 = g_4\Delta_3$) and for $g_4 > 0$ (since

$$\Delta_3 = -g_4 g_1^2 + \frac{\alpha^2 R^2 e^{2RV}}{24} \times \left(-24\delta + \alpha R^4 e^{RV} - 2e^{(1+R)V}(1 + R + (2+R)\sqrt{2+2R+R^2})\right) < 0$$

for $|\delta|, |\alpha|, |\beta| < \varepsilon < 0.001$).

Fix α and β. Since the sign of any of the Hurwitz determinants of the functions $G_\delta(s)$ does not depend on δ for all $|\delta| < \varepsilon$ (Δ_1 does not depend on δ, the sign of

$$\Delta_2 = \frac{\alpha R e^{RV}}{24}\left(-24\delta + \alpha R^4 e^{RV} - 2e^{(1+R)V}(1 + R + (2+R)\sqrt{2+2R+R^2})\right)$$

does not depend on δ for $|\delta| < \varepsilon$, $\Delta_3 < 0$ for $|\delta|, |\alpha|, |\beta| < \varepsilon$, and the sign of $\Delta_4 = g_4\Delta_3$ does not depend on δ since g_4 does not depend on δ), then [13, Chap. XV, Thm. 5] $G_\delta(s)$ has the same number of roots in the RHP for any $|\delta| < \varepsilon$. As δ changes from $-\varepsilon$ to ε, the roots of $G_\delta(s)$ in the RHP form the set R (separated from the imaginary axis), and the rest of the roots form the set L. The continuous homotopy $g(s) + \gamma\delta(s)s^4$ (for $\gamma \in [0, 1]$) preserves the rotation on ∂U and continuously moves the roots of $g(s) + \delta(s)s^4$ to the roots of $g(s)$. For any $\gamma \in [0, 1]$ the roots of $g(s)$ in U belong to the set $R \cup L$ (if $r \in U$ is a root of $g(s) + \gamma\delta(s)s^4$, then it is also a root of $G_{\gamma\delta(r)}(s)$: since $\delta(s)$ is a regular function, $|\delta(r)| < \varepsilon$). Since $g(s)$ has a root $r \in R$ for $\gamma = 0$, then the homotopic prototype of r for $\gamma = 1$ also belongs to R. This completes the proof of Theorem 1.

6.2. Proof of Theorem 2. The proof of Theorem 2 closely follows the proof of Theorem 1.

Since $f(W) = f'(W) = f''(W) = f'''(W) = 0$ for $A = A^*$, $B = B^*$ and $C = C^*$, then W is a quadruple root of $f(z)$. To prove that all other roots have lesser real parts than V, the following three steps are accomplished. First, we prove that there are exactly four roots in the domain $Q = \{z : |\Im z| \leq 2\pi/(1 + R)\}$. Next, we prove that there are no roots of $f(z)$ in the domain $P = \{z : \Re z \geq -1.268,\ |z| \geq 2\pi/(1 + R)\}$. Since the quadruple root W belongs to Q, that would mean the absence of other zeros in Q. The final step is to combine two previous steps and to conclude that there are no zeros (but W) in the domain $R = \{z : |\Re z| \geq -1.268\}$.

The first step uses theory of rotation of planar vector fields [21]. We will analyze the rotation of the vector field $(\Re f(z), \Im f(z))$ on the boundaries of rectangular domains

$$Q_N = \{z : |\Im z| \leq 2\pi/(1 + R),\ |\Re z| \leq N\}, \qquad N \to \infty.$$

Consider the right vertical boundary of Q_N. For $N \to \infty$ the member $e^{(1+R)z}z^2$ becomes the dominant member in $f(z)$ and, as such, determines the rotation of $(\Re f(z), \Im f(z))$ on the segment $(x = N,\ y \in [-2\pi/(1+R), 2\pi/(1+R)])$. The rotation of $e^{(1+R)z}z^2$ on this segment tends (as $N \to \infty$) to the rotation of $e^{(1+R)z}$, which is 4π. On the left vertical boundary, $-B$ is the dominant member, so the rotation tends to zero as $N \to \infty$. On the upper boundary (where $y \equiv 2\pi/(1 + R)$) the function $f(x + iy) = g_1(x) + ig_2(x)$ has the form

$$g_1(x) = C + e^{(1+R)x}\left(\frac{-4\pi^2}{(1+R)^2} + x^2\right) + Be^{\frac{Rx}{2}}\cos\left(\frac{\pi R}{1+R}\right) + Ae^{Rx}\cos\left(\frac{2\pi R}{1+R}\right),$$

$$g_2(x) = \frac{4e^{(1+R)x}\pi x}{1+R} + Be^{\frac{Rx}{2}}\sin\left(\frac{\pi R}{1+R}\right) + Ae^{Rx}\sin\left(\frac{2\pi R}{1+R}\right).$$

The rotation tends to 2π as $N \to \infty$: to prove it, it is sufficient to show that (i) $\Im g(x) < 0$ for $x \le -1$; (ii) $\Im g(x) > 0$ for $x > 1$; (iii) $\Re g(x) < 0$ for $-1 \le x \le 1$. That will imply that $g(x) \neq 0$ for all x (and the rotation is defined correctly) and the rotation is indeed equal to 2π (given the behavior of $g_1(x) + ig_2(x)$ on infinity).

The case (i) is verified by direct analysis.

Consider case (ii). If $R < 1/3$, then the statement is verified by direct analysis. Otherwise, since $|A|, |B| \le 7$ for $R > 1/3$, we then have that the first member is larger than $17e^{Rx}$ and the third member is positive, whereas the second member is less than $15e^{Rx}$.

In case (iii) $e^{(1+R)x}(x^2 - 4\pi^2/(1+R)^2) \le -1.2$ for $-1 \le x \le 1$. If $R > 1/3$, then the last member in

$$q = C + Be^{\frac{Rx}{2}} \cos\left(\frac{\pi R}{1+R}\right) + Ae^{Rx} \cos\left(\frac{2\pi R}{1+R}\right)$$

is negative, and it is sufficient to prove that $Q = C + Be^{Rx/2}\cos(\pi R/(1+R)) \le 1$, which is verified directly. In the same way we verify that q is negative for $R < 1/3$.

The same analysis is applicable for the lower boundary ($y \equiv \pi$): the rotation there tends to π.

Therefore, the rotation of $(\Re f(z), \Im f(z))$ on the boundaries of Q_N is 8π for sufficiently large N. Hence $f(z)$ has four zeros (counted with their multiplicity) inside Q. This completes the first step.

The second step is proven as follows. For $0 < R \le 1$ the analysis shows that $AR^2 < 1.61$, $BR^2 < -1.8$, $CR^2 < 0.82$. Hence

$$|Ae^{Rz} + Be^{\frac{Rz}{2}} + C| \le |Ae^{Rz}| + |Be^{\frac{Rz}{2}}| + |C| \le |e^{Rz}|\frac{1.61}{R^2} + |e^{\frac{Rz}{2}}|\frac{1.8}{R^2} + \frac{0.82}{R^2}$$

on the domain $\{x \ge -1.268,\ y \ge 2\pi/R\}$. This expression is less than $|e^{Rz}|8/R^2$ (since $|e^{Rz}| \ge e^{-1.3}$). On the other hand, $|e^{(R+1)z}z^2| \ge e^{-1.3}(4\pi^2/R^2)|e^{Rz}| \ge |e^{Rz}|(10/R^2)$, which exceeds the previous expression for $R \le 1$. This means there are no roots in the domain $\{x \ge -1.268,\ y \ge 2\pi/R\}$.

Now consider the domain $D_1 = \{0 \le x \le N,\ 2\pi/(1+R) \le y \le 2\pi/R\}$. It is sufficient to prove the rotation of f on the boundaries of D is the same as the rotation of $e^{(R+1)z}z^2$ (which has no zeros inside D_1).

Consider all four boundaries of D_1 separately. On the upper boundary $|e^{(R+1)z}z^2| > |Ae^{Rz} + Be^{Rz/2} + C|$ as has been proven already, and the same is true for the right boundary for sufficiently large N. On the left boundary the direct analysis of both $Ae^{Rx}\cos(Ry) + Be^{Rx/2}\cos(Ry/2) + C$ and $Ae^{Rx}\sin(Ry) + Be^{Rx/2}\sin(Ry/2) + C$ shows that these functions are least 1.5 times less than the main member $|e^{(R+1)z}z^2| = |y|^2$. On the lower boundary we use the already established (in the first step) existence of a homotopy between $(\Re f, \Im f)$ and $(\Re g, \Im g)$ on this boundary.

Consider now the domain $D_2 = \{-1.268 \le x \le 0,\ 2\pi/(1+R) \le y \le 2\pi/R\}$. It is sufficient to prove the rotation of f on the boundaries of D is the same as the rotation of $e^{(R+1)z}z^2$ (which has no zeros inside D_1).

Consider all four boundaries of D_2 separately. The upper boundary and the right boundary are analyzed in the same way as for D_1. On the left boundary the function $|Ae^{Rz} + Be^{Rz/2} + C|^2 - |e^{(R+1)z}z^2|^2$ is negative for $R < 1/2$, which is verified directly. Let $R > 1/2$. Then for $2\pi/(1+R) \le y \le (8/3)\pi/(1+R)$ the imaginary parts of both $Ae^{Rz} + Be^{Rz/2} + C$ and $e^{(R+1)z}z^2$ are negative, which permits a linear homotopy between them. For $4\pi/(1+R)y > (8/3)\pi/(1+R)$ we prove directly that $|Ae^{Rz} + Be^{Rz/2} + C|^2 - |e^{(R+1)z}z^2|^2$ is negative, and for $y > 4\pi/(1+R)$ we prove it by proving that the upper estimate of this difference, which is

$|Ae^x - Be^R x/2 + C - e^{(R+1)x}(x^2+y^2)$, is negative for $0 < R \le 1$. On the lower boundary we use the already established (in the first step) existence of a homotopy between $(\Re f, \Im f)$ and $(\Re g, \Im g)$ on this boundary. This proves the second statement of the theorem.

To prove the third statement of the theorem, denote $A = A^* + a$, $B = B^* + b$, $C = C^* + c$, where $|a| < \varepsilon$, $|b| < \varepsilon$, $|c| < \varepsilon$, $\varepsilon \ll 1$. Consider the Taylor series of $f(z)$ in the neighborhood of W:

$$f(z) = f_0 + f_1(z-W) + f_2(z-W)^2 + f_3(z-W)^3 + f_4(z-W)^4 + f_5(z-W)^5 + O(z-W)^6.$$

Then the function

$$g(s) = g_0 s^5 + g_1 s^4 + g_2 s^3 + g_3 s^2 + g_4 s + g_5$$

has the same roots as $f(z-W)$, where

$$g_0 = \alpha + \beta + \gamma + \varepsilon_1 + o(\varepsilon),\ g_1 = (5/R)\alpha + (10/R)\beta + \varepsilon_2,\ g_2 = (20/R^2)\alpha + (80/R^2)\beta,$$
$$g_3 = (60/R^3)\alpha + (480/R^3)\beta,$$
$$g_4 = (120/R^4)\alpha + (1920/R^4)\beta,\quad g_5 = (120/R^5)\alpha + (3840/R^5)\beta + \gamma,$$

and

$$\alpha = e^{WR}(R^5/120)a,\quad \beta = e^{WR/2}(R^5/3840)b,\quad \gamma = c,$$

$$\varepsilon_1 = \frac{e^{W(1+R)}\left(4 + 6R + 2R^2 + (4+7R)\sqrt{12 + 24R + 18R^2 + 6R^3 + R^4}\right)}{240},$$

$$\varepsilon_2 = \frac{e^{W(1+R)}\sqrt{12 + 24R + 18R^2 + 6R^3 + R^4}}{24}.$$

The technique used in the proof of Theorem 1 is applicable here as well, and we have to prove only that all the Routh–Hurwitz determinants $\Delta_1, \ldots, \Delta_5$ of the polynomial $G_\delta(s) = g(s) + \delta s^5$ cannot be positive for small $|\alpha|, |\beta|, |\gamma| < \varepsilon$.

Let $g_5 = 0$. Then one of the Hurwitz polynomials of $G_\delta(s)/s$ (if $g_4 \ne 0$) and $G_\delta(s)/s^2$ (if $g_4 = 0$) is negative. Let $g_5 \ne 0$. Since $\Delta_5 = g_5 \Delta_4$, it is sufficient to prove that both $g_5 > 0$ and $\Delta_4 > 0$ cannot hold simultaneously for small $\alpha, \beta, \gamma, \delta$.

Consider the Hessian of Δ_4. It has one negative eigenvalue, tending to $-e^{W(1+R)}h \times (28800R^{10})^{-1}$, where h is equal to

$$3070080000 + 9203328000R + 11381760000R^2 + 7430400000R^3 + 2772720000R^4$$
$$+ 590256000R^5 + 59112000R^6 + 208R^{10} + 1104R^{11} + 2272R^{12} + 2304R^{13} + 1238R^{14}$$
$$+ 350R^{15} + 49R^{16}$$
$$+ (472320000 + 944064000R + 589536000R^2 + 117792000R^3 + 32R^{10}$$
$$+ 104R^{11} + 100R^{12} + 28R^{13}) \times \sqrt{12 + 24R + 18R^2 + 6R^3 + R^4}.$$

The other two zero eigenvalues define the two-dimensional invariant subspace M for the kernel of the Hessian. Since $g_5 > 0$ and

$$g_5 = \frac{10\sqrt{12 + 24R + 18R^2 + 6R^3 + R^4}\left(1920\beta + \gamma R^5\right)}{R^4\left(-4 - 6R - 2R^2 + (3R-4)\sqrt{12 + 24R + 18R^2 + 6R^3 + R^4}\right)}$$

on M, then $1920\beta + \gamma R^5 < 0$ on M. Since the restriction of Δ_4 on M has the form $\Delta_4 = (1920\beta + \gamma R^5)g$, where g is analytical on M, then it is sufficient to prove that g is

positive on M. The Hessian of g has one positive eigenvalue, equal to $e^{W(1+R)}g_1/g_2$, where

$$\begin{aligned}
g_1 &= 4146934579200 + 29659181875200R + 97501976985600R^2 + 194586181632000R^3 \\
&\quad + 262598757580800R^4 + 252803933798400R^5 + 178538876928000R^6 \\
&\quad + 93651451084800R^7 + 36470878617600R^8 + 10381801881600R^9 \\
&\quad + 2074990272768R^{10} + 265720094208R^{11} + 17063053824R^{12} \\
&\quad + 39632640R^{13} + 60500352R^{14} + 64471296R^{15} + 49433920R^{16} \\
&\quad + 27682752R^{17} + 11348884R^{18} + 3362664R^{19} + 693152R^{20} + 90860R^{21} + 5929R^{22} \\
&\quad + (993735475200 + 5950734336000R + 15851814912000R^2 + 24608489472000R^3 \\
&\quad + 24505461964800R^4 + 16245389721600R^5 + 7204036608000R^6 \\
&\quad + 2086060032000R^7 + 366206976000R^8 + 30582374400R^9 \\
&\quad + 1198808R^{10} + 883200R^{11} + 2803200R^{12} + 5034240R^{13} + 5649792R^{14} \\
&\quad + 4125504R^{15} + 1976000R^{16} + 608480R^{17} + 112480R^{18} + 9856R^{19}) \\
&\quad \times\sqrt{12 + 24R + 18R^2 + 6R^3 + R^4}, \\
g_2 &= 4320R^{10}(2368 + 4896R + 1248R^2 - 3600R^3 - 2568R^4 + 324R^5 + 860R^6 \\
&\quad + 342R^7 + 54R^8 + (960 + 240R - 816R^2 - 120R^3 + 328R^4 - 18R^5 - 54R^6 - 27R^7) \\
&\quad \times\sqrt{12 + 24R + 18R^2 + 6R^3 + R^4},
\end{aligned}$$

while the second eigenvalue is zero. Restricting the function g on the one-dimensional subspace L, corresponding to the zero eigenvalue, we see that g is equal to $\overline{g} = \gamma^3 R^5 p_1/p_2$, where

$$\begin{aligned}
p_1 &= -(2 + R)(1 + R + \sqrt{12 + 24R + 18R^2 + 6R^3 + R^4}) \\
&\quad \times(4 + 6R + 2R^2 + (4 + 7R)\sqrt{12 + 24R + 18R^2 + 6R^3 + R^4})^2 \\
&\quad \times(936 + 4224R + 7614R^2 + 7008R^3 + 3516R^4 + 952R^5 + 127R^6 \\
&\quad + (144 + 406R + 357R^2 + 95R^3)\sqrt{12 + 24R + 18R^2 + 6R^3 + R^4})^2, \\
p_2 &= 3072(4 + 6R + 2R^2 + (4 - 3R)\sqrt{12 + 24R + 18R^2 + 6R^3 + R^4}) \\
&\quad \times(936 + 3408R + 5274R^2 + 4368R^3 + 2046R^4 \\
&\quad + 530R^5 + 68R^6 + (144 + 338R + 255R^2 + 61R^3)\sqrt{12 + 24R + 18R^2 + 6R^3 + R^4})^3.
\end{aligned}$$

This expression is positive if $\gamma < 0$. The latter relation follows from the fact that $(1920\beta + \gamma R^5) < 0$ on M (and on L as well) and from the representation of $(1920\beta + \gamma R^5)$ on L in the form

$$\begin{aligned}
z_1 &= 3\gamma R^5(2 + R)\left(52 + 70R + 8R^2 - 28R^3 - 14R^4 - 3R^5\right. \\
&\quad \left. + (8 + 7R + R^2)\sqrt{12 + 24R + 18R^2 + 6R^3 + R^4}\right), \\
z_2 &= 936 + 3408R + 5274R^2 + 4368R^3 + 2046R^4 + 530R^5 + 68R^6 \\
&\quad + (144 + 338R + 255R^2 + 61R^3)\sqrt{12 + 24R + 18R^2 + 6R^3 + R^4}.
\end{aligned}$$

This proves the third statement of the theorem.

Acknowledgments. I am grateful to the anonymous reviewers for their helpful comments.

REFERENCES

[1] E. ALTMAN, F. BACCELLI, AND J. C. BOLOT, *Discrete-time analysis of adaptive rate control mechanisms*, in Proceedings of the 5th International Conference on Data Communication Systems and their Performance, Raleigh, NC, October 1993.

[2] R. BELLMAN AND K. COOKE, *Differential-Difference Equations*, Academic Press, New York, 1963.

[3] L. BENMOHAMED AND S. M. MEERKOV, *Feedback control of congestion in packet switching networks: The case of a single congested node*, IEEE/ACM Transactions on Networking, 1 (1993), pp. 693–708.

[4] J. C. BOLOT AND A. U. SHANKAR, *Dynamic behavior of rate-based flow control mechanism*, Computer Communications Review, 30 (1990), pp. 35–49.

[5] J. C. BOLOT AND A. U. SHANKAR, *Analysis of a fluid approximation to flow control dynamics*, in Proc. IEEE INFOCOM'92, Florence, 1992, pp. 2398–2407.
[6] D. M. CHIU AND R. JAIN, *Analysis of the increase and decrease algorithms for congestion avoidance in computer networks*, Computer Networks and ISDN Systems, 17 (1989), pp. 1–14.
[7] M. DE PRYCKER, *Asynchronous Transfer Mode: Solution for Broadband ISDN*, Ellis Horwood, New York, 1991.
[8] A. ELWALID, *Analysis of adaptive rate-based control for high-speed wide-area networks*, in Proc. ICC'95 Conf., Seattle, 1995, pp. 1948–1953.
[9] K. W. FENDICK, D. MITRA, I. MITRANI, M. A. RODRIGUES, J. B. SEERY, AND A. WEISS, *An approach to high-performance, high-speed data networks*, IEEE Communications Magazine, October 1991, pp. 74–82.
[10] K. W. FENDICK, M. A. RODRIGUES, AND A. WEISS, *Analysis of a rate-based feedback control strategy for long haul data transport*, Performance Evaluation, 16 (1992), pp. 67–84.
[11] K. W. FENDICK AND M. A. RODRIGUES, *An adaptive framework for dynamic access to bandwidth at high speeds*, in Proc. SIGCOMM'93 Conf., San Francisco, 1993, pp. 127–136.
[12] ———, *Asymptotic analysis of adaptive rate control for diverse sources with delayed feedback*, IEEE Trans. Inform. Theory, 40 (1994), pp. 2008–2025.
[13] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea House, New York, 1964.
[14] R. IZMAILOV, *The peak effect in stationary linear systems with scalar inputs and outputs*, Automat. Remote Control, 48 (1987), pp. 1018–1024.
[15] ———, *Adaptive feedback control algorithms for large data transfers in high-speed networks*, in Proc. 33rd IEEE Conf. Decision and Control, Lake Buena Vista, FL, 1994, pp. 2093–2096.
[16] ———, *Adaptive feedback control algorithms for large data transfers in high-speed networks*, IEEE Trans. Automat. Control, 40 (1995), pp. 1469–1471.
[17] V. JACOBSON, *Congestion avoidance and control*, Computer Communications Review, 18 (1988), pp. 314–329.
[18] M. KANAKIA, P. P. MISHRA, AND A. REIBMAN, *An adaptive congestion control scheme for real-time packet video transport*, in Proc. SIGCOMM'93 Conf., San Francisco, 1993, pp. 20–31.
[19] S. KESHAV, *A control-theoretic approach to flow control*, in Proc. SIGCOMM'91 Conf., Zürich, 1991, pp. 3–15.
[20] V. KOLMANOVSKII AND A. MYSHKIS, *Applied Theory of Functional Differential Equations*, Kluwer Academic Publishers, Amsterdam, 1992.
[21] M. A. KRASNOSEL'SKII AND P. P. ZABREIKO, *Plane Vector Fields*, Academic Press, New York, 1966.
[22] P. P. MISHRA AND M. KANAKIA, *A hop-by-hop rate-based congestion control scheme*, in Proc. SIGCOMM'92 Conf., Baltimore, 1993, pp. 112–123.
[23] D. MITRA, *Asymptotically optimal design of congestion control for high speed data networks*, IEEE Transactions on Communications, 40 (1992), pp. 301–311.
[24] A. MUKHERJEE AND J. C. STRIKWERDA, *Analysis of dynamic congestion control protocols—a Fokker-Planck approximation*, in Proc. SIGCOMM'91 Conf., Zürich, 1991, pp. 159–169.
[25] K. K. RAMAKRISHNAN AND R. JAIN, *A binary feedback scheme for congestion avoidance in computer networks*, ACM Transactions on Computer Systems, 8 (1990), pp. 158–181.
[26] G. RAMAMURTHY AND B. SENGUPTA, *A predictive hop-by-hop congestion control policy for high speed networks*, in Proc. IEEE INFOCOM'93, San Francisco, 1993, pp. 1033–1041.
[27] N. YIN AND M. G. HLUCHYJ, *On closed-loop rate control for ATM cell relay networks*, in Proc. IEEE INFOCOM'94, Toronto, 1994, pp. 99–108.
[28] L. A. ZADEH AND C. A. DESOER, *Linear System Theory*, McGraw-Hill, New York, 1963.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1781–1797, September 1996

015

HEAVY TRAFFIC ANALYSIS OF A CONTROLLED MULTICLASS QUEUEING NETWORK VIA WEAK CONVERGENCE METHODS*

HAROLD J. KUSHNER† AND L. FELIPE MARTINS‡

Abstract. The workload formulation due to Harrison and coworkers of multiclass queueing networks has been fundamental to its analysis. Until recently, there was no actual theory which started with the physical queue and showed that under heavy traffic conditions, the optimal costs could be approximated by those for an optimization problem using the "limit" workload equations. Recently, this was done via viscosity solution methods by Martins, Shreve, and Soner for one important class. For this same class of problems (and including the cases not treated there), we use weak convergence methods to show that the sequence of optimal costs for the original network converges to the optimal cost for the workload limit problem. The proof is simpler and allows weaker (and non-Markovian) conditions. It uses current techniques in weak convergence analysis. It seems to be the first analysis of such multiclass "workload" problems by weak convergence methods. The general structure of the development seems applicable to the analysis of more complex systems.

Key words. controlled queues, weak convergence, heavy traffic analysis, multiclass queues

AMS subject classifications. 90B15, 90B22, 93E20, 93E25, 60F17

1. Introduction.

Multiclass queueing network problems where some processor is required to serve more than one class of customers have been the subject of much recent interest. The choice of which customer to serve at any time (or whether the processor should remain idle, even if customers are present in its queue) has not been easy. The basic papers [2, 3, 10, 11] cast the problem in the "heavy traffic/workload" formulation, and various rules were derived for the control policies in the limit. These papers developed key ideas and intuition, but did not actually present a proof of the validity of the "approximating" workload equations for getting the controls. In [8], a proof was provided in a heavy traffic context for one important case.

The problem in [8], in what has been named the "criss-cross" system, has two processors. Processor 1 receives customers of two classes (named 1 and 2) from outside the system. Each class has its own interarrival time distribution and service time requirements. When service is completed on a class 1 customer, that customer leaves the system. When service is completed on a class 2 customer, that customer goes on to processor 2 and is renamed a class 3 customer.

In [8], the interarrival and service times were exponentially distributed and mutually independent so that a Markovian framework could be used. The analysis was of a heavy traffic type, in that one studied the limit of a sequence of problems (indexed by N), for which the difference between the mean rate of offered traffic and the service capacity went to zero as $N \to \infty$. The paper used viscosity solution methods for the convergence analysis for the Bellman equation. They established the important fact that the sequence of optimal value functions for the physical processes (indexed by N) converged to a function which solved the Bellman equation in the viscosity solution sense, and in certain cases they exhibited the control form which was optimal for the limit problem and nearly optimal for the physical process for large N. A key technique was the conversion of the physical problem into the so-called workload formulation of [2, 3, 10, 11]. The limit problem was actually a stochastic singular control problem.

*Received by the editors October 14, 1994; accepted for publication (in revised form) June 30, 1995.

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (hjk@dam.brown.edu). The research of this author was supported in part by AFOSR grant F-49620-92-0081, NSF grant ECS-9302137, and ARO grant DAAL03-92-G-0115.

‡Department of Mathematics, Cleveland State University, Cleveland, OH 44115 (35586288%taonode@vmcms.csuohio.edu).

Our approach is also of a heavy traffic type, but we use simple adaptations of the weak convergence methods of the type used in [4, 6, 7, 9] for various control problems with a single customer class. We work with the basic system of [8], also with the workload formulation, and show that the limit of the optimal cost functions for the physical system is the optimal cost for an appropriately defined limit of the workload equations for the physical systems. A crucial innovation in the analysis here is that we introduce the workload formulation *before* we pass to the limit.

The weak convergence methods are capable of treating a more general problem than that in [8]. Markovian assumptions are not needed, and we treat all of the cases discussed in [8], even the cases not solved there. We do not obtain the actual controls. However, the methods of proof can be adapted to work on variations of the problem for the type of queue structure at hand: for example, with correlated service or interarrival times, batch arrivals or services, state dependencies in the service and arrival data, and appropriate nonlinear cost functions. We believe that the proofs used here point the way toward possible proofs for more general queues with more complex flow and class structure, and some specific remarks to that end are at the end of the paper. Numerical methods of the Markov chain approximation type exist for singular control problems more general than the limit workload problem obtained for our cases [5, 6]. These can be of use on those problems where the actual solution cannot be obtained analytically, and the proof technique shows how to adapt these for use on the actual physical problem.

Heretofore, it was not clear how to use weak convergence methods on such multiclass problems, particularly with the workload formulation. Even when one could write the equations for the "physical" workload processes, it was not clear how to show that this sequence was tight, so that limits of weakly convergent subsequences could be chosen and worked with. It had appeared that for such problems, where one could not easily get convergent subsequences (in the weak convergence sense), that viscosity solutions were more natural. The relatively simple technique used here shows a direction for using the powerful weak convergence methods in many such problems. Basically, we just use ideas that are well understood from other problems in heavy traffic limits, and the proofs are relatively simple. An advantage of our approach is that it does not require knowledge of the solution of the limit problem. Obtaining this solution was the difficult part of the work in [8].

The problem is formulated, assumptions stated, and useful representations of the input and output processes are given in §2. The dynamical equations for the queues and workloads are developed in §3. Section 4 discusses the so-called workload cost transformation and states the workload limit equations. The convergence theorem and proof are in §5.

2. Assumptions and problem formulation. The set of all interarrival and service intervals are mutually independent, and the members of each class of arrival and service times are identically distributed. This assumption is made to simplify the notation. (As is common with weak convergence-type analyses, various forms of correlations can be introduced without affecting the end result, except for the variances of the Wiener processes.) We consider a sequence of problems indexed by N, which is a measure of the traffic intensity, and assume that the intensity approaches unity as $N \to \infty$. The interarrival intervals for classes $i = 1, 2$ are denoted by $\alpha_{i,j}^N$, $j = 1, 2, \ldots$, and have mean values $\overline{\alpha}_i^N$. The service times for classes $i = 1, 2, 3$ are denoted by $\Delta_{i,j}^N$, $j = 1, 2, \ldots$, and have mean values $\overline{\Delta}_i^N$. Let the arrival rate for class $i = 1, 2$ be denoted by $\lambda_i^N = [\overline{\alpha}_i^N]^{-1}$, and the mean service rates by $\mu_i^N = [\overline{\Delta}_i^N]^{-1}$.

Suppose that there are positive λ_i, μ_i such that $\lambda_i^N \to \lambda_i$ and $\mu_i^N \to \mu_i$. The basic heavy traffic assumption is that for $\lambda_i/\mu_i = p_i$, $i = 1, 2$, $p_1 + p_2 = 1$, and there are $b_i > 0$ such that

$$\lim_N \sqrt{N}\left(\frac{\lambda_1^N}{\mu_1^N} + \frac{\lambda_2^N}{\mu_2^N} - 1\right) = -b_1, \tag{2.1}$$

$$\lim_N \sqrt{N}\left(\frac{\lambda_2^N}{\mu_3^N} - 1\right) = -b_2. \tag{2.2}$$

From (2.1) and (2.2) it follows that the difference between service capacity and mean total input rate is positive for large N and is of order $O(1/\sqrt{N})$ for each processor. We suppose that the third absolute moments of the interarrival and service intervals are bounded uniformly in N, j, i, and that for finite $\sigma_{i,A}$, $\sigma_{i,D}$

$$E\left[(\alpha_{i,j}^N - \overline{\alpha}_i^N)/\overline{\alpha}_i^N\right]^2 \to [\sigma_{i,A}]^2,$$

$$E\left[(\Delta_{i,j}^N - \overline{\Delta}_i^N)/\overline{\Delta}_i^N\right]^2 \to [\sigma_{i,D}]^2.$$

We refer to the queue of customers of class i as queue i. Define $Z_i^N(t)$ to be the number of customers in queue i at time Nt divided by $\sqrt{N}$. Define $A_i^N(t)$ ($D_i^N(t)$, resp.) to be the number of arrivals (and completed services, resp.) at queue i by time Nt, divided by $\sqrt{N}$. Finally, define

$$S_i^{N,A}(t) = A_i^N(t)/\sqrt{N}$$

and

$$S_i^{N,D}(t) = D_i^N(t)/\sqrt{N}.$$

In this "$1/N$" time scale, the input-output equation for queue i is[1]

$$Z_i^N(t) = Z_i^N(0) + A_i^N(t) - D_i^N(t). \tag{2.3}$$

By the definition, $NS_i^{N,A}(t)$ is the number of arrivals of class i by time Nt. It can be written as $\max\{n : \sum_{j=1}^n \alpha_{i,j}^N \le Nt\}$. For use below, define $\bar{S}_i^{N,A}(t)$ by

$$N\bar{S}_i^{N,A}(t) = \min\left\{n : \sum_{j=1}^n \alpha_{i,j}^N \ge Nt\right\}.$$

Define $\bar{S}_i^{N,D}(t)$ analogously. Then $\bar{S}_i^{N,A}(t) - S_i^{N,A}(t)$ is either zero or $1/N$, and similarly for D replacing A.

Controls. Let $c_i > 0$, $\beta > 0$. The cost function of interest in this paper is

$$E\int_0^\infty e^{-\beta t}\sum_{i=1}^3 c_i Z_i^N(t)dt.$$

For the physical system, the only control problem concerns what to do at processor 1. There, at any time we must choose between serving class 1 or 2 or to not serve either. In the physical

[1]Notice that, contrary to frequent practice in heavy traffic analysis, $D_i^N(t)$ represents the *actual* number of class i customers on which service was completed by time t, not the number that would have been served if the processor kept processing and turning out "fictitious" outputs during its idle times.

system, it is obvious that processor 2 should always work when there is work for it to do, and processor 1 should always work when $Z_1^N(t) > 0$. However, to facilitate the analysis below, it will be convenient to allow somewhat more general controls by adding the possibility of shutting processor 2 off when there is work for it to do or of shutting processor 1 off when $Z_1^N(t) > 0$. Obviously, use of the extended controls will not reduce the cost.

We now define a model for the allocation of time (or for planned idling) which will be convenient for the weak convergence analysis. The basic control is the amount of time allocated for work on each class, and for idling. These will be simply nondecreasing and nonanticipative processes which do not violate the physical feasibility constraint that we can use only the time that is available.

We define the controls by starting with a particular prior allocation of service time and then modifying that to get whatever policy is actually desired. The prior allocation is $p_i Nt$ to the physical queue, $i = 1, 2$, and Nt to queue 3, in the real time interval $[0, Nt]$. This allocation is equivalent to an allocation of $p_i t$, $i = 1, 2$, in the rescaled time used in (2.3). The control for processor 1 then involves the reallocation of time between queues 1 and 2, as well as the choice of whether or not to actually use the finally allocated time or to let the processor idle (even if customers are present). We define the reallocation of service time between queues 1 and 2 by introducing a function $G_i^N(\cdot)$ such that $\sqrt{N}G_i^N(t)$ is the service time in the real time interval $[0, Nt]$ allocated originally to queue i but which is reallocated to queue j, $j \neq i$, in that interval. All the allocations and reallocations are assumed to be nonanticipative. Analogously, let $\sqrt{N}L_i^N(t)$, $i = 1, 2, 3$, denote the total service time finally allocated to queue i in the real time interval $[0, Nt]$ but not used because either queue i was empty or because the control required that there be no service. In the latter case, the allocated time cannot be allocated to another queue but remains unused.

Useful representations of the arrival and departure processes. Define the processes

$$\hat{A}_i^N(n) = \frac{1}{\sqrt{N}} \sum_{j=1}^{n} \left(1 - \frac{\alpha_{i,j}^N}{\bar{\alpha}_i^N}\right),$$

$$\hat{D}_i^N(n) = \frac{1}{\sqrt{N}} \sum_{j=1}^{n} \left(1 - \frac{\Delta_{i,j}^N}{\bar{\Delta}_i^N}\right),$$

and let $\hat{\mathcal{B}}_i^{N,A}(n)$ (resp., $\hat{\mathcal{B}}_i^{N,D}(n)$) be the minimal σ-algebra which measures the family $\{\hat{A}_i^N(j), j \leq n\}$ (resp., $\{\hat{D}_i^N(j), j \leq n\}$). For future use, we note that $\hat{A}_i^N(n)$ is a $\hat{\mathcal{B}}_i^{N,A}(n)$-martingale, and for each t, $N\bar{S}^{N,A}(t)$ is a $\hat{\mathcal{B}}_i^{N,A}(n)$-stopping time, with analogous assertions holding for $\hat{D}_i^N(n)$ and $N\bar{S}^{N,D}(t)$.

For $i = 1, 2$, the definitions of the terms involved allow us to write

$$A_i^N(t) = \frac{1}{\sqrt{N}} \sum_{j=1}^{NS_i^{N,A}(t)} 1 = W_i^{N,A}(t) + \frac{1}{\sqrt{N}} \sum_{j=1}^{NS_i^{N,A}(t)} \frac{\alpha_{i,j}^N}{\bar{\alpha}_i^N}, \tag{2.4}$$

where we define

$$W_i^{N,A}(t) = \frac{1}{\sqrt{N}} \sum_{j=1}^{NS_i^{N,A}(t)} \left[1 - \frac{\alpha_{i,j}^N}{\bar{\alpha}_i^N}\right].$$

By the definitions of $\bar{S}_i^{N,A}(t)$ and $\hat{A}_i^N(n)$, we have (note that the upper limit of the sum is $N\bar{S}_i^{N,A}(t)$ here)

$$W_i^{N,A}(t) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N\bar{S}_i^{N,A}(t)} \left[1 - \frac{\alpha_{i,j}^N}{\bar{\alpha}_i^N}\right] - r_i^{N,A}(t) \tag{2.4a}$$

$$= \hat{A}_i^N(N\bar{S}_i^{N,A}(t)) - r_i^{N,A}(t),$$

where $r_i^{N,A}(t)$ is either zero or $1/\sqrt{N}$ times the last term in the sum in (2.4a).

By the definition of λ_i^N and the fact that

$$\sum_{j=1}^{NS_i^{N,A}(t)} \alpha_{i,j}^N = Nt - \rho_i^N(t),$$

where $\rho_i^N(t)$ is Nt minus the time of the last arrival of a class i input before time Nt, we can write

$$A_i^N(t) = W_i^{N,A}(t) + \frac{1}{\sqrt{N}}\lambda_i^N[Nt - \rho_i^N(t)].$$

All the terms called ρ^N which are used below are either of similar origin (i.e., residual times) to the $\rho_i^N(t)$ above or are bounded by a constant times the sum of such expressions. *We will omit the subscripts and use the same symbol for all of them.*

We can similarly write, for $i = 1, 2$,

$$D_i^N(t) = \frac{1}{\sqrt{N}} \sum_{j=1}^{NS_i^{N,D}(t)} 1 = W_i^{N,D}(t) + \frac{1}{\sqrt{N}} \sum_{j=1}^{NS_i^{N,D}(t)} \frac{\Delta_{i,j}^N}{\bar{\Delta}_i^N}, \tag{2.5}$$

where we define

$$W_i^{N,D}(t) = \frac{1}{\sqrt{N}} \sum_{j=1}^{NS_i^{N,D}(t)} \left[1 - \frac{\Delta_{i,j}^N}{\bar{\Delta}_i^N}\right].$$

Recall that we defined $[\bar{\Delta}_i^N]^{-1} = \mu_i^N$. We can write (for $j \neq i, i = 1, 2$)

$$D_i^N(t) = W_i^{N,D}(t) + \frac{1}{\sqrt{N}}\mu_i^N\left[p_i Nt + \sqrt{N}G_j^N(t) - \sqrt{N}G_i^N(t) - \sqrt{N}L_i^N(t) - \rho^N(t)\right].$$

In getting this expression, we use the fact that the term in the square bracket in the above equation is $\sum_{j=1}^{NS_i^{N,D}(t)} \Delta_{i,j}^N$, the total time allocated to processing completed customers of class i by time Nt: the prior allocation plus the *net* reallocation minus the unused allocation of time minus the time ($\rho^N(t)$) spent on the current incompleted customer (if any).

Analogously to the "arrival" case, the definitions of $\bar{S}_i^{N,D}(t)$ and $\hat{D}_i^N(n)$ imply that (note that the upper limit of the sum is $N\bar{S}_i^{N,D}(t)$ here)

$$W_i^{N,D}(t) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N\bar{S}_i^{N,D}(t)} \left[1 - \frac{\Delta_{i,j}^N}{\bar{\Delta}_i^N}\right] - r_i^{N,D}(t) \tag{2.5a}$$

$$= \hat{D}_i^N(N\bar{S}_i^{N,D}(t)) - r_i^{N,D}(t),$$

where $r_i^{N,D}(t)$ is either zero or $1/\sqrt{N}$ times the last term in the sum in (2.5a).

3. Dynamical equations and workload formulation. We now put the representations of the last section together. Define $\tilde{W}_i^N(t) = W_i^{N,A}(t) - W_i^{N,D}(t)$, $i = 1, 2$. Then we can write

$$\begin{aligned} Z_1^N(t) &= Z_1^N(0) + \tilde{W}_1^N(t) + \sqrt{N}t[\lambda_1^N - \mu_1^N p_1] \\ &\quad + \mu_1^N[G_1^N(t) - G_2^N(t) + L_1^N(t)] + \rho^N(t)/\sqrt{N}. \end{aligned} \tag{3.1a}$$

Analogously,

$$\begin{aligned} Z_2^N(t) &= Z_2^N(0) + \tilde{W}_2^N(t) + \sqrt{N}t[\lambda_2^N - \mu_2^N p_2] \\ &\quad + \mu_2^N[G_2^N(t) - G_1^N(t) + L_2^N(t)] + \rho^N(t)/\sqrt{N}. \end{aligned} \tag{3.1b}$$

The term $D_3^N(\cdot)$ can be treated analogously to what was done for the $D_1^N(\cdot)$, except that there is no reallocation now. Using such a representation, defining $\tilde{W}_3^N(t) = W_2^{N,D}(t) - W_3^{N,D}(t)$, and using the fact that $A_3^N(t) = D_2^N(t)$, we can write

$$\begin{aligned} Z_3^N(t) &= Z_3^N(0) + \tilde{W}_3^N + \sqrt{N}t[\mu_2^N p_2 - \mu_3^N] \\ &\quad + \mu_2^N[G_1^N(t) - G_2^N(t) - L_2^N(t)] + \mu_3^N L_3^N(t) + \rho^N(t)/\sqrt{N}. \end{aligned} \tag{3.1c}$$

The workload formulation. Define the workloads[2] for the Nth queueing system by

$$WL_1^N(t) = \frac{Z_1^N(t)}{\mu_1^N} + \frac{Z_2^N(t)}{\mu_2^N}, \quad WL_2^N(t) = \frac{Z_2^N(t) + Z_3^N(t)}{\mu_3^N}.$$

It was shown in [3, 2, 10, 11] that the workload concept is fundamental to the formulation of multiclass queueing problems and to the derivation of heavy traffic approximations to them. Define

$$W_1^N(t) = \frac{\tilde{W}_1^N(t)}{\mu_1^N} + \frac{\tilde{W}_2^N(t)}{\mu_2^N}, \quad W_2^N(t) = \frac{\tilde{W}_2^N(t) + \tilde{W}_3^N(t)}{\mu_3^N}.$$

The workload satisfies the equations

$$\begin{aligned} WL_1^N(t) &= WL_1^N(0) + W_1^N(t) + [L_1^N(t) + L_2^N(t)] \\ &\quad + \sqrt{N}t\left[\frac{\lambda_1^N}{\mu_1^N} - p_1\right] + \sqrt{N}t\left[\frac{\lambda_2^N}{\mu_2^N} - p_2\right] + \rho^N(t)/\sqrt{N}, \end{aligned} \tag{3.2}$$

$$\begin{aligned} WL_2^N(t) &= WL_2^N(0) + W_2^N(t) + L_3^N(t) \\ &\quad + \sqrt{N}t\left[\frac{\lambda_2^N}{\mu_3^N} - 1\right] + \rho^N(t)/\sqrt{N}. \end{aligned} \tag{3.3}$$

Representing the idle time terms in the above equations in terms of reflection terms and control functions. The idle time terms $L_1^N(\cdot) + L_2^N(\cdot)$ and $L_3^N(\cdot)$ in (3.2) and (3.3), respectively, might increase at times t at which the corresponding workloads (1 and 2, resp.)

[2]There is a problem in the literature with the use of W for different things. W is often used for both a Wiener and a workload process. Partially compromising with these traditions leads to our use of WL for the workload, and W for either a Wiener or an "almost" Wiener process.

are positive. Let $Y_i^N(\cdot), i = 1, 2$, denote the part of these terms which increases only when the corresponding workload (1 or 2, resp.) is zero. Define $F_i^N(\cdot)$ by

$$L_1^N(t) + L_2^N(t) = Y_1^N(t) + F_1^N(t), \quad L_3^N(t) = Y_2^N(t) + F_2^N(t). \tag{3.4}$$

To get some intuitive feeling for the formulation, it is worthwhile to discuss the physical meaning of the F^N-terms. The discussion which follows, however, is not used subsequently. We do not require any detailed knowledge of the optimal controls to do the proof.

Under an optimal control, $L_2^N(\cdot)$ would not increase at t unless $Z_1^N(t) = 0$, since we would not allocate service time to queue 2 when it is empty and queue 1 is not empty. However, it is conceivable that under an optimal control, service time will be allocated to queue 1 when it is empty but queue 2 is not. This would happen, for example, if the cost of waiting at queue 3 is relatively large and $Z_3^N(t) > 0$. In this case, we might prefer not to add new customers to queue 3 until it drops below a certain level. Thus, under an optimal control, the term $F_1^N(\cdot)$ represents the part of $L_1^N(\cdot)$ which increases when queue 2 is not empty. Of course, if we are not using an optimal control, then $F_1^N(\cdot)$ represents the increase in $L_1^N(\cdot) + L_2^N(\cdot)$ at those times that either $Z_1^N(t)$ or $Z_2^N(t)$ are positive. Under an optimal control, $F_2^N(\cdot)$ represents the part of $L_3^N(\cdot)$ which increases when $Z_2^N(t) \neq 0$ and $Z_3^N(t) = 0$, and we expect that it would be close to zero for large N. If we shut processor 2 off when $Z_3^N(t) > 0$, then $F_2^N(\cdot)$ would also increase at those times, although this is not an optimal procedure.

The cost function of interest. The $G^N(\cdot) = (G_i^N(\cdot), i = 1, 2)$ and $L^N(\cdot)$ determine the $F^N(\cdot) = (F_i^N(\cdot), i = 1, 2)$ and $(G^N(\cdot), F^N(\cdot))$ determines $L^N(\cdot)$. It will be convenient to consider the pair $(G^N(\cdot), F^N(\cdot))$ as the control, although the (G, L)-terms will not show up in the limit workload equations, and only the F-terms remain. For $c_i > 0, \beta > 0$, and $Z^N(0) = z$, define the cost

$$V^N(z, G^N, F^N) = E \int_0^\infty e^{-\beta t} \sum_{i=1}^3 c_i Z_i^N(t) dt,$$

$$\bar{V}^N(z) = \inf_{F,G} V^N(z, G, F).$$

Now, using (2.1), (2.2), and the definitions of the $F^N(\cdot), Y^N(\cdot)$, rewrite the workload equations as

$$\begin{aligned} WL_1^N(t) &= WL_1^N(0) + W_1^N(t) + Y_1^N(t) + F_1^N(t) - b_1 t + \rho^N(t)/\sqrt{N} + \delta_1^N(t), \\ WL_2^N(t) &= WL_2^N(0) + W_2^N(t) + Y_2^N(t) + F_2^N(t) - b_2 t + \rho^N(t)/\sqrt{N} + \delta_2^N(t), \end{aligned} \tag{3.5}$$

where the $\delta_i^N \to 0$ and are due to the use of (2.1), (2.2) and are ignored henceforth. Note that the only explicit influence of $G^N(\cdot)$ on the workload process is via the $F^N(\cdot)$-terms.

4. The workload cost transformation and limit problem. The full exploitation of the workload concept occurs via the following transformation of the cost function which allows one to view the workload equations as the primary system equations, at least in the heavy traffic limit [2, 3, 8, 10, 11]. Define $g(z) = \sum c_i z_i$, $w_1^N(z) = z_1/\mu_1^N + z_2/\mu_2^N$, $w_2^N(z) = (z_2 + z_3)/\mu_3^N$. For each $w = (w_1, w_2)$ with $w_i \geq 0$, define

$$\bar{g}^N(w) = \min\{g(z) : w_1^N(z) = w_1, w_2^N(z) = w_2, z_i \geq 0\}. \tag{4.1}$$

We can thus write

$$g(z) = \bar{g}^N(w^N(z)) + \tilde{g}^N(z), \ \tilde{g}^N(z) \geq 0. \tag{4.2}$$

Let $\bar{z}^N(w)$ denote the minimizing value of z in (4.1). We also define the "limit quantities" analogously: $w_1(z) = z_1/\mu_1 + z_2/\mu_2$, $w_2(z) = (z_2 + z_3)/\mu_3$, and

$$\bar{g}(w) = \min\{g(z) : w_1(z) = w_1, w_2(z) = w_2, z_i \geq 0\}.$$

We note that there are $\epsilon_i^N \to 0$ as $N \to \infty$ such that

$$|w^N(z) - w(z)| \leq \epsilon_1^N |z|, \quad |\bar{g}^N(w) - \bar{g}(w)| \leq \epsilon_2^N |w|,$$

and $\bar{g}(w)$, $\bar{g}^N(w)$, $\tilde{g}(z)$, $\tilde{g}^N(z)$ have at most (uniform) linear growth.

From the reformulation in terms of workload, it is suggested that the individual classes have essentially disappeared, and that there is now one equation in (3.5) for each processor. Suppose that, given the current values $WL^N(t) = w(Z^N(t)) = w$, $Z^N(t) = z$, we are able to "instantaneously and freely" rearrange the queues in such a way that the workload is conserved but the minimal cost point $\bar{g}^N(w)$ is attained. Then, ignoring the small terms $\rho^N(\cdot)/\sqrt{N}$ and the possible policy dependence of the $W_i^N(\cdot)$, the only control appears to be the $F^N(\cdot)$. This formal point of view will be validated in the limit as $N \to \infty$. (It will also be shown in the proof that the $W_i^N(\cdot)$ are asymptotically independent of the policy, for reasonable policies.)

The limit problem. The correct equations for the limit workload problem will turn out to be

$$\begin{aligned} WL_1(t) &= WL_1(0) + W_1(t) + Y_1(t) + F_1(t) - b_1 t, \\ WL_2(t) &= WL_2(0) + W_2(t) + Y_2(t) + F_2(t) - b_2 t, \end{aligned} \tag{4.3}$$

where the $W_i(\cdot)$ are Wiener processes with variances

$$E[W_1(1)]^2 = \frac{\lambda_1}{\mu_1^2}(\sigma_{1,A}^2 + \sigma_{1,D}^2) + \frac{\lambda_2}{\mu_2^2}(\sigma_{2,A}^2 + \sigma_{2,D}^2), \tag{4.4a}$$

$$E[W_2(1)]^2 = \frac{\lambda_2}{\mu_3^2}(\sigma_{2,A}^2 + \sigma_{3,D}^2), \tag{4.4b}$$

$$EW_1(1)W_2(1) = \frac{\lambda_2}{\mu_2\mu_3}\sigma_{2,A}^2. \tag{4.4c}$$

In (4.3), the $Y_j(\cdot)$ are the minimal nondecreasing processes such that the $WL_i(\cdot)$ are nonnegative and the $F_i(\cdot)$, $i = 1, 2$ are arbitrary, nonanticipative, nondecreasing controls (singular controls). For $WL(0) = w$, the associated cost is

$$V(w, F) = E\int_0^\infty e^{-\beta t}\bar{g}(WL(t))dt, \tag{4.5}$$

and we define

$$\bar{V}(w) = \inf_F V(w, F).$$

The problem can be broken into the two following fundamental cases, according to the value of the minimizer in (4.1) [8].

Case 1. $c_1\mu_1^N + c_3\mu_2^N \leq c_2\mu_2^N$. Then the optimizing value in (4.1) is $\bar{z}_2(w) = 0$, and

$$\bar{g}^N(w) = c_1\mu_1^N w_1 + c_3\mu_3^N w_2.$$

To approximate this value as closely as possible requires that queue 2 have priority over queue 1 and $F^N(t) = 0$. There is no control problem (see [8]).

Case 2. $c_1\mu_1^N + c_3\mu_2^N > c_2\mu_2^N$. In this case, the minimizer of the linear program (4.1) depends on the region of the nonnegative quadrant to which (w_1, w_2) belongs:

(4.6a) If $\mu_3^N w_2 > \mu_2^N w_1$ or, equivalently, $z_3 > \mu_2^N z_1/\mu_1^N$,

then the optimizing value $\bar{z}(w)$ is

(4.7a) $$\bar{z}_1(w) = 0,\ \bar{z}_2(w) = \mu_2^N w_1,\ \bar{z}_3(w) = \mu_3^N w_2 - \mu_2^N w_1,$$

and we have

(4.8a) $$\bar{g}^N(w) = [c_2\mu_2^N - c_3\mu_2^N]w_1 + c_3\mu_2^N w_2.$$

(4.6b) If $\mu_3^N w_2 \le \mu_2^N w_1$ or, equivalently, $z_3 \le \mu_2^N z_1/\mu_1^N$,

then the optimizing values are given by

(4.7b) $$\bar{z}_3(w) = 0, \quad \bar{z}_1(w) = \frac{\mu_1^N}{\mu_2^N}[\mu_2^N w_1 - \mu_3^N w_2], \quad \bar{z}_2(w) = \mu_3^N w_2,$$

and we have

(4.8b) $$\bar{g}^N(w) = c_1\mu_1^N w_1 + \frac{\mu_3^N}{\mu_2^N}[c_2\mu_2^N - c_1\mu_1^N]w_2.$$

Since the μ_i^N depend on N, it is conceivable that one could switch infinitely often between Cases 1 and 2 as $N \to \infty$. In order to avoid this minor annoyance, we suppose that only one of the cases holds for all N.

Note that Case 2 is equivalent to $\bar{z}_1(w)\bar{z}_3(w) = 0$. Under (4.6a), $\bar{z}_1(w) = 0$ and $\bar{z}_3(w) > 0$. To approximate this for the physical system, queue 1 has priority, and we continue to give queue 1 priority until queue 3 is zero, when (4.6b) takes over. To attain the minimizer in (4.7b) requires that queue 1 again have priority, but (loosely speaking) we must avoid "starving" queue 3. Thus, when queue 3 is zero, "briefly" serve queue 2. This rough description ignores some details, such as the possible introduction of idleness in processor 1. Full details are in the proof in the next section, and we see that the main purpose of the $G^N(\cdot)$ is to bring us close to this strategy for large N.

5. The main theorem: Uniqueness of the solution to (4.3). We use $D^k[0, \infty)$, the space of CADLAG functions, for appropriate integers k, with the Skorohod topology as the canonical sample space. Given the distribution of $(F(\cdot), W(\cdot), WL(0))$, with $(F(\cdot), WL(0))$ nonanticipative with respect to $W(\cdot)$, the distribution of the process $(WL(\cdot), F(\cdot), W(\cdot), Y(\cdot))$ is determined uniquely.

The following lemmas will be useful in the proof of the main theorem.

LEMMA 5.1. *Given any* ϵ, *for each* $w = WL(0)$ *there are a* $T(\epsilon) > 0$ *and an* ϵ*-optimal control* $F^\epsilon(\cdot)$ *for* (4.3), (4.5) *such that* $F^\epsilon(\cdot)$ *is constant after time* $T(\epsilon)$.

Proof. It is easily verified that the policy with the "zero control" $F(t) \equiv 0$ has finite cost. In general, we have

$$F_j(t) + Y_j(t) = WL_j(t) - WL_j(0) + b_j t - W_j(t).$$

For any policy with finite cost this implies that

(5.1) $$E\int_0^\infty e^{-\beta t}(F_j(t) + Y_j(t))\,dt < \infty.$$

Now fix T and define the policy $F^T(\cdot)$ by

$$F^T(t) = F(t) \text{ if } t \le T$$

and

$$F^T(t) = F(T) \text{ if } t \ge T.$$

We can suppose without loss of generality that we have processes $WL^T(\cdot)$, $WL(\cdot)$ defined on the same sample space and driven by the *same* Wiener process $W(\cdot)$ and with controls $F^T(\cdot)$, $F(\cdot)$, respectively. Denote the corresponding reflection terms by $Y^T(\cdot)$, $Y(\cdot)$. We clearly have

$$E\int_0^\infty e^{-\beta t}(F_j^T(t) + Y_j^T(t))\,dt < \infty. \tag{5.2}$$

Then, for $t > T$,

$$WL_j(t) - WL_j^T(t) = F_j(t) - F_j(T) + Y_j(t) - Y_j^T(t) \ge 0,$$

and (5.1) and the Lipschitz continuity of $\bar g(\cdot)$ imply that

$$|\bar g(WL(t)) - \bar g(WL^T(t))| \le \gamma \sum_j [F_j(t) - F_j(T) + Y_j(t) - Y_j^T(t)], \quad t > T,$$

where γ is the Lipschitz constant. This gives

$$\begin{aligned}
&E\int_0^\infty e^{-\beta t}|\bar g(WL(t)) - \bar g(WL^T(t))|\,dt \\
&= E\int_T^\infty e^{-\beta t}|\bar g(WL(t)) - \bar g(WL^T(t))|\,dt \\
&\le \gamma E\int_T^\infty e^{-\beta t}\sum_j [F_j(t) - F_j(T) + Y_j(t) - Y_j^T(t)]\,dt \\
&\le \gamma E\int_T^\infty e^{-\beta t}\sum_j (F_j(t) + Y_j(t))\,dt,
\end{aligned}$$

but, by (5.1) and the monotone convergence theorem, the right-hand term above goes to zero as $T \to \infty$. This proves that the cost of the control $F^T(\cdot)$ converges to the cost of the control $F(\cdot)$ as T goes to infinity. The lemma follows if we take $F^\epsilon(\cdot)$ to be a $\epsilon/2$-optimal policy, and $T = T(\epsilon)$ suitably large. □

LEMMA 5.2. *Fix $\epsilon > 0$ and $WL(0) = w$. For each w, there are $T(\epsilon) > 0$, an integer K, $\Delta > 0$, $0 < \Delta_1 < \Delta$, $\rho > 0$, and an ϵ-optimal control $F^\epsilon(\cdot)$ for* (4.3), (4.5) *satisfying the following properties:*

(a) *$F^\epsilon(\cdot)$ is constant in each of the intervals $[k\Delta, (k+1)\Delta)$ and on $[T(\epsilon), \infty)$, and only one of the two components can increase at each $k\Delta$.*

(b) *$F^\epsilon(n\Delta + \Delta) - F^\epsilon(n\Delta) \equiv \delta F^\epsilon(n\Delta)$ takes values in the discrete set $\{j\rho : j = 0, 1, \dots, K\}$.*

(c) *For $n = 0, \dots$, $j = 0, \dots, K$, $i = 1, 2$, there are functions $q_{nji}(\cdot)$ such that the choice of $F^\epsilon(\cdot)$ according to the probability law* (5.3) *gives a (possibly randomized) ϵ-optimal control whose sample paths satisfy* (a) *and* (b)*:*

$$\begin{aligned}
&q_{nji}(\delta F^\epsilon(m\Delta), m < n, W(l\Delta_1), l\Delta_1 \le n\Delta) \\
&= P\{\delta F_i^\epsilon(n\Delta) = j\rho | \delta F^\epsilon(m\Delta), m < n, W(l\Delta_1), l\Delta_1 \le n\Delta\},
\end{aligned} \tag{5.3}$$

and $q_{nji}(\cdot)$ is continuous in the W-variables for each choice of the other variables.

Comment on the proof. Given the previous lemma and the weak sense uniqueness of the solution to (4.3), the proof is identical to the proof of the analogous result in [9, Thms. 15, 16], and is omitted. (The solution is also pathwise unique, but that fact is not needed in the proof.)

THEOREM 5.3. *Let* $Z^N(0) = z^N \to z$ *as* $N \to \infty$. *Then, under the conditions of* §2, *as* $N \to \infty$,

$$\bar{V}^N(z^N) \to \bar{V}(w(z)). \tag{5.4}$$

Proof. Part 1. In all of the weak convergence analysis, we use the Skorohod topology on $D^k[0, \infty)$ [1] for appropriate integers k.

In this part of the proof we prove, for a class of "good" policies, certain bounds and the weak convergence of the processes associated with the limit Wiener processes. These will be used to prove both the lower bound (5.7) and the upper bound (5.12).

Let $T < \infty$ and $\epsilon > 0$. Owing to the fact that the interarrival and service intervals have uniformly bounded third absolute moments, the $\rho^N(\cdot)$ processes are tight and $\lim_N E|\rho^N(t)|/\sqrt{N} = 0$ in all cases. Also, $\{\rho^N(t)/\sqrt{N}, N, t\}$ is uniformly integrable. Let $\bar{G}^N(\cdot)$, $\bar{F}^N(\cdot)$ be a sequence of controls for which $\liminf_{N\to\infty} V^N(z^N, \bar{G}^N, \bar{F}^N) < \infty$. Let us abuse terminology and let N index an "infimizing sequence in the "liminf" above.

Recall from formulas (2.4a) and (2.5a) that

$$W_i^{N,A}(t) = \hat{A}_i^N(N\bar{S}_i^{N,A}(t)) - r_i^{N,A}(t),$$
$$W_i^{N,D}(t) = \hat{D}_i^N(N\bar{S}_i^{N,D}(t)) - r_i^{N,D}(t),$$

where the arrival/departure r-terms are bounded by

$$\frac{1}{\sqrt{N}}(1 + \text{ time between last arrival/departure before } Nt \text{ and first after } Nt).$$

The r-terms are of the same type as, and will be absorbed into, the ρ-terms.

By the comments in §2 concerning the martingale properties of the $\hat{A}_i^N(n)$ and $\hat{D}_i^N(n)$ and stopping time properties of the $N\bar{S}_i^{N,A}(t)$ and $N\bar{S}_i^{N,D}(t)$, the processes $\hat{A}_i^N(N\bar{S}_i^{N,A}(t))$ and $\hat{D}_i^N(N\bar{S}_i^{N,D}(t))$ are continuous-parameter martingales (with respect to the "natural σ-algebras).

The above comments and representations in terms of the martingales (modulo small errors) imply that the mean square value of $W_i^{N,A}(t)$ is bounded above (modulo $O(1/N)$) by a constant times $ES_i^{N,A}(t)$.

We have $\limsup_N [ES_i^{N,A}(t)/\lambda_i^N t] \le 1$. Also, no matter what policy is being used, we have $S_i^{N,D}(t) \le S_i^{N,A}(t)$ for all t. This then implies a similar bound for $E[W_i^{N,D}(t)]^2$. Thus, for some constant C_1 independent of N,

$$\limsup_N EW_i^{N,A}(t)^2 \le C_1 t, \quad \limsup_N EW_i^{N,D}(t)^2 \le C_1 t. \tag{5.5a}$$

A similar argument and Doob's inequality yields

$$\limsup_n E \sup_{s\le t} W_i^{N,A}(s)^2 \le 4C_1 t, \quad \limsup_N E \sup_{s\le t} W_i^{N,D}(s)^2 \le 4C_1 t. \tag{5.5b}$$

We claim that

$$\limsup_{N\to\infty} E\bar{F}_i^N(t) < \infty, \text{ each } t < \infty. \tag{5.6}$$

To prove this, we note first that $EY_i^N(t) \leq \text{constant(t)} + E\max_{s\leq t}|W_i^N(s)|$ which, together with (5.5), yields $\sup_N EY_i^N(t) < \infty$ for each t. Thus, failure of the assertion (5.6) yields that $EWL^N(s)$ (and hence $EZ^N(s)$) becomes arbitrarity large on some interval after t, as $N \to \infty$. This implies that the costs go to infinity as $N \to \infty$, a contradiction.

These arguments also imply that $\limsup_N E\sup_{s\leq t}|Z^N(s)| < \infty$. Equivalently, by the definition of $Z_i^N(t)$,

$$E\sup_{s\leq t}\frac{\text{number of class } i \text{ arrivals by } Nt \text{ not served by } Nt}{\sqrt{N}} < \infty.$$

This can be written as (using $S_3^{N,A}(\cdot) = S_2^{N,D}(\cdot)$)

$$\limsup_{N\to\infty} E\sup_{s\leq t}\sqrt{N}(S_i^{N,A}(s) - S_i^{N,D}(s)) < \infty,\ i = 1, 2, 3.$$

A functional law of large numbers implies that $S_i^{N,A}(\cdot)$ converges weakly to the constant process with values $\lambda_i t$. Thus, for $i = 1, 2$, $S_i^{N,D}(\cdot)$ converges weakly to the process with values $\lambda_i t$, and $S_3^{N,D}(\cdot)$ converges weakly to the process with values $\lambda_2 t$.

Using the cited convergences of the $S_i^{N,A}(\cdot)$, $S_i^{N,D}(\cdot)$ and the martingale properties of the $\hat{A}_i^N(N\bar{S}_i^{N,A}(\cdot))$ and $\hat{D}_i^N(N\bar{S}_i^{N,D}(\cdot))$ and standard arguments such as those associated with equation (2.6) of [9] or with the B^ϵ-terms in (3.10) of [7, Lem. 5.2], it follows that the $W_i^{N,A}(\cdot)$, $W_i^{N,D}(\cdot)$ are tight, with all weak limits being Wiener processes. Using the definitions of the $W_i^N(\cdot)$, it now follows that these latter sequences are tight and all limits are Wiener processes. The calculation of the variance parameters (4.4) follow from the corresponding values for the prelimit processes.

Part 2. We next prove the lower bound

$$\liminf_N \bar{V}^N(z_N) \geq \bar{V}(w(z)). \tag{5.7}$$

We will use a weak convergence argument. This will be a little indirect, since in general it is not possible to prove tightness of $\bar{F}^N(\cdot)$. In [5, Chap. 11.1.2] and [6, §4] a very useful time rescaling idea was introduced which greatly simplified the treatment of weak convergence issues for "singular" or "reflected" cases. That method will be used here.

Define $T^N(t) = t + \bar{F}_1^N(t) + \bar{F}_2^N(t)$, and its inverse $\hat{T}^N(t) = \inf\{s : T^N(s) > t\}$. Define the "hat" processes by the time transformation: $\hat{Z}^N(t) = Z^N(\hat{T}^N(t))$, $\hat{F}^N(t) = \bar{F}^N(\hat{T}^N(t))$, etc. Note that the transformation "stretches out" time and that $\hat{F}_i^N(\cdot)$ are Lipschitz continuous with constant ≤ 1. (In the references, the Y^N-processes were included in the $T^N(t)$, but that is not needed here.) Then

$$\hat{WL}_i^N(t) = WL_i^N(0) + \hat{W}_i^N(t) + \hat{Y}_i^N(t) + \hat{F}_i^N(t) - b_i\hat{T}^N(t) + \hat{\rho}^N(t)/\sqrt{N},$$

where $\hat{Y}_i^N(\cdot)$ is the reflection process at the boundary segment $w_i = 0$.

The set

$$\{\hat{WL}^N(\cdot), \hat{F}^N(\cdot), \hat{W}^N(\cdot), \hat{T}^N(\cdot), \hat{Y}^N(\cdot), W^N(\cdot)\}$$

is tight, all weak limits are continuous, and $\hat{\rho}^N(\cdot)/\sqrt{N}$ converges to the zero process. Let $\hat{WL}(\cdot)$, $\hat{F}(\cdot)$, $\hat{W}(\cdot)$, $\hat{T}(\cdot)$, $\hat{Y}(\cdot)$, $W(\cdot)$ denote the limit of an arbitrary weakly convergent subsequence. We have $\hat{W}(t) = W(\hat{T}(t))$. The limit processes satisfy

$$\hat{WL}_i(t) = WL_i(0) + \hat{W}_i(t) + \hat{Y}_i(t) + \hat{F}_i(t) - b_i\hat{T}(t). \tag{5.8}$$

By standard weak convergence and martingale arguments it can be shown that the other processes in (5.8) are nonanticipative with respect to the (stretched out) Wiener process $\hat{W}(\cdot)$.

Using the fact that $w(\hat{Z}^N(t)) = \hat{WL}^N(t)$ and $g(z) \geq \bar{g}^N(w^N(z))$ and changing the time scale yield

$$
\begin{aligned}
V^N(z^N, \bar{G}^N, \bar{F}^N) &= E\int_0^\infty e^{-\beta t} g(Z^N(t))dt \\
&= E\int_0^\infty e^{-\beta \hat{T}^N(t)} g(\hat{Z}^N(t))d\hat{T}^N(t) \\
&\geq E\int_0^\infty e^{-\beta \hat{T}^N(t)} \bar{g}^N(\hat{WL}^N(t))d\hat{T}^N(t) \\
&= E\int_0^\infty e^{-\beta \hat{T}^N(t)} \bar{g}(\hat{WL}^N(t))\, d\hat{T}^N(t) \\
&\quad + E\int_0^\infty e^{-\beta \hat{T}^N(t)} [\bar{g}^N(\hat{WL}^N(t)) - \bar{g}(\hat{WL}^N(t))]\, d\hat{T}^N(t).
\end{aligned} \tag{5.9}
$$

There is $\epsilon_N \to 0$ and a constant K such that the last term is bounded above by

$$
\epsilon_N E\int_0^\infty e^{-\beta \hat{T}^N(t)} |\hat{WL}^N(t)| d\hat{T}^N(t) \leq K\epsilon_N V^N(z^N, \bar{G}^N, \bar{F}^N),
$$

which goes to zero. Also, by the weak convergence and Fatou's lemma,

$$
\liminf_N E\int_0^\infty e^{-\beta \hat{T}^N(t))} \bar{g}(\hat{WL}^N(t))d\hat{T}^N(t) \geq E\int_0^\infty e^{-\beta \hat{T}(t))} \bar{g}(\hat{WL}(t))d\hat{T}(t). \tag{5.10}
$$

Now define the inverse $T(t) = \inf\{s : \hat{T}(s) > t\}$. (See [5, p. 312] or [6, Thm. 5.3] for a similar transformation and application.) Since $\hat{T}^N(t) \leq t, \hat{T}(t) \leq t$. Since $\limsup_N E\bar{F}_i^N(t) < \infty$ for each t, $\hat{T}(t)$ goes to infinity with probability one as $t \to \infty$. Hence, $T(t) \to \infty$ as $t \to \infty$ with probability one. It is also right continuous. Define the rescaled processes by $WL(t) = \hat{WL}(T(t))$, $F(t) = \hat{F}(T(t))$, etc. The rescaled processes satisfy (4.3), the other rescaled processes are nonanticipative with respect to the Wiener process $W(\cdot)$, and $Y(\cdot)$ is the reflection process. We now see that the right side of (5.10) equals

$$
E\int_0^\infty e^{-\beta t} \bar{g}(WL(t))dt = V(w(z), F).
$$

By the minimality of $\bar{V}(w(z))$,

$$
V(w(z), F) \geq \bar{V}(w(z)). \tag{5.11}
$$

Now (5.9)–(5.11) and the fact that $\bar{G}^N(\cdot)$, $\bar{F}^N(\cdot)$ are arbitrary controls with finite cost yields (5.7).

Part 3. Now we need to get the reverse inequality to (5.7), namely,

$$
\limsup_N \bar{V}^N(z^N) \leq \bar{V}(w(z)). \tag{5.12}
$$

This will be done via a standard "comparison control" method (see [5, 6]) and using the time transformation idea of Part 2. Fix $\epsilon > 0$ and let $F^\epsilon(\cdot)$ be an ϵ-optimal policy for (4.3), (4.5) defined by the $q_{nji}(\cdot)$ conditional probability functions in Lemma 5.2, and with the properties

given in Lemma 5.2. Recall that the conditional distribution of the jumps of $F^\epsilon(\cdot)$ at time $n\Delta$ is given by the functions q_{nij} of (5.3). We write

$$WL(t) = w(z) + W(t) + F^\epsilon(t) + Y(t) - bt. \tag{5.13}$$

We will adapt the $F^\epsilon(\cdot)$ strategy to the physical process by appropriately adapting the $q_{nji}(\cdot)$ rules, and then use the optimality of $\bar{V}^N(z^N)$ to get the desired result.

Denote the *desired* adaptation of $F^\epsilon(\cdot)$ to the physical system by $F^N_\epsilon(\cdot)$. The subscript ϵ will be used to denote the various processes ($WL^N_\epsilon(\cdot)$, $W^N_\epsilon(\cdot)$, etc.) associated with the adaptation of the $F^\epsilon(\cdot)$ to the physical system. Define a set of random variables $\delta\bar{F}^N_\epsilon(n\Delta) = (\delta\bar{F}^N_{\epsilon,1}(n\Delta), \delta\bar{F}^N_{\epsilon,2}(n\Delta))$ in such a way that

$$\begin{aligned} &q_{nji}(\delta\bar{F}^N_\epsilon(m\Delta), m < n, W^N_\epsilon(l\Delta_1), l\Delta_1 \le n\Delta) \\ &\quad = P\{\delta\bar{F}^N_{\epsilon,i}(n\Delta) = j\rho | \delta\bar{F}^N_\epsilon(m\Delta), m < n, W^N_\epsilon(l\Delta_1), l\Delta_1 \le n\Delta\}. \end{aligned} \tag{5.14}$$

We would like the conditional distribution of the jumps of $F^N_\epsilon(\cdot)$ at time $n\Delta$ to be exactly equal to $\delta\bar{F}^N_\epsilon(n\Delta)$. However, the $F^N_\epsilon(\cdot)$ cannot be realized as a "pure jump" process, as can the $F^\epsilon(\cdot)$, since the only physical influence that we have over the $F^N_\epsilon(\cdot)$ is via manipulation of the idle times and sequencing of the customer classes at processor 1. But there will be $\delta'_N \to 0$ such that for each n, the $\delta\bar{F}^N_\epsilon(n\Delta)$ are realized by $F^N_\epsilon(\cdot)$ on $[n\Delta, n\Delta + \delta'_N]$, with probability close to one as $N \to \infty$. By "realizing" $\delta\bar{F}^N_{\epsilon,i}(n\Delta)$ on an interval $[n\Delta, t]$, $t \le n\Delta + \Delta$, we mean that $F^N_{\epsilon,i}(t) - F^N_{\epsilon,i}(n\Delta) = \delta\bar{F}^N_{\epsilon,i}(n\Delta)$.

The $F^N_\epsilon(\cdot)$ which will be constructed below will be bounded by the bound on the $F^\epsilon(\cdot)$ and will not increase after time $T(\epsilon)$. Thus, for each $T_1 < \infty$,

$$\sup_N E \sup_{t \le T_1} |Z^N(t)|^2 < \infty. \tag{5.15a}$$

It is easy to verify by a direct calcualation that there are real numbers k_i such that for all $t < \infty$

$$\sup_N E|WL^N(t)| \le k_1 + k_2 t. \tag{5.15b}$$

Equations (5.15a) and (5.15b) imply that as long as $F^N_\epsilon(\cdot)$ is uniformly bounded, nonanticipative, and constant after some $T(\epsilon)$ (no matter how the terms are realized), we have that

$$\lim_T \limsup_N E \int_T^\infty e^{-\beta t} |WL^N_\epsilon(t)| dt = 0, \tag{5.15c}$$

and then it follows that

$$\lim_T \limsup_N E \int_T^\infty e^{-\beta t} |Z^N_\epsilon(t)| dt = 0. \tag{5.15d}$$

Keep in mind that the control constructed below is used only to prove (5.12), and is not a practical control.

In the first two cases (0a, 0b) and (1a, 1b) below, we modify the problem as follows. If $Z^N_{\epsilon,3} = 1/\sqrt{N}$ and $Z^N_{\epsilon,2}(t) > 0$ and we are serving queue 2, then if the current service at queue 3 is completed before the current service at queue 2, extend the service at queue 3 until the next customer arrives there so that queue 3 is not empty in this event. This does not affect the truth of (5.15a)–(5.15d). Also, if (5.12) holds with this modification it holds without it, since the modification does not decrease the costs. The modification serves to assure that $F^N_{\epsilon,2}(\cdot)$

will not increase in this interval due to queue 3 waiting for a customer when a nonempty queue 2 is being served. Recall that (Lemma 5.2) at most one component of $\bar{F}^{\epsilon}(\cdot)$ can increase at a time, hence at most one component of $\delta\bar{F}^N_{\epsilon}(n\Delta)$ can be positive for each n. Let ϵ_N be a sequence of positive numbers going to zero and satisfying $\epsilon_N\sqrt{N} \to \infty$.

First, suppose that at $t = n\Delta$, the sample that (5.14) tells us to realize is $\delta\bar{F}^N_{\epsilon}(n\Delta) = 0$. Then follow the steps (0a, 0b) below until time $(n\Delta + \Delta)$. Processor 2 continues to work when there is work to do.

(0a) Serve queue 1 until either $\{Z^N_{\epsilon,1}(t) = 0$ *and* $Z^N_{\epsilon,2}(t) > 0\}$ or $\{Z^N_{\epsilon,3}(t) \le \epsilon_N$ *and* $Z^N_{\epsilon,2}(t) > 0\}$ and then go to (0b).

(0b) Serve queue 2 until either $Z^N_{\epsilon,2}(t) = 0$ or $\{Z^N_{\epsilon,1}(t) \ge \epsilon_N$ *and* $Z^N_{\epsilon,3}(t) > \epsilon_N\}$. Then go to (0a).

By (5.15a) and the above rules, for any $\epsilon' > 0$

$$\lim_N P\{Z^N_{\epsilon,1}(t)Z^N_{\epsilon,3}(t) \ge \epsilon'\} = 0, \quad t \in (n\Delta, n\Delta + \Delta]. \tag{5.16}$$

By (5.16), the fact that $|\tilde{g}^N(z)| \le k_0 + k_1|z|$ for some real k_i, the fact that $\tilde{g}^N(z) \to 0$ uniformly on any compact set as $z_1 z_3 \to 0$, and the uniform integrability (5.15d), we have

$$\begin{gathered} \lim_N E\tilde{g}^N(Z^N_{\epsilon}(t)) = 0, \quad t \in (n\Delta, n\Delta + \Delta], \\ \sup_{t,N} E\tilde{g}^N(z^N(t)) < \infty. \end{gathered} \tag{5.17}$$

Note that the procedure (0a, 0b) does not increase the F^N_{ϵ}-term. While the $G^N_{\epsilon}(\cdot)$-terms do not appear explicitly above, they represent the part of the given rule which guarantees (5.16).

Now suppose that the realization that we aim for at time $n\Delta$ is $\delta\bar{F}^N_{\epsilon,1}(n\Delta) > 0$. Then follow the steps (1a, 1b) below until time $(n\Delta + \Delta)$ is reached. Processor 2 continues to work when there is work to do.

(1a) Processor 1 serves queue 1 only (or idles when $Z^N_{\epsilon,1}(t) = 0$) until either $Z^N_{\epsilon,3}(t) \le \epsilon_N$ or $\delta\bar{F}^N_{\epsilon,1}(n\Delta)$ is realized. In the second case, go to (0a). In the first case, if $Z^N_{\epsilon,2}(t) > 0$ go to (1b). Otherwise serve queue 1 or idle (if $Z^N_{\epsilon,1}(t) = 0$) until $Z^N_{\epsilon,2}(t) > 0$, and then go to (1b).

(1b) Processor 1 serves queue 2 or idles (if $Z^N_{\epsilon,2}(t) = 0$) until either $Z^N_{\epsilon,3}(t) \ge \epsilon_N$ or $\delta\bar{F}^N_{\epsilon,1}(n\Delta)$ is realized, whichever comes first. In the first case, go to (1a), and in the second go to (0a).

The increase in $F^N_{\epsilon,1}(\cdot)$ is due to the idling of processor 1 when either queue 1 or 2 is nonzero. There are $\delta'_N \to 0$ such that the probability that the desired magnitude $\delta\bar{F}^N_{\epsilon,1}(n\Delta)$ is realized on $[n\Delta, n\Delta + \delta'_N]$ goes to unity as $N \to \infty$. Note that (5.16), (5.17) continue to hold, and that $F^N_{\epsilon,2}(\cdot)$ does not change on the interval.

Next, suppose that our aim is to realize $\delta\bar{F}^N_{\epsilon,2}(n\Delta) > 0$. Then follow (2a) below until time $(n\Delta + \Delta)$ is reached.

(2a) Shut down processor 2 and serve both queues 1 and 2, with queue 1 given priority. If $\delta\bar{F}_{\epsilon,2}(n\Delta)$ is realized then go to (0a).

Note that $F^N_{\epsilon,1}(\cdot)$ doesn't increase during this last procedure, and that (5.16), (5.17) hold. The increase in $F^N_{\epsilon,2}(\cdot)$ is due to the idling of processor 2 and the fact that queue 2 continues to be served when queue 1 is empty.

We now use the time transformation $\hat{T}^N(t)$ of Part 2 (to be called $\hat{T}^N_{\epsilon}(t)$ here). Abusing notation, let us henceforth use N to index a weakly convergent subsequence of the

$$\{\hat{WL}^N_{\epsilon}(\cdot), \hat{F}^N_{\epsilon}(\cdot), \hat{W}^N_{\epsilon}(\cdot), \hat{T}^N_{\epsilon}(\cdot), \hat{Y}^N_{\epsilon}(\cdot), W^N_{\epsilon}(\cdot)\}.$$

Denote the limit by $(\hat{WL}_{\epsilon}(\cdot), \hat{F}_{\epsilon}(\cdot), \hat{W}_{\epsilon}(\cdot), , \hat{T}_{\epsilon}(\cdot), \hat{Y}_{\epsilon}(\cdot), W_{\epsilon}(\cdot)$. As in Part 2 of the proof,

$$WL_{\epsilon}(t) = w(z) + \hat{W}_{\epsilon}(\cdot) + \hat{F}_{\epsilon}(\cdot) - b\hat{T}_{\epsilon}(t) + \hat{Y}_{\epsilon}(\cdot). \tag{5.18}$$

We have

$$V^N(z^N, G_\epsilon^N, F_\epsilon^N) = E\int_0^\infty e^{-\beta t} g(Z_\epsilon^N(t))dt$$
$$= E\int_0^\infty e^{-\beta t}\bar{g}^N(WL_\epsilon^N(t))dt + E\int_0^\infty e^{-\beta t}\tilde{g}^N(Z_\epsilon^N(t))dt.$$

By (5.17) the term with $\tilde{g}^N$ goes to zero as $N \to \infty$. We can write

$$E\int_0^\infty e^{-\beta t}\bar{g}^N(WL_\epsilon^N(t))dt = E\int_0^\infty e^{-\beta \hat{T}_\epsilon^N(t)}\bar{g}^N(\hat{WL}_\epsilon^N(t))d\hat{T}_\epsilon^N(t). \tag{5.19}$$

Now (5.15a), (5.15b), and the fact $F_\epsilon^N(\cdot)$ stops increasing after $T(\epsilon)$ can be used to show that

$$\lim_T \limsup_N E\int_T^\infty e^{-\beta \hat{T}_\epsilon^N(t)}|\hat{WL}_\epsilon^N(t)|dt = 0. \tag{5.20}$$

Now using (5.15a), (5.20), the weak convergence, the linear growth of $\bar{g}(\cdot)$ and $\bar{g}^N(\cdot)$ (uniformly in N), the convergence $\bar{g}^N(\cdot)$ to $\bar{g}(\cdot)$ (uniformly on compacta), and the fact that for large t, $d\hat{T}_\epsilon^N(t) = dt$, we see that the right side of (5.19) converges to

$$E\int_0^\infty e^{-\beta \hat{T}_\epsilon(t)}\bar{g}(\hat{WL}_\epsilon(t))d\hat{T}_\epsilon(t). \tag{5.21}$$

Define $T_\epsilon(t) = \inf\{s : \hat{T}_\epsilon(s) > t\}$. Then defining the inverses $WL_\epsilon(t) = \hat{WL}(T_\epsilon(t))$, etc., analogous to what was done in Part 2, yields

$$WL_\epsilon(t) = w(z) + W_\epsilon(t) + F_\epsilon(t) - bt + Y_\epsilon(t),$$

where $Y_\epsilon(\cdot)$ is the reflection term and the processes are nonanticipative with respect to the Wiener process $W_\epsilon(\cdot)$. By the inverse transformation, the right side of (5.21) equals

$$E\int_0^\infty e^{-\beta t}\bar{g}(WL_\epsilon(t))dt. \tag{5.22}$$

By the minimality of $\bar{V}^N(z^N)$, we have

$$\limsup_N \bar{V}^N(z^N) \le \lim_N V^N(z^N, F_\epsilon^N, G_\epsilon^N) = (5.22). \tag{5.23}$$

By the method of construction of the $\delta\bar{F}_\epsilon^N(\cdot)$ in terms of the $W_\epsilon^N(\cdot)$ via (5.14) and the rules (0a, 0b), (1a, 1b) and (2a), we see that the distribution of $(F_\epsilon(\cdot), W_\epsilon(\cdot))$ is the same as that of $(F^\epsilon(\cdot), W(\cdot))$ of (5.13). Using this and the weak sense uniqueness of the solution to (4.3), (5.22) equals $V(w(z), F^\epsilon)$. By the ϵ-optimality of $F^\epsilon(\cdot)$, $V(w(z), F^\epsilon) \le \bar{V}(w(z)) + \epsilon$. Since ϵ is arbitrary, (5.11) follows. □

Comments on extensions. The scheme used in the paper follows a simple pattern which should be applicable to more general problems. First, $g^N(\cdot)$ is minimized subject to the "workload" constraints with minimum denoted by $\bar{g}^N(\cdot)$. Then the dynamical equation for the workload process is obtained in terms of a control (the F^N-term) and a reflection (the Y^N-term). A weak convergence analysis is applied to the workload formulation, under heavy traffic conditions and cost rate $\bar{g}^N(\cdot)$. This yields (5.7). To get (5.12), use the workload

equations to get uniform integrability, and obtain appropriate estimates for the "errors" $\tilde{g}^N(\cdot)$ and $\bar{g}^N(\cdot) - \bar{g}(\cdot)$. The appropriate form of the ϵ-optimal comparison control $F^\epsilon(\cdot)$ will exist under quite broad conditions, as shown in the references. The main problem is getting the right rule for the realization of the $F^\epsilon(\cdot)$ on the physical process, something akin to our (0a, 0b), (1a, 1b), (2a). A more general method for doing this without having to compute the minimizing $z(w)$ explicitly in each case is needed.

REFERENCES

[1] S. N. ETHIER AND T. G. KURTZ, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1986.

[2] J. M. HARRISON AND L. M. WEIN, *Scheduling networks of queues: Heavy traffic analysis of a two-station closed network*, Oper. Res., 38 (1990), pp. 1052–1064.

[3] ———, *Scheduling networks of queues: Heavy traffic analysis of a simple open network*, Queueing Systems, 5 (1989), pp. 265–280.

[4] H. J. KUSHNER, *Control of trunk line systems in heavy traffic*, SIAM J. Control Optim., 33 (1995), pp. 765–803.

[5] H. J. KUSHNER AND P. DUPUIS, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer, New York, Berlin, 1992.

[6] H. J. KUSHNER AND L. F. MARTINS, *Numerical methods for stochastic singular control problems*, SIAM J. Control Optim., 29 (1991), pp. 1443–1475.

[7] H. J. KUSHNER AND K. M. RAMACHANDRAN, *Optimal and approximately optimal control policies for queues in heavy traffic*, SIAM J. Control Optim., 27 (1989), pp. 1293–1318.

[8] L. F. MARTINS, S. E. SHREVE, AND H. M. SONER, *Heavy traffic convergence of a controlled, multi-class queueing system*, SIAM J. Control Optim., 34 (1996), to appear.

[9] L. F. MARTINS AND H. J. KUSHNER, *Routing and singular control for queueing networks in heavy traffic*, SIAM J. Control Optim., 28 (1990), pp. 1209–1233.

[10] L. M. WEIN, *Scheduling networks of queues: Heavy traffic analysis of a two-station network with controllable inputs*, Oper. Res., 38 (1990), pp. 1065–1078.

[11] ———, *Scheduling networks of queues: Heavy traffic analysis of a multistation network with controllable inputs*, Oper. Res., 40 (1992), pp. S312–S334.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1798–1813, September 1996

016

STABILIZABILITY DOES NOT IMPLY HOMOGENEOUS STABILIZABILITY FOR CONTROLLABLE HOMOGENEOUS SYSTEMS*

RODOLPHE SEPULCHRE† AND DIRK AEYELS‡

Abstract. This paper presents an example of a homogeneous planar controllable system which is stabilizable while not stabilizable by homogeneous feedback. Addition of an integrator to the system provides a three-dimensional *affine* system with the same properties.

Key words. continuous feedback stabilization, homogeneous systems, homogeneous feedbacks

AMS subject classifications. 34D20, 93D15, 93D20

1. Introduction.

This paper deals with the local stabilization problem for nonlinear systems that are small-time locally controllable (STLC, [9]) and homogeneous with respect to some dilation [8]. By stabilization, we mean that there exists a continuous feedback $u(x, y) \in C^1(\mathbb{R}^2\backslash\{0\}, \mathbb{R})$ that renders the null solution of the closed-loop system (locally) asymptotically stable. This particular class of systems has recently been considered by several authors (e.g., [12], [8], [7], [2]). These systems appear naturally as local approximations of general controllable systems where the higher-order terms have been neglected. The associated homogeneous approximation might exhibit important control properties of the original system: for instance, local controllability of the homogeneous approximation implies local controllability of the original system. (See, for instance, [8]. Note that the converse is not true in general [10].) It seems then an attractive idea—in order to construct a stabilizing feedback for a nonlinear system—to consider a homogeneous approximation and to try to design a *homogeneous* stabilizing feedback for this approximation. First of all this would be a simpler problem to solve—in particular for small-dimensional systems—as one may take full profit of the symmetry properties of the homogeneous vectorfield [12]. But more important, for the closed-loop *homogeneous* system there exists a *homogeneous* Liapunov function [15]. This in turn implies robustness of the stabilization law with respect to higher-order terms (see also [8] for an independent proof). In summary, the stabilizing homogeneous control law of the homogeneous approximation would serve as a stabilizing feedback for the original system.

It is well known that with respect to the linearization of a nonlinear system this procedure works perfectly well. Also for the special class of *planar affine* controllable systems Kawski [11] has shown that every *planar affine* controllable system admits a homogeneous approximation which is stabilizable by homogeneous feedback. It has, however, been shown that the restriction to homogeneous feedback introduces extra necessary conditions for the stabilization problem [12], [7]. This leads us to the fundamental question of whether the approach sketched above works in general or—if not—for systems that are affine in the control (see also [7]).

The goal of this paper is to show that for general controllable homogeneous systems, the existence of a stabilizing feedback does not necessarily imply the existence of a *homogeneous*

*Received by the editors May 6, 1994; accepted for publication (in revised form) July 17, 1995. This research was supported by the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture, and by the EC-Science Project SC1-0433-C(A).

†Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 (sepulchr@bessel.ece.ucsb.edu). Permanent address: CESAME, Université Catholique de Louvain, Bâtiment Euler, Avenue Georges Lemaître 4-6, B1348 Louvain-La-Neuve, Belgium. The research of this author was supported in part by National Science Foundation grant ECS-9203491, Air Force Office of Scientific Research grant F-49620-92-J-0495, the Belgian American Educational Foundation, and the North Atlantic Treaty Organization.

‡Department of Systems Dynamics, Universiteit Gent, Technologiepark-Zwijnaarde 9, 9052 Gent, Belgium (dirk.aeyels@rug.ac.be).

stabilizing feedback. More precisely we will show that the system

$$(\Sigma_1) \qquad \begin{cases} \dot{x} &= x + u, \\ \dot{y} &= 3y + xu^2, \end{cases} \qquad (x, y) \in \mathbb{R}^2,\ u \in \mathbb{R},$$

is stabilizable while not stabilizable by homogeneous feedback.

The restriction to homogeneous stabilization is also a limitation for *affine* controllable systems of dimension larger than two. Indeed we will prove that the addition of an integrator to (Σ_1) results in a three-dimensional affine controllable system

$$(\Sigma_1') \qquad \begin{cases} \dot{x} &= x + z, \\ \dot{y} &= 3y + xz^2, \\ \dot{z} &= u, \end{cases} \qquad (x, y, z) \in \mathbb{R}^3,\ u \in \mathbb{R},$$

which also is stabilizable while not stabilizable by homogeneous feedback.

In §2, we review some definitions and provide basic properties for (Σ_1) and (Σ_1'). Section 3 is devoted to the explicit construction of a stabilizing feedback for (Σ_1). We also analyze the regularity of the proposed feedback. In §4, we combine the result of §3 and a recent result of Rosier [14] in order to establish the stabilizability of (Σ_1'). Conclusions are given in §5.

2. Preliminaries. We first recall the general definition of a homogeneous (control) system as given in [6]: the control system

$$\dot{x} = f(x, u) \tag{1}$$

with $f = (f_i)_{i=1,n} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ a map of class C^1 satisfying

$$\forall i \in \{1, \dots, n\},\ \forall x = (x_1, \dots, x_n)^T \in \mathbb{R}^n,\ \forall \epsilon \geq 0,\ \forall u \in \mathbb{R},$$
$$f_i(\epsilon^{r_1} x_1, \dots, \epsilon^{r_n} x_n, \epsilon^{r_{n+1}} u) = \epsilon^{\tau + r_i} f_i(x_1, \dots, x_n, u)$$

for some $r_i > 0$ and some $\tau \in (-\min_j\{r_j\}, +\infty)$ is said to be *homogeneous of degree* τ *with respect to the dilation* $\delta_\epsilon^r(x, u) = (\epsilon^{r_1} x_1, \dots, \epsilon^{r_n} x_n, \epsilon^{r_{n+1}} u)$. The system (1) is *stabilizable by homogeneous feedback* if there exists a continuous feedback law $u(x) \in C^1(\mathbb{R}^n \setminus \{0\}, \mathbb{R})$ such that

1. $u(\epsilon^{r_1} x_1, \dots, \epsilon^{r_n} x_n) = \epsilon^{r_{n+1}} u(x_1, \dots, x_n)\ \forall x \in \mathbb{R}^n,\ \forall \epsilon \geq 0$;
2. the point $x = 0$ is a globally asymptotically stable equilibrium of the closed-loop system $\dot{x} = f(x, u(x))$.

The following proposition collects some basic properties of the system (Σ_1) and of its dynamic extension (Σ_1').

PROPOSITION 1. (i) (Σ_1) *(resp.,* (Σ_1')*) is homogeneous of degree* 0 *with respect to the dilation* $\delta_\epsilon^r(x, y, u) = (\epsilon x, \epsilon^3 y, \epsilon u)$ *(resp.,* $\delta_\epsilon^r(x, y, z, u) = (\epsilon x, \epsilon^3 y, \epsilon z, \epsilon u)$*);*

(ii) (Σ_1) *and* (Σ_1') *are locally controllable (i.e., STLC);*

(iii) (Σ_1) *and* (Σ_1') *are not stabilizable by homogeneous feedback.*

Proof. (i) The proof follows from the definition.

(ii)(Σ_1') is locally controllable: using the standard notation

$$f(x, y, z) := (x + z, 3y + xz^2, 0)^T,\ g(x, y, z) := (0, 0, 1)^T,$$

we compute that the nonvanishing brackets at the origin are $g(0) = (0, 0, 1)^T$, $[f, g](0) = (-1, 0, 0)^T$, and $[[[f, [f, g]], g], g](0) = (0, -2, 0)^T$. They are independent and satisfy the so-called Hermes condition. Using a result of Coron (see §5 of [5]), this also implies that (Σ_1) is locally controllable.

(iii) Suppose that $u(x, y)$ is a homogeneous stabilizing feedback for (Σ_1). Then there exists a homogeneous Liapunov function $V(x, y)$ such that [15]

$$\nabla V(x, y).h(x, y) < 0 \quad \forall (x, y) \neq (0, 0),$$

where $h(x, y) := (x + u(x, y), 3y + xu^2(x, y))^T$ and with the additional property (due to homogeneity of V, say, of degree k) that

$$\nabla V(x, y).\lambda(x, y) = kV(x, y) > 0 \quad \forall (x, y) \neq (0, 0),$$

where $\lambda(x, y) = (x, 3y)^T$ is the Euler vector field associated with the dilation $\delta_\epsilon^r(x, y) = (\epsilon x, \epsilon^3 y)$. As a consequence, we have that

$$\nabla V(x, y).(h(x, y) - \lambda(x, y)) < 0 \quad \forall (x, y) \neq (0, 0),$$

which means that $u(x, y)$ also stabilizes the system $f(x, u) := (u, xu^2)^T$. But this is a contradiction since the latter system does not satisfy Brockett's necessary condition for continuous stabilization [3]. A similar argument proves that (Σ_1') also is not stabilizable by homogeneous feedback. □

3. Main result. This section is devoted to an explicit construction of a stabilizing feedback for (Σ_1) and to an analysis of its properties.

THEOREM 1. *There exists a stabilizing feedback $u(x, y) \in C(\mathbb{R}^2, \mathbb{R})$ such that the origin of (Σ_1) is locally asymptotically stable.*

Proof. The proof is constructive. For each $\epsilon > 0$ sufficiently small, we construct in the plane a simple closed curve ∂V_ϵ surrounding the origin. The family of curves $\{\partial V_\epsilon\}_{0<\epsilon\leq\bar{\epsilon}}$ implicitly defines the ϵ-level sets of a continuous positive definite function $V(x, y)$. We design a continuous function $u(x, y)$ such that

$$D^+V(x, y) \leq 0 \quad \forall (x, y) \in U, \tag{2}$$

where $D^+(.)$ denotes the right derivative (also called the Dini derivative; see [13]) of V along the trajectories of the closed-loop system.

Our construction is illustrated in Figure 1. We divide the plane in sectors delimited by the following curves:

1. $\mathcal{A} \equiv \{(x, y) \mid y + \alpha x^3 = 0\}$,
2. $\mathcal{B} \equiv \{(x, y) \mid y + \beta x^{3+\nu} = 0\}$,
3. $\mathcal{C} \equiv \{(x, y) \mid y + \gamma x^{3+\nu} = 0\}$,
4. $\mathcal{D} \equiv \{(x, y) \mid y + \delta x^{3+\nu} = 0\}$,
5. $\mathcal{E} \equiv \{(x, y) \mid y - 3x^3 = 0\}$,

where the positive constants $0 < \nu < 1$ and $\alpha > \beta > \gamma > \delta > 0$ will be suitably chosen in the following. (The value of α and the ratios β/γ and γ/δ will be fixed in the course of the proof.)

Let $\epsilon > 0$ sufficiently small. In each sector we will define ∂V_ϵ as a smooth curve in x and y and design the control function in such a way that (2) is satisfied in the considered region. Each sector being symmetric with respect to the origin, the construction is given in a half-sector and can be extended to the full sector by means of a symmetry argument. (As a consequence, the stabilizing control function will be odd.) The construction holds for arbitrary small ϵ, and therefore (2) ensures stability of the origin. In addition, we will show that the largest invariant set contained in the set $\{(x, y) \mid D^+V(x, y) = 0\}$ is the origin. Asymptotic stability of the origin follows by LaSalle's theorem.

I. Let S_1 be the connected region of the right-half plane delimited by $\mathcal{D}$ and $\mathcal{E}$. Let $D \in \mathcal{D}$, with $x_D = d\epsilon$, $d > 1$, and $E \in \mathcal{E}$, with $x_E = \epsilon$. Define ∂V_ϵ in S_1 by connecting D

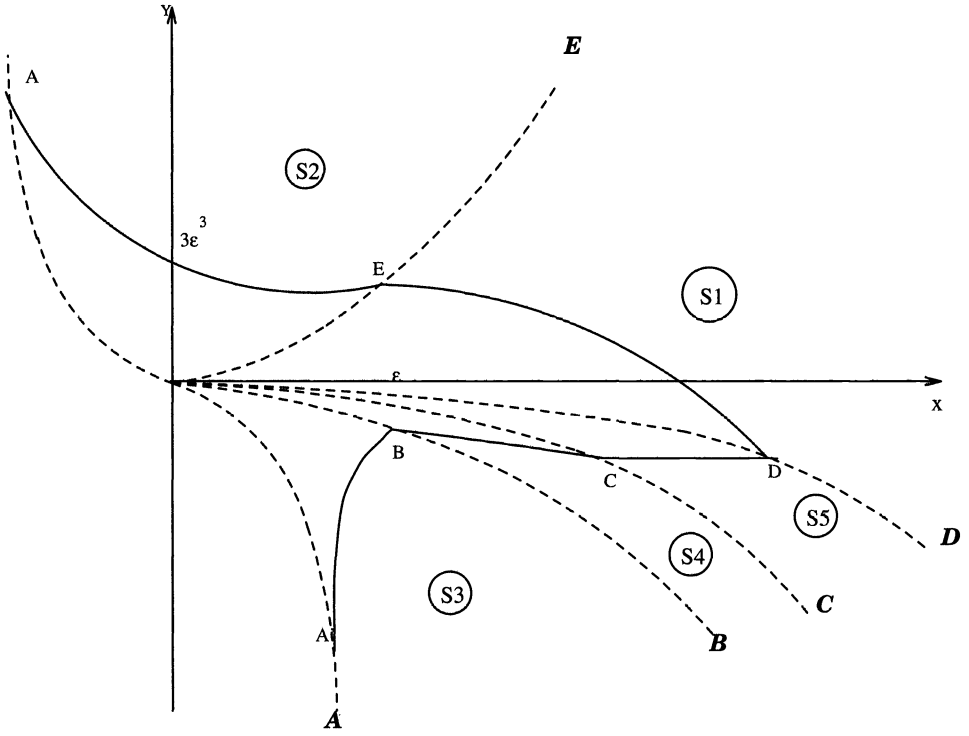


FIG. 1. *Half ϵ-level set of $V(x, y)$ for small $\epsilon > 0$.*

and E along the smooth curve

$$\Gamma_1 \equiv \{(x, y) \mid y + nx^3 - (n+3)\epsilon^3 = 0,\ x \in [\epsilon, d\epsilon]\}, \tag{3}$$

with n chosen such that D belongs to $\mathcal{D}$, i.e.,

$$n = n(\epsilon) = \frac{3 + \delta d^{3+\nu}\epsilon^{\nu}}{d^3 - 1}. \tag{4}$$

On Γ_1, the constraint (2) can be rewritten as follows:

$$\dot{\Gamma}_1 = 3(n+3)\epsilon^3 + ux(u + 3nx) \leq 0, \quad \epsilon \leq x \leq d\epsilon. \tag{5}$$

Since $x > 0$ on Γ_1, the quantity $ux(u+3nx)$ is minimized with the choice $u(x, y) = -3nx/2$. For simplicity, we design a control independent of ϵ and we choose $u(x, y) = -3\bar{n}x/2$ with

$$\bar{n} := n(0) = \frac{3}{d^3 - 1}.$$

Choosing for instance $\bar{n} = 10/3$ (which fixes the value of $d = (19/10)^{1/3}$), we obtain a smooth control function $u(x, y)$ in S_1:

$$u(x, y) := f_1(x, \epsilon) = -5x. \tag{6}$$

Substituting (6) in (5) we obtain on Γ_1

$$\dot{\Gamma}_1 \leq -6\epsilon^3 - 12(n - \bar{n})\epsilon^3. \tag{7}$$

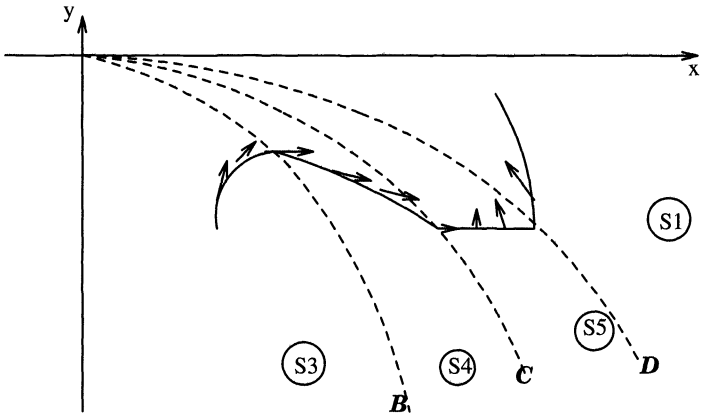


FIG. 2. *Change of sign in* $u(x, y)$.

Since $(n - \bar{n})$ is positive for each $\epsilon \geq 0$, the right-hand side of the above inequality is negative and (2) is satisfied in S_1.

II. Next let S_2 be the connected region of the upper half-plane delimited by $\mathcal{E}$ and $\mathcal{A}$. Let $A \in \mathcal{A}$, with $x_A = -\epsilon/2$. Define ∂V_ϵ in S_2 by connecting E and A along the smooth curve

$$\Gamma_2 \equiv \{(x, y) \mid y + 3(x - 2\epsilon)^3 = 0\}, \qquad x \in [x_A, x_E]. \tag{8}$$

Note that this determines the value of α ($\alpha = 375$). On Γ_2, the constraint (2) can be rewritten as follows:

$$xu^2 + 9(x - 2\epsilon)^2(u + 2\epsilon) \leq 0, \qquad x_A \leq x \leq x_E. \tag{9}$$

Define for $(x, y) \in \Gamma_2$ the control function

$$\begin{aligned} u(x, y) &= -5\epsilon, && 0 \leq x \leq x_E, \\ &= Kx - 5\epsilon, && x_A \leq x \leq 0, \end{aligned} \tag{10}$$

where K is a positive constant still to be chosen. It is readily checked that this choice satisfies the constraint (9) on Γ_2 and is compatible with (6) in $E = \Gamma_1 \cap \Gamma_2$. On the other hand, by choosing K sufficiently large, it is possible to obtain in A

$$u(x_A, y_A) = -a\epsilon, \tag{11}$$

with $a > 0$ as large as desired. By smoothing the control (10) around $x = 0$ and by repeating the construction for each $\epsilon > 0$, we design a continuous function

$$u(x, y) := f_2(x, \epsilon)$$

in S_2 such that f_2 is continuously differentiable as a function of x and ϵ, and such that (2) is fulfilled in this region.

III. Consider the lower right half-plane, i.e., the fourth quadrant. By construction, the control is positive along $\mathcal{A}$ (using (11) and oddness of the control) and negative along $\mathcal{D}$ (by (6)). By continuity, the control must change sign in the fourth quadrant. Since $\dot{y} = 3y + xu^2$, the change of sign of $u(x, y)$ leads to a region in the fourth quadrant where $\dot{y}$ is nonpositive (indeed $\dot{y} < 0$ in the neighborhood of $u = 0$). This is illustrated in Figure 2.

We will design the control in such a way that this region is precisely S_4, i.e., the connected region of the fourth quadrant delimited by $\mathcal{B}$ and $\mathcal{C}$. In particular, the constraint $\dot{y} = 0$ determines the control value on $\mathcal{B}$ and $\mathcal{C}$ according to

$$u(x, y) = \sqrt{3\beta x^{2+\nu}} \quad \forall (x, y) \in \mathcal{B},\ x \geq 0 \tag{12}$$

and

$$u(x, y) = -\sqrt{3\gamma x^{2+\nu}} \quad \forall (x, y) \in \mathcal{C},\ x \geq 0. \tag{13}$$

IV. Consider S_3, the connected region of the fourth quadrant delimited by $\mathcal{A}$ and $\mathcal{B}$. Let $A' \in \mathcal{A}$, with $x_{A'} = \epsilon/2$, and $B \in \mathcal{B}$, with $x_B = \epsilon$. Define ∂V_ϵ in sector$(\mathcal{A}, \mathcal{B})$ by connecting A' and B along the smooth curve

$$\Gamma_3 \equiv \{(x, y) \mid y - m(x - \epsilon)^3 + \beta\epsilon^{3+\nu} = 0\}, \qquad x \in [x_{A'}, x_B], \tag{14}$$

where m is chosen in such a way that $A' \in \Gamma_3$, i.e.,

$$m = \alpha - 8\beta\epsilon^{\nu}.$$

Note that m is positive for small ϵ. On Γ_3, the constraint (2) can be rewritten as follows:

$$xu^2 - 3m(x - \epsilon)^2(\epsilon + u) - 3\beta\epsilon^{3+\nu} \geq 0, \qquad x_{A'} < x < x_B. \tag{15}$$

The left-hand side of (15) is a quadratic function in u. For x positive, the expression has two real roots and is positive, provided that

$$u \geq \frac{3m(x - \epsilon)^2 + \sqrt{\rho}}{2x} \tag{16}$$

with $\rho := ((3m(x - \epsilon)^2)^2 + 4x(3m\epsilon(x - \epsilon)^2 + 3\beta\epsilon^{3+\nu})$ being positive in the considered interval $[x_{A'}, x_B] = [\epsilon/2, \epsilon]$. Note that (16) is satisfied in A', provided that a is sufficiently large. On the other hand, (16) or equivalently (15) reduces in B to the constraint $\dot{y} \geq 0$, which is satisfied by construction with the choice (12). As a consequence, we can design a continuous control function along Γ_3, which interpolates between the value $a\epsilon$ in A' and the value at B specified by (12) and satisfies (16). Repeating the argument for each $\epsilon > 0$, we design a continuous function

$$u(x, y) := f_3(x, \epsilon)$$

in $S_3 \cap U$ such that f_3 is continuously differentiable as a function of x and ϵ and such that (2) is satisfied.

V. Consider S_5, the connected region of the fourth quadrant delimited by $\mathcal{C}$ and $\mathcal{D}$. Let $C \in \mathcal{C}$, with $x_C = c\epsilon, c \in (1, d)$. Define ∂V_ϵ in S_5 by connecting C and D along the horizontal line

$$\Gamma_5 \equiv \{(x, y) \mid y = y_C,\ x \in [c\epsilon, d\epsilon]\}. \tag{17}$$

This specifies the ratio γ/δ:

$$\frac{\gamma}{\delta} = \left(\frac{d}{c}\right)^{3+\nu}. \tag{18}$$

The constraint (2) is therefore satisfied in S_5 if $\dot{y} \geq 0$. We thus design a continuous function

$$u(x, y) := f_5(x, \epsilon) \tag{19}$$

in S_5, monotonically decreasing along Γ_5 from $u(x_C, y_C) = -\sqrt{3\gamma x_C^{2+\nu}}$ until $u(x_D, y_D) = -5d\epsilon$, such that (2) is satisfied and such that f_5 is continuously differentiable as a function of x and ϵ.

VI. Finally consider S_4, the connected region of the fourth quadrant delimited by $\mathcal{B}$ and $\mathcal{C}$. Let the control function be a continuous function

$$u(x, y) := f_4(x, \epsilon)$$

in S_4 such that f_4 is continuously differentiable as a function of x and ϵ and monotonically decreases as a function of x from $u(x, y) = \sqrt{3\beta x^{2+\nu}}$ in B to $u(x, y) = -\sqrt{3\gamma x^{2+\nu}}$ in C. It must be shown that B and C can be connected along a curve segment in such a way that (2) is satisfied for some $\beta > \gamma > 0$ and $c \in (1, d)$. For this we note that in S_4 and for x positive, we have

$$u(x, y) \geq -\sqrt{3\gamma x^{2+\nu}}$$

and therefore

$$\dot{x} \geq (1 - \sqrt{3\gamma x^{\nu}})x, \tag{20}$$

$$\dot{y} \geq 3y. \tag{21}$$

Define a constant $\mu \in (0, 1)$. Then (20) can be rewritten as

$$\dot{x} \geq (1 - \sqrt{3\gamma x^{\nu}})x > \mu x, \tag{22}$$

provided that

$$x \in [0, ((1 - \mu)/\sqrt{3\gamma})^{2/\nu}),$$

which is satisfied for $x \in [x_B, x_C]$ when ϵ is small. Consider the following curve in S_4:

$$\Gamma_4 \equiv \{(x, y) \mid x = x_B e^{\mu t},\ y = y_B e^{3t},\ t \geq 0,\ x \leq x_C\}. \tag{23}$$

Since for small ϵ the trajectories of the closed-loop system satisfy (21), (22) in S_4, we obtain

$$\dot{\Gamma}_4(p) < 0 \quad \forall p \in \Gamma_4. \tag{24}$$

It is therefore sufficient to impose that C belongs to Γ_4 in order to obtain a closed curve ∂V_ϵ in such a way that (2) holds in S_4. This leads to the following requirement:

$$\left(\frac{y_C}{y_B}\right)^{\mu} = \left(\frac{x_C}{x_B}\right)^{3} \tag{25}$$

$$\Leftrightarrow \left(\frac{\gamma}{\beta} c^{3+\nu}\right)^{\mu} = c^3 \tag{26}$$

$$\Leftrightarrow c = \left(\frac{\gamma}{\beta}\right)^{\frac{\mu}{3-(3+\nu)\mu}}. \tag{27}$$

This fixes the ratio γ/β and shows that by choosing μ such that

$$\frac{3}{3+\nu} < \mu < 1,$$

we obtain the desired connection between B and C with c as close to 1 as desired. (Note that this flexibility allows to satisfy $c < d$.) The connection between B and C closes the boundary of V_ϵ. Moreover we have shown that the constraint (2) can be satisfied with a strict inequality in any point of ∂V_ϵ except in B and C. Note that by construction $x\dot{x} > 0$ and $\dot{y} = 0$ at any point (except the origin) of the curves $\mathcal{B}$ and $\mathcal{C}$ (see Figure 2). As a consequence these curves do not contain any invariant set outside the origin.

To prove asymptotic stability of the null solution, it remains to show that the different level sets ∂V_ϵ do not intersect and continuously cover a neighborhood of the origin. An argument for this part of the proof is postponed in Lemma 2. □

The next corollary shows that the above construction can be extended to an arbitrarily large bounded set containing the origin (semiglobal stabilization).

COROLLARY 1. *The stabilization of* Σ_1 *is semiglobal.*

Proof. We denote the open ball centered at the origin and of radius r by $B(0, r)$. Let K be the compact to be contained in the region of attraction of the origin, and choose $\bar{r}$ such that $K \subset B(0, \bar{r})$. Note that for each ϵ, the ball $B(0, r(\epsilon))$ is included in V_ϵ with $r(\epsilon)$ defined by

$$r(\epsilon) := \min\left\{\frac{\epsilon}{2}, \frac{3\epsilon^3}{2}, \frac{\beta\epsilon^{3+\nu}}{2}\right\}.$$

As a consequence, it is always possible to choose $\bar{\epsilon}$ large enough such that $B(0, \bar{r}) \subset V_{\bar{\epsilon}}$. The construction of Theorem 1 holds only locally around the origin, i.e., for sufficiently small ϵ. However, note that the constants d and c define the ratios β/γ and γ/δ according to (27) and (18). The remaining freedom can be used in order that the construction of Theorem 1 holds for $\epsilon \leq \bar{\epsilon}$.

First remark that the parameter m defined in (14) must be a positive value in order that the construction hold. This leads to the constraint

$$\beta < \frac{375}{8\bar{\epsilon}^\nu}, \tag{28}$$

which provides a maximum admissible value for the constant β.

Inequality (22) also imposes a maximum value for the constant γ with the constraint

$$\sqrt{3\gamma(c\bar{\epsilon})^\nu} \leq 1 - \mu. \tag{29}$$

Note that inequalities (28) and (29) can be satisfied by choosing δ sufficiently small. □

The remainder of the section is devoted to a regularity analysis of the control function constructed in Theorem 1. Note that for each $i \in \{1, \ldots, 5\}$, the curve Γ_i is defined in S_i by an implicit equation in (x, y, ϵ), which we denote by $G_i(x, y, \epsilon) = 0$. The following lemma shows that this equation implicitly defines a function $\epsilon_i(x, y)$ with some regularity properties. The proof of this technical lemma is given in the appendix.

LEMMA 1. *In each sector* S_i, $i \in \{1, \ldots, 5\}$, *there exists a continuous function* $\epsilon_i(x, y) \in C^1(S_i \backslash \{0\}, \mathbb{R})$ *such that*

(i)

$$\epsilon_i(0, 0) = 0 \; \forall (x, y) \in S_i \backslash \{0\} : G_i(x, y, \epsilon_i(x, y)) = 0, \tag{30}$$

(ii)

$$\forall (x, y) \in S_i \backslash \{0\} \cap U : \left| \frac{\partial \epsilon_i}{\partial x} \right| \leq C, \quad \left| \frac{\partial \epsilon_i}{\partial y} \right| \leq C \mid y \mid^{\frac{1}{3+\nu}-1}, \tag{31}$$

(iii)

$$\forall (x, y) \in S_i \backslash \{0\} : \frac{\partial \epsilon_i}{\partial y}(x, y) \neq 0, \tag{32}$$

with U some neighborhood of $(0, 0)$ and C some positive constant.

The proof of Lemma 1 can be used to complete the proof of Theorem 1, i.e., to show that the level sets ∂V_ϵ, $0 \leq \epsilon \leq \bar{\epsilon}$, do not intersect and cover an arbitrary large neighborhood of the origin. The proof of the following lemma is given in the appendix.

LEMMA 2. *Define the function*

$$V(x, y) := \epsilon_i(x, y), \qquad (x, y) \in S_i.$$

Then there exists a neighborhood U of the origin such that $V \in C^0(U, \mathbb{R})$, $V(0, 0) = 0$, and V is "radially" increasing in U. Moreover, the constants β, γ, and δ of Theorem 1 can be chosen such that U contains an arbitrary large compact set.

The existence of a continuous stabilizing feedback which is continuously differentiable outside the origin is asserted from Theorem 1 by a result of Coron [4]. The next corollary specifies the regularity that we can obtain with our particular construction. This regularity will be used in the next section.

COROLLARY 2. *The continuous stabilizing feedback of Theorem 1 can be chosen in $C^1(\mathbb{R}^2 \backslash \{0\}, \mathbb{R})$. Moreover we have for some constant C*

$$\forall (x, y) \in U \backslash \{(0, 0)\} : \left| \frac{\partial u}{\partial x} \right| \leq C, \quad \left| \frac{\partial u}{\partial y} \right| \leq C \mid y \mid^{\frac{1}{3+\nu}-1}. \tag{33}$$

Proof. First consider a point p in the interior of some sector S_i. The result is then a consequence of the chain rule and of the regularity of f_i and ϵ_i for each $i \in \{1, \ldots, 5\}$:

$$\frac{\partial u}{\partial x}(x, y) = \frac{\partial f_i}{\partial x}(x, \epsilon) + \frac{\partial f_i}{\partial \epsilon}(x, \epsilon) \frac{\partial \epsilon_i}{\partial x}(x, y), \tag{34}$$

$$\frac{\partial u}{\partial y}(x, y) = \frac{\partial f_i}{\partial \epsilon}(x, \epsilon) \frac{\partial \epsilon_i}{\partial y}(x, y), \tag{35}$$

with $G_i(x, y, \epsilon_i(x, y)) = 0$. The right-hand members of (34) and (35) are continuous and bounded in a neighborhood of the origin. As a consequence, $u(x, y)$ is C^1 in the interior of each sector S_i. Define for each i the constant

$$C_i := \sup_{S_i \cap U} \left\{ \left| \frac{\partial f_i}{\partial x} \right|, \left| \frac{\partial f_i}{\partial \epsilon} \right| \right\} < +\infty. \tag{36}$$

It follows from (34) and (35) that

$$\left| \frac{\partial u}{\partial x}(x, y) \right| \leq \max_i C_i \left(1 + \left| \frac{\partial \epsilon}{\partial x}(x, y) \right| \right), \tag{37}$$

$$\left| \frac{\partial u}{\partial y}(x, y) \right| \leq \max_i C_i \left| \frac{\partial \epsilon}{\partial y}(x, y) \right|. \tag{38}$$

Using Lemma 1(ii), we obtain the estimate (33) for any point in the interior of $S_i \cap U$, $i \in \{1, \ldots, 5\}$.

Next consider a point $(x, y) \in S_i \cap S_j$ different from the origin. We will further specify the choice of the control in the neighborhood of the curves $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ in order to achieve differentiability outside the origin. (The control is already smooth in the neighborhood of $\mathcal{E}$.) Since the procedure is relatively straightforward, we give a complete argument only around $\mathcal{A}$ and omit the details for the other curves.

Define the local coordinate $\theta = y^{1/3}/x$, $x > 0$. We claim that for a sufficiently small constant $\xi > 0$, we can choose the control law around $\mathcal{A}\backslash\{0\}$ according to the particular form

$$u(x, y) = (g_1(\theta) + C_1)x, \qquad \theta \in [\bar{\theta} - \xi, \bar{\theta} + \xi], \tag{39}$$

with $\bar{\theta} := -\alpha^{1/3}$, C_1 a constant, and $g_1(\theta)$ smooth. For $-\infty < \theta \leq \bar{\theta}$, there is no restriction to impose on $f_2(x, \epsilon)$ to be of the form (39), provided that $g_1(\theta)$ is sufficiently large. We can choose g_1 smooth, strictly increasing for $\theta < \bar{\theta}$, and maximum for $\bar{\theta}$. For $\bar{\theta} \leq \theta \leq \bar{\theta} + \xi$, the only constraint imposed by the construction of Theorem 1 is again that $g_1(\theta)$ be sufficiently large in the considered interval. In particular, (11) imposes $g_1(\bar{\theta}) + C \geq 2a$. We can choose g_1 smooth and strictly decreasing for $\theta > \bar{\theta}$.

Note that the definition (39) is not in contradiction with the continuous differentiability of $f_2(x, \epsilon)$ (resp., $f_3(x, \epsilon)$) as a function of ϵ and x in S_2 (resp., S_3). For f_2, this is immediate since the expression (53) gives

$$\epsilon_2(x, y) = \frac{1}{2}(3^{-1/3}\theta + 1)x$$

and therefore

$$\frac{\partial \theta}{\partial \epsilon} = \frac{2\epsilon}{3^{-1/3}x}.$$

It follows that the expression

$$\frac{\partial f_2}{\partial \epsilon}(x, y) = x g_1'(\theta)\frac{\partial \theta}{\partial \epsilon}$$

is well defined and smooth. For f_3, using the chain rule and (56), we obtain

$$\frac{\partial f_3}{\partial \epsilon}(x, y) = g_1'(\theta)\frac{1}{3x^2\theta^2}(K_1(x)\epsilon^2 + K_2(x)\epsilon^{\nu+2}),$$

where θ stands for $y^{1/3}/x$. The above expression is bounded for θ in a neighborhood of $\bar{\theta}$.

A similar procedure can be used to express the control law in the neighborhood of the curves $\mathcal{B}$ and $\mathcal{C}$. Define the local coordinate $\eta := y/x^{3+\nu}$, $x > 0$. Then it is sufficient to choose (locally) the functions f_3, f_4, and f_5 of the particular form

$$u(x, y) = (g_2(\eta))x^{\frac{2+\nu}{2}}, \qquad \eta \in [-\beta - \xi, -\gamma + \xi], \tag{40}$$

with g_2 smooth. We can be in accordance with the construction of Theorem 1 by choosing g_2 smooth, strictly decreasing, and satisfying $g_2(\beta) = \sqrt{3\beta}$ and $g_2(\gamma) = -\sqrt{3\gamma}$.

Finally it is clear that f_5 can be chosen in such a way that the control is smooth along $\mathcal{D}$: this follows from the fact that the partial derivatives of f_1 satisfy

$$\frac{\partial f_1}{\partial \epsilon} = 0, \qquad \frac{\partial f_1}{\partial x} = -5$$

and that the function f_5 is only constrained to be decreasing in x.

From the particular form of the expressions (39) and (40), it is clear that the estimate (33) also holds at the intersection of the different sectors. □

4. Adding an integrator. In this section, we show that the stabilizability properties of Σ_1 are basically unchanged under the addition of an integrator. Proposition 1 states that (Σ_1') is controllable and not stabilizable by homogeneous feedback. To prove that (Σ_1') is stabilizable by (nonhomogeneous) continuous feedback, we apply to (Σ_1) the following result of Rosier.

PROPOSITION 2 (see [14]). *Let $F \in C^0(\mathbb{R}^{n+1}, \mathbb{R}^n)$ with $F(0,0) = 0$. Assume that there exists a function $u \in C^1(U \setminus \{0\}, \mathbb{R}) \cap C^0(U, \mathbb{R})$ (U is some neighborhood of 0 in $\mathbb{R}^n$) with $u(0) = 0$ such that the system*

$$\dot{x} = F(x, u(x))$$

is locally asymptotically stable and such that the following holds:

$$\sup_{y \in [[0, u(x)]]} |\nabla u(x) \cdot F(x, y)| \to 0 \text{ as } x \to 0$$

([[0, $u(x)$]] denoting the set $[\min(0, u(x)), \max(0, u(x))]$*). Then the system*

$$\begin{cases} \dot{x} = F(x, y), \\ \dot{y} = v \end{cases}$$

is locally asymptotically stabilizable around $(0, 0)$.

THEOREM 2. *There exists a control function $u(x, y, z) \in C^1(U \setminus \{0\}, \mathbb{R}) \cap C^0(U, \mathbb{R})$ (U is some neighborhood of 0 in $\mathbb{R}^3$) such that the origin of (Σ_1') is locally asymptotically stable.*

Proof. Let $u(x, y)$ be the stabilizing feedback designed in Theorem 1 with the regularity properties characterized by (33). Throughout the proof, the notation $z(x, y)$ stands for any (scalar) value in $[[0, u(x, y)]]$. Using the above criterion, it is sufficient to prove that

$$\left| \frac{\partial u}{\partial x}(x + z(x, y)) + \frac{\partial u}{\partial y}(3y + xz^2(x, y)) \right| \to 0, \quad |(x, y)| \to 0. \tag{41}$$

Note that by (33), $|\partial u / \partial x|$ is bounded in U. Since $u(x, y)$ is continuous at the origin and $u(0, 0) = 0$, we conclude that

$$\left| \frac{\partial u}{\partial x}(x + z(x, y)) \right| \to 0, \qquad |(x, y)| \to 0. \tag{42}$$

Now we will prove that

$$\left| \frac{\partial u}{\partial y}(3y + xz^2(x, y)) \right| \to 0, \qquad |(x, y)| \to 0. \tag{43}$$

For a point (x, y) in $U \cap S_1$, (43) is immediate because in this region

$$\frac{\partial u}{\partial y} \equiv 0.$$

On the other hand, we have by (33)

$$\left| \frac{\partial u}{\partial y} \right| \leq C |y|^{\frac{1}{3+\nu} - 1}. \tag{44}$$

It is therefore sufficient to prove that in $U \cap S_i$, $i \neq 1$,

$$|3y + xz^2(x, y)| < K_3 |y|^{\frac{3}{3+\nu}}. \tag{45}$$

Inequalities (44) and (45) show that near the origin,

$$\left|\frac{\partial u}{\partial y}(3y + xz^2(x, y))\right| < K_2 K_3 \mid y \mid^{\frac{4}{3+\nu}-1},$$

which establishes (43) provided that $0 < \nu < 1$.

It remains to verify that (45) holds in each sector S_i, $i \neq 1$. First consider $S_2 \cup S_3$. We have in this region that $\mid x \mid \leq \mid \epsilon(x, y) \mid$, and we can also assume

$$\mid u(x, y) \mid \leq a \mid \epsilon(x, y) \mid .$$

Therefore we obtain

$$\mid xz^2(x, y) \mid \leq \mid xu^2(x, y) \mid < a^2 \mid \epsilon^3(x, y) \mid,$$

with a defined by (11). Now we easily check that in S_2

$$\mid y \mid = \mid 3(x - 2\epsilon)^3 \mid \geq 3 \mid \epsilon^3(x, y) \mid$$

and that in S_3

$$\mid y \mid = \mid m(x - \epsilon)^3 - \beta\epsilon^{3+\nu} \mid \geq (\beta\epsilon^{\nu} + m/8) \mid \epsilon^3(x, y) \mid .$$

We conclude that in $S_2 \cup S_3$

$$\mid 3y + xz^2(x, y) \mid < 3 \mid y \mid + a^2 \mid \epsilon^3(x, y) \mid < K_4 \mid y \mid$$

for some constant K_4. For small $\mid y \mid$, this establishes (45) in the considered region.
Next consider S_4. In this region, we have by construction

$$\mid 3y + xu^2(x, y) \mid \leq \mid 3y \mid,$$

and therefore (45) immediately follows. Finally consider S_5. We have in this region that $\mid x \mid \leq d \mid \epsilon(x, y) \mid$ and also

$$\mid u(x, y) \mid < 3n/2 \mid x \mid .$$

Therefore we get for some constant K_5 that

$$\mid xz^2(x, y) \mid \leq \mid xu^2(x, y) \mid < K_5 \mid \epsilon^3(x, y) \mid .$$

Now since in S_5

$$\mid y \mid = K_6 \mid \epsilon^{3+\nu} \mid$$

holds for some constant K_6, we conclude that (45) also holds in this region. This ends the proof of (43), and (41) follows from (42) and (43). □

5. Conclusions. It has been shown that the existence of a continuous stabilizing feedback for a homogeneous system does not imply the existence of a continuous stabilizing feedback that preserves the homogeneity of the system in the closed loop. Our example is an analytic planar controllable system. The stabilizing feedback that we have exhibited is semiglobal and has continuous partial derivatives outside the origin. The addition of an integrator to the original system provides an affine three-dimensional homogeneous system that is controllable, not stabilizable by homogeneous feedback, and stabilizable by continuous feedback. The paper extends partial results presented in [1].

Remark. Independently, Rosier [14] has recently provided an example of planar system which is stabilizable by continuous feedback while not stabilizable by continuous *homogeneous* feedback. Contrary to our result, this example is not analytic, and addition of an integrator results in a nonstabilizable system.

6. Appendix.

6.1. Proof of Lemma 1. We prove that Lemma 1 holds in each sector S_i, $i \in \{1, \ldots, 5\}$. Consider S_1. We compute successively

$$G_1(x, y, \epsilon) = y + nx^3 - (n+3)\epsilon^3, \tag{46}$$

$$\frac{\partial G_1}{\partial x} = 3nx^2, \quad \frac{\partial G_1}{\partial y} = 1, \tag{47}$$

$$\frac{\partial G_1}{\partial \epsilon} = \frac{\partial n}{\partial \epsilon}(x^3 - \epsilon^3) - 3(n+3)\epsilon^2$$

$$= -K_1\epsilon^2 - K_2(x)\epsilon^{\nu+2}, \tag{48}$$

with K_1 some positive constant and $K_2(x) \leq d^3 - 4 < 0$ for $\epsilon \leq x \leq d\epsilon$. Since $\partial G_1/\partial \epsilon$ does not vanish for $\epsilon \neq 0$, the implicit function theorem asserts that $\epsilon_1(x, y)$ exists and has continuous partial derivatives in $S_1 \backslash \{0\}$. On the other hand we have in $S_1 \backslash \{0\}$

$$\left|\frac{\partial \epsilon}{\partial x}\right| = \left|\frac{\partial G_1}{\partial x}\left(\frac{\partial G_1}{\partial \epsilon}\right)^{-1}\right|$$

$$= \left|\frac{3nx^2}{K_1\epsilon^2 + K_2(x)\epsilon^{2+\nu}}\right|$$

$$< \left|\frac{3nd}{K_1 + K_2(x)\epsilon^{\nu}}\right|, \tag{49}$$

which shows that $\partial \epsilon/\partial x$ is bounded in a neighborhood the origin. Analogously, we compute

$$\left|\frac{\partial \epsilon}{\partial y}\right| = \left|\frac{\partial G_1}{\partial y}\left(\frac{\partial G_1}{\partial \epsilon}\right)^{-1}\right|$$

$$= \left|\frac{1}{K_1\epsilon^2 + K_2(x)\epsilon^{2+\nu}}\right| \tag{50}$$

$$< \left|\frac{1}{K_1\epsilon^2}\right|. \tag{51}$$

For $y \in S_1$ and ϵ small, we have by construction

$$| y | < 3\epsilon^3$$

and therefore

$$\left|\frac{1}{K_1\epsilon^2}\right| < C_1 |y^{-2/3}| \tag{52}$$

for some constant C. Equations (51) and (52) end the proof of Lemma 1(ii). Lemma 1(iii) follows from (50).

In S_2 we can obtain the following explicit expression for $\epsilon_2(x, y)$:

$$\epsilon_2(x, y) = 1/2\left(x + \left(\frac{y}{3}\right)^{1/3}\right). \tag{53}$$

Lemma 1 is therefore readily checked.

In S_3 the proof follows the same lines as in S_1. We compute successively

$$G_3(x, y, \epsilon) = y - m(x - \epsilon)^3 + \beta\epsilon^{3+\nu}, \tag{54}$$

$$\frac{\partial G_3}{\partial x} = -3m(x - \epsilon)^2, \quad \frac{\partial G_3}{\partial y} = 1, \tag{55}$$

$$\frac{\partial G_3}{\partial \epsilon} = -\frac{\partial m}{\partial \epsilon}(x - \epsilon)^3 + 3m(x - \epsilon)^2 + (3 + \nu)\beta\epsilon^{2+\nu} \tag{56}$$

$$= K_1(x)\epsilon^2 + K_2(x)\epsilon^{\nu+2}, \tag{57}$$

with $K_1(x) \geq 0$ and $K_2(x) \geq 3$ for $\epsilon/2 \leq x \leq \epsilon$. Since $\partial G_3/\partial\epsilon$ does not vanish for $\epsilon \neq 0$, the implicit function theorem asserts that $\epsilon_3(x, y)$ exists and has continuous partial derivatives in $S_3\backslash\{0\}$.

As in S_1, we compute

$$\left|\frac{\partial \epsilon}{\partial x}\right| = \left|\frac{3m(x - \epsilon)^2}{3m(x - \epsilon)^2 + K_2(x)\epsilon^{\nu+2}}\right| \tag{58}$$

$$< \left|\frac{3m/4}{3m/4 + K_2(x)\epsilon^{\nu}}\right|, \tag{59}$$

which shows that $\partial\epsilon/\partial x$ is bounded in a neighborhood the origin and

$$\left|\frac{\partial \epsilon}{\partial y}\right| = \left|\frac{1}{3m(x - \epsilon)^2 + K_2(x)\epsilon^{\nu+2}}\right|. \tag{60}$$

Using (54) and (60), Lemma 1(ii) is proven if the inequality

$$\frac{|\, m(x - \epsilon)^3 + \beta\epsilon^{3+\nu}\,|^{\frac{2+\nu}{3+\nu}}}{|\, 3m(x - \epsilon)^2 + K_2(x)\epsilon^{nu+2}\,|} < C_3 \tag{61}$$

holds in S_3 for some constant C_3. The left-hand side can be unbounded only when ϵ tends to zero. For $x = \epsilon$, the left-hand side is a constant, while for $x \neq \epsilon$, the left-hand side tends to $0(\epsilon^{\nu})$ for ϵ sufficiently small. Finally Lemma 1(iii) follows from (60).

In S_4 we can also obtain an explicit expression for $\epsilon_4(x, y)$:

$$\epsilon_4(x, y) = \left(\frac{-y^{\mu}}{\beta^{\mu}x^3}\right)^{\xi}, \tag{62}$$

with $\xi := 1/((3 + \nu)\mu - 3) > 0$. We easily verify that

$$\left|\frac{\partial \epsilon}{\partial x}\right| = K_4\left|\frac{y}{x^{3+\nu}}\right|^{\mu\xi} \leq K_4\,|\gamma|^{\mu\xi},$$

$$\left|\frac{\partial \epsilon}{\partial y}\right| = K_4'\left|\frac{y^{3+\nu}}{x^3}\right|^{\mu\xi}\left|y^{\frac{1}{3+\nu}-1}\right| \leq K_4''\left|y^{\frac{1}{3+\nu}-1}\right|$$

for some constants K_4, K_4', and K_4''.

Finally Lemma 1 is readily checked in S_5 with the following explicit expression for $\epsilon_5(x, y)$:

$$\epsilon_5(x, y) = \frac{1}{c}\left(\frac{-y}{\gamma}\right)^{\frac{1}{3+\nu}}. \tag{63}$$

Lemma 1 is proven in each sector. □

6.2. Proof of Lemma 2. It has already been proven that for a fixed $\epsilon > 0$ the level set ∂V_ϵ defines a simple closed curve surrounding the origin. Morevover, any compact set K can be included in ∂V_ϵ by a suitable choice of the constants ϵ, β, γ, and δ (see the proof of Corollary 1).

Here we will prove that the function V is "radially" increasing in the following sense: in each sector S_i, we will consider a set of local "polar" coordinates (ρ, θ) such that, for a fixed θ, ϵ_i is strictly increasing as a function of ρ.

The argument is straightforward in the sectors where we have an explicit expression for $V(x, y)$. In S_2, the natural coordinates are $\theta = x/y^{1/3}$ and $\rho = y^{1/3}$. Then the expression (53) gives

$$V(x, y) = \frac{1}{2}(\theta + 3^{-1/3})\rho.$$

Noting that $\theta \in [-\alpha^{-1/3}, 3^{-1/3}]$ and that $\alpha > 3$, we obtain $\partial V/\partial\rho > 0$ in $S_2\backslash\{0\}$. In S_4 and S_5, natural coordinates are given by $\theta = -y/x^{3+\nu}$ and $\rho = x$, since the expressions (62) and (63) respectively give

$$\epsilon_4(x, y) = \left(\frac{\theta}{\beta}\right)^{\mu\xi} \rho$$

and

$$\epsilon_5(x, y) = \frac{1}{c}\left(\frac{\theta}{\gamma}\right)^{\frac{1}{3+\nu}} \rho,$$

from which it directly follows that $\partial V/\partial\rho > 0$ in $S_4 \cup S_5\backslash\{0\}$.

Now consider S_1. Natural polar coordinates are given by $\theta = y/x^3$ and $\rho = x$. Then we have

$$\tilde{G}_1(\rho, \theta, \epsilon) := \frac{G_1(x, y, \epsilon)}{x^3} = \theta + n - (n+3)\left(\frac{\epsilon}{\rho}\right)^3,$$

$$\frac{\partial \tilde{G}_1}{\partial \rho} = 3(n+3)\epsilon^3\rho^{-4} > 0,$$

$$\frac{\partial \tilde{G}_1}{\partial \epsilon} = \rho^{-3}\frac{\partial G_1}{\partial \epsilon} < 0,$$

where the inequalities hold for any (x, y) in $S_1\backslash\{0\}$. By the implicit function theorem, we conclude that $\partial V/\partial\rho > 0$ in $S_1\backslash\{0\}$.

Finally consider S_3. We can cover the sector with the following definition of "radial" curves: pick any (x_1, y_1) in S_3. If (x_1, y_1) is such that $-\alpha \le y_1/x_1^3 \le -\beta$, we define a radial curve through (x_1, y_1) by $\theta = \{(x, y) \in S_3 \mid y/x^3 = y_1/x_1^3\}$. If (x_1, y_1) is such that $y_1/x_1^3 \ge -\beta$, then we choose $s \in [0, \nu]$ such that $y_1/x_1^{3+s} = -\beta$ and define a radial curve through (x_1, y_1) by $\theta = \{(x, y) \in S_3 \mid y/x^{3+s} = y_1/x_1^{3+s}\}$. It is clear that with this definition, each point of $S_3\backslash\{0\}$ belongs to one (and only one) "radial" curve. We will show that V is strictly increasing as a function of $\rho := x$ on each "radial" curve.

Pick any "radial" curve in S_3. Then we have for some constant $s \in [0, \nu]$ and $C = y/x^{3+s} < 0$

$$\tilde{G}_3(\rho, \epsilon) := \frac{G_3(x, y, \epsilon)}{x^{3+s}} = C + \frac{1}{\rho^{3+s}}(-m(\rho - \epsilon)^3 + \beta\epsilon^{3+\nu}),$$

$$\frac{\partial \tilde{G}_3}{\partial \rho} = (3+s)C\frac{1}{\rho} + \frac{1}{\rho^{3+s}}\frac{\partial G_3}{\partial x} < 0,$$

$$\frac{\partial \tilde{G}_3}{\partial \epsilon} = \frac{1}{\rho^{3+s}}\frac{\partial G_3}{\partial \epsilon} > 0,$$

where the inequalities hold in $S_3 \setminus \{0\}$. By the implicit function theorem, we conclude that $\partial V / \partial \rho > 0$ in S_3, which ends the proof of the lemma. □

REFERENCES

[1] D. AEYELS, R. SEPULCHRE, AND I. MAREELS, *On the stabilization of planar homogeneous systems*, in Proc. of the Math. Theory of Networks and Systems 1993, Germany, U. Helmke, R. Mennichen, and J. Sauer, eds., Akademie Verlag, Berlin, 1994, pp. 27–30.

[2] A. BACCIOTTI, *Local stabilization of nonlinear systems*, Ser. Adv. Math. Appl. Sci. 8, World Scientific, River Edge, NJ, 1992.

[3] R. W. BROCKETT, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millmann, and H. J. Susssmann, eds., Progr. Math. 27, Birkhäuser, Boston, 1983, pp. 181–191.

[4] J. M. CORON, *Linearized control systems and applications to smooth stabilization*, SIAM J. Control Optim., 32 (1994), pp. 358–386.

[5] ———, *Links between local controllability and local continuous stabilization*, in IFAC Nonlinear Control Systems Design Symposium, 1992, M. Fliess, ed., Pergamon Press, Oxford, New York, Seoul, Tokyo, 1993, pp. 165–171.

[6] J. M. CORON AND L. PRALY, *Adding an integrator for the stabilization problem*, Systems Control Lett., 17 (1991), pp. 89–104.

[7] W. P. DAYAWANSA, *Recent advances in the stabilization problem for low dimensional systems*, in Proc. IFAC Nonlinear Control Systems Design Symposium, 1992, Bordeaux, France, pp. 1–8.

[8] H. HERMES, *Nilpotent and high order approximations of vector fields systems*, SIAM Rev., 33 (1991), pp. 238–264.

[9] M. KAWSKI, *High order local controllability*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Pure Appl. Math., 133, John Wiley and Sons, New York, 1990, pp. 431–467.

[10] ———, *Control variations with an increasing number of switchings*, Bull. AMS, 2 (1988), pp. 149–152.

[11] ———, *Stabilization of nonlinear systems in the plane*, Systems Control Lett., 2 (1989), pp. 169–175.

[12] ———, *Homogeneous stabilizing feedback laws*, Control Theory Adv. Tech., 6 (1990), pp. 497–516.

[13] N. ROUCHE, P. HABETS, AND M. LALOY, *Stability Theory by Liapunov's Direct Method*, Springer-Verlag, New York, 1977.

[14] L. ROSIER, *Etude de quelques problèmes de stabilisation*, Ph.D. thesis, Ecole Nationale Supérieure de Cachan, France, 1993.

[15] ———, *Homogeneous Liapunov function for continuous vector fields*, Systems Control Lett., 19 (1992), pp. 467–473.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 5, pp. 1814–1830, September 1996

017

MODIFIED PROJECTION-TYPE METHODS FOR MONOTONE VARIATIONAL INEQUALITIES*

MICHAEL V. SOLODOV† AND PAUL TSENG‡

Abstract. We propose new methods for solving the variational inequality problem where the underlying function F is monotone. These methods may be viewed as projection-type methods in which the projection direction is modified by a strongly monotone mapping of the form $I-\alpha F$ or, if F is affine with underlying matrix M, of the form $I+\alpha M^T$, with $\alpha \in (0, \infty)$. We show that these methods are globally convergent, and if in addition a certain error bound based on the natural residual holds locally, the convergence is linear. Computational experience with the new methods is also reported.

Key words. monotone variational inequalities, projection-type methods, error bound, linear convergence

AMS subject classifications. 49M45, 90C25, 90C33

1. Introduction.

We consider the monotone variational inequality problem of finding an $x^* \in X$ satisfying

$$F(x^*)^T(x-x^*) \geq 0 \ \ \forall x \in X, \tag{1.1}$$

where X is a closed convex set in $\Re^n$ and F is a monotone and continuous function from $\Re^n$ to $\Re^n$. This problem, which we abbreviate as VI(X, F), is well known in optimization (see [1, 6, 15]) and, in the special case where F is affine and X is the nonnegative orthant, reduces to the classical monotone linear complementarity problem (LCP) (see [7, 36]).

Many methods have been proposed to solve VI(X, F). The simplest of these is the projection method [46] (also see [1, 2, 3, 8, 27]) which, starting with any $x \in \Re^n$, iteratively updates x according to the formula

$$x^{\text{new}} := [x-\alpha F(x)]^+,$$

where $[\cdot]^+$ denotes the orthogonal projection map onto X and α is a judiciously chosen positive stepsize. However, the projection method requires the restrictive assumption that F or F^{-1} be strongly monotone for convergence. The extragradient method [22] (also see [47, 20, 21, 31] for extensions) overcomes this difficulty by the ingenious technique of updating x according to the double projection formula:

$$x^{\text{new}} := \left[x-\alpha F\left([x-\alpha F(x)]^+\right)\right]^+.$$

This method, by virtue of its using only function evaluations and projection onto X, is easy to implement, uses little storage, and can readily exploit any sparsity or separable structure in F or in X, such as those arising in the applications considered in [3, 9, 38, 45]. Moreover, its convergence requires only that a solution exists [20], while its only drawback is its, at best, linear convergence. In contrast, the methods in [4, 8, 12, 27, 32, 33, 34, 38, 40, 50, 54] require restrictive assumptions on the problem (such as F or F^{-1} being strongly monotone or F being affine; for some of the methods, it is further required that F be continuously differentiable with nonsingular Jacobian or X be bounded and polyhedral), while the matrix-splitting methods in

*Received by the editors June 1, 1994; accepted for publication (in revised form) July 21, 1995.

†Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706 (solodov@cs.wisc.edu). The work of this author was supported by Air Force Office of Scientific Research grant F49620-94-1-0036 and National Science Foundation grant CCR-9101801.

‡Department of Mathematics, Box 354350, University of Washington, Seattle, WA 98195 (tseng@math.washington.edu). The research of this author was supported by National Science Foundation grant CCR-9311621.

[10, 34, 49, 51] are applicable only when F is affine (and these methods also have, at best, linear convergence). And all these methods require more computation per iteration than the extragradient method. For the special case where X is the nonnegative orthant (the monotone nonlinear complementarity problem) or a box, many other solution methods exist, but these methods tend to be ill suited for large sparse problems and are not practically extendable to more general X. Thus, it can be said that, unless F has a special structure (F or F^{-1} is strongly monotone or F is affine) and X has a special structure (X is polyhedral or, better still, just a box), the extragradient method is a very practical method (and sometimes the only practical method) for solving VI(X, F). And, even when F is affine, there are situations where the extragradient method may be practical. As a case in point, suppose X is the Cartesian product of simplices and ellipsoids and F is affine with an underlying matrix M that is asymmetric, positive semidefinite, sparse, and having no particular structure (so M^{-1} may be dense and impractical to compute). The extragradient method can be practically implemented to solve this special case of VI(X, F) since it requires only projection onto the simplices and ellipsoids (for which many efficient methods exist [42, 53]) and multiplication of x by the sparse matrix M. In contrast, the matrix-splitting methods in [10, 34, 49, 51] require solving a nontrivial strongly monotone variational inequality problem over X at each iteration. And even on structured problems such as the discrete-time deterministic optimal control problem [45], the extragradient method may yet be practical since it is linearly convergent like the methods in [5, 10, 49, 55], while its iterations are simpler.

In this paper, we propose a new class of methods for solving VI(X, F) that are as versatile and capable of exploiting problem structure as the extragradient method and, yet, are even simpler than the latter and have a scaling feature absent in the latter. And our preliminary computational experience suggests that the new methods are practical alternatives to the extragradient method. The idea of the new methods is to choose an $n \times n$ symmetric positive definite matrix P and, starting with any $x \in \Re^n$, to iteratively update x according to the formula

$$x^{\text{new}} := x - \gamma P^{-1}\left(T_\alpha(x) - T_\alpha\left([x - \alpha F(x)]^+\right)\right), \tag{1.2}$$

where γ is a positive stepsize and either $T_\alpha \equiv I - \alpha F$ or, if F is affine with underlying matrix M, $T_\alpha \equiv I + \alpha M^T$, with $\alpha \in (0, \infty)$ chosen so T_α is strongly monotone. These methods are like the projection method except the projection direction $[x - \alpha F(x)]^+ - x$ is modified by T_α and P^{-1}. Like the extragradient method, these methods use two function evaluations per iteration and, as we shall show (see Theorems 2.1 and 3.2), their convergence requires only that a solution exists. Unlike the extragradient method, these methods require only one projection per iteration, rather than two, and they have an additional parameter, the scaling matrix P, that can be chosen to help accelerate the convergence (see §2 and §4 for examples and further discussions). Thus, the new methods require less work per iteration than does the extragradient method (assuming P is chosen so P^{-1} is easily computed and stored), with the savings being the greatest when the projection is expensive. Our computational experience (§4) suggests that the new methods are practical alternatives to the extragradient method, especially when F is affine or when projection onto X is expensive.

Although we will also present computational results to illustrate the practical behavior of the new methods, the focus of our paper is on laying the theoretical foundations for these methods. In particular, we will present various convergence and rate of convergence results for the new methods. Central to our rate of convergence analysis is the following growth condition on the 2-norm of the projection residual function $r : \Re^n \mapsto \Re^n$, given by

$$r(x) = x - [x - F(x)]^+,$$

near the solution set S of VI(X, F) (i.e., S comprises all $x^* \in X$ satisfying (1.1)): There exist positive constants μ and δ (depending on F, X only) such that

$$d(x, S) \leq \mu \|r(x)\| \quad \forall x \text{ with } \|r(x)\| \leq \delta, \tag{1.3}$$

where $\|\cdot\|$ denotes the 2-norm and $d(\cdot, S)$ denotes the 2-norm distance to S. (It is well known that an $x^* \in \Re^n$ solves VI(X, F) if and only if $r(x^*) = 0$.) This growth condition on $\|r(\cdot)\|$ (also called error bound) has been used in the rate of convergence analysis of various methods [25, 26, 51] and is known to hold whenever X is polyhedral and either F is affine (see [26, 43]) or F has certain strong monotonicity structure (see [51, Thm. 2]). Moreover, under additional assumptions on F, this condition holds with $\delta = \infty$ (see [23, 24, 28, 39]). Our rate of convergence analysis, similar to that in [51], entails (roughly) showing that $d(x, S)^2$ decreases by an amount in the order of $\|r(x)\|^2$ per iteration, so $\|r(x)\|$ must eventually decrease below δ, at which time (1.3) yields that $d(x, S)^2$ decreases at a linear rate. The analysis is also similar in spirit to those for feasible descent methods (see [25, 26, 28]) but uses $d(\cdot, S)^2$, rather than the objective function, as the merit function.

Our main results are as follows: In §2, we consider the special case of VI(X, F) where F is affine. We show that, for suitable choices of the stepsize γ, the iterates generated by (1.2) with $T_\alpha \equiv I + \alpha M^T$ and $\alpha = 1$ converge to a solution of VI(X, F) and, under the assumption of (1.3) for some μ and δ, the convergence is linear (see Algorithm 2.1 and Theorem 2.1). We then extend this method by replacing the projection direction with a more general matrix-splitting direction (see Algorithm 2.2 and Theorem 2.2). Also, we consider a modification of this method whereby one of the "$[x - F(x)]^+$" terms is replaced with $x - F(x)$ and an extra projection step is taken (see Algorithm 2.3 and Theorem 2.3). In §3, we consider the general case of VI(X, F) and we analogously analyze the convergence of iterates generated by (1.2) with $T_\alpha \equiv I - \alpha F$ (see Algorithms 3.1, 3.2 and Theorems 3.1, 3.2). In §4, we report our preliminary computational experience with the new methods on sparse linear programs (LPs), dense monotone LCPs, and linearly constrained variational inequality problems. In §5, we give some concluding remarks.

Subsequent to the writing of this paper, we learned of the recently proposed methods of He [18, 19] which may be viewed as special cases of Algorithm 2.1 in §2, with specific choices of the scaling matrix P. He's convergence and rate of convergence results for his methods are similar to ours for Algorithm 2.1 (Theorem 2.1), although He's rate of convergence results further require X to be an orthant. In an earlier work [17], He proposed a related method which may be viewed as a version of Algorithm 2.3 in §2 with $P = I$ and X an orthant, but using a different choice of the stepsize γ_i. During the finalization of this paper, we learned of a very recently proposed method of Sun [48] which may be viewed as a version of Algorithm 3.2 in §3 with $P = I$, but using a different choice of the stepsize γ_i (see Remark 4.3 therein). The convergence analysis in [48] applies to the more general problem where F is pseudomonotone and continuous, though no convergence rate result was given.

A few words about our notation. We denote by $\Re^n$ the space of n-dimensional real column-vectors and by superscript T the transpose (of vectors and matrices). We denote by $\|\cdot\|$ the 2-norm (i.e., $\|x\| = (x^T x)^{\frac{1}{2}}$ for all vectors x) and, for any $n \times n$ symmetric positive definite matrix P, by $\|\cdot\|_P$ the 2-norm in $\Re^n$ scaled by P (i.e., $\|x\|_P = (x^T P x)^{\frac{1}{2}}$ for all $x \in \Re$) and by $P^{-1/2}$ the (unique) $n \times n$ symmetric positive definite matrix whose product with itself is P^{-1}. We denote by I either the identity matrix or the identity map and, by R-linear convergence and Q-linear convergence, we mean linear convergence in the root sense and in the quotient sense, respectively, as defined in [37].

2. Algorithms for F affine. In this section we consider the case of VI(X, F) where F is monotone and affine, i.e.,

$$F(x) = Mx + q$$

for some $n \times n$ positive semidefinite (not necessarily symmetric) matrix M and some $q \in \Re^n$. We present and analyze three methods for solving this special case of VI(X, F). The first method is our basic method (1.2) with $T_\alpha \equiv I + \alpha M^T$ and, for simplicity, $\alpha = 1$. The second method is an extension of the first method in which the projection direction is replaced with a matrix-splitting direction. The third method is a modification of the first method in which the projection operation is removed from one part and added to another part of the method.

We describe the first method formally below.

ALGORITHM 2.1. *Choose any $n \times n$ symmetric positive definite matrix P and any $x^0 \in \Re^n$. Also choose a $\theta \in (0, 2)$. For $i = 0, 1, \ldots$, compute x^{i+1} from x^i according to*

$$x^{i+1} = x^i - \gamma_i P^{-1}(I + M^T)r(x^i), \tag{2.1}$$

where

$$\gamma_i = \theta \| P^{-1/2}(I + M^T)r(x^i)\|^{-2} \|r(x^i)\|^2. \tag{2.2}$$

The parameters P and θ are key to the performance of Algorithm 2.1. We can choose P so that P^{-1} is easily computed and stored (e.g., $P = I$) or so that $\|P^{-1/2}(I + M^T)\|$ is small (e.g., $P = (I + M^T)(I + M)$) so γ_i is large. Below we show that this simple method is convergent and, when the error bound (1.3) holds, is linearly convergent. The proof is based on showing that $(I + M^T)r(x)$ makes an acute angle with $x - x^*$ for any solution x^*, so the distance from x to the solution set S, measured in the scaled 2-norm $\| \cdot \|_P$, decreases when x is moved opposite the direction $P^{-1}(I + M^T)r(x)$.

THEOREM 2.1. *Assume that $F(x) = Mx + q$ for some $n \times n$ positive semidefinite matrix M and some $q \in \Re^n$, and that the solution set S of* VI(X, F) *is nonempty. Then any sequence $\{x^i\}$ generated by Algorithm* 2.1 *converges to an element of S and, if* (1.3) *holds for some μ and δ, the convergence is R-linear.*

Proof. Let x^* be any element of S. For each $i \in \{0, 1, \ldots\}$, we have from (2.1) that

$$\begin{aligned} &\|x^{i+1} - x^*\|_P^2 \\ &= \|x^i - x^* - \gamma_i P^{-1}(I + M^T)r(x^i)\|_P^2 \\ &= \|x^i - x^*\|_P^2 - 2\gamma_i (x^i - x^*)^T (I + M^T)r(x^i) + \gamma_i^2 \|P^{-1/2}(I + M^T)r(x^i)\|^2. \end{aligned} \tag{2.3}$$

We bound below the next-to-last term in (2.3). Let $z^i = [x^i - Mx^i - q]^+$ (so $r(x^i) = x^i - z^i$). By properties of the projection operator, we have

$$0 \le (y - z^i)^T (Mx^i + q + z^i - x^i) \ \ \forall y \in X.$$

Similarly, since x^* is a solution of VI(X, F), we have

$$0 \le (y - x^*)^T (Mx^* + q) \ \ \forall y \in X.$$

Taking $y = x^*$ in the first inequality and taking $y = z^i$ in the second inequality and then adding the two resulting inequalities yields

$$\begin{aligned} 0 &\le (x^* - z^i)^T (M(x^i - x^*) + z^i - x^i) \\ &= (x^* - x^i)^T M(x^i - x^*) + (x^i - x^*)^T (I + M^T)(x^i - z^i) - \|x^i - z^i\|^2 \\ &\le (x^i - x^*)^T (I + M^T)(x^i - z^i) - \|x^i - z^i\|^2 \\ &= (x^i - x^*)^T (I + M^T)r(x^i) - \|r(x^i)\|^2, \end{aligned}$$

where the second inequality follows from the positive semidefinite property of M. Using this to bound the next-to-last term in (2.3) yields the key relation

$$\|x^{i+1}-x^*\|_P^2 \le \|x^i-x^*\|_P^2 - 2\gamma_i\|r(x^i)\|^2 + \gamma_i^2\|P^{-1/2}(I+M^T)r(x^i)\|^2$$

$$= \|x^i-x^*\|_P^2 - \theta(2-\theta)\|P^{-1/2}(I+M^T)r(x^i)\|^{-2}\|r(x^i)\|^4 \tag{2.4}$$

$$\le \|x^i-x^*\|_P^2 - \theta(2-\theta)\|P^{-1/2}(I+M^T)\|^{-2}\|r(x^i)\|^2, \tag{2.5}$$

where the equality follows from (2.2). The remaining argument is patterned after the proof of [44, Thm. 1] and of [51, Thm. 1].

Since (2.5) holds for all i, it follows that $\|x^i-x^*\|_P$ is nonincreasing with i and that $\|r(x^i)\| \to 0$ as $i \to \infty$. This shows that $\{x^i\}$ is bounded and, by continuity of $r(\cdot)$, each cluster point x^∞ satisfies $r(x^\infty)=0$ and hence is in S. Then, we can choose x^* in (2.5) to be x^∞ and conclude that $\|x^i-x^\infty\|_P \to 0$ as $i \to \infty$, i.e., $\{x^i\}$ converges to x^∞.

Assume that (1.3) holds for some μ and δ. Let $\psi(x) = \min_{x^* \in S}\|x-x^*\|_P^2$ (so $\psi(x) \le \|P\|d(x,S)^2$). Since (2.4) holds for all i and all $x^* \in S$, by choosing (for each i) x^* to be the element of S closest to x^i in the norm $\|\cdot\|_P$, we obtain for all i

$$\psi(x^{i+1}) \le \|x^{i+1}-x^*\|_P^2$$

$$\le \|x^i-x^*\|_P^2 - \theta(2-\theta)\|P^{-1/2}(I+M^T)r(x^i)\|^{-2}\|r(x^i)\|^4$$

$$= \psi(x^i) - \theta(2-\theta)\|P^{-1/2}(I+M^T)r(x^i)\|^{-2}\|r(x^i)\|^4 \tag{2.6}$$

$$\le \psi(x^i) - \eta\|r(x^i)\|^2, \tag{2.7}$$

where we let $\eta = \theta(2-\theta)\|P^{-1/2}(I+M^T)\|^{-2}$. Since $\|r(x^i)\| \to 0$, we have $\|r(x^i)\| \le \delta$ for all i greater than some $\bar{i}$, in which case (1.3) yields $d(x^i,S) \le \mu\|r(x^i)\|$. Using this to bound the right-hand side of the above inequality yields

$$\psi(x^{i+1}) \le \psi(x^i) - \frac{\eta}{\mu^2}d(x^i,S)^2 \le \psi(x^i) - \frac{\eta}{\mu^2\|P\|}\psi(x^i)$$

for all $i > \bar{i}$, so $\{\psi(x^i)\}$ converges Q-linearly to zero and, by (2.7), $\{r(x^i)\}$ converges R-linearly to zero. Since by (2.1), (2.2), and (2.6) we have

$$\|x^{i+1}-x^i\|_P = \theta\|P^{-1/2}(I+M^T)r(x^i)\|^{-1}\|r(x^i)\|^2 \le \theta^{1/2}(2-\theta)^{-1/2}(\psi(x^i)-\psi(x^{i+1}))^{1/2}$$

for all i, it follows from $\{\psi(x^i)\}$ converging Q-linearly to zero that $\{\|x^{i+1}-x^i\|_P\}$ converges R-linearly to zero and hence $\{x^i\}$ converges R-linearly. □

The above proof shows that we can alternatively choose $\gamma_i = \bar{\gamma}$ for all i in Algorithm 2.1, where $\bar{\gamma}$ is any scalar satisfying

$$0 < \bar{\gamma} < 2\|P^{-1/2}(I+M^T)\|^{-2}.$$

However, this constant stepsize choice is impractical since it is conservative and difficult to compute.

Algorithm 2.1 can be further extended by replacing the projection term $[x-(Mx+q)]^+$ in the definition of $r(x)$ with a more general matrix-splitting term. In particular, consider the following method.

ALGORITHM 2.2. *Choose any $n \times n$ symmetric positive definite matrix P and any $x^0 \in \Re^n$. Also choose an $n \times n$ positive definite matrix B and a $\theta \in (0,2)$. For $i=0,1,\ldots$, compute x^{i+1} from x^i according to*

$$x^{i+1} = x^i - \gamma_i P^{-1}(B+M^T)(x^i-z^i) \tag{2.8}$$

where z^i is the unique solution of the nonlinear equations

$$z^i = [z^i - (B(z^i - x^i) + Mx^i + q)]^+ \tag{2.9}$$

and γ_i is given by

$$\gamma_i = \theta \| P^{-1/2}(B + M^T)(x^i - z^i)\|^{-2}(x^i - z^i)^T B(x^i - z^i). \tag{2.10}$$

Notice that if we choose $B = I$, then Algorithm 2.2 reduces to Algorithm 2.1. In general, we should choose B to be close to M (so that z^i is close to S for fast convergence) and yet to have enough structure (e.g., lower/upper triangular or tridiagonal or block diagonal) so that z^i is easily computable. We have the following result whose proof is similar to that of Theorem 2.1 and thus is omitted.

THEOREM 2.2. *Assume that $F(x) = Mx + q$ for some $n \times n$ positive semidefinite matrix M and some $q \in \Re^n$, and that the solution set S of* VI(X, F) *is nonempty. Then any sequence $\{x^i\}$ generated by Algorithm* 2.2 *converges to an element of S and, if* (1.3) *holds for some μ and δ, the convergence is R-linear.*

We note that Algorithm 2.2 is closely related to the following iterative method proposed in [11]:

$$x^{i+1} = \arg\min_{x \in X} \left\{ \psi_i(x) := (x - x^i)^T(Mx + q) + \frac{\nu}{2}\|x - x^i\|^2 \right\}, \tag{2.11}$$

where ν is a positive scalar. For the specific choice of $B = M + M^T + \nu I$, we have from (2.9) that

$$\begin{aligned} z^i &= [z^i - (B(z^i - x^i) + Mx^i + q)]^+ \\ &= [z^i - \big((M + M^T + \nu I)(z^i - x^i) + Mx^i + q\big)]^+ \\ &= [z^i - \big((M + M^T)z^i - M^T x^i + q + \nu(z^i - x^i)\big)]^+ \\ &= [z^i - \nabla\psi_i(z^i)]^+, \end{aligned}$$

so that

$$z^i = \arg\min_{x \in X} \psi_i(x).$$

Thus (2.9) generalizes (2.11). We note that in [11] no convergence result is given for (2.11). Theorem 2.2 shows that if the step (2.8) is added, the resulting method (2.8)–(2.10) converges to a solution of VI(X, F) and, if (1.3) holds (as in the case where X is also polyhedral), the convergence is R-linear.

Additional modifications of the preceding methods are possible. For example, we can pass each iterate through a nearest-point projection (with respect to the norm $\|\cdot\|_P$) on to X. For Algorithm 2.1, this modification would entail replacing (2.1) with

$$x^{i+1} = [x^i - \gamma_i P^{-1}(I + M^T)r(x^i)]_P^+,$$

where $[y]_P^+$ denotes the point in X whose distance to y (measured in the norm $\|\cdot\|_P$) is minimal. To see that this does not affect the convergence (and, in fact, accelerates convergence) of the methods, we use the following fact about nearest-point projection:

$$\|[y]_P^+ - x^*\|_P^2 \le \|y - x^*\|_P^2 - \|y - [y]_P^+\|_P^2 \tag{2.12}$$

for all $y \in \Re^n$ and all $x^* \in X$ (see, e.g., [31, Appendix]).

We next present a modification, rather than an extension, of Algorithm 2.1, in which we expand out $(I + M^T)r(x^i) = x^i - [x^i - (Mx^i + q)]^+ + M^T r(x^i)$ and replace the "$[x^i - (Mx^i + q)]^+$" term with $x^i - (Mx^i + q)$. In contrast to Algorithm 2.1, an extra projection onto X is needed.

ALGORITHM 2.3. *Choose any $n \times n$ symmetric positive definite matrix P and any $x^0 \in X$. Also choose a $\theta \in (0, 2)$. For $i = 0, 1, \ldots$, compute x^{i+1} from x^i according to*

$$x^{i+1} = \left[x^i - \gamma_i P^{-1}(Mx^i + q + M^T r(x^i))\right]_P^+, \tag{2.13}$$

where γ_i is given by

$$\gamma_i = \theta \| P^{-1/2}(Mx^i + q + M^T r(x^i))\|^{-2} \|r(x^i)\|^2. \tag{2.14}$$

The convergence properties of Algorithm 2.3 are stated in the following theorem, whose proof is similar to that of Theorem 2.1.

THEOREM 2.3. *Assume that $F(x) = Mx + q$ for some $n \times n$ positive semidefinite matrix M and some $q \in \Re^n$, and that the solution set S of* VI(X, F) *is nonempty. Then any sequence $\{x^i\}$ generated by Algorithm* 2.3 *converges to an element of S.*

Proof. Let x^* be any element of S. For each $i \in \{0, 1, \ldots\}$, we have from (2.13) and (2.12) (with $y = x^i - \gamma_i P^{-1}(Mx^i + q + M^T r(x^i))$) that

$$\begin{aligned} \|x^{i+1} - x^*\|_P^2 &\le \|x^i - x^* - \gamma_i P^{-1}(Mx^i + q + M^T r(x^i))\|_P^2 \\ &= \|x^i - x^*\|_P^2 - 2\gamma_i (x^i - x^*)^T (Mx^i + q + M^T r(x^i)) \\ &\quad + \gamma_i^2 \| P^{-1/2}(Mx^i + q + M^T r(x^i))\|^2. \end{aligned} \tag{2.15}$$

We bound below the next-to-last term in (2.15). Let $z^i = [x^i - Mx^i - q]^+$ (so $r(x^i) = x^i - z^i$). By properties of the projection operator, we have

$$0 \le (y - z^i)^T (Mx^i + q + z^i - x^i) \quad \forall y \in X.$$

Similarly, since x^* is a solution of VI(X, F), we have

$$0 \le (y - x^*)^T (Mx^* + q) \quad \forall y \in X.$$

Taking $y = x^i$ in the first inequality and taking $y = z^i$ in the second inequality and then adding the two resulting inequalities yield

$$\begin{aligned} 0 &\le (x^i - z^i)^T (Mx^i + q + z^i - x^i) + (z^i - x^*)^T (Mx^* + q) \\ &= (x^i - x^*)^T (Mx^i + q + M^T (x^i - z^i)) + (x^i - x^*)^T M (x^* - x^i) - \|x^i - z^i\|^2 \\ &\le (x^i - x^*)^T (Mx^i + q + M^T (x^i - z^i)) - \|x^i - z^i\|^2 \\ &= (x^i - x^*)^T (Mx^i + q + M^T r(x^i)) - \|r(x^i)\|^2, \end{aligned}$$

where the second inequality follows from the positive semidefinite property of M. Using this to bound the next-to-last term in (2.3) yields

$$\begin{aligned} \|x^{i+1} - x^*\|_P^2 &\le \|x^i - x^*\|_P^2 - 2\gamma_i \|r(x^i)\|^2 + \gamma_i^2 \| P^{-1/2}(Mx^i + q + M^T r(x^i))\|^2 \\ &= \|x^i - x^*\|_P^2 - \theta(2 - \theta) \| P^{-1/2}(Mx^i + q + M^T r(x^i))\|^{-2} \|r(x^i)\|^4, \end{aligned}$$

where the equality follows from (2.14). The remainder of the proof is similar to that of Theorem 2.1, but using the above relation instead of (2.5). □

Notice that Algorithm 2.3 requires two projections per iteration, the same as the extragradient method. However, unlike the extragradient method, Algorithm 2.3 does not appear to have linear convergence, even if (1.3) holds for some μ and δ.[1]

3. Algorithms for F nonaffine. In this section we consider the general case of VI(X, F) where F is monotone and continuous. We present and analyze two versions of our basic method (1.2) with $T_\alpha \equiv I - \alpha F$ and with α chosen so T_α is strongly monotone. The first version, which uses a fixed α, is simpler but requires F furthermore to be Lipschitz continuous. The second version, which chooses α dynamically, is more intricate but is more practical and solves the general problem.

We describe the first method formally below. For this method to be applicable, we require F furthermore to be Lipschitz continuous.

ALGORITHM 3.1. *Choose any $n \times n$ symmetric positive definite matrix P and any $x^0 \in \Re^n$. Also choose any $\theta \in (0, 2)$ and any $\alpha \in (0, 1/\lambda)$, where λ is a constant satisfying*

$$(x - z)^T(F(x) - F(z)) \leq \lambda \|x - z\|^2 \quad \forall x, z \in \Re^n. \tag{3.1}$$

(We can, for example, take λ to be the Lipschitz constant of F.) For $i = 0, 1, \ldots$, compute x^{i+1} from x^i according to

$$x^{i+1} = x^i - \gamma_i P^{-1}(x^i - z^i - \alpha F(x^i) + \alpha F(z^i)), \tag{3.2}$$

where z^i and γ_i are given by, respectively,

$$z^i = [x^i - \alpha F(x^i)]^+, \tag{3.3}$$

$$\gamma_i = \theta(1 - \alpha\lambda)\|P^{-1/2}(x^i - z^i - \alpha F(x^i) + \alpha F(z^i))\|^{-2}\|x^i - z^i\|^2. \tag{3.4}$$

Algorithm 3.1 requires less computation per iteration than the extragradient method (in particular, it avoids performing an extra projection step). Also, unlike the extragradient method, Algorithm 3.1 allows scaling of direction by P^{-1} without having to accordingly scale the norm with respect to which projection is taken. In the case where F is affine, i.e., $F(x) = Mx + q$ for some $n \times n$ positive semidefinite matrix M and some $q \in \Re^n$, the formula (3.2) reduces to

$$x^{i+1} = x^i - \gamma_i P^{-1}(I - \alpha M)(x^i - z^i),$$

which is reminiscent of (2.1). If in addition M is skew symmetric (i.e., $M^T = -M$) so that (3.1) holds with $\lambda = 0$, we can choose α arbitrarily large and can reasonably choose P to be $P = (I - \alpha M)(I - \alpha M^T)$. In fact, for the choice of $\alpha = 1$ (and using $M^T = -M$), the formula (3.2) reduces precisely to (2.1).

We show in the following theorem that Algorithm 3.1 has convergence properties similar to that of Algorithm 2.1. The proof of this theorem is patterned after that of Theorem 2.1.

THEOREM 3.1. *Assume that F is monotone and Lipschitz continuous and that the solution set S of* VI(X, F) *is nonempty. Then any sequence $\{x^i\}$ generated by Algorithm* 3.1 *converges to an element of S and, if* (1.3) *holds for some μ and δ, the convergence is R-linear.*

Proof. Let x^* be any element of S. For each $i \in \{0, 1, \ldots\}$, we have from (3.2) that

$$\begin{aligned} \|x^{i+1} - x^*\|_P^2 &= \|x^i - x^* - \gamma_i P^{-1}(x^i - z^i - \alpha F(x^i) + \alpha F(z^i))\|_P^2 \\ &= \|x^i - x^*\|_P^2 - 2\gamma_i(x^i - x^*)^T(x^i - z^i - \alpha F(x^i) + \alpha F(z^i)) \\ &\quad + \gamma_i^2\|P^{-1/2}(x^i - z^i - \alpha F(x^i) + \alpha F(z^i))\|^2. \end{aligned} \tag{3.5}$$

[1] Subsequent to the writing of the original paper, we were informed by B. He that, for the method he proposed in [17] (which may be viewed as Algorithm 2.3 with $P = I$ and X an orthant, but using a different choice of γ_i), he could show a linear convergence result analogous to that shown in [18] (see [19, eq. (8)]).

We bound below the next-to-last term in (3.5). By (3.3) and properties of the projection operator, we have

$$0 \le (y - z^i)^T(\alpha F(x^i) + z^i - x^i) \quad \forall y \in X.$$

Similarly, since x^* is a solution of VI(X, F), we have

$$0 \le (y - x^*)^T F(x^*) \quad \forall y \in X.$$

Taking $y = x^*$ in the first inequality and taking $y = z^i$ in the second inequality and then adding the two resulting inequalities yield

$$\begin{aligned}
0 &\le (x^* - z^i)^T(\alpha F(x^i) + z^i - x^i) + \alpha(z^i - x^*)^T F(x^*) \\
&= \alpha(x^* - z^i)^T(F(z^i) - F(x^*)) + (x^* - x^i)^T(\alpha F(x^i) - \alpha F(z^i) + z^i - x^i) \\
&\quad + \alpha(x^i - z^i)^T(F(x^i) - F(z^i)) - \|x^i - z^i\|^2 \\
&\le (x^* - x^i)^T(\alpha F(x^i) - \alpha F(z^i) + z^i - x^i) + \alpha(x^i - z^i)^T(F(x^i) - F(z^i)) - \|x^i - z^i\|^2 \\
&\le (x^* - x^i)^T(\alpha F(x^i) - \alpha F(z^i) + z^i - x^i) - (1 - \alpha\lambda)\|x^i - z^i\|^2,
\end{aligned}$$

where the second inequality follows from the monotone property of F and the last inequality follows from (3.1). Using this to bound the next-to-last term in (3.5) yields

$$\begin{aligned}
&\|x^{i+1} - x^*\|_P^2 \\
&\le \|x^i - x^*\|_P^2 - 2\gamma_i(1 - \alpha\lambda)\|x^i - z^i\|^2 + \gamma_i^2\|P^{-1/2}(x^i - z^i - \alpha F(x^i) + \alpha F(z^i))\|^2 \\
&= \|x^i - x^*\|_P^2 - \theta(2 - \theta)(1 - \alpha\lambda)^2\|P^{-1/2}(x^i - z^i - \alpha F(x^i) + \alpha F(z^i))\|^{-2}\|x^i - z^i\|^4,
\end{aligned} \tag{3.6}$$

where the equality follows from (3.4).

The remainder of the proof is similar to that of Theorem 2.1, but using (3.6) instead of (2.5). For the R-linear convergence result, we also use the observations (see (3.3) and (3.4)) that

$$\begin{aligned}
\|x^{i+1} - x^i\|_P &= \theta(1 - \alpha\lambda)\|P^{-1/2}(x^i - z^i - \alpha F(x^i) + \alpha F(z^i))\|^{-1}\|x^i - z^i\|^2 \\
&\ge \theta(1 - \alpha\lambda)\|P^{-1/2}\|^{-1}(1 + \alpha L)^{-1}\|x^i - z^i\| \\
&\ge \theta(1 - \alpha\lambda)\|P^{-1/2}\|^{-1}(1 + \alpha L)^{-1}\min\{1, \alpha\}\|r(x^i)\|
\end{aligned}$$

for all i, where L denotes the Lipschitz constant of F and the last inequality follows from [13, Lem. 1]. Thus, the rightmost term in (3.6) is bounded above by a positive constant times $-\|r(x^i)\|^2$ and, whenever this term converges R-linearly to zero as $i \to \infty$, so does $\|x^{i+1} - x^i\|_P^2$; hence $\{x^i\}$ converges R-linearly. □

Algorithm 3.1 is a conceptual method since in practice F need not be Lipschitz continuous or the constant λ may be difficult to estimate or a stepsize of less than $1/\lambda$ may be too conservative. Below we present a practical version of Algorithm 3.1 that chooses α dynamically according to a novel Armijo–Goldstein-type rule. This practical version has all the convergence properties of Algorithm 3.1 and requires F to be only monotone and continuous for convergence (see Theorem 3.2).

ALGORITHM 3.2. *Choose any $n \times n$ symmetric positive definite matrix P and any $x^0 \in \Re^n$ and $\alpha_{-1} \in (0, \infty)$. Also choose any $\theta \in (0, 2)$, $\rho \in (0, 1)$, and $\beta \in (0, 1)$. For $i = 0, 1, \ldots$, compute (x^{i+1}, α_i) from (x^i, α_{i-1}) according to: Choose α_i to be the largest $\alpha \in \{\alpha_{i-1}, \alpha_{i-1}\beta, \alpha_{i-1}\beta^2, \ldots\}$ satisfying*

$$\alpha(x^i - z^i(\alpha))^T(F(x^i) - F(z^i(\alpha))) \le (1 - \rho)\|x^i - z^i(\alpha)\|^2, \tag{3.7}$$

and let

$$x^{i+1} = x^i - \gamma_i P^{-1}(x^i - z^i(\alpha_i) - \alpha_i F(x^i) + \alpha_i F(z^i(\alpha_i))), \tag{3.8}$$

where $z^i(\alpha)$ *and* γ_i *are given by, respectively,*

$$z^i(\alpha) = [x^i - \alpha F(x^i)]^+ \quad \forall \alpha \in (0, \infty), \tag{3.9}$$

$$\gamma_i = \theta\rho \| P^{-1/2}(x^i - z^i(\alpha_i) - \alpha_i F(x^i) + \alpha_i F(z^i(\alpha_i)))\|^{-2} \|x^i - z^i(\alpha_i)\|^2. \tag{3.10}$$

The motivation for taking trial values of α starting at α_{i-1} comes from our empirical experience that, for $i > 0$, $\alpha = \alpha_{i-1}$ either satisfies or comes close to satisfying (3.7), so in general only a few trial values of α are needed to find α_i. The condition (3.7) may be viewed as a local approximation to the condition $\alpha < 1/\lambda$ used in Algorithm 3.1. (If we let $\lambda_i = (x^i - z^i(\alpha))^T(F(x^i) - F(z^i(\alpha)))/\|x^i - z^i(\alpha)\|^2$, then (3.7) reduces to $\alpha \leq (1-\rho)/\lambda_i$.) We had also considered choosing α_i to be the largest $\alpha \in \{\sigma, \sigma\beta, \sigma\beta^2, \ldots\}$ satisfying (3.7), where $\sigma \in (0, \infty)$. It can be checked that the convergence results below still hold for this alternative stepsize rule, but this rule is not as practical since it typically needs many more trial values of α to find α_i. Lastly, as with Algorithm 2.1 for the affine case, we can pass each iterate generated by Algorithm 3.2 through a nearest-point projection (with respect to the norm $\|\cdot\|_P$) onto X and the convergence results below would still hold.

Below we present the convergence results for Algorithm 3.2. The proof is patterned after that for Theorem 3.1 and, for simplicity, we supply only the key steps.

THEOREM 3.2. *Assume that* F *is monotone and continuous and that the solution set* S *of* VI(X, F) *is nonempty. Then any sequence* $\{x^i\}$ *generated by Algorithm* 3.2 *converges to an element of* S *and, if* (1.3) *holds for some* μ *and* δ *and* F *is Lipschitz continuous on* $S + \epsilon B$ *for some* $\epsilon > 0$ *(where* $B = \{x \mid \|x\| \leq 1\}$*), the convergence is R-linear.*

Proof. First, we claim that, for each i, (3.7) holds for all α sufficiently small, so α_i is well defined. To see this, note that $z^i(\alpha) \to [x^i]^+$ as $\alpha \to 0$, so if $x^i \notin X$, then the right-hand side of (3.7) would tend to a positive limit while the left-hand side of (3.7) would tend to zero as $\alpha \to 0$, implying the claim. If $x^i \in X$ (so $x^i = [x^i]^+$), then since F is continuous and $z^i(\alpha) \to [x^i]^+ = x^i$ as $\alpha \to 0$, we have

$$\|F(x^i)\| \|F(x^i) - F(z^i(\alpha))\| \leq (1-\rho)\|r(x^i)\|^2$$

for all $\alpha \in (0, 1]$ sufficiently small. For any such α, we have

$$\begin{aligned}
\alpha(x^i - z^i(\alpha))^T(F(x^i) - F(z^i(\alpha))) &= \alpha([x^i]^+ - [x^i - \alpha F(x^i)]^+)^T(F(x^i) - F(z^i(\alpha))) \\
&\leq \alpha^2 \|F(x^i)\| \|F(x^i) - F(z^i(\alpha))\| \\
&\leq \alpha^2(1-\rho)\|r(x^i)\|^2 \\
&\leq (1-\rho)\|x^i - z^i(\alpha)\|^2,
\end{aligned}$$

where the first inequality uses the Cauchy–Schwartz inequality and the nonexpansive property of $[\cdot]^+$; the last inequality uses $\alpha \in (0, 1]$ and [13, Lem. 1]. Thus, the claim holds.

To show that $\{x^i\}$ converges to an element of S, let x^* be any element of S. For each $i \in \{0, 1, \ldots\}$, we have by an argument analogous to the proof of Theorem 3.1, but with (3.1)–(3.4) replaced by (3.7)–(3.10) (and taking $\alpha = \alpha_i$ in (3.7)), that (cf. (3.6))

$$\begin{aligned}
\|x^{i+1} - x^*\|_P^2 \leq \|x^i - x^*\|_P^2 &- \theta(2-\theta)\rho^2 \|x^i - z^i(\alpha_i)\|^4 \\
&\cdot \|P^{-1/2}(x^i - z^i(\alpha_i) - \alpha_i F(x^i) + \alpha_i F(z^i(\alpha_i)))\|^{-2}.
\end{aligned} \tag{3.11}$$

Thus $\{x^i\}$ is bounded and $\{\|x^i - z^i(\alpha_i)\|\} \to 0$. Also, $\{\alpha_i\}$ is nonincreasing, so it has a limit α_∞. We claim that $\{x^i\}$ has at least one cluster point in S. In the case where $\alpha_\infty > 0$, this

follows from $\{\|x^i - z^i(\alpha_i)\|\} \to 0$ and the the continuity of F and the projection operator, which imply that every cluster point x^∞ of $\{x^i\}$ satisfies

$$x^\infty = [x^\infty - \alpha_\infty F(x^\infty)]^+,$$

and hence is in S. In the case where $\alpha_\infty = 0$, we argue by contradiction by supposing that every cluster point of $\{x^i\}$ is not in S. Since $\alpha_\infty = 0$, there must exist a subsequence K of $\{0, 1, \ldots, \}$ satisfying $\alpha_i < \alpha_{i-1}$ for all $i \in K$, and, by passing to a subsequence if necessary, we can assume that $\{x^i\}_{i\in K}$ converges to some $x^\infty \notin S$. Since $x^\infty \notin S$, it follows from the continuity of F and our earlier argument showing that (3.7) holds for all α sufficiently small that, for all $i \in K$ sufficiently large (so that x^i is near x^∞ and α_{i-1} is sufficiently small), $\alpha = \alpha_{i-1}$ satisfies (3.7). This implies we would choose $\alpha_i = \alpha_{i-1}$ for all $i \in K$ sufficient large, contradicting our hypothesis on K. Thus, $\{x^i\}$ has at least one cluster point, say x^∞, that is in S. Letting $x^* = x^\infty$ in (3.11), we obtain that the sequence $\{\|x^i - x^\infty\|\}$ is nonincreasing. Since this sequence has a subsequence converging to zero, the entire sequence must converge to zero.

In the case where (1.3) holds for some μ and δ and F is Lipschitz continuous (with constant L) on $S + \epsilon B$ for some $\epsilon > 0$, we note that since $\{x^i\}$ converges to an element of S, we have x^i and $z^i(\alpha)$ inside $S+\epsilon B$ for all $\alpha \in (0, \alpha_{-1}]$ and all i exceeding some $\bar{i}$. For all such i, (3.7) holds for all $\alpha \in (0, (1-\rho)/L)$, so our choice of α_i implies $\alpha_i \geq \min\{\alpha_{\bar{i}}, \beta(1-\rho)/L\}$. Thus, $\{\alpha_i\}$ is bounded away from zero. The R-linear convergence of $\{x^i\}$ then follows from an argument analogous to the the proof of Theorem 3.1. □

4. Computational experience. To better understand the behavior of the new methods in practice, we implemented Algorithm 2.1 in Fortran to solve sparse LPs and dense monotone LCPs and implemented Algorithms 2.1 and 3.2 in Matlab to solve linearly constrained variational inequality problems (using the quadratic-program solver qp.m from the Matlab optimization toolbox to perform the projection). For a benchmark, we compared the performance of these implementations with analogous implementations of the extragradient method as described in [31]. (We have included LPs and dense monotone LCPs in our tests not because they are problems which the new methods are designed to solve but because these problems are well-known special cases of VI(X, F) and tests on them give us a better overall understanding of the new methods.) Though our results are preliminary, they suggest that the new methods are practical alternatives to the extragradient method, especially when F is affine or when projection onto X is expensive. We describe the test details below.

All Fortran codes were compiled by the DEC Fortran-77 compiler Version 4.2 using the default optimization option and were run on a Decstation 5000 under the operating system Ultrix Version 4.2A. All Matlab codes were run on the same Decstation 5000 under Matlab Version 4.2a.

Our first set of tests was conducted on sparse LP of the form $\min\{\, c^T y \mid Ay = b,\ y \geq 0\,\}$, where A is an $m \times l$ matrix, $b \in \Re^m$, and $c \in \Re^l$. We reformulated the LP as a VI(X, F) with

$$X = \Re^l_+ \times \{0\}^m, \quad F(x) = Mx + q, \quad M = \begin{bmatrix} 0 & -A^T \\ A & 0 \end{bmatrix}, \quad q = \begin{bmatrix} c \\ -b \end{bmatrix}.$$

Then we applied Algorithm 2.1 and the extragradient method to this VI(X, F). The first six test problems were randomly generated, with the entries of c uniformly generated from [1, 100], with the number of nonzeros per column of A fixed at 5% and the nonzeros uniformly generated from [−5, 5], and with $b = A\bar{x}$, where $\bar{x} = (10/l, \ldots, 10/l)$. The seventh to ninth test problems were taken from the Netlib library (see [14]). The performance of Algorithm 2.1 is sensitive to the choice of P and θ, and in our implementation of Algorithm 2.1, we chose P to be the diagonal part of $(I + M^T)(I + M)$ (which made P^{-1} easy to compute and

TABLE 1
Results for Algorithm 2.1 and extragradient method on LP.

Problem			Algorithm 2.1[2]				Extragradient[3]			
			($\epsilon = 10^{-2}$)		($\epsilon = 10^{-3}$)		($\epsilon = 10^{-2}$)		($\epsilon = 10^{-3}$)	
Name	m	l	iter.[4]	CPU[5]	iter.[4]	CPU[5]	iter.[4]	CPU[5]	iter.[4]	CPU[5]
RanLP1	100	200	738	2.6	2776	12.3	1009	5.0	5056	31.5
RanLP2	100	300	599	4.0	2811	15.1	867	6.3	8380	65.2
RanLP3	100	400	697	5.1	3326	25.0	762	8.0	3058	31.8
RanLP4	200	400	790	22.6	3174	81.9	759	28.4	3005	107.6
RanLP5	200	600	691	14.7	2301	85.0	748	23.4	2980	82.8
RanLP6	200	800	875	25.3	3215	97.3	861	38.0	4496	177.2
Adlittle	56	138	56219	123.9	73804	163.9	–[6]	–	–	–
Scorpion	388	466	1609	13.01	6058	56.8	3372	36.7	14277	159.3
Bandm	305	472	1202607	12837	–	–	–	–	–	–

still yielded fast convergence) and chose $\theta = .7$ (which yielded much faster convergence than with $\theta = 1$). The parameters in the extragradient method were similarly tuned to optimize the method's performance. The test results are summarized in Table 1. In general, Algorithm 2.1 required fewer iterations and less time than the extragradient method, with the improvement most pronounced when $l \leq 2m$. However, both methods did very poorly on the Netlib problems, which suggests that these methods are not well suited for solving small to medium-sized LP. For large-sized LP, these methods may yet be practical since they have low storage requirements and can exploit sparsity structure in the problem.

Our second set of tests was conducted on dense monotone LCP, corresponding to VI(X, F) with

$$X = \Re^n_+, \qquad F(x) = Mx + q$$

for some $n \times n$ positive semidefinite matrix M and some $q \in \Re^n$. The first three (respectively, fourth to sixth) test problems were randomly generated with

$$M = \omega E E^T + E - E^T,$$

where $\omega = 0$ (respectively, $\omega = 1$) and every entry of the $n \times n$ matrix E was uniformly generated from $[-5, 5]$, and with $q = -M\bar{x} + \bar{y}$, where each entry of $\bar{x}$ has equal probability of being 0 or being uniformly generated from [5, 10] and each entry of $\bar{y}$ is 0 if the corresponding entry of $\bar{x}$ is 0 and otherwise has equal probability of being 0 or being uniformly generated from [5, 10] (so $\bar{x}$ is a solution). The seventh to ninth test problems were deterministically generated with

$$M = EE^T,$$

where the (i, j)th entry of the $n \times n$ matrix E is $5(i-j)/n$ for all i and j, and with $q = -M\bar{x} + \bar{y}$, where the first $n/2$ entries of $\bar{x}$ are 0 and the rest are 7.5 and the first $n/4$ entries of $\bar{y}$ are 5 and

[2]Algorithm 2.1 with P being the diagonal part of $(I + M^T)(I + M)$ and $\theta = .7$.

[3]The extragradient method as described in [31], with $\beta = .7$ and initial $\alpha = 1$.

[4]For all methods, $x^0 = 0$ and the termination criterion is $\|r(x)\| \leq \epsilon$.

[5]Time (in seconds) obtained using the intrinsic function SECNDS and with the codes compiled by the DEC Fortran-77 compiler and run on a Decstation 5000; does not include time to read problem data.

[6]$\|r(x)\| \approx 2 \cdot 10^{-2}$ after 50955000 iterations.

TABLE 2
Results for Algorithm 2.1 and extragradient method on LCP.

		Algorithm 2.1[7]				Extragradient[8]			
Problem		($\epsilon = 10^{-2}$)		($\epsilon = 10^{-3}$)		($\epsilon = 10^{-2}$)		($\epsilon = 10^{-3}$)	
Name	n	iter.[9]	CPU[10]	iter.[9]	CPU[10]	iter.[9]	CPU[10]	iter.[9]	CPU[10]
RanLCP1	100	5721	113.0	11600	233.7	36611	739.5	71491	1462.3
RanLCP2	200	59744	5474.5	144157	15028.2	48013	7456.4	198282	18831.7
RanLCP3	300	37769	8415.2	171963	46096.1	316489	13201.0	–	–
RanLCP4	100	2378	45.7	3149	60.4	7802	148.2	10369	195.0
RanLCP5	200	1133	112.6	1412	138.1	3425	341.6	4276	444.3
RanLCP6	300	748	200.4	944	246.2	2394	517.9	3033	713.4
DetLCP1	100	32	2.1	36	2.2	136	2.9	157	2.9
DetLCP2	200	37	16.5	42	16.7	156	18.1	178	19.6
DetLCP3	300	40	50.9	45	52.8	167	36.4	189	43.6
HPEasy	100	79	2.9	109	3.5	423	8.0	531	9.7
HPHard	100	64	2.7	85	3.8	855	16.2	1115	20.9
Lemke	100	1057	21.3	1107	22.0	1508	27.5	2261	44.3

the rest are 0 (so $\bar{x}$ is a solution). The remaining test problems were borrowed from [16, §5]. In particular, the tenth (respectively, eleventh) test problem was randomly generated with

$$M = AA^T + B + D,$$

where every entry of the $n \times n$ matrix A and of the $n \times n$ skew-symmetric matrix B is uniformly generated from $(-5, 5)$ and every diagonal entry of the $n \times n$ diagonal B is uniformly generated from $(0, 0.3)$ (so M is positive definite), and with every entry of q uniformly generated from $(-500, 500)$ (respectively, $(-500, 0)$). The twelfth test problem is one for which Lemke's method is known to run in exponential time, with the (i, j)th entry of M equal to 2 (respectively, 1 and 0) if $j > i$ (respectively, $j = i$ and $j < i$) for all i and j (so M is positive semidefinite), and with every entry of q equal to -1. In our implementation of Algorithm 2.1, we chose P to be $(I + M^T)(I + M)$ and chose $\theta = 1$ (so $\gamma_i = 1$ for all i). The performance of Algorithm 2.1 also benefited substantially from a priori scaling of M and q and, in our test, we scaled M and q by multiplying both with $10 \cdot (\text{maximum magnitude of entries of } M \text{ and } q)^{-1}$. (We did not need to scale M and q for the extragradient method since the scaling is done automatically via its stepsize parameter α.) The test results are summarized in Table 2. In general, Algorithm 2.1 required fewer iterations and less time than the extragradient method, though both had difficulty on skew-symmetric problems (the first three test problems). On the other hand, we caution that the performance of Algorithm 2.1 strongly depends on the scaling of M and q and finding a suitable choice of scaling can be difficult in general.

Our third set of tests was conducted on VI(X, F), where X is not an orthant or a box. The first test problem, used first by Mathiesen [35], and later in [41, 54], has

$$F(x_1, x_2, x_3) = \begin{bmatrix} .9(5x_2 + 3x_3)/x_1 \\ .1(5x_2 + 3x_3)/x_2 - 5 \\ -3 \end{bmatrix},$$

$$X = \left\{(x_1, x_2, x_3) \in \Re^3_+ \mid x_1 + x_2 + x_3 = 1, x_1 - x_2 - x_3 \leq 0\right\}.$$

[7]Algorithm 2.1 with $P = (I + M^T)(I + M)$ and $\theta = 1$.

[8]The extragradient method as described in [31], with $\beta = .7$ and initial $\alpha = 1$.

[9]For all methods, $x^0 = 0$ and the termination criterion is $\|r(x)\| \leq \epsilon$.

[10]Time (in seconds) obtained using the intrinsic function SECNDS and with the codes compiled by the DEC Fortran-77 compiler and run on a Decstation 5000; does not include time to read problem data.

TABLE 3

Results for Algorithms 2.1 and 3.2 and extragradient method on linearly constrained variational inequality problems.

		Algorithm 2.1[11]		Algorithm 3.2[12]		Extragradient[13]	
Name	n	iter.(nf/np)[14]	CPU[15]	iter.(nf/np)[14]	CPU[15]	iter.(nf/np)[14]	CPU[15]
Mathiesen	3	–	–	25(56/31)	3.9	260(524/524)	66.1
		–	–	18(40/22)	2.7	13(30/30)	3.2
KojimaSh	4	–	–	38(85/47)	3.9	16(36/36)	2.4
Nash5	5	–	–	74(155/81)	6.6	43(89/89)	5.5
Nash10	10	–	–	93(192/99)	10.6	84(172/172)	13.4
HPHard	20	38(38/38)	31.5	286(579/293)	264.3	248(499/499)	395.2
qHPHard	20	–	–	274(555/281)	251.6	239(481/481)	380.4

We had trouble finding more test problems from the literature, so we created five additional test problems of our own, in which $X = \{x \in \Re^n_+ \mid x_1 + \cdots + x_n = n\}$ and F and n are specified as follows: For the first three problems, F is the function from, respectively, the Kojima–Shindo nonlinear complementarity problem (NCP) (with $n = 4$) and the Nash–Cournot NCP (with $n = 5$ and $n = 10$) [41, pp. 321–322]; for the fourth problem, F is affine and is generated as in the problem HPHard of Table 2, but with $n = 20$; for the fifth problem, we took the F from the fourth problem and added to its ith component the linear/quadratic term $\max\{0, x_i\}^2$ for $i = 1, \ldots, \lfloor n/2 \rfloor$. In our implementation of Algorithm 3.2, we chose $P = I$, $\alpha_{-1} = 1$, $\theta = 1.5$, $\rho = .1$, and $\beta = .3$. On the Mathiesen problem, we used the same x^0 as in [54]; on the other problems, we used $x^0 = (1, \ldots, 1)$. (The F from the Mathiesen problem and from the Nash–Cournot NCP are defined on the positive orthant only.) The test results are summarized in Table 3. In general, Algorithm 3.2 requires more iterations and function evaluations, but fewer projections, than the extragradient method. (The performance of Algorithm 3.2 is also less sensitive to the starting point than the extragradient method. Surprisingly, both methods solved problems, such as the Kojima–Shindo problem, for which F is not monotone.) Thus, on problems where projection onto X is expensive, Algorithm 3.2 may be more practical than the extragradient method, as is reflected in its lower CPU times on all problems except Nash5. But if F is affine, Algorithm 2.1 may be more practical than either method (compare their CPU times on HPHard). In general, the performance of Algorithm 3.2 is insensitive to x^0 or ρ or α_{-1}, as long these parameters are reasonably chosen. We had also tried alternative choices for P and more conservative choices for θ and β (e.g., $\theta = 1$ and $\beta = .7$), but the results were typically worse.

5. Concluding remarks. We have presented new iterative methods for solving monotone variational inequality problems and have established their convergence and rate of convergence under mild assumptions on the problem. Preliminary computational experience with the new methods suggests the new methods are practical alternatives to the extragradient method.

[11] Algorithm 2.1 with $P = (I + M^T)(I + M)$ and $\theta = 1.5$.

[12] Algorithm 3.2 with $P = I$, $\alpha_{-1} = 1$, $\theta = 1.5$, $\rho = .1$ and $\beta = .3$.

[13] The extragradient method as described in [31], with $\beta = .7$ and initial $\alpha = 1$.

[14] For all methods, the termination criterion is $\|r(x)\| \leq 10^{-4}$. (nf denotes the total number of times F is evaluated, and np denotes the total number of times a projection onto X is performed.) On the Mathiesen problem, we ran each method twice with $x^0 = (.1, .8, .1)$ and $x^0 = (.4, .3, .3)$, respectively; on the other problems, we used $x^0 = (1, \ldots, 1)$.

[15] Time (in seconds) obtained using the intrinsic Matlab function etime and with the codes run on a Decstation 5000; does not include time to read problem data.

We mention in passing that Algorithms 2.1 and 3.1 may be generated by the following general approach: We set y in the inequality

$$0 \leq (y - z)^T(\alpha F(x) + z - x) \ \ \forall y \in X,$$

where $z = [x - \alpha F(x)]^+$, to x^*; and we set y in the inequality

$$0 \leq \alpha(y - x^*)F(x^*) \ \ \forall y \in X,$$

where $x^* \in S$, to z. Then we add the two inequalities and, by using the monotone property of F and, if necessary, the affine property of F, we reduce the resulting inequality to the form

$$0 \leq (x - x^*)^T T(x) + (\text{an expression involving } \alpha, F, x, \text{ and } z \text{ only})$$

for some mapping T (depending on F and α) from $\Re^n$ to $\Re^n$. Provided that the rightmost term is negative, the method then updates x according to the formula

$$x^{\text{new}} := x - \gamma T(x).$$

Algorithm 2.3, as well as the extragradient method, may be similarly generated except we set y in the first inequality to x instead. Then, we need x to be in X which is why an extra projection onto X is needed. (We can also set y in the second inequality to x, but this does not appear to yield anything useful.)

REFERENCES

[1] A. A. AUSLENDER, *Optimisation: Méthodes Numériques*, Masson, Paris, 1976.

[2] A. B. BAKUSINSKII AND B. T. POLYAK, *On the solution of variational inequalities*, Soviet Math. Dok., 15 (1974), pp. 1705–1710.

[3] D. P. BERTSEKAS AND E. M. GAFNI, *Projection methods for variational inequalities with application to the traffic assignment problem*, Math. Programming Study, 17 (1982), pp. 139–159.

[4] J. F. BONNANS, *Local analysis of Newton-type methods for variational inequalities and nonlinear programming*, Appl. Math. Optim., 29 (1994), pp. 161–186.

[5] H.-G. CHEN, *Forward-Backward Splitting Techniques: Theory and Applications*, Ph.D. thesis, Department of Applied Mathematics, University of Washington, Seattle, WA, December 1994.

[6] R. W. COTTLE, F. GIANNESSI, AND J.-L. LIONS, eds., *Variational Inequalities and Complementarity Problems: Theory and Applications*, Wiley, New York, 1980.

[7] R. W. COTTLE, J.-S. PANG, AND R. E. STONE, *The Linear Complementarity Problem*, Academic Press, New York, 1992.

[8] S. C. DAFERMOS, *An iterative scheme for variational inequalities*, Math. Programming, 26 (1983), pp. 40–47.

[9] S. C. DAFERMOS AND S. C. MCKELVEY, *Partitionable variational inequalities with applications to network and economic equilibria*, J. Optim. Theory Appl., 73 (1992), pp. 243–268.

[10] J. ECKSTEIN AND M. C. FERRIS, *Operator Splitting Methods for Monotone Affine Variational Inequalities, with a Parallel Application to Optimal Control*, Tech. Report, Thinking Machines Corporation, Cambridge, MA, and Computer Sciences Department, University of Wisconsin, Madison, WI, December 1994.

[11] M. C. FERRIS AND O. L. MANGASARIAN, *Error bounds and strong upper semicontinuity for monotone affine variational inequalities*, Ann. Oper. Res., 47 (1993), pp. 293–305.

[12] M. FUKUSHIMA, *Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems*, Math. Programming, 53 (1992), pp. 99–110.

[13] E. M. GAFNI AND D. P. BERTSEKAS, *Two-metric projection methods for constrained optimization*, SIAM J. Control Optim., 22 (1984), pp. 936–964.

[14] D. GAY, *Electronic mail distribution of linear programming test problems*, Math. Programming Soc. COAL Newslett., December 1985.

[15] R. GLOWINSKI, J.-L. LIONS, AND R. TRÉMOLIÉRES, *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.

[16] P. HARKER AND J.-S. PANG, *A damped-Newton method for the linear complementarity problem*, in Computational Solution of Nonlinear Systems of Equations, Lectures in Appl. Math. 26, G. Allgower and K. Georg, eds., American Mathematical Society, Providence, RI, 1990, pp. 265–284.

[17] B. He, *A projection and contraction method for a class of linear complementarity problems and its application in convex quadratic programming*, Appl. Math. Optim., 25 (1992), pp. 247–262.
[18] ———, *A new method for a class of linear variational inequalities*, Math. Programming, 66 (1994), pp. 137–144.
[19] ———, *Solving a class of linear projection equations*, Numer. Math., 68 (1994), pp. 71–80.
[20] A. N. Iusem, *An iterative algorithm for the variational inequality problem*, Mat. Apl. Comput., 13 (1994), pp. 103–114.
[21] E. N. Khobotov, *A modification of the extragradient method for the solution of variational inequalities and some optimization problems*, Zh. Vychisl. Mat. i Mat. Fiz., 27 (1987), pp. 1462–1473.
[22] G. M. Korpelevich, *The extragradient method for finding saddle points and other problems*, Matecon, 12 (1976), pp. 747–756.
[23] X.-D. Luo and P. Tseng, *On global projection-type error bound for the linear complementarity problem*, Linear Algebra Appl., to appear.
[24] Z.-Q. Luo, O. L. Mangasarian, J. Ren, and M. V. Solodov, *New error bounds for the linear complementarity problem*, Math. Oper. Res., 19 (1994), pp. 880–892.
[25] Z.-Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM J. Control Optim., 30 (1992), pp. 408–425.
[26] ———, *Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem*, SIAM J. Optim., 2 (1992), pp. 43–54.
[27] T. L. Magnanti and G. Perakis, *On the Convergence of Classical Variational Inequality Algorithms*, Working Paper, Operations Research Center, Masachusetts Institute of Technology, Cambridge, MA, 1993.
[28] O. L. Mangasarian, *Convergence of iterates of an inexact matrix splitting algorithm for the symmetric monotone linear complementarity problem*, SIAM J. Optim., 1 (1991), pp. 114–122.
[29] O. L. Mangasarian and J. Ren, *New improved error bounds for the linear complementarity problem*, Math. Programming, 66 (1994), pp. 241–255.
[30] O. L. Mangasarian and M. V. Solodov, *Nonlinear complementarity as unconstrained and constrained minimization*, Math. Programming, 62 (1993), pp. 277–297.
[31] P. Marcotte, *Application of Khobotov's algorithm to variational inequalities and network equilibrium problems*, Inform. Systems Oper. Res., 29 (1991), pp. 258–270.
[32] P. Marcotte and J.-P. Dussault, *A note on a globally convergent Newton method for solving monotone variational inequalities*, Oper. Res. Lett., 6 (1987), pp. 35–42.
[33] ———, *A sequential linear programming algorithm for solving monotone variational inequalities*, SIAM J. Control Optim., 27 (1989), pp. 1260–1278.
[34] P. Marcotte and J.-H. Wu, *On the convergence of projection methods: Application to the decomposition of affine variational inequalities*, J. Optim. Theory Appl., 85 (1995), pp. 347–362.
[35] L. Mathiesen, *An algorithm based on a sequence of linear complementarity problems applied to a Walras' equilibrium model: An example*, Math. Programming, 37 (1987), pp. 1–18.
[36] K. G. Murty, *Linear Complementarity, Linear and Nonlinear Programming*, Helderman-Verlag, Berlin, 1988.
[37] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[38] J.-S. Pang, *Asymmetric variational inequality problems over product sets: Applications and iterative methods*, Math. Programming, 31 (1985), pp. 206–219.
[39] ———, *A posteriori error bounds for the linearly-constrained variational inequality problem*, Math. Oper. Res., 12 (1987), pp. 474–484.
[40] J.-S. Pang and D. Chan, *Iterative methods for variational and complementarity problems*, Math. Programming, 24 (1982), pp. 284–313.
[41] J.-S. Pang and S. A. Gabriel, *NE/SQP: A robust algorithm for the nonlinear complementarity problem*, Math. Programming, 60 (1993), pp. 295–337.
[42] P. M. Pardalos and N. Kovoor, *An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds*, Math. Programming, 46 (1986), pp. 235–238.
[43] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Programming Study, 14 (1981), pp. 206–214.
[44] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[45] R. T. Rockafellar and R. J.-B. Wets, *Generalized linear-quadratic problems of deterministic and stochastic optimal control in discrete time*, SIAM J. Control Optim., 28 (1990), pp. 810–822.
[46] M. Sibony, *Méthodes itératives pour les équations et inéquations aux dérivées partielles nonlinéares de type monotone*, Calcolo, 7 (1970), pp. 65–183.
[47] D. Sun, *A new step-size skill for solving a class of nonlinear projection equations*, J. Comput. Math., 13 (1995), pp. 357–368.
[48] ———, *A class of iterative methods for solving nonlinear projection equations*, J. Optim. Theory Appl., 91 (1996), to appear.
[49] P. Tseng, *Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming*, Math. Programming, 48 (1990), pp. 249–263.

[50] P. TSENG, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM J. Control Optim., 29 (1991), pp. 119–138.

[51] ———, *On linear convergence of iterative methods for the variational inequality problem*, J. Comput. Appl. Math., 60 (1995), pp. 237–252.

[52] N. YAMASHITA AND M. FUKUSHIMA, *On stationary points of the implicit Lagrangian for nonlinear complementarity problems*, J. Optim. Theory Appl., 84 (1995), pp. 653–663.

[53] Y. YE, *A new complexity result on minimization of a quadratic function with a sphere constraint*, in Recent Advances in Global Optimization, C. Floudas and P. M. Pardalos, eds., Princeton University Press, Princeton, NJ, 1992.

[54] L. ZHAO AND S. DAFERMOS, *General economic equilibrium and variational inequalities*, Oper. Res. Lett., 10 (1991), pp. 369–376.

[55] C. ZHU AND R. T. ROCKAFELLAR, *Primal-dual projected gradient algorithms for extended linear-quadratic programming*, SIAM J. Optim., 3 (1993), pp. 751–783.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 1831–1847, November 1996

INFINITE-DIMENSIONAL HAMILTON–JACOBI EQUATIONS AND DIRICHLET BOUNDARY CONTROL PROBLEMS OF PARABOLIC TYPE*

PIERMARCO CANNARSA† AND MARIA ELISABETTA TESSITORE‡

Abstract. The paper is concerned with an infinite-dimensional Hamilton–Jacobi equation. This equation is related to boundary control problems of Dirichlet type for semilinear parabolic systems.

The viscosity solution approach is adapted to the equation under consideration, using the properties of fractional powers of generators of analytic semigroups. An existence-and-uniqueness result for such problem is obtained.

Key words. boundary control, viscosity solutions, Hamilton–Jacobi equation, parabolic equations, Dirichlet boundary conditions

AMS subject classifications. 49C20, 35K22, 35K55

1. Introduction.

In this paper we study the existence and uniqueness of viscosity solutions to the infinite-dimensional Hamilton–Jacobi equation

$$\lambda v(x) + \langle Ax + \Phi(x), Dv(x)\rangle + H(A^\beta x, Dv(x)) = 0, \quad x \in X, \tag{1.1}$$

where X is a real Hilbert space, $\lambda > 0$, and $H : X \times X \to \mathbb{R}$ is continuous. Moreover, $A : D(A) \subset X \to X$ is a closed linear operator with a compact and dense inclusion $D(A) \subset X$. Also, we assume A to be positive and self-adjoint. We denote by A^β the fractional power of A. Finally $\Phi : D(A^\beta) \to D(A^{-\beta})$ is Lipschitz continuous.

There is an increasing interest and a growing literature on Hamilton–Jacobi equations in infinite dimensions. These equations were first studied by V. Barbu and G. Da Prato (see, e.g., [2]), setting the problem in classes of convex functions and using semigroup and perturbation methods.

The viscosity solution approach was then adapted to infinite-dimensional equations by M. G. Crandall and P. L. Lions in a sequence of papers [9]. This approach was introduced in [8] (see also [7]) for finite-dimensional problems. It allows one to obtain uniqueness and comparison results for weak solutions of nonlinear first-order PDEs. Additional contributions to the viscosity solution method were obtained by M. Soner [18], H. Ishii [14], and D. Tataru [19], [20]. The latter two authors treated equations with a maximal monotone operator A, possibly multivalued. On the other hand, due to the presence of the unbounded term A^β inside the Hamiltonian H, the results proved in these papers do not apply to (1.1) except for the case of $\beta = 0$.

In this paper we study the above equation for $\beta \in (0, 1)$. We are interested in this problem because it is related to boundary control of parabolic equations under Dirichlet boundary conditions. We now briefly describe such a problem, and more details are given in §2.

It is well known that a possible abstract formulation for modelling parabolic systems controlled at the boundary is given by

$$\begin{cases} x'(t) + Ax(t) + F(x(t)) = A^\beta B\gamma(t), \\ x(0) = x_0, \end{cases} \tag{1.2}$$

where $x_0 \in X$ and $\gamma : [0, +\infty) \to U$ is measurable, U being another Hilbert space. Moreover, $B : U \to X$ is a bounded operator, $F : X \to X$ is Lipschitz, and A is maximal accretive. In

*Received by the editors February 18, 1994; accepted for publication (in revised form) May 15, 1995.

†Dipartimento di Matematica, Università di Roma "Tor Vergata," Via O. Raimondo, 00173 Roma, Italy. The research of this author was supported in part by the Institute for Mathematics and Its Application, with funds provided by the National Science Foundation, and in part by the Italian National Project MURST.

‡Dipartimento di Matematica, Università di Roma "La Sapienza," Piazzale A. Moro 2, 00185 Roma, Italy.

our setup, Neumann-type boundary conditions correspond to values of β in the interval $\left(\frac{1}{4}, 1\right)$, whereas Dirichlet data restrict the range of β to $\left(\frac{3}{4}, 1\right)$. Therefore, all the results that can be obtained for (1.1) with $\beta \in \left(\frac{3}{4}, 1\right)$ apply to both Dirichlet and Neumann boundary control problems.

Denoting by $x(\cdot; x_0, \gamma)$ the mild solution of (1.2), one then seeks to minimize a suitable cost functional over all controls γ. In this paper, we consider the functional

$$J(x_0; \gamma) = \int_0^\infty e^{-\lambda t} L(x(t; x_0, \gamma), \gamma(t))dt, \tag{1.3}$$

where $\lambda > 0$ and $L : X \times U \to \mathbb{R}$ is a given running cost.

Boundary control plays a central role in the theory of distributed parameter systems. There is a vast literature dealing with linear quadratic problems; see, for instance, [1], [11], [15], [3]. In this theory, the main tool for constructing optimal boundary controls is represented by the Riccati equation. The technique used to study this equation for Neumann boundary conditions differs substantially from the one used for Dirichlet conditions. In particular, the way to solve Riccati equations for Neumann data does not apply to Dirichlet data; see, for instance, [12], [13], [10], [15]. In fact, the latter problem requires a much more careful choice of weighted norms and function spaces; see, e.g., [3].

For boundary control problems that are not linear quadratic, the role of the Riccati equation is played by the dynamic programming equation

$$\lambda u(x) + \langle Ax + F(x), Du(x)\rangle + H(x, A^\beta Du(x)) = 0, \quad x \in X, \tag{1.4}$$

where $H : X \times X \to \mathbb{R}$ is defined as

$$H(x, p) = \sup_{\gamma \in U} \left[-\langle B\gamma, p\rangle - L(x, \gamma)\right]. \tag{1.5}$$

The value function of problem (1.3), defined as

$$u(x_0) = \inf \left\{ \int_0^\infty e^{-\lambda t} L(x(t; x_0, \gamma), \gamma(t))dt \mid \gamma : [0, +\infty) \to U \right\}, \tag{1.6}$$

is characterized as the unique solution of (1.4).

In [5], the viscosity solution approach has been adapted to (1.4) for $\beta \in \left(\frac{1}{4}, \frac{1}{2}\right)$. Therefore the results of [5] yield an existence-and-uniqueness theorem for the dynamic programming equation of boundary control problems of Neumann type. On the other hand, similarly to the linear quadratic case, the method of [5] does not apply to (1.4) if $\beta \geq \frac{1}{2}$ and, in particular, to boundary conditions of Dirichlet type.

In this paper we transform the state equation (1.2) by the change of variable $y = A^{-\beta}x$. Accordingly, the dynamic programming equation (1.4) is tranformed into (1.1) with H defined as in (1.5) and

$$\Phi(x) = A^{-\beta}F(A^\beta x).$$

Following the approach of [14], in §3 we give a definition of solution to (1.1) which requires that the equation be satisfied in a suitable viscosity sense only on $D(A)$. Using this definition, we obtain a comparison result for Hölder continuous viscosity solutions of (1.1); see Theorem 3.2. In §4 we prove a Hölder continuity result for the function $v(x) = u(A^\beta x)$. Moreover, we show that v is the unique viscosity solution of (1.1); see Corollary 4.3. In particular, our results characterize the value function u in (1.6) as well.

We conclude this introduction with some comments on possible extensions and applications of our approach. The assumption that $A = A^*$ has been made just to simplify the exposition. Using similar ideas one can treat systems governed by operators that are not necessarily self-adjoint. On the other hand, to prove that the function v is a viscosity solution of (1.1), we need to assume that $-A$ generates an analytic semigroup of compact operators. Therefore, the results of this paper concerning existence typically apply to parabolic boundary control problems in bounded space domains.

Finally, the techniques of this paper can also be used to study boundary control problems of Dirichlet type with finite horizon. In this case, the dynamic programming equation is an evolution equation. For the corresponding Cauchy problem one can prove existence and uniqueness results. The analogous equation for Neumann boundary control is treated in [6].

2. Preliminaries. Let X and U be two real Hilbert spaces and let $\tilde{U} \subset U$ be closed and bounded. We set $R = \sup_{\gamma \in \tilde{U}} |\gamma|$.

Let $x_0 \in X$ and consider the problem of minimizing the functional

$$J(x_0; \gamma) = \int_0^\infty e^{-\lambda t} L(x(t; x_0, \gamma), \gamma(t)) dt \tag{2.1}$$

over all measurable functions $\gamma : [0, \infty) \to \tilde{U}$ (usually called controls). Here $x(\cdot; x_0, \gamma)$ is the mild solution of

$$\begin{cases} x'(t) + Ax(t) + F(x(t)) = A^\beta B\gamma(t), \\ x(0) = x_0 \end{cases} \tag{2.2}$$

that is the solution of the integral equation

$$x(t) = e^{-tA}x_0 - \int_0^t e^{-(t-s)A} F(x(s)) ds + A^\beta \int_0^t e^{-(t-s)A} B\gamma(s) ds. \tag{2.3}$$

In (2.2), A^β denotes the fractional powers of the operator A; see [17]. The discount factor λ is positive, and L satisfies the following assumptions:

$$\begin{aligned} &\text{(i)} \quad L \in C(X \times \tilde{U}), \quad |L(x, \gamma)| \le C_L \quad \forall (x, \gamma) \in X \times \tilde{U}, \\ &\text{(ii)} \quad |L(x, \gamma) - L(y, \gamma)| \le K_L |x - y| \quad \forall \gamma \in \tilde{U}, \ x, y \in X, \end{aligned} \tag{2.4}$$

for some $C_L > 0$ and $K_L > 0$. Moreover we assume

$$\begin{aligned} &\text{(i)} \quad A : D(A) \subset X \to X \text{ is a closed linear operator} \\ &\qquad \text{such that } A = A^* \text{ and } \langle Ax, x \rangle \ge \omega |x|^2 \text{ for some } \omega > 0 \text{ and all } x \in D(A); \\ &\text{(ii)} \quad \text{the inclusion } D(A) \subset X \text{ is dense and compact;} \\ &\text{(iii)} \quad F : X \to X, \ |F(x) - F(y)| \le K_F |x - y|, \ |F(x)| \le C_F \\ &\qquad \forall x, y \in X; \\ &\text{(iv)} \quad \beta \in (0, 1); \\ &\text{(v)} \quad \text{there exists } \rho > 0 \text{ such that } B \in \mathcal{L}(U, D(A^\rho)) \end{aligned} \tag{2.5}$$

for some constants $K_F, C_F > 0$.

We note that (i) and (ii) imply that $-A$ is the infinitesimal generator of an analytic semigroup satisfying $||e^{-tA}|| \le e^{-\omega t}$ for some $\omega > 0$ and all $t \ge 0$. In assumption (v)

above, we have denoted by $\mathcal{L}(U, D(A^\rho))$ the Banach space of all bounded linear operators $B : U \to D(A^\rho)$, where $D(A^\rho)$ is equipped with the graph norm.

It is well known that, under the above assumptions, problem (2.3) has a unique solution in $L^2(0, T; X)$ for any $T > 0$. We define the value function of problem (2.1), (2.2) as

$$u(x_0) = \inf \left\{ \int_0^\infty e^{-\lambda t} L(x(t; x_0, \gamma), \gamma(t)) dt \;\middle|\; \gamma : [0, +\infty) \to \tilde{U} \text{ measurable} \right\}. \tag{2.6}$$

Control processes as above are very important for applications. In fact, (2.2) describes the evolution of a system which is governed by a parabolic PDE and controlled by Dirichlet-type boundary data. We explain this fact below. Let $\Omega \subset \mathbb{R}^n$ be open and bounded with smooth boundary. Consider the following problem:

$$\begin{cases} \dfrac{\partial x}{\partial t}(t, \xi) = \Delta_\xi x(t, \xi) + f(x(t, \xi)) & \text{in } (0, \infty) \times \Omega, \\ x(0, \xi) = x_0(\xi) & \text{on } \Omega, \\ x(t, \xi) = \gamma(t, \xi) & \text{on } (0, \infty) \times \partial\Omega, \end{cases} \tag{2.7}$$

where $x_0 \in L^2(\Omega)$, $\gamma \in L^2(0, \infty; L^2(\partial\Omega))$, and $f : \mathbb{R} \to \mathbb{R}$.

Problem (2.7) may be rewritten in abstract form as follows. Let $X = L^2(\Omega)$ and $U = L^2(\partial\Omega)$, and define an unbounded operator A in X by

$$\begin{aligned} D(A) &= H^2(\Omega) \cap H_0^1(\Omega), \\ Ax &= -\Delta x. \end{aligned}$$

Next, we define the Dirichlet map $\mathbf{D} : U \to X$ as

$$\mathbf{D}\gamma = x \quad \Leftrightarrow \quad \begin{cases} \Delta x = 0 & \text{in } \Omega, \\ x = \gamma & \text{on } \partial\Omega. \end{cases}$$

Formally, (2.7) may be written as

$$\begin{cases} x'(t) + Ax(t) + F(x(t)) = A\mathbf{D}\gamma(t), \\ x(t_0) = x_0, \end{cases} \tag{2.8}$$

where

$$F(x)(\xi) = -f(x(\xi)) \quad \forall x \in X.$$

The right-hand side of (2.8) is not well defined because the range of $\mathbf{D}$ is not contained in $D(A)$. However, we note that $\mathbf{D}$ has some regularizing effect. Indeed, by classical results (see, e.g., [16]), $\mathbf{D} : L^2(\partial\Omega) \to H^{\frac{1}{2}}(\Omega)$, which may be expressed in abstract form using the fractional powers of A. In fact,

$$D(A^\theta) = \begin{cases} H^{2\theta}(\Omega) & \text{if } 0 \le \theta < \frac{1}{4}, \\ \left\{ x \in H^{2\theta}(\Omega) \;:\; x = 0 \text{ on } \partial\Omega \right\} & \text{if } \frac{1}{4} < \theta \le 1. \end{cases}$$

Hence $\mathbf{D} : U \to D(A^\alpha)$ for all $\alpha \in \left[0, \frac{1}{4}\right)$. Consequently, having fixed $\beta \in \left(\frac{3}{4}, 1\right]$, (2.8) can be written as

$$\begin{cases} x'(t) + Ax(t) + F(x(t)) = A^\beta \mathbf{D}_\beta \gamma(t), \\ x(t_0) = x_0, \end{cases} \tag{2.9}$$

where $\mathbf{D}_\beta = A^{1-\beta}\mathbf{D} \in \mathcal{L}(U, X)$. Moreover $\mathbf{D}_\beta$ satisfies (2.5)(v) for any $\rho < \beta - \frac{3}{4}$.

Using the same technique described above, one can show that Neumann-type boundary control problems may be formulated in the same abstract form (2.2). In this case β may be taken in the interval $\left(\frac{1}{4}, 1\right]$. Note that if $\beta \in \left(\frac{3}{4}, 1\right)$ we are modelling the Neumann and the Dirichlet boundary conditions at the same time.

We now return to the analysis of problem (2.6). We transform (2.2) by the change of variable

$$y(t) = A^{-\beta}x(t). \tag{2.10}$$

More precisely, let $y_0 \in X$, and denote by $y(\cdot; y_0, \gamma)$ the solution of

$$\begin{cases} y'(t) + Ay(t) + A^{-\beta}F(A^{\beta}y(t)) = B\gamma(t), \\ y(0) = y_0 \in X. \end{cases} \tag{2.11}$$

Again the above equation has to be understood in mild form

$$y(t) = e^{-tA}y_0 - A^{-\beta}\int_0^t e^{-(t-s)A}F(A^{\beta}y(s))ds + \int_0^t e^{-(t-s)A}B\gamma(s)ds. \tag{2.12}$$

The solution of (2.12) turns out to be continuous, as we show below.

We recall that, since operator $-A$ is the generator of an analytic semigroup in X, for every $\theta \in [0, 1]$ there exists a constant $M_\theta > 0$ such that

$$|A^{\theta}e^{-tA}x| \le \frac{M_\theta}{t^{\theta}}|x| \quad \forall t > 0, \forall x \in X. \tag{2.13}$$

Moreover let $\gamma \in (0, 1]$ and $\alpha \in (0, \gamma)$. Then, a well-known interpolation inequality (see, e.g., [17]) states that for every $\sigma > 0$ there exists $C_\sigma > 0$ such that

$$|A^{\alpha}x| \le \sigma|A^{\gamma}x| + C_\sigma|x| \quad \forall x \in D(A^{\gamma}), \tag{2.14}$$

and there exists $C_{\alpha\gamma} > 0$ such that

$$|A^{\alpha}x| \le C_{\alpha\gamma}|A^{\gamma}x|^{\frac{\alpha}{\gamma}}|x|^{1-\frac{\alpha}{\gamma}} \quad \forall x \in D(A^{\gamma}). \tag{2.15}$$

PROPOSITION 2.1. *Assume that* (2.5) *holds. Let* $\gamma : [0, \infty) \to U$ *be a bounded measurable control, and fix* $T > 0$. *Then for any* $y_0 \in X$ *there exists a unique solution of* (2.12) *and*

$$y \in C([0, T]; X) \cap L^1(0, T; D(A^{\beta})). \tag{2.16}$$

Moreover, if $y_0 \in D(A^{\frac{1}{2}})$, *then*

$$y \in C([0, T]; D(A^{\frac{1}{2}})) \cap L^2(0, T; D(A)) \cap W^{1,2}(0, T; X). \tag{2.17}$$

Finally, if $y_0 \in D(A)$, *then*

$$y \in C([0, T]; D(A)). \tag{2.18}$$

Proof. The argument is well known. We sketch the proof for the reader's convenience. First we show that (2.12) has a unique solution $y \in L^1(0, T; D(A^{\beta}))$. Fix $y_0 \in X$, and let $T_1 = \frac{1}{2K_F}$. Define the map Φ on $L^1(0, T_1; D(A^{\beta}))$ by

$$\Phi y(t) = e^{-tA}y_0 - A^{-\beta}\int_0^t e^{-(t-s)A}F(A^{\beta}y(s))ds + \int_0^t e^{-(t-s)A}B\gamma(s)ds$$

for any $0 \le t \le T_1$. Let us prove that

$$\Phi : L^1(0, T_1; D(A^\beta)) \to L^1(0, T_1; D(A^\beta)).$$

Indeed, recalling (2.13), we have

$$\int_0^{T_1} |A^\beta \Phi y(t)| dt \le \int_0^{T_1} |A^\beta e^{-tA} y_0| dt$$

$$+ \int_0^{T_1} \left| \int_0^t e^{-(t-s)A} F(A^\beta y(s)) ds \right| dt + \int_0^{T_1} \left| A^\beta \int_0^t e^{-(t-s)A} B\gamma(s) ds \right| dt$$

$$\le M_\beta \int_0^{T_1} \frac{|y_0|}{t^\beta} dt + C_F \int_0^{T_1} \int_0^t \left(|A^\beta y(s)| + 1\right) ds dt + M_\beta \int_0^{T_1} \int_0^t \frac{|B\gamma(s)|}{(t-s)^\beta} ds dt$$

$$\le M_\beta |y_0| T_1^{1-\beta} + C_F T_1 ||y||_{L^1(0,T_1;D(A^\beta))} + C_F T_1^2 + M_\beta R ||B|| T_1^{1-\beta},$$

recalling that $|\gamma(s)| \le R$. Hence $\Phi y \in L^1(0, T_1; D(A^\beta))$.

Next we prove that Φ is a contraction. For any $y, z \in L^1(0, T_1; D(A^\beta))$ we have

$$\int_0^{T_1} \left| A^\beta (\Phi y(s) - \Phi z(s)) \right| ds$$

$$\le K_F \int_0^{T_1} \int_0^t \left| A^\beta (y(s) - z(s)) \right| ds dt = K_F T_1 ||y(s) - z(s)||_{L^1(0,T_1;D(A^\beta))}.$$

By the contraction map theorem it follows that (2.12) has a unique solution $y \in L^1(0, T_1; D(A^\beta))$. Then by classical results (see, e.g., [17]), $y(t) \in C([0, T_1]; X)$. Therefore, iterating this procedure, we can cover the interval $[0, T]$ with a finite number of steps.

As for (2.17), the maximal regularity result $y \in L^2(0, T; D(A)) \cap W^{1,2}(0, T; X)$ is well known; see, e.g., [3]. The fact that $y \in C([0, T]; D(A^{\frac{1}{2}}))$ is also a well-known consequence of the maximal regularity result.

Finally, if $y_0 \in D(A)$, then recalling assumption (2.5)(v) and writing Ay as

$$Ay(t) = Ae^{-tA} y_0 - A^{1-\beta} \int_0^t e^{-(t-s)A} F(A^\beta y(s)) ds + A^{1-\rho} \int_0^t A^\rho e^{-(t-s)A} B\gamma(s) ds$$

$$= Ay_1(t) + A^{1-\beta} y_2(t) + A^{1-\rho} y_3(t)$$

we easily see that, since e^{-tA} is a strongly continuous semigroup, then Ay_1 is continuous. In addition $A^{1-\beta} y_2$ is continuous since we know that $y_2 \in C([0, T]; D(A^{\frac{1}{2}}))$ and $1 - \beta < \frac{1}{2}$. Moreover since

$$|A^{1-\rho}(y_3(t_1) - y_3(t_2))| \le \left| A^{1-\rho} \int_0^{t_2} e^{-(t_2-s)A} [e^{-(t_1-t_2)A} - I] A^\rho B\gamma(s) ds \right|$$

$$+ \left| A^{1-\rho} \int_{t_2}^{t_1} e^{-(t_1-s)A} A^\rho B\gamma(s) ds \right|$$

if $t_1 \ge t_2$, exploiting inequality (2.13) and the boundedness of γ, we derive that $A^{1-\rho} y_3(t)$ is continuous. □

By inserting the change of variable (2.10) in the cost functional (2.1), we obtain a new optimal control problem whose value function v is given by

$$v(y_0) = \inf_{\gamma(t)\in\tilde{U}} \int_0^\infty e^{-\lambda t} L(A^\beta y(t; y_0, \gamma), \gamma(t))dt. \tag{2.19}$$

It is easy to realize that value functions v and u are related by the formula

$$u(x) = v(A^{-\beta}x) \quad \forall x \in X. \tag{2.20}$$

In particular, u is uniquely determined once v has been characterized. Therefore, we will study problem (2.11)–(2.19) instead of (2.2)–(2.6).

We will show that v is the unique solution of the following Hamilton–Jacobi–Bellman equation:

$$\lambda v(x) + H(A^\beta x, Dv(x)) + \langle Ax + A^{-\beta}F(A^\beta x), Dv(x)\rangle = 0, \tag{2.21}$$

where

$$H(x, p) = \sup_{\gamma\in\tilde{U}} \left[-\langle B\gamma, p\rangle - L(x, \gamma)\right]. \tag{2.22}$$

Clearly, one needs a suitable notion of weak solution of problem (2.21), since v is not everywhere differentiable and the coefficients of the equation are discontinuous. In the subsequent discussion, we use viscosity solutions to overcome these difficulties.

3. Definition of viscosity solution and comparison result. In this section we study the Hamilton–Jacobi equation

$$\lambda u(x) + H(A^\beta x, Du(x)) + \langle Ax + A^{-\beta}F(A^\beta x), Du(x)\rangle = 0. \tag{3.1}$$

We assume that (2.5) holds and that $H : X \times X \to \mathbb{R}$ is a function, not necessarily given by (2.22), satisfying

$$|H(A^\beta x, p) - H(A^\beta y, q)| \le K_H \left(|A^\beta(x - y)| + |p - q|\right) \text{ for some } K_H > 0. \tag{3.2}$$

Let $w, \varphi : D(A^{\frac{1}{2}}) \to \mathbb{R}$ be given. For any $\delta > 0$ we define $M_\delta^+(w, \varphi)$ to be the set of all points $x \in D(A^{\frac{1}{2}})$ such that

$$w(x) - \varphi(x) - \frac{\delta}{2}|A^{\frac{1}{2}}x|^2 \ge w(y) - \varphi(y) - \frac{\delta}{2}|A^{\frac{1}{2}}y|^2 \tag{3.3}$$

for all $y \in D(A^{\frac{1}{2}})$. Similarly, we denote by $M_\delta^-(w, \varphi)$ the set of all points $x \in D(A^{\frac{1}{2}})$ such that

$$w(x) - \varphi(x) + \frac{\delta}{2}|A^{\frac{1}{2}}x|^2 \le w(y) - \varphi(y) + \frac{\delta}{2}|A^{\frac{1}{2}}y|^2 \tag{3.4}$$

for all $y \in D(A^{\frac{1}{2}})$.

DEFINITION 3.1. *We say that a bounded continuous function $w : X \to \mathbb{R}$ is a viscosity subsolution of* (3.1) *if w is sequentially weakly upper semicontinuous, and, for every test function $\varphi \in C^1(D(A^{\frac{1}{2}}))$ and $\delta > 0$,*

(i) $M_\delta^+(w, \varphi) \subset D(A)$;

(ii) $$\lambda w(x) + H(A^\beta x, D\varphi(x) + \delta Ax) + \langle Ax + A^{-\beta}F(A^\beta x), D\varphi(x)\rangle$$
$$+\delta|Ax|^2 + \delta\langle Ax, A^{-\beta}F(A^\beta x)\rangle \le 0 \quad \forall x \in M_\delta^+(w, \varphi). \tag{3.5}$$

We say that w is a viscosity supersolution of (3.1) *if w is sequentially weakly lower semicontinuous, and, for every test function $\varphi \in C^1(D(A^{\frac{1}{2}}))$ and $\delta > 0$,*

$$\text{(i)} \quad M_\delta^-(w,\varphi) \subset D(A);$$

$$\text{(ii)} \quad \lambda w(x) + H(A^\beta x, D\varphi(x) - \delta Ax) + \left\langle Ax + A^{-\beta}F(A^\beta x), D\varphi(x)\right\rangle \tag{3.6}$$
$$-\delta|Ax|^2 - \delta\left\langle Ax, A^{-\beta}F(A^\beta x)\right\rangle \geq 0 \quad \forall x \in M_\delta^-(w,\varphi).$$

We say that w is a viscosity solution of (3.1) *if it is both a viscosity subsolution and a supersolution of* (3.1).

Now we give a comparison result between viscosity subsolutions and supersolutions of (3.1).

THEOREM 3.2. *Assume that* (2.5) *and* (3.2) *hold true and suppose $\beta \in \left(\frac{3}{4}, 1\right)$. Define $\alpha_\beta \in (0,1)$ as*

$$\alpha_\beta = \frac{4\beta - 3}{2\beta - 1}. \tag{3.7}$$

Let u and v be, respectively, a viscosity subsolution and supersolution of the Hamilton–Jacobi equation (3.1). *If u and v are Hölder continuous of exponent $\alpha > \alpha_\beta$, then*

$$u(x) \leq v(x) \quad \forall x \in X. \tag{3.8}$$

Proof. For simplicity we take $\lambda = 1$. For ε and δ positive, we define a function $\phi : D(A^{\frac{1}{2}}) \times D(A^{\frac{1}{2}}) \to \mathbb{R}$ as

$$\phi(x,y) = u(x) - v(y) - \frac{1}{2\varepsilon}\left\langle A^{\frac{1}{2}}(x-y), x-y\right\rangle - \frac{\delta}{2}\left[\langle Ax, x\rangle + \langle Ay, y\rangle\right]. \tag{3.9}$$

Note that ϕ is weakly upper semicontinuous. Let $(x_{\varepsilon,\delta}, y_{\varepsilon,\delta}) \in D(A^{\frac{1}{2}}) \times D(A^{\frac{1}{2}})$ be such that

$$\phi(x_{\varepsilon,\delta}, y_{\varepsilon,\delta}) = \max_{D(A^{\frac{1}{2}})\times D(A^{\frac{1}{2}})} \phi(x,y).$$

First of all we prove that

$$|x_{\varepsilon,\delta} - y_{\varepsilon,\delta}| \leq C_1 \varepsilon^{\frac{1}{2-\alpha}}, \tag{3.10}$$

where $C_1 > 0$ and α is the Hölder exponent of u and v. Since

$$\phi(x_{\varepsilon,\delta}, x_{\varepsilon,\delta}) + \phi(y_{\varepsilon,\delta}, y_{\varepsilon,\delta}) \leq 2\phi(x_{\varepsilon,\delta}, y_{\varepsilon,\delta}),$$

from the Hölder continuity of u and v we derive

$$\frac{1}{\varepsilon}|x_{\varepsilon,\delta} - y_{\varepsilon,\delta}|^2 \leq C|x_{\varepsilon,\delta} - y_{\varepsilon,\delta}|^\alpha \tag{3.11}$$

for some positive constant C. Therefore (3.10) holds.

Now let us consider

$$\varphi(x) = v(y_{\varepsilon,\delta}) + \frac{1}{2\varepsilon}\left\langle A^{\frac{1}{2}}(x - y_{\varepsilon,\delta}), x - y_{\varepsilon,\delta}\right\rangle + \frac{\delta}{2}\left\langle Ay_{\varepsilon,\delta}, y_{\varepsilon,\delta}\right\rangle,$$
$$\psi(y) = u(x_{\varepsilon,\delta}) - \frac{1}{2\varepsilon}\left\langle A^{\frac{1}{2}}(x_{\varepsilon,\delta} - y), x_{\varepsilon,\delta} - y\right\rangle - \frac{\delta}{2}\left\langle Ax_{\varepsilon,\delta}, x_{\varepsilon,\delta}\right\rangle.$$

Note that $\varphi, \psi \in C^1(D(A^{\frac{1}{2}}))$. Also, $x_{\varepsilon,\delta} \in M_\delta^+(u,\varphi)$ and $y_{\varepsilon,\delta} \in M_\delta^-(v,\psi)$ by construction. Since u is a viscosity subsolution, using φ as a test function, we have

$$\begin{aligned} &u(x_{\varepsilon,\delta}) + H\left(A^\beta x_{\varepsilon,\delta}, \frac{A^{\frac{1}{2}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})}{\varepsilon} + \delta A x_{\varepsilon,\delta}\right) + \delta |A x_{\varepsilon,\delta}|^2 \\ &+\delta \left\langle A x_{\varepsilon,\delta}, A^{-\beta} F(A^\beta x_{\varepsilon,\delta})\right\rangle + \left\langle A x_{\varepsilon,\delta} + A^{-\beta} F(A^\beta x_{\varepsilon,\delta}), \frac{A^{\frac{1}{2}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})}{\varepsilon}\right\rangle \le 0. \end{aligned} \tag{3.12}$$

Since v is a viscosity supersolution, using ψ as a test function, we have

$$\begin{aligned} &v(y_{\varepsilon,\delta}) + H\left(A^\beta y_{\varepsilon,\delta}, \frac{A^{\frac{1}{2}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})}{\varepsilon} - \delta A y_{\varepsilon,\delta}\right) - \delta |A y_{\varepsilon,\delta}|^2 \\ &-\delta \left\langle A y_{\varepsilon,\delta}, A^{-\beta} F(A^\beta y_{\varepsilon,\delta})\right\rangle + \left\langle A y_{\varepsilon,\delta} + A^{-\beta} F(A^\beta y_{\varepsilon,\delta}), \frac{A^{\frac{1}{2}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})}{\varepsilon}\right\rangle \ge 0. \end{aligned} \tag{3.13}$$

Subtracting (3.13) from (3.12), we obtain

$$\begin{aligned} &u(x_{\varepsilon,\delta}) - v(y_{\varepsilon,\delta}) + \delta\left[|A x_{\varepsilon,\delta}|^2 + |A y_{\varepsilon,\delta}|^2\right] + \frac{1}{\varepsilon}|A^{\frac{3}{4}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^2 \\ &\le H\left(A^\beta y_{\varepsilon,\delta}, \frac{A^{\frac{1}{2}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})}{\varepsilon} - \delta A y_{\varepsilon,\delta}\right) - H\left(A^\beta x_{\varepsilon,\delta}, \frac{A^{\frac{1}{2}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})}{\varepsilon} + \delta A x_{\varepsilon,\delta}\right) \\ &-\delta\left[\left\langle A x_{\varepsilon,\delta}, A^{-\beta} F(A^\beta x_{\varepsilon,\delta})\right\rangle + \left\langle A y_{\varepsilon,\delta}, A^{-\beta}[F(A^\beta y_{\varepsilon,\delta})]\right\rangle\right] \\ &+\left\langle A^{-\beta}\left[F(A^\beta y_{\varepsilon,\delta}) - F(A^\beta x_{\varepsilon,\delta})\right], \frac{A^{\frac{1}{2}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})}{\varepsilon}\right\rangle. \end{aligned} \tag{3.14}$$

Recalling assumption (3.2) on H and assumption (2.5) on F, the above inequality yields

$$\begin{aligned} &u(x_{\varepsilon,\delta}) - v(y_{\varepsilon,\delta}) + \delta\left[|A x_{\varepsilon,\delta}|^2 + |A y_{\varepsilon,\delta}|^2\right] + \frac{1}{\varepsilon}|A^{\frac{3}{4}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^2 \\ &\le K_H \delta\left[|A x_{\varepsilon,\delta}| + |A y_{\varepsilon,\delta}|\right] + K_H |A^\beta(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})| \\ &+\delta C_F ||A^{-\beta}||[|A x_{\varepsilon,\delta}| + |A y_{\varepsilon,\delta}|] + K_F |A^\beta(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})| \frac{|A^{\frac{1}{2}-\beta}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|}{\varepsilon}. \end{aligned} \tag{3.15}$$

Now we estimate the right-hand side of (3.15). Recalling the interpolation inequality (2.15), we get

$$K_H |A^\beta(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})| \le C_2 |A(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^{4\beta-3} |A^{\frac{3}{4}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^{4-4\beta} \tag{3.16}$$

for some $C_2 > 0$. Moreover recall the well-known inequality

$$ab \le \frac{\sigma^p}{p} a^p + \frac{1}{\sigma^q q} b^q \tag{3.17}$$

for every $a, b \in \mathbb{R}^+$, $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, and $\sigma > 0$. Choosing $\sigma = (\frac{\delta}{16})^{(4\beta-3)/2}$ and $p = \frac{2}{4\beta-3}$ in (3.17) and applying it to (3.16) we derive

$$
\begin{aligned}
&C_2|A(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^{4\beta-3}|A^{\frac{3}{4}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^{4-4\beta} \\
&\leq \frac{\delta}{16}|A(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^2 + \frac{C_3}{\delta^{\frac{4\beta-3}{5-4\beta}}}\left|A^{\frac{3}{4}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})\right|^{\frac{8-8\beta}{5-4\beta}},
\end{aligned} \tag{3.18}
$$

where C_3 is some positive constant. On the other hand, again applying (3.17) with $p = \frac{5-4\beta}{4-4\beta}$ and $\sigma = \left(\frac{1}{4\varepsilon}\right)^{(4-4\beta)/(5-4\beta)}$, we find

$$
\frac{C_3}{\delta^{\frac{4\beta-3}{5-4\beta}}}\left|A^{\frac{3}{4}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})\right|^{\frac{8-8\beta}{5-4\beta}} \leq \frac{1}{4\varepsilon}|A^{\frac{3}{4}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^2 + \frac{C_4\varepsilon^{4-4\beta}}{\delta^{4\beta-3}} \tag{3.19}
$$

for $C_4 > 0$. From estimates (3.18) and (3.19), inequality (3.16) can be rewritten as

$$
K_H|A^\beta(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})| \leq \frac{\delta}{8}\left[|Ax_{\varepsilon,\delta}|^2 + |Ay_{\varepsilon,\delta}|^2\right] + \frac{1}{4\varepsilon}|A^{\frac{3}{4}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^2 + \frac{C_4\varepsilon^{4-4\beta}}{\delta^{4\beta-3}}. \tag{3.20}
$$

Moreover

$$
K_H\delta\left[|Ax_{\varepsilon,\delta}| + |Ay_{\varepsilon,\delta}|\right] \leq \frac{\delta}{8}\left[|Ax_{\varepsilon,\delta}|^2 + |Ay_{\varepsilon,\delta}|^2\right] + C_5\delta, \tag{3.21}
$$

where C_5 is a positive constant. On the other hand we get

$$
\delta C_F||A^{-\beta}||[|Ax_{\varepsilon,\delta}| + |Ay_{\varepsilon,\delta}|] \leq \frac{\delta}{8}\left[|Ax_{\varepsilon,\delta}|^2 + |Ay_{\varepsilon,\delta}|^2\right] + C_6\delta \tag{3.22}
$$

for $C_6 > 0$. Finally from estimate (3.11), it follows that

$$
K_F|A^\beta(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|\frac{|A^{\frac{1}{2}-\beta}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|}{\varepsilon} \leq \frac{C_7|A^\beta(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|}{\varepsilon^{\frac{1-\alpha}{2-\alpha}}}, \tag{3.23}
$$

where $C_7 > 0$. Applying the interpolation inequality (2.15) and inequality (3.17) to (3.23) as we did in (3.16) we find

$$
\frac{C_7|A^\beta(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|}{\varepsilon^{\frac{1-\alpha}{2-\alpha}}} \leq \frac{\delta}{16}|A(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^2 + \frac{C_8|A^{\frac{3}{4}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^{\frac{8-8\beta}{5-4\beta}}}{\delta^{\frac{4\beta-3}{5-4\beta}}\varepsilon^{\frac{2-2\alpha}{(2-\alpha)(5-4\beta)}}}
$$

for some positive constant C_8. Again, using (3.17) in the last term of the above inequality we rewrite (3.23) as

$$
\begin{aligned}
&K_F|A^\beta(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|\frac{|A^{\frac{1}{2}-\beta}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|}{\varepsilon} \\
&\leq \frac{\delta}{16}|A(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^2 + \frac{1}{4\varepsilon}|A^{\frac{3}{4}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^2 + \frac{C_9\varepsilon^{4-4\beta}}{\delta^{4\beta-3}\varepsilon^{\frac{2-2\alpha}{2-\alpha}}},
\end{aligned} \tag{3.24}
$$

with C_9 positive constant. Substituting estimates (3.20)–(3.22) and (3.24) in inequality (3.15) we get

$$
u(x_{\varepsilon,\delta}) - v(y_{\varepsilon,\delta}) + \frac{\delta}{2}\left[|Ax_{\varepsilon,\delta}|^2 + |Ay_{\varepsilon,\delta}|^2\right] + \frac{1}{2\varepsilon}|A^{\frac{3}{4}}(x_{\varepsilon,\delta} - y_{\varepsilon,\delta})|^2 \leq C_{10}\delta + \frac{C_9\varepsilon^\gamma}{\delta^{4\beta-3}}, \tag{3.25}
$$

where $C_{10} > 0$ and $\gamma = 4 - 4\beta - \frac{2-2\alpha}{2-\alpha}$ is positive as $\alpha > \alpha_\beta$. Therefore, if $x \in D(A^{\frac{1}{2}})$ we have

$$u(x) - v(x) = \phi(x, x) + \delta\langle Ax, x\rangle \leq \phi(x_{\varepsilon,\delta}, y_{\varepsilon,\delta}) + \delta\langle Ax, x\rangle$$

$$\leq u(x_{\varepsilon,\delta}) - v(y_{\varepsilon,\delta}) + \delta\langle Ax, x\rangle \leq C_{10}\delta + \frac{C_9\varepsilon^\gamma}{\delta^{4\beta-3}} + \delta\langle Ax, x\rangle.$$

Letting $\varepsilon \to 0$ and then $\delta \to 0$ we conclude that

$$u(x) \leq v(x) \quad \forall x \in D(A^{\frac{1}{2}}).$$

Since $D(A^{\frac{1}{2}})$ is dense in X, we have $u(x) \leq v(x)$ for every $x \in X$. □

4. Properties of the value function and existence results. In this section we prove that the value function v of problem (2.11)–(2.19) is the unique viscosity solution of

$$\lambda v(x) + H(A^\beta x, Dv(x)) + \left\langle Ax + A^{-\beta}F(A^\beta x), Dv(x)\right\rangle = 0, \tag{4.1}$$

where

$$H(x, p) = \sup_{\gamma \in \tilde{U}} \left[-\langle B\gamma, p\rangle - L(x, \gamma)\right].$$

We first show a Hölder continuity result for v. We will exploit the technique of [5].

PROPOSITION 4.1. *Assume* (2.5), (2.4). *Then the value function v defined in* (2.19) *is Hölder continuous in X with any exponent $\alpha \in (0, 1]$ satisfying $\alpha < \frac{\lambda}{K_F}$. Moreover for any $\theta \in [0, 1 - \beta)$ there exists a constant $C_{\alpha\theta} > 0$ such that*

$$|v(x) - v(y)| \leq C_{\alpha\theta}|A^{-\theta}(x - y)|^\alpha \tag{4.2}$$

for all $x, y \in X$.

Proof. Let $x_0, y_0 \in X$ and $\gamma(t) \in \tilde{U}$ be given. Let us set $x(\cdot) = x(\cdot; x_0, \gamma)$ and $y(\cdot) = y(\cdot; y_0, \gamma)$. Then

$$x(t) = e^{-tA}x_0 - A^{-\beta}\int_0^t e^{-(t-s)A}F(A^\beta x(s))ds + \int_0^t e^{-(t-s)A}B\gamma(s)ds$$

and

$$y(t) = e^{-tA}y_0 - A^{-\beta}\int_0^t e^{-(t-s)A}F(A^\beta y(s))ds + \int_0^t e^{-(t-s)A}B\gamma(s)ds$$

for any $t \geq 0$. Now we estimate $|A^\beta(x(t) - y(t))|^\alpha$. From assumption (2.5) and from inequality (2.13) we have

$$|A^\beta(x(t) - y(t))| \leq \frac{M_\theta}{t^{\beta+\theta}}|A^{-\theta}(x_0 - y_0)| + K_F\int_0^t |A^\beta(x(s) - y(s))|ds \tag{4.3}$$

for any $\theta \in [0, 1 - \beta)$. Now set $\eta(t) = \int_0^t |A^\beta(x(s) - y(s))|ds$. Integrating the above inequality we get

$$\eta(t) \leq \frac{M_\theta}{1 - (\beta + \theta)}|A^{-\theta}(x_0 - y_0)|t^{1-\beta-\theta} + K_F\int_0^t \eta(s)ds.$$

By Gronwall's lemma we obtain an estimate on $\eta(t)$. Applying this estimate to the right-hand side of (4.3) we derive

$$|A^\beta(x(t)-y(t))| \le \left(\frac{C}{t^{\beta+\theta}} + Ce^{K_F t} t^{1-\beta-\theta}\right) |A^{-\theta}(x_0-y_0)|. \tag{4.4}$$

Then, for every $\alpha \in (0, \frac{\lambda}{K_F})$, $\alpha \le 1$, we have

$$|A^\beta(x(t)-y(t))|^\alpha \le 2^\alpha \left(\frac{C}{t^{\alpha(\beta+\theta)}} + Ce^{K_F \alpha t} t^{(1-\beta-\theta)\alpha}\right) |A^{-\theta}(x_0-y_0)|^\alpha. \tag{4.5}$$

Moreover, by (2.4),

$$\begin{aligned} &\left|L(A^\beta x(t), \gamma(t)) - L(A^\beta y(t), \gamma(t))\right| \\ &\le (2C_L)^{1-\alpha} \left|L(A^\beta x(t), \gamma(t)) - L(A^\beta y(t), \gamma(t))\right|^\alpha \le \tilde{L}|A^\beta(x(t)-y(t))|^\alpha, \end{aligned} \tag{4.6}$$

where $\tilde{L} = (2C_L)^{1-\alpha} K_L^\alpha$. Now choose T such that

$$\frac{2e^{-\lambda T} C_L}{\lambda} \le |A^{-\theta}(x_0-y_0)|^\alpha.$$

From the definition of value function and from the dynamic programming principle it follows that there exists a control $\gamma(\cdot)$ such that

$$v(y_0) > \int_0^T e^{-\lambda t} L(A^\beta y(t), \gamma(t))dt + e^{-\lambda T} v(y(T)) - |A^{-\theta}(x_0-y_0)|^\alpha.$$

Here, we may suppose that $|A^{-\theta}(x_0-y_0)|^\alpha > 0$, since (4.2) is trivial if $|A^{-\theta}(x_0-y_0)|^\alpha = 0$.

From the dynamic programming principle and from the above estimate it follows that

$$\begin{aligned} &v(x_0) - v(y_0) \le |A^{-\theta}(x_0-y_0)|^\alpha \\ &+ \int_0^T e^{-\lambda t} \left|L(A^\beta x(t), \gamma(t)) - L(A^\beta y(t), \gamma(t))\right| dt + e^{-\lambda T}[v(x(T)) - v(y(T))] \\ &\le 2|A^{-\theta}(x_0-y_0)|^\alpha + \tilde{L} \int_0^T e^{-\lambda t} |A^\beta(x(t)-y(t))|^\alpha dt. \end{aligned} \tag{4.7}$$

Substituting (4.5) in (4.7), we get

$$\begin{aligned} &v(x_0) - v(y_0) \le 2|A^{-\theta}(x_0-y_0)|^\alpha \\ &+\tilde{L}2^\alpha |A^{-\theta}(x_0-y_0)|^\alpha \int_0^T \left(\frac{Ce^{-\lambda t}}{t^{\alpha(\beta+\theta)}} + Ce^{(K_F\alpha-\lambda)t} t^{(1-\beta-\theta)\alpha}\right) dt. \end{aligned}$$

The result follows since $\alpha < \frac{\lambda}{K_F}$ and $\theta \in [0, 1-\beta)$. □

We now give an existence result for (4.1).

THEOREM 4.2. *Assume that* (2.5) *and* (2.4) *hold true. Then the value function* v *is a viscosity solution of* (2.21) *in the sense of Definition* 3.1.

Proof. Recalling the compactness assumption (2.5)(ii), from (4.2) we conclude that v is sequentially weakly continuous in X. It remains to prove that v satisfies (3.5) and (3.6). We show this fact in the next two steps.

Step I: Proof of (3.5). Let $\gamma \in \tilde{U}$ be a constant control, $\varphi \in C^1(D(A^{\frac{1}{2}}))$, and $x \in M_\delta^+(v, \varphi)$. Moreover let $y(\cdot) = y(\cdot; x, \gamma)$. Then, recalling Proposition 2.1, we obtain

$$v(x) - \varphi(x) - \frac{\delta}{2}\langle Ax, x\rangle \geq v(y(t)) - \varphi(y(t)) - \frac{\delta}{2}\langle Ay(t), y(t)\rangle \tag{4.8}$$

for every $t \geq 0$. From (4.8) and from the dynamic programming principle it follows that

$$\begin{aligned} &\frac{\varphi(x) - \varphi(y(t))}{t} + \frac{\delta}{2}\frac{\langle Ax, x\rangle - \langle Ay(t), y(t)\rangle}{t} \\ &\leq \frac{v(x) - v(y(t))}{t} \leq \frac{1}{t}\int_0^t e^{-\lambda s} L(A^\beta y(s), \gamma) ds + \frac{e^{-\lambda t} - 1}{t} v(y(t)). \end{aligned} \tag{4.9}$$

Note that, since by Proposition 2.1 $y \in L^2(0, T; D(A))$, we get

$$\varphi(x) - \varphi(y(t)) = -\int_0^t \big\langle D\varphi(y(s)), -Ay(s) - A^{-\beta} F(A^\beta y(s)) + B\gamma \big\rangle ds \tag{4.10}$$

and

$$\begin{aligned} \frac{1}{2}(\langle Ax, x\rangle - \langle Ay(t), y(t)\rangle) &= -\int_0^t \big\langle Ay(s), -Ay(s) - A^{-\beta} F(A^\beta y(s)) + B\gamma \big\rangle ds \\ &= \int_0^t \big[|Ay(s)|^2 + \big\langle Ay(s), A^{-\beta} F(A^\beta y(s)) - B\gamma \big\rangle\big] ds. \end{aligned} \tag{4.11}$$

Therefore, exploiting (4.10) and (4.11), (4.9) can be rewritten as

$$\begin{aligned} &\frac{1}{t}\int_0^t \big\langle D\varphi(y(s)), Ay(s) + A^{-\beta} F(A^\beta y(s)) - B\gamma \big\rangle ds \\ &+ \frac{\delta}{t}\int_0^t \big[|Ay(s)|^2 + \big\langle Ay(s), A^{-\beta} F(A^\beta y(s)) - B\gamma \big\rangle\big] ds \\ &\leq \frac{1}{t}\int_0^t e^{-\lambda s} L(A^\beta y(s), \gamma) ds + \frac{e^{-\lambda t} - 1}{t} v(y(t)). \end{aligned} \tag{4.12}$$

By Proposition 2.1 $y \in L^2(0, T; D(A))$ and so

$$\frac{1}{t}\int_0^t \big\langle D\varphi(y(s)), Ay(s) + A^{-\beta} F(A^\beta y(s)) - B\gamma \big\rangle ds \leq \frac{\delta}{4t}\int_0^t |Ay(s)|^2 ds + C_\delta \tag{4.13}$$

and

$$\begin{aligned} &\frac{\delta}{t}\int_0^t \big\langle Ay(s), A^{-\beta} F(A^\beta y(s)) - B\gamma \big\rangle ds \\ &\leq \frac{\delta}{t}\int_0^t ||A^{-\beta}|| C_F |Ay(s)| ds + \frac{\delta}{t}\int_0^t |Ay(s)|||B||R ds \leq \frac{\delta}{4t}\int_0^t |Ay(s)|^2 ds + C_\delta \end{aligned} \tag{4.14}$$

for some positive C_δ.

From (4.12)–(4.14), since v and L are bounded it follows that

$$\frac{1}{t}\int_0^t |Ay(s)|^2 ds \leq C_\delta$$

for C_δ positive. Hence, there exists a sequence $\{t_n\}$, $t_n \downarrow 0$ such that

$$|Ay(t_n)| \le C_\delta.$$

Taking a subsequence we have that $Ay(t_n) \rightharpoonup z$ and $y(t_n) \to x$. Therefore we get $y(t_n) = A^{-1}Ay(t_n) \rightharpoonup A^{-1}z = x$, and so $x \in D(A)$. This proves (3.5)(i).

In order to show that (ii) holds, we recall that if $x \in D(A)$ and $\gamma(\cdot) = \gamma$, then $y \in C([0,T]; D(A))$; see Proposition 2.1. Then, passing to the limit as $t \downarrow 0$ in (4.12), we derive

$$\begin{aligned}&\langle D\varphi(x), Ax + A^{-\beta}F(A^\beta x)\rangle + \left[-\langle D\varphi(x) + \delta Ax, B\gamma\rangle - L(A^\beta x, \gamma)\right]\\ &+\delta|Ax|^2 + \delta\langle Ax, A^{-\beta}F(A^\beta x)\rangle + \lambda v(x) \le 0.\end{aligned}$$

Taking the supremum over $\gamma \in \tilde{U}$ we obtain (3.5)(ii).

Step II: Proof of (3.6). Let $\varphi \in C^1(D(A^{\frac{1}{2}}))$ and $x \in M_\delta^-(v,\varphi)$. For every $n \in \mathbb{N}$ there exists a control $\gamma_n(\cdot)$ such that

$$v(x) + \frac{1}{n^2} \ge \int_0^{\frac{1}{n}} e^{-\lambda s} L(A^\beta y_n(s), \gamma_n(s))ds + e^{-\frac{\lambda}{n}} v\left(y_n\left(\frac{1}{n}\right)\right), \tag{4.15}$$

where $y_n(\cdot) = y_n(\cdot; x, \gamma_n)$. Moreover we get

$$v(x) - \varphi(x) + \frac{\delta}{2}\langle Ax, x\rangle \le v\left(y_n\left(\frac{1}{n}\right)\right) - \varphi\left(y_n\left(\frac{1}{n}\right)\right) + \frac{\delta}{2}\left\langle Ay_n\left(\frac{1}{n}\right), y_n\left(\frac{1}{n}\right)\right\rangle.$$

From (4.15) and from the above inequality we obtain

$$\begin{aligned}&n\left[\varphi(x) - \varphi\left(y_n\left(\frac{1}{n}\right)\right)\right] + \frac{\delta n}{2}\left[\left\langle Ay_n\left(\frac{1}{n}\right), y_n\left(\frac{1}{n}\right)\right\rangle - \langle Ax, x\rangle\right]\\ &\ge n\int_0^{\frac{1}{n}} e^{-\lambda s} L(A^\beta y_n(s), \gamma_n(s))ds + n(e^{-\frac{\lambda}{n}} - 1)v\left(y_n\left(\frac{1}{n}\right)\right) + \omega\left(\frac{1}{n}\right).\end{aligned} \tag{4.16}$$

Here and in the rest of the proof we denote by $\omega(t)$ a function such that $\omega(t) \downarrow 0$ as $t \downarrow 0$. Similarly to Step I we have

$$\varphi(x) - \varphi\left(y_n\left(\frac{1}{n}\right)\right) = \int_0^{\frac{1}{n}} \langle D\varphi(y_n(s)), Ay_n(s) + A^{-\beta}F(A^\beta y_n(s)) - B\gamma_n(s)\rangle ds \tag{4.17}$$

and

$$\begin{aligned}&\frac{1}{2}\left[\left\langle Ay_n\left(\frac{1}{n}\right), y_n\left(\frac{1}{n}\right)\right\rangle - \langle Ax, x\rangle\right]\\ &= \int_0^{\frac{1}{n}} \langle Ay_n(s), -Ay_n(s) - A^{-\beta}F(A^\beta y_n(s)) + B\gamma_n(s)\rangle ds\\ &= \int_0^{\frac{1}{n}} \left[-|Ay_n(s)|^2 - \langle Ay_n(s), A^{-\beta}F(A^\beta y_n(s))\rangle + \langle Ay_n(s), B\gamma_n(s)\rangle\right] ds.\end{aligned} \tag{4.18}$$

Therefore inequality (4.16) can be rewritten as

$$
\begin{aligned}
&n\int_0^{\frac{1}{n}} \left\langle D\varphi(y_n(s)),\, Ay_n(s) + A^{-\beta}F(A^{\beta}y_n(s)) - B\gamma_n(s)\right\rangle ds \\
&+n\delta\int_0^{\frac{1}{n}} \left[-|Ay_n(s)|^2 - \left\langle Ay_n(s),\, A^{-\beta}F(A^{\beta}y_n(s))\right\rangle + \langle Ay_n(s),\, B\gamma_n(s)\rangle\right] ds \qquad (4.19)\\
&\geq n\int_0^{\frac{1}{n}} e^{-\lambda s}L(A^{\beta}y_n(s),\gamma_n(s))ds + n(e^{-\frac{\lambda}{n}}-1)v\left(y_n\left(\frac{1}{n}\right)\right) + \omega\left(\frac{1}{n}\right).
\end{aligned}
$$

Again, reasoning as in Step I, from the above estimate we derive

$$
n\int_0^{\frac{1}{n}} |Ay_n(s)|^2 ds \leq C_{\delta};
$$

hence, there exists a sequence $\{s_n\}$, $s_n \downarrow 0$, such that

$$
|Ay_n(s_n)| \leq C, \quad y_n(s_n) \to x \quad \text{and} \quad Ay_n(s_n) \rightharpoonup z. \qquad (4.20)
$$

As in Step I we conclude that $x \in D(A)$. Therefore (3.6)(i) holds.

In order to show that (ii) is verified, we have to estimate the terms contained in (4.19). First we note that, by easy computations exploiting estimate (2.13), as $x \in D(A)$

$$
\lim_{n\to\infty} \sup_{0\leq t\leq \frac{1}{n}} |A^{\alpha}(y_n(t) - x)| = 0, \qquad (4.21)
$$

where $\alpha \in [0, 1)$. Since $\varphi \in C^1(D(A^{\frac{1}{2}}))$ and $x \in D(A)$, by (4.21) we obtain

$$
-n\int_0^{\frac{1}{n}} \langle D\varphi(y_n(s)), B\gamma_n(s)\rangle\, ds = -n\int_0^{\frac{1}{n}} \langle D\varphi(x), B\gamma_n(s)\rangle ds + \omega\left(\frac{1}{n}\right). \qquad (4.22)
$$

Moreover, by (4.21)

$$
n\int_0^{\frac{1}{n}} \langle D\varphi(y_n(s)), A^{-\beta}F(A^{\beta}y_n(s))\rangle ds = \langle D\varphi(x), A^{-\beta}F(A^{\beta}x)\rangle + \omega\left(\frac{1}{n}\right). \qquad (4.23)
$$

On the other hand we recall that by assumption (2.5)(v), there exists ρ such that $A^{\rho}B$, is bounded. Hence, (4.21) yields

$$
\begin{aligned}
n\int_0^{\frac{1}{n}} \langle Ay_n(s), B\gamma_n(s)\rangle\, ds &= n\int_0^{\frac{1}{n}} \left\langle A^{1-\rho}x,\, A^{\rho}B\gamma_n(s)\right\rangle ds + \omega\left(\frac{1}{n}\right) \\
&= n\int_0^{\frac{1}{n}} \langle Ax, B\gamma_n(s)\rangle\, ds + \omega\left(\frac{1}{n}\right).
\end{aligned} \qquad (4.24)
$$

In addition, from (4.21) we have

$$
-n\int_0^{\frac{1}{n}} \left\langle Ay_n(s),\, A^{-\beta}F(A^{\beta}y_n(s))\right\rangle ds = -\left\langle Ax,\, A^{-\beta}F(A^{\beta}x)\right\rangle + \omega\left(\frac{1}{n}\right). \qquad (4.25)
$$

Finally, again from (4.21)

$$
n\int_0^{\frac{1}{n}} e^{-\lambda s}L(A^{\beta}y_n(s),\gamma_n(s))ds = n\int_0^{\frac{1}{n}} L(A^{\beta}x,\gamma_n(s))ds + \omega\left(\frac{1}{n}\right). \qquad (4.26)
$$

Since v is continuous, substituting estimates (4.22)–(4.26) into (4.19), we derive

$$\lambda v(x) + n\int_0^{\frac{1}{n}} \left[-\langle D\varphi(x) - \delta Ax, B\gamma_n(s)\rangle - L(A^\beta x, \gamma_n(s))\right] ds + \left\langle A^{-\beta}F(A^\beta x), D\varphi(x)\right\rangle$$

$$+n\int_0^{\frac{1}{n}} \left[\langle D\varphi(y_n(s)), Ay_n(s)\rangle - \delta|Ay_n(s)|^2\right] ds - \delta\left\langle Ax, A^{-\beta}F(A^\beta x)\right\rangle \geq \omega\left(\frac{1}{n}\right).$$

On the other hand, recalling the definition of the Hamiltonian (2.22),

$$n\int_0^{\frac{1}{n}} \left[-\langle D\varphi(x) - \delta Ax, B\gamma_n(s)\rangle - L(A^\beta x, \gamma_n(s))\right] ds \leq H(A^\beta x, D\varphi(x) - \delta Ax).$$

Therefore, for some sequence $\{s_n\}$, $0 \leq s_n \leq \frac{1}{n}$, as in (4.20) it follows that

$$\lambda v(x) + H(A^\beta x, D\varphi(x) - \delta Ax) + \left\langle A^{-\beta}F(A^\beta x), D\varphi(x)\right\rangle - \delta\left\langle Ax, A^{-\beta}F(A^\beta x)\right\rangle$$

$$\geq -\langle D\varphi(y_n(s_n)), Ay_n(s_n)\rangle + \delta|Ay_n(s_n)|^2 + \omega\left(\frac{1}{n}\right).$$

By (4.20), taking the $\liminf_{n\to\infty}$ of the right-hand side of the above inequality, we derive that (3.6)(ii) holds. □

Combining Theorem 4.2 with Theorem 3.2 we obtain the following existence and uniqueness result for the Hamilton–Jacobi equation (4.1).

COROLLARY 4.3. *Assume that* (2.5) *and* (2.4) *hold true and let* $\lambda_F = \min\{1, \frac{\lambda}{K_F}\}$. *Fix*

$$\beta \in \left(\frac{3}{4}, \frac{3-\lambda_F}{4-2\lambda_F}\right). \tag{4.27}$$

Then the value function v *defined in* (2.19) *is the unique viscosity solution of the Hamilton–Jacobi equation* (4.1) *satisfying a Hölder condition with exponent* $\alpha \in (\alpha_\beta, 1)$, *where* α_β *is defined in* (3.7).

Proof. Applying Theorem 4.2, we obtain that v is a viscosity solution of equation (4.1). From Proposition 4.1 it follows that v is Hölder continuous of exponent α for any $0 < \alpha < \lambda_F$. From (4.27) it is easily seen that $\lambda_F > \alpha_\beta$. The proof of existence is thus complete. As for uniqueness we note that assumption (2.4) implies that the Hamiltonian (2.22) satisfies hypothesis (3.2). Therefore uniqueness follows from Theorem 3.2. □

REFERENCES

[1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, Heidelberg, 1976.

[2] V. BARBU AND G. DA PRATO, *Hamilton-Jacobi Equations in Hilbert Spaces*, Pitman, Boston, 1982.

[3] A. BENSOUSSAN, G. DA PRATO, M. C. DELFOUR, AND S. K. MITTER, *Representation and Control of Infinite Dimensional Systems*, Birkhäuser, Boston, 1992.

[4] P. CANNARSA AND H. FRANKOWSKA, *Value function and optimality conditions for semilinear control problems*, Appl. Math. Optim., 26 (1992), pp. 139–169.

[5] P. CANNARSA, F. GOZZI, AND H. M. SONER, *A dynamic programming approach to nonlinear boundary control problems of parabolic type*, J. Funct. Anal., 117 (1993), pp. 25–61.

[6] P. CANNARSA AND M. E. TESSITORE, *Cauchy problem for the dynamic programming equation of boundary control*, in Boundary Control and Variation, Lecture Notes in Pure and Appl. Math. 163, J. P. Zolesio, ed., Marcel Dekker, New York, 1994.

[7] M. G. CRANDALL, L. C. EVANS, AND P. L. LIONS, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[8] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[9] M. G. CRANDALL AND P. L. LIONS, *Hamilton–Jacobi equations in infinite dimensions. Part* I: *Uniqueness of viscosity solutions*, J. Funct. Anal., 62 (1985), pp. 379–396; *Part* II: *Existence of viscosity solutions*, J. Funct. Anal., 65 (1986), pp. 368–405; *Part* III, J. Funct. Anal., 68 (1986), pp. 214–247; *Part* IV: *Hamiltonians with unbounded linear terms*, J. Funct. Anal., 90 (1990), pp. 237–283; *Part* V: *Unbounded linear terms and B–continuous solutions*, J. Funct. Anal., 97/2 (1991), pp. 417–465; *Part* VI: *Nonlinear A and Tataru's Method Refined*, preprint; *Part* VII: *The HJB Equation Is Not Always Satisfied*, preprint.

[10] G. DA PRATO AND A. ICHIKAWA, *Riccati equations with unbounded coefficient*, Ann. Mat. Pura Appl., 140 (1985), pp. 209–221.

[11] H. O. FATTORINI, *Boundary control systems*, SIAM J. Control Optim., 6 (1968), pp. 349–385.

[12] F. FLANDOLI, *Riccati equations arising in a boundary control problem with distributed parameters*, SIAM J. Control Optim., 22 (1984), pp. 76–86.

[13] ———, *A counterexample in the boundary control of parabolic system*, Appl. Math. Lett., 3 (1990), pp. 47–50.

[14] H. ISHII, *Viscosity solutions for a class of Hamilton–Jacobi equations in Hilbert spaces*, J. Funct. Anal., 105 (1992), pp. 301–341.

[15] I. LASIECKA AND R. TRIGGIANI, *Differential and Algebraic Riccati Equations with Application to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*, Lecture Notes in Control and Information Sci. 164, Springer-Verlag, Berlin, 1991.

[16] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications* II, Dunod, Paris, 1968.

[17] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, Heidelberg, Berlin, 1983.

[18] H. M. SONER, *On the Hamilton–Jacobi–Bellman equations in Banach spaces*, J. Optim. Theory Appl., 57 (1988), pp. 345–392.

[19] D. TATARU, *Viscosity solutions of Hamilton–Jacobi equations with unbounded linear terms*, J. Math. Anal. Appl., 163 (1992), pp. 345–392.

[20] ———, *Viscosity solutions for the dynamic programming equation*, Appl. Math. Optim., 25 (1992), pp. 109–126.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 1848–1873, November 1996

002

VALUE ITERATION IN A CLASS OF COMMUNICATING MARKOV DECISION CHAINS WITH THE AVERAGE COST CRITERION*

ROLANDO CAVAZOS-CADENA†

Abstract. Markov decision processes with denumerable state space and discrete time parameter are considered. The performance index of a control policy is the (lim sup expected) *average cost criterion*, and the the main structural restrictions on the model are the following: (i) under the action of any stationary policy, the state space is a communicating class; (ii) the cost function has an almost monotone—or penalized—structure [V. S. Borkar, *SIAM J. Control Optim.*, 21 (1983), pp. 652–666; 22 (1984), pp. 965–978]; and (iii) some stationary policy induces an ergodic chain with finite average cost. In this context it is shown that the *value iteration* scheme can be used to construct convergent approximations of a solution to the optimality equation, as well as a sequence of stationary policies whose limit points are optimal.

Key words. Markov decision chains, average cost criterion, almost monotone cost function, value iteration scheme, pointwise convergence

AMS subject classifications. Primary, 90C40, 93E20; Secondary, 60J05

1. Introduction.

This paper concerns Markov decision processes with denumerable state space and discrete time parameter. The performance of a control strategy is measured by the (lim sup expected) *average cost criterion*, and in addition to standard continuity–compactness conditions, the class of models under consideration is essentially determined by the following restrictions: (i) under the action of each stationary policy, the state space is a communicating class; (ii) some stationary policy induces an ergodic chain with finite average cost; and (iii) the cost function is "almost monotone" in the following sense: outside a finite set of states, the cost is sufficiently large. These restrictions are basically those used by Borkar in [1, 2] to establish the existence of optimal stationary policies and a solution to the *average cost optimality equation* (ACOE) (see also [3, 4, 29]). However, in contrast to the context in [1, 2], the cost function is allowed to depend on the actions. Within this framework *the main objective of this paper* is to show that the *value iteration* (VI) scheme can be used to produce

(1) convergent approximations of a solution to the ACOE,

(2) a sequence of stationary policies whose limit points are *average optimal.*

The main result in this note, stated in Theorem 3.1 below, provides a solution to these problems. Since the VI method has been studied extensively in the literature, it is convenient to begin by pointing out the main differences between the results in this note and other theorems already available. First, the VI scheme has been widely studied under strong stability assumptions on the transition structure of the model—such as the diverse variants of the simultaneous Doeblin condition (SDC) [28]—under which the relative value functions and differential costs produced by the VI scheme are shown to converge, at a geometric rate, to a solution of the ACOE whenever the cost function is *bounded.* However, SDC is quite restrictive and is seldom satisfied, except for models with *finite state space* [12, 14, 20, 26, 30], so the results obtained within such a framework cannot be used in many important applications. On the other hand, Hordjik, Schweitzer, and Tijms obtained convergence results for the VI method in [17] under the so-called *Lyapunov function condition* (LFC) [16, 28], a stability assumption that is substantially weaker than SDC and allows unbounded costs but implies that the Markov chain induced by each stationary policy is ergodic, a feature that is frequently absent in simple

*Received by the editors January 30, 1991; accepted for publication (in revised form) June 14, 1995. This research was supported in part by PSF Organization grant 160-350/90–94–2 and MAXTOR Foundation for Applied Probability and Statistics (MAXFAPS) grant 01–01–56/2–94.

†Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro, Buenavista, Saltillo, COAH, 25315, México.

but interesting applications, as in Example 9.1. The present paper is more on the line of the work by Sennott [24], where the assumptions on the recurrence structure of the model are quite general. The approach used in [24] is, essentially, based on the results in [17]. In the latter paper it was postulated that "the first error function" produced by the VI scheme is bounded—an assumption that is very strong and is closely related to SDC [7]—and such a restriction was translated to the context in [24] as Assumption 5, which, generally, is very difficult to verify; the techniques in [24] were extended to more general contexts in [19]. On the other hand, this note extends ideas recently used in [7, 10], where LFC was assumed but the condition on the boundedness of the first error function was avoided. Roughly, the usual way to study the VI scheme consists of analyzing the differences $V_n(\cdot) - (n+1)g$, where V_n is the nth VI function and g is the optimal average cost. Here, following [7, 10], the analysis relies on a direct study of the differential cost function $V_n - V_{n-1}$, and this allows one to develop an approach to solve problems (1) and (2) under, essentially, *nothing more* but the conditions in [1, 2]; however, it should be mentioned that, unfortunately, the required argumentation becomes substantially more complicated within this framework.

The remainder of the paper is organized as follows: in §2 the decision model is introduced, whereas the VI scheme is presented in §3, with the main result being stated in the form of Theorem 3.1. The necessary technical preliminaries to establish this theorem have been divided into four parts presented in §§4–7, whereas the proof of Theorem 3.1 is given in §8, with an example presented in §9. Finally, the paper concludes with some brief comments in §10.

Notation. $\mathbb{R}$ denotes the set of real numbers, and $\mathbb{N}$ stands for the set of nonnegative integers. If $a, b \in \mathbb{R}$, set $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. On the other hand, a Cartesian product of topological spaces is always endowed with the corresponding product topology, and for topological spaces A and $\mathbb{K}$, the set of all transition kernels on A given $\mathbb{K}$ is denoted by $\mathbb{P}(A|\mathbb{K})$, i.e., $\pi(\cdot|\cdot) \in \mathbb{P}(A|\mathbb{K})$, if, for each $k \in \mathbb{K}$, $\pi(\cdot|k)$ is a probability measure on the Borel subsets of A, and for each Borel subset B of A, the mapping $k \mapsto \pi(B|k)$, $k \in \mathbb{K}$, is measurable. Finally, for an event W, the corresponding indicator function is denoted by $I[W]$.

2. Decision model. Let (S, A, C, p) be a Markov decision process (MDP) where the *state space* S is a denumerable set endowed with the discrete topology and the *action set* A is a compact metric space. On the other hand, C is the cost function and p is the transition law. This model represents a dynamical system evolving as follows: at each time $t \in \mathbb{N}$ the state of the system is observed—say, $X_t = x \in S$—and an action $A_t = a \in A$ is chosen. As a consequence, (i) a cost $C(x, a)$ is incurred, and (ii) regardless of the states observed and actions applied before t, the state of the system at time $t+1$ will be $X_{t+1} = y \in S$ with probability $p_{xy}(a)$; this is the Markov property of the process. Note that it is assumed that all actions in A are available at each state x; as noted by Borkar in [2], this assumption does not imply any loss of generality.

Assumption 2.1. (i) (Continuity.) For each $x, y \in S$, the mappings $a \mapsto C(x, a)$ and $a \mapsto p_{xy}(a)$ are continuous on A.

(ii) The cost function is nonnegative: $C(x, a) \geq 0$, $x \in S$, $a \in A$.

Policies. Let $t \in \mathbb{N}$ be arbitrary. The space of state–action histories up to time t is denoted by H_t and is given by $H_0 := S$ and $H_t := H_{t-1} \times (A \times S)$, $t \geq 1$, whereas $h_t := (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t)$ stands for a generic element of H_t. Define the information vector up to time t as follows:

$$I_0 := X_0 \quad \text{and} \quad I_t := (X_0, A_0, \ldots, X_{t-1}, A_{t-1}, X_t), \quad t \geq 1. \tag{2.1}$$

A *policy* is a sequence $\pi = \{\pi_t \in \mathbb{P}(A|H_t) | t \in \mathbb{N}\}$; when the system is in progress under π and $I_t = h_t$ has been observed, $\pi_t(B|h_t)$ is the probability of choosing action A_t within (the

Borel subset) $B \subset A$. The space of all policies is denoted by $\mathbb{P}$. Now set $\mathbb{F} := \Pi_{x \in S} A$—i.e., $\mathbb{F}$ consists of all functions $f : S \to A$—and note that $\mathbb{F}$ is a compact metric space in the product topology [11]. A policy π is stationary if there exists $f \in \mathbb{F}$ such that $\pi_t(\{f(x_t)\}|h_t) = 1$ for all $t \in \mathbb{N}$ and $h_t \in H_t$. Next, define $\mathbb{M} := \Pi_{t \in \mathbb{N}} \mathbb{F}$, which is also a compact metric space and consists of all sequences $\{f_t | t \in \mathbb{N}\}$ of members of $\mathbb{F}$. A policy π is Markovian if there exists $\{f_t\} \in \mathbb{M}$ such that $\pi_t(\{f_t(x_t)\}|h_t) = 1$ is always valid. The class of stationary (resp., Markovian) policies is naturally identified with $\mathbb{F}$ (resp., $\mathbb{M}$) and, with these conventions, $\mathbb{F} \subset \mathbb{M} \subset \mathbb{P}$.

On the other hand, given the initial state $X_0 = x$ and the policy $\pi \in \mathbb{P}$ being used, the distribution of the state–action process $\{(X_t, A_t)\}$ is uniquely determined via the Ionescu–Tulcea theorem; see [14, 15, 20] for details. This distribution is denoted by P_x^π, whereas E_x^π stands for the corresponding expectation operator. It is known that under the action of any policy $f \in \mathbb{F}$ the state process $\{X_t\}$ is a Markov chain with stationary transition mechanism [14, 20, 22].

Assumption 2.2 (communication). Under the action of any policy $f \in \mathbb{F}$ the state space is a communicating class. More explicitly, given $x, y \in S$ and $f \in \mathbb{F}$, there exists a positive integer $m = m(x, y, f)$ such that $P_x^f[X_m = y] > 0$.

Optimality criterion. The (lim sup expected) *average cost* at state $x \in S$ under policy $\pi \in \mathbb{P}$ is defined by

$$J(x, \pi) := \limsup_{n \to \infty} \frac{1}{n+1} E_x^\pi \left[\sum_{t=0}^{n} C(X_n, A_n) \right], \tag{2.2}$$

while

$$J(x) := \inf_{\pi \in \mathbb{P}} J(x, \pi) \tag{2.3}$$

is the *optimal average cost* at state x. A policy π is *average optimal* (AO) if $J(x, \pi) = J(x)$ for all $x \in S$.

Since $C \geq 0$, the expectation in (2.2) is well defined, but under Assumptions 2.1 and 2.2 alone, $J(\cdot)$ may be identically infinite. Assumption 2.3 below avoids this situation and, in addition, allows one to establish the existence of a solution to the ACOE yielding AO stationary policies. This restriction is a version of a condition introduced by Borkar in [1, 2] for cost functions depending only on the state; see also [3, 4], where the latter restriction is avoided.

Assumption 2.3. (i) (Almost monotone (penalized) costs.) For each $b > 0$ there exists a *finite* set $G(b) \subset S$ such that

$$C(x, a) > b \quad \text{for all} \quad (x, a) \in (S \setminus G(b)) \times A.$$

(ii) (Finite average cost.) There exists a policy $f \in \mathbb{F}$ such that $J(x, f) < \infty$ for all $x \in S$.

The following basic result is the most important consequence of Assumptions 2.1–2.3. Throughout the remainder of the paper, $z \in S$ is a *fixed state*.

LEMMA 2.1. *There exist* $g \in \mathbb{R}$ *and* $h : S \to \mathbb{R}$ *satisfying* (i)–(iv) *below.*

(i) $J(x) = g$ *for all* $x \in S$.

(ii) h *is bounded below and* $h(z) = 0$.

(iii) *The ACOE holds:*

$$g + h(x) = \inf_{a \in A} \left[C(x, a) + \sum_y p_{xy}(a) h(y) \right], \quad x \in S. \tag{2.4}$$

(iv) *An optimal stationary policy exists. Indeed, for each* $x \in S$, *the term in brackets in* (2.4) *has a minimizer* $f^*(x) \in A$ *and the corresponding policy* f^* *is optimal.*

This lemma was essentially obtained in [1–4]. A short proof is given, since it plays a basic role in this note.

Proof of Lemma 4.1. From Assumptions 2.1–2.3 it can be shown that conditions 1–3 and 3* in [23] hold true; see [6] or Remark 6.3ii in [5]. Thus, the conclusion follows from Sennott's results in [23]. □

Note that g in Lemma 2.1 is *the optimal average* cost at each state, so it is uniquely determined. As will be shown in §4, a function h satisfying (2.4) is also unique if it is bounded below and is normalized by the condition $h(z) = 0$.

As already mentioned, the main objective of the paper is to show that the value iteration scheme can be used to produce convergent approximations of the pair $(g, h(\cdot))$ in (2.4). The result in this direction is stated in the next section as Theorem 3.1 and requires the next additional condition.

Assumption 2.4. For all $x \in S$ and $a \in A$, $p_{xx}(a) > 0$.

Throughout the remainder of this paper, Assumptions 2.1–2.4 are supposed to hold true, even without explicit reference. On the other hand, it is interesting to observe that Assumption 2.4 will be used only in one place, namely, in the proof of Theorem 7.1(i) in §7.

Remark 2.1. The restriction imposed in Assumption 2.4 does not imply any loss of generality. In fact, assume that $M = (S, A, C, p)$ satisfies Assumptions 2.1–2.3, and define a new transition law p^* as follows: for all $x, y \in S$ and $a \in A$,

$$p^*_{xy}(a) := \alpha\delta_{xy} + (1-\alpha)p_{xy}(a),$$

where $\alpha \in (0, 1)$ is a given number and $\delta_{xy} := 0$ (resp., 1) if $x \neq y$ (resp., $x = y$); the transformation $p \mapsto p^*$ was introduced by Schweitzer in [27]. Now set $M^* = (S, A, C, p^*)$, which clearly satisfies Assumptions 2.1 and 2.4; note that $p^*_{xx}(a)$ is always $\geq \alpha$. Now let $P^{*\pi}_x$ and $J^*(x, \pi)$ be associated with model M^* in the same way as P^π_x and $J(x, \pi)$ are related to M. In this case it is not difficult to verify that for all $x, y \in S$, $f \in \mathbb{F}$, and $m \in \mathbb{N}$, $P^{*f}_x[X_m = y] \geq (1-\alpha)^m P^f_x[X_m = y]$ and $J^*(x, f) = J(x, f)$, and these facts immediately yield that M^* also satisfies Assumptions 2.2 and 2.3. In short, if M satisfies Assumptions 2.1–2.3, then the modified model M^* satisfies Assumptions 2.1–2.4. Furthermore, M and M^* are *equivalent* MDPs in the following sense: let (g^*, h^*) be the pair given in Lemma 2.1 applied to M^*. Then (i) $g^* = g$; (ii) $h^* = h/\alpha$; and (iii) a policy $f^* \in \mathbb{F}$ is obtained as in Lemma 2.1(iv) applied to model M if and only if f^* satisfies Lemma 2.1(iv) applied to model M^*; see [20, pp. 371–372].

This section concludes with some remarks about Markov chains that will be used later; details can be found, for instance, in [18, Chap. 1]. First, recall that a function $\mu_f : S \to \mathbb{R}$ is an *invariant* distribution of the Markov chain induced by a stationary policy $f \in \mathbb{F}$ or, simply, of the transition matrix $P_f \equiv [P_f(x, y)] := [p_{xy}(f(x))]$ if

$$\text{(2.5)} \qquad \mu_f(y) \geq 0, \quad \sum_x \mu_f(x) = 1, \quad \text{and} \quad \mu_f(y) = \sum_x \mu_f(x) p_{xy}(f(x)), \quad y \in S.$$

Also, by Assumption 2.2, if the Markov chain induced by f has an invariant distribution, then it is *unique* [18, pp. 39–42] and it follows that

$$\text{(2.6)} \qquad \begin{aligned} J(x, f) &= \lim_{n\to\infty} \frac{1}{n+1} E^f_x \left[\sum_{t=0}^{n} C(X_t, A_t) \right] \\ &= \sum_y \mu_f(y) C(y, f(y)), \quad x \in S. \end{aligned}$$

Next, for each $x \in S$ define the first passage time T_x by

$$T_x := \min\{n > 0 | X_n = x\}, \tag{2.7}$$

where, as usual, the minimum of the empty set is ∞. By notational convenience, the time of the first arrival to the distinguished state z in a positive time is simply written as T, i.e.,

$$T_z \equiv T. \tag{2.8}$$

Remark 2.2. Let $f \in \mathbb{F}$ be fixed.

(i) P_f has an invariant distribution if and only if $E_y^f[T_y] < \infty$ for some state y; in this case Assumption 2.2 yields [18] that

(2.9)

$$E_x^f[T_x] < \infty \quad \text{and} \quad \frac{1}{E_x^f[T_x]} = \lim_{n\to\infty} \frac{1}{n+1} \sum_{t=0}^{n} P_y^f[X_t = x] = \mu_f(x) > 0 \quad \text{for all } x, y \in S.$$

(ii) If the Markov chain induced by f has an invariant distribution, then successive visits to a fixed state y determine a (possibly delayed) renewal process. In this case, if $J(x, f) < \infty$,

$$\begin{aligned} \infty > J(x, f) &= \lim_{n\to\infty} \frac{1}{n+1} E_x^f \left[\sum_{t=0}^{n} C(X_t, A_t) \right] \\ &= \frac{E_y^f[\sum_{n=0}^{T_y - 1} C(X_t, A_t)]}{E_y^f[T_y]} \\ &= \frac{E_z^f[\sum_{n=0}^{T-1} C(X_t, A_t)]}{E_z^f[T]}; \end{aligned} \tag{2.10}$$

see (2.7) and (2.8). These equalities follow from the theory of (delayed) renewal reward processes as presented in [22].

(iii) Suppose that P_f has an invariant distribution and that $J(\cdot, f)$ is finite. In this case (2.6) holds and, by Assumption 2.2, there is a positive probability of reaching a given state $y \in S$ between successive visits to state z, and it follows that

$$S = \bigcup_{k=0}^{\infty} S_k,$$

where

$$S_0 := \{z\} \quad \text{and} \quad S_k := \{y \in S | P_z^f[X_k = y, T > k] > 0\}.$$

The following fact, whose proof can be seen in [23, Prop. 4], will be useful:

$$E_y^f \left[\sum_{t=0}^{T-1} C(X_t, A_t) \right] < \infty, \quad y \in S; \tag{2.11}$$

see (2.7) and (2.8). This is true for arbitrary (nonnegative) cost function, and setting $C(\cdot, \cdot) \equiv 1$ it follows that

$$E_y^f[T] < \infty, \quad y \in S. \tag{2.12}$$

3. Main result. In this section the main result of this note is stated as Theorem 3.1 below. First, the necessary notions are introduced.

DEFINITION 3.1 (the VI scheme). *The sequence* $\{V_k : S \to \mathbb{R} | k = -1, 0, 1, 2, \ldots\}$ *of VI functions is recursively defined as follows:* $V_{-1} \equiv 0$ *and, for* $k \geq 0$,

$$V_k(x) := \inf_{a \in A} \left[C(x, a) + \sum_y p_{xy}(a) V_{k-1}(y) \right], \quad x \in S.$$

It is known that, for each $k \in \mathbb{N}$, there exists a policy π^k such that [14, 20, 22, $\cdots$] for all $x \in S$,

$$\begin{aligned} V_k(x) &= E_x^{\pi^k} \left[\sum_{t=0}^{k} C(X_t, A_t) \right] \\ &= \min_{\pi \in \mathbb{P}} E_\pi \left[\sum_{t=0}^{k} C(X_t, A_t) \right]. \end{aligned} \tag{3.1}$$

Also, since the cost function is nonnegative,

$$0 \leq V_n(x) \leq V_{n+k}(x), \quad x \in S, \ n, k \in \mathbb{N}. \tag{3.2}$$

On the other hand, observe that if $f \in \mathbb{F}$ is as in Assumption 2.3(ii), then $J(x, f) < \infty$ immediately implies that $\infty > E_x^f[\sum_{t=0}^n C(X_t, A_t)] \geq V_n(x)$ for all state x and $n \in \mathbb{N}$, so the VI functions are always finite.

DEFINITION 3.2. (i) *The relative value functions* $\{R_n : S \to \mathbb{R}\}$ *are defined by*

$$R_n(x) := V_n(x) - V_n(z), \quad x \in S, \quad n = -1, 0, 1, \ldots.$$

(ii) *The nth differential cost at state* $x \in S$ *is given by* $g_n(x) := V_n(x) - V_{n-1}(x)$, $n \in \mathbb{N}$.

The main result in this note is the following.

THEOREM 3.1. *Suppose that Assumptions* 2.1–2.4 *hold true, and let* g *and* $h(\cdot)$ *be as in Lemma* 4.1. *Then*

(i) *for all* $x \in S$, $\lim_{n \to \infty} g_n(x) = g$;

(ii) *for each* $x \in S$, $\lim_{n \to \infty} R_n(x) = h(x)$.

Furthermore,

(iii) *for each* $n \in \mathbb{N}$, *there exists a policy* $f_n \in \mathbb{F}$ *such that, for each* $x \in S$, $f_n(x)$ *is a minimizer of the mapping* $a \mapsto C(x, a) + \sum_y R_n(y) p_{xy}(f_n(a))$, $a \in A$. *Moreover, every limit point of* $\{f_n\}$ *is AO.*

We have not been able to find a simple and direct proof of this result. Indeed, the method used below to establish Theorem 3.1 is somewhat technical and consists of four steps, which are presented in §§4–7.

Remark 3.1. Theorem 3.1 is an extension of results in [8, 9, 25]. In these papers it was shown that, as $n \to \infty$, the sequence $\{(g_n(z), R_n(\cdot)\}$ converges to $(g, h(\cdot))$ in the Cesàro sense, i.e.,

$$\lim_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} g_n(z) = g \tag{3.3}$$

and

$$\lim_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} R_n(x) = h(x), \quad x \in S.$$

These convergences were obtained in [8] under the Lyapunov function condition, in [9] under conditions slightly stronger than Assumptions 2.1–2.3, and in [25] under conditions involving the behavior of the optimal *discounted* value function as the discount factor increases to 1. The assumptions in [25] are satisfied within the framework described in §2; see, for instance, [6]. On the other hand, (3.3) was obtained by Gosh and Marcus in [13] as a tool to establish the existence of strong average optimal stationary policies under Assumptions 2.1–2.3.

4. Preliminaries: First part. The starting point on the way to the proof of Theorem 3.1 is the following result, which establishes uniqueness of the function $h(\cdot)$ in Lemma 2.1.

THEOREM 4.1. *Let* $h_1, h_2 : S \to \mathbb{R}$ *satisfy* (a)–(c) *below.*

(a) h_1 *and* h_2 *are* bounded below.

(b) $h_1(z) = h_2(z) = 0$.

(c) h_1 *and* h_2 *satisfy the ACOE, i.e.,*

$$g + h_i(x) = \min_{a \in A}\left[C(x,a) + \sum_y p_{xy}(a)h_i(y)\right], \quad x \in S, \quad i = 1, 2. \tag{4.1}$$

Then $h_1(\cdot) = h_2(\cdot)$.

The proof of this theorem is based on the following lemma, which is the most important technical tool in this note.

LEMMA 4.1. *Let* $W : S \to \mathbb{R}$ *be a* bounded below *function satisfying* $W(z) = 0$, *and suppose that the stationary policy* $f \in \mathbb{F}$ *is such that*

$$g + W(x) \geq C(x, f(x)) + \sum_y p_{xy}(f(x))W(y), \quad x \in S. \tag{4.2}$$

Then assertions (i)–(vi) *below occur.*

(i) $g = J(x, f)$, $x \in S$, *i.e., f is AO.*

(ii) *Every state x is positive recurrent* [18] *with respect to* P_f. *Hence* P_f *has an invariant distribution* μ_f; *see* (2.5).

(iii) *For all* $x \in S$,

$$\frac{1}{n+1}E_x^f\left[\sum_{t=0}^{n} W(X_{n+1})\right] \to 0 \quad as \quad n \to \infty.$$

(iv) *For all* $a \in A$ *and* $x \in S$,

$$g + W(x) \leq C(x,a) + \sum_y p_{xy}(a)W(y); \tag{4.3}$$

in particular, equality holds in (4.2) *and the pair* $(g, W(\cdot))$ *satisfies the ACOE:*

$$g + W(x) = \min_{a \in A}\left[C(x,a) + \sum_y p_{xy}(a)W(y)\right], \quad x \in S.$$

(v) *Let* $D : S \to [0, \infty)$ *be such that* (a) $D(z) = 0$ *and* (b) $D(x) = E_x^f[D(X_1)]$, $x \in S$. *Then*

$$D(x) = 0 \quad \textit{for all state } x.$$

(vi) $W(x) = E_x^f[\sum_{t=0}^{T-1}(C(X_t, A_t) - g)]$ *for all* $x \in S$ (*see* (2.7) *and* (2.8)).

Proof. To begin, note that, since $W(\cdot)$ is bounded below,

$$\liminf_{n\to\infty} \frac{E_x^f[W(X_{n+1})]}{n+1} \geq 0. \tag{4.4}$$

(i) A simple induction argument using (4.2) shows that for all $x \in S$ and $n \in \mathbb{N}$,

$$g + \frac{1}{n+1}W(x) \geq \frac{1}{n+1}E_x^f\left[\sum_{t=0}^{n} C(X_t, A_t)\right] + \frac{1}{n+1}E_x^f[W(X_{n+1})], \tag{4.5}$$

and taking limit superior in both sides of this inequality, it follows, via (4.4), that for all $x \in S$

$$g \geq \limsup_{n\to\infty} \frac{1}{n+1}E_x^f\left[\sum_{t=0}^{n} C(X_t, A_t)\right] = J(x, f),$$

which yields that $g = J(x, f)$, $x \in S$, since g is the optimal average cost at every state; see (2.2), (2.3), and Lemma 2.1(i).

(ii) By Assumption 2.2, it is sufficient to prove that *some* state is positive recurrent [18] with respect to P_f. This fact will be established by contradiction. Thus, *suppose that*

every state is transient or null recurrent with respect to P_f. (4.6)

Under this assumption, for each *finite* set $G \subset S$ [18],

$$\lim_{n\to\infty} \frac{1}{n+1}E_x^f\left[\sum_{t=0}^{n} I[X_t \in G]\right] = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{1}{n+1}E_x^f\left[\sum_{t=0}^{n} I[X_t \in S \setminus G]\right] = 1. \tag{4.7}$$

Combining these relations with (2.2) it follows immediately that

$$J(x, f) = \limsup_{n\to\infty} \frac{1}{n+1}E_x^f\left[\sum_{t=0}^{n} C(X_t, A_t) I[X_t \in S \setminus G]\right]. \tag{4.8}$$

To conclude, choose $G = G(g+1)$, the finite set in Assumption 2.3(i) with $b = g + 1$. In this case $C(X_t, A_t)I[X_t \in S \setminus G] \geq (g+1)I[X_t \in S \setminus G]$, and (4.7) and (4.8) together yield that $J(x, f) \geq g + 1$, which contradicts part (i). In short, (4.6) leads to a contradiction, and this establishes part (ii).

(iii) Since the Markov chain induced by f has an invariant distribution, it follows that

$$g = J(x, f) = \lim_{n\to\infty} \frac{1}{n+1}E_x^f\left[\sum_{t=0}^{n} C(X_t, A_t)\right];$$

see (2.6). This immediately yields, in combination with (4.5), that

$$\begin{aligned} g &\geq \limsup_{n\to\infty} \left\{\frac{1}{n+1}E_x^f\left[\sum_{t=0}^{n} C(X_t, A_t)\right] + \frac{1}{n+1}E_x^f[W(X_{n+1})]\right\} \\ &= g + \limsup_{n\to\infty} \frac{1}{n+1}E_x^f[W(X_{n+1})], \end{aligned}$$

so

$$\limsup_{n} \frac{1}{n+1}E_x^f[W(X_{n+1})] \leq 0,$$

and the desired conclusion follows from combining this inequality with (4.4).

(iv) The proof is by contradiction. Let $(\tilde{x}, \tilde{a}) \in S \times A$ be arbitrary but fixed, and *suppose* that

$$\Delta := g + W(\tilde{x}) - C(\tilde{x}, \tilde{a}) - \sum_y p_{\tilde{x}y}(\tilde{a})W(y) > 0. \tag{4.9}$$

Define a new policy $\tilde{f} \in \mathbb{F}$ and a function $\psi : S \to \mathbb{R}$ by

$$\tilde{f}(x) = \begin{cases} f(x) & \text{if } x \neq \tilde{x}, \\ \tilde{a} & \text{if } x = \tilde{x} \end{cases}$$

and

$$\psi(x) = \begin{cases} \Delta & \text{if } x = \tilde{x}, \\ 0 & \text{if } x \neq \tilde{x}. \end{cases} \tag{4.10}$$

Combining these definitions with (4.2) and (4.9) it follows that for all $x \in S$,

$$\begin{aligned} g + W(x) &\geq C(x, \tilde{f}(x)) + \psi(x) + \sum_y P_{\tilde{x}y}(\tilde{f}(x))W(y) \\ &\geq C(x, \tilde{f}(x)) + \sum_y P_{\tilde{x}y}(\tilde{f}(x))W(y) \end{aligned}$$

so that, by parts (i) and (ii), $\tilde{f}$ is AO and the Markov chain induced by $\tilde{f}$ has an invariant distribution $\mu_{\tilde{f}}$. Next observe that for all $x \in S$ and $n \in \mathbb{N}$ (see the proof of part (i)),

$$g + \frac{1}{n+1}W(x) \geq \frac{1}{n+1}E_x^{\tilde{f}}\left[\sum_{t=0}^{n}(C(X_t, A_t) + \psi(X_t))\right] + \frac{1}{n+1}E_x^{\tilde{f}}[W(X_{n+1})]. \tag{4.11}$$

By (2.6) with $f = \tilde{f}$, as $n \to \infty$,

$$\frac{1}{n+1}E_x^{\tilde{f}}\left[\sum_{t=0}^{n} C(X_t, A_t)\right] \to \sum_y \mu_{\tilde{f}}(y)C(y, \tilde{f}(y)) = J(x, \tilde{f})$$

and similarly (see (2.9) and (4.10))

$$\frac{1}{n+1}E_x^{\tilde{f}}\left[\sum_{t=0}^{n} \psi(X_t)\right] \to \sum_y \mu_{\tilde{f}}(y)\psi(y) = \mu_{\tilde{f}}(\tilde{x})\Delta.$$

Using these convergences and part (iii) with $\tilde{f}$ instead of f, it follows, after letting n increase to ∞ in (4.11), that

$$g \geq J(x, \tilde{f}) + \mu_{\tilde{f}}(\tilde{x})\Delta > J(x, \tilde{f}),$$

where the strict inequality follows from (2.9) and (4.9). However, this contradicts the equality $g = J(x, \tilde{f})$, and it follows that Δ in (4.9) is ≤ 0, establishing (4.3), since the pair $(\tilde{x}, \tilde{a}) \in S \times A$ was arbitrary. Clearly, (4.3) implies that equality holds in (4.2), and, finally, (4.2) and (4.3) together yield that the pair $(g, W(\cdot))$ satisfies the ACOE.

(v) It is sufficient to prove that $D(y) \leq 0$ for $y \neq z$ since, by assumption, D is a nonnegative function and $D(z) = 0$. Note that $D(x) = E_x^f[D(X_1)]$ is equivalent to $D(x) = E_x^f[D(X_1)I[T > 1]]$; this is due to the condition $D(z) = 0$ and the definition of T (see (2.7) and (2.8)). Then, from a simple induction argument using the Markov property, it follows that

$$D(x) = E_x^f[D(X_n)I[T > n]], \quad x \in S, \ n \in \mathbb{N}.$$

Next, set $x = z$ to obtain

$$0 = D(z) = E_z^f[D(X_n)I[T > n]] \geq D(y)P_z^f[X_n = y, T > n], \quad n \in \mathbb{N}, \tag{4.12}$$

where $y \neq z$ is an arbitrary state and the nonnegativity of $D(\cdot)$ was used to obtain the inequality. Since there exists a positive integer n with $P_z^f[X_n = y, T > n] > 0$ (see Remark 2.2(iii)), (4.12) yields that $0 \geq D(y)$, and as already mentioned, this establishes part (v), since $y \in S \setminus \{z\}$ was arbitrary.

(vi) Recalling that equality holds in (4.2) and using that $W(z) = 0$, it follows that for each $x \in S$

$$\begin{aligned} W(x) &= C(x, f(x)) - g + \sum_{y \neq z} p_{xy}(f(x))W(y) \\ &= E_x^f[(C(X_0, A_0) - g)I[T > 0] + W(X_1)I[T > 1]]; \end{aligned} \tag{4.13}$$

note that $T = T_z$ is always ≥ 1 by (2.7) and (2.8). This implies, via an induction argument, that

$$W(x) = E_x^f\left[\sum_{t=0}^{n}(C(X_t, A_t) - g)I[T > t] + W(X_{n+1})I[T > n+1]\right], \quad x \in S,\ n \in \mathbb{N}. \tag{4.14}$$

Next observe that, since P_f has an invariant distribution and $J(\cdot, f) \equiv g(< \infty)$ (by parts (i) and (ii)),

(a) (2.11), (2.12), and the dominated converge theorem together imply that

$$\begin{aligned} \lim_{n\to\infty} E_x^f\left[\sum_{t=0}^{n}(CX_t, A_t) - g)I[T > t]\right] &= E_x^f\left[\sum_{t=0}^{\infty}(C_t, A_t) - g)I[T > t]\right] \\ &= E_x^f\left[\sum_{t=0}^{T-1}(C_t, A_t) - g)\right]; \end{aligned}$$

(b) $E_x^f[W(X_n)I[T > n]] \geq E_x^f[bI[T > n]] = bP_x^f[T > n] \to bP_x^f[T = \infty] = 0$ as $n \to \infty$, where b is a lower bound of $W(\cdot)$, and the last equality is a consequence of (2.12). Then, letting n increase to ∞ in (4.14) and using (a) and (b) it follows that

$$W(x) \geq E_x^f\left[\sum_{t=0}^{\infty}(C(X_t, A_t) - g)I[T > t]\right] = E_x^f\left[\sum_{t=0}^{T-1}(C(X_t, A_t) - g)\right] =: M(x). \tag{4.15}$$

On the other hand, the Markov property and the definition of $M(\cdot)$ together yield that

$$\begin{aligned} M(x) &= C(x, f(x)) - g + \sum_{y \neq z} p_{xy}(f(x))M(y) \\ &= E_x^f[(C(X_0, A_0) - g)I[T > 0] + M(X_1)I[T > 1]], \end{aligned}$$

which combined with (4.13) implies that

$$W(x) - M(x) = E_x^f[(W(X_1) - M(X_1))I[T_z > 1]], \quad x \in S. \tag{4.16}$$

To conclude set $D(x) := W(x) - M(x)$, $x \in S$, and observe that $D(\cdot) \geq 0$ by (4.15). Now recall that $J(\cdot, f) = g$, so (4.15) and the third equality in (2.10) together yield that $M(z) = 0$. Since $W(z) = 0$ (by assumption), it follows that $D(z) = W(z) - M(z) = 0$. Then (4.16) can be written as

$$D(x) = E_x^f[D(X_1)], \quad x \in S,$$

and from an application of part (v) it follows that $W(x) - M(x) = D(x) = 0$ for all $x \in S$, and the conclusion follows from combining this with the definition of $M(\cdot)$ in (4.15). □

This lemma will be now used to establish Theorem 4.1.

Proof of Theorem 4.1. Let $x \in S$ be arbitrary. Since h_i is bounded below, Assumption 2.1 yields that the mapping $a \mapsto C(x, a) + \sum_y p_{xy}(a)h_i(y)$ is lower semicontinuous and then has a minimizer $f_i(x)$, since A is compact [11], so

$$g + h_i(x) = C(x, f_i(x)) + \sum_y p_{xy}(f_i(x))h_i(y), \quad i = 1, 2 \quad x \in S. \tag{4.17}$$

Then, for $i = 1, 2$, f_i is AO and has an invariant distribution μ_{f_i}, by parts (i) and (ii) of Lemma 4.1. Now define $\hat{h} : S \to \mathbb{R}$ by

$$\hat{h}(x) := h_1(x) \wedge h_2(x), \quad x \in S; \tag{4.18}$$

note that $\hat{h}(\cdot)$ is bounded below and that (4.17) implies that for all $x \in S$,

$$g + h_i(x) \geq C(x, a) + \sum_y p_{xy}(f_i(x))\hat{h}(y), \quad i = 1, 2.$$

Then

$$g + \hat{h}(x) \geq C(x, \hat{f}(x)) + \sum_y p_{xy}(\hat{f}(x))\hat{h}(y), \tag{4.19}$$

where the policy $\hat{f} \in \mathbb{F}$ is given by

$$\hat{f} := \begin{cases} f_1(x) & \text{if } h_1(x) \leq h_2(x), \\ f_2(x) & \text{if } h_2(x) < h_1(x). \end{cases}$$

Using Lemma 4.1(iv) with $W = \hat{h}$ and $\hat{f}$ instead of f, (4.19) implies that

$$g + \hat{h}(x) \leq C(x, f_i(x)) + \sum_y p_{xy}(f_i(x))\hat{h}(y), \quad x \in S, \ i = 1, 2, \tag{4.20}$$

and it will be proven below that the equality holds, i.e.,

$$g + \hat{h}(x) = C(x, f_i(x)) + \sum_y p_{xy}(f_i(x))\hat{h}(y). \tag{4.21}$$

The conclusion follows from this assertion, since by Lemma 4.1(vi), (4.17) and (4.21) together yield that

$$h_i(x) = E_x^{f_i}\left[\sum_{t=0}^{T-1}(C(X_t, A_t) - g)\right] = \hat{h}(x), \quad x \in S, \ i = 1, 2;$$

recall that $h_1(z) = h_2(z) = 0$ and then $\hat{h}(z) = 0$ by (4.18). Thus, to complete the proof it is sufficient to establish (4.21). With this in mind let $\tilde{x} \in S$ and $i \in \{1, 2\}$ be fixed, and note that $\sum_y p_{xy}(f_i(x))\hat{h}(y) \leq \sum_y p_{xy}(f_i(x))h_i(y) < \infty$ by (4.17) and (4.18). Now define $\psi : S \to \mathbb{R}$ by

$$\psi(x) := \begin{cases} g + \hat{h}(x) - C(x, f_i(x)) - \sum_y p_{xy}(f_i(x))\hat{h}(y) & \text{if } x = \tilde{x}, \\ 0 & \text{if } x \neq \tilde{x}, \end{cases} \tag{4.22}$$

so (4.20) implies that for all state x

$$g + \hat{h}(x) \le C(x, f_i(x)) + \psi(x) + \sum_y p_{xy}(f_i(x))\hat{h}(y),$$

with equality for $x = \tilde{x}$. An induction argument yields that

(4.23)

$$g + \frac{1}{n+1}\hat{h}(x) \le \frac{1}{n+1}E_x^{f_i}\left[\sum_{t=0}^{n} C(X_t, f_i(X_t)) + \psi(X_t)\right] + \frac{1}{n+1}E_x^{f_i}[\hat{h}(X_{n+1})], \quad x \in S.$$

Next observe that, by (4.17) and Lemma 4.1(iii) with $W(\cdot) = h_i(\cdot)$ and $f = f_i$, $E_x^{f_i}[h_i(X_{n+1})]/(n+1) \to 0$ as $n \to \infty$. Since $\hat{h} \le h_i$ this implies (recall that $\hat{h}$ is bounded below) that $\lim_{n\to\infty} E_x^{f_i}[\hat{h}(X_{n+1})]/(n+1) = 0$. With this in mind, take limit as $n \to \infty$ in both sides of (4.23) to obtain, via (2.6), (2.9), and (4.22), that

$$g \le \lim_{n\to\infty} \frac{1}{n+1}E_x^{f_i}\left[\sum_{t=0}^{n} C(X_t, f_i(X_t)) + \psi(X_t)\right] = \sum_y \mu_{f_i}(y)C(y, f_i(y)) + \mu_{f_i}(\tilde{x})\psi(\tilde{x}).$$

Since $\sum_y \mu_{f_i}(y)C(y, f_i(y)) = g$ (f_i is AO) and $\mu_{f_i}(\tilde{x}) > 0$ (by (2.9)), the last displayed inequality renders that $\psi(\tilde{x}) \ge 0$ and then $\psi(\tilde{x}) = 0$, since by (4.20) and (4.22), $\psi(\cdot) \le 0$. Therefore, (4.21) holds for $x = \tilde{x}$, and the conclusion follows, since $\tilde{x}$ and i were arbitrary. □

5. Preliminaries: Second part. This section is the second step in the journey to the proof of Theorem 3.1. The present objective is to establish Theorem 5.1, which concerns the asymptotic behavior of the differential costs introduced in Definition 3.2. First, it is convenient to introduce some useful notation.

DEFINITION 5.1. *The functions $U, L : S \to \mathbb{R}$ are defined as follows: for each $x \in S$,*

$$U(x) := \limsup_{n\to\infty} g_n(x) \quad \textit{and} \quad L(x) := \liminf_{n\to\infty} g_n(x).$$

Note that (3.2) and Definition 3.2 together imply that U and L are nonnegative functions. The next theorem is the main result of this section.

THEOREM 5.1. *Let $g \in \mathbb{R}$ and $f^* \in \mathbb{F}$ be as in Lemma* 2.1. *Then*

$$\frac{U(x) - L(x)}{E_x^{f^*}[T_x]} + L(x) \le g \quad \textit{for all } x \in S. \tag{5.1}$$

In particular, U and L are finite *functions.*

Proof. For each $k \in \mathbb{N}$ let policy π^k be as in (3.1). Now let the state x and the nonnegative integer n be arbitrary but fixed, and define a new policy π^{*n} as follows: for $t \in \mathbb{N}$ and $h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t) \in H_t$,

(a) suppose that $x_s \ne x$ for $1 \le s \le t$. In this case

$$\pi_t^{*n}(\{f^*(x_t)\}|h_t) := 1.$$

(b) if $x_s \ne x$ for $1 \le s < k$ and $x_k = x$ for some integer $k \le t$, set

$$\pi_t^{*n}(\cdot|\cdot) := \pi_{t-k}^{n-k}(\cdot|\cdot) \quad \text{if} \quad k \le n$$

and

$$\pi_t^{*n}(\{f^*(x_t)\}|h_t) := 1 \quad \text{if} \quad k > n.$$

In words, a controller using π^{*n} operates as follows: actions are chosen according to f^* until state x is reached in a positive time; if it occurs after time n, policy f^* continues in use, whereas if $T_x = k \leq n$ the controller switches to policy π^{n-k} as if the process had started again. Then it is clear that for all $x \in S$,

$$E_x^{\pi^{*n}}\left[\sum_{t=0}^{n} C(X_t, A_t)I[T_x > n]\right] = E_x^{f^*}\left[\sum_{t=0}^{n} C(X_t, A_t)I[T_x > n]\right] \tag{5.2}$$

and

$$E_x^{\pi^{*n}}\left[\sum_{t=0}^{T_x-1} C(X_t, A_t)I[T_x \leq n]\right] = E_x^{f^*}\left[\sum_{t=0}^{T_x-1} C(X_t, A_t)I[T_x \leq n]\right]. \tag{5.3}$$

Now let $k \leq n$ and recall that $X_k = x$ on $[T_x = k]$ and that T_x is I_k-measurable. Since on $[T_x = k]$ policy π^{*n} coincides with π^{n-k} from time k onward, the Markov property yields that

$$\begin{aligned} E_x^{\pi^{*n}}\left[\sum_{t=T_x}^{n} C(X_t, A_t)I[T_x = k]|I_k\right] &= I[T_x = k]E_{X_k}^{\pi^{*n}}\left[\sum_{t=T_x}^{n} C(X_t, A_t)\right] \\ &= I[T_x = k]E_x^{\pi^{n-k}}\left[\sum_{t=0}^{n-k} C(X_t, A_t)\right] \\ &= I[T_x = k]V_{n-k}(x) \qquad \text{(see (3.1))}. \end{aligned}$$

Therefore, $E_x^{\pi^{*n}}[\sum_{t=T_x}^{n} C(X_t, A_t)I[T_x = k]] = P_x^{\pi^{*n}}[T_x = k]V_{n-k}(x) = P_x^{f^*}[T_x = k]V_{n-k}(x)$ for all $k \leq n$, where the second equality follows from the definition of π^{*n}; this implies that

$$E_x^{\pi^{*n}}\left[\sum_{t=T_x}^{n} C(X_t, A_t)I[T_x \leq n]\right] = E_x^{f^*}[V_{n-T_x}(x)I[T_x \leq n]].$$

Combining this equality with (3.1), (5.2), and (5.3) it follows that

$$\begin{aligned} V_n(x) &\leq E_x^{\pi^{*n}}\left[\sum_{t=0}^{n} C(X_t, A_t)\right] \\ &= E_x^{\pi^{*n}}\left[\sum_{t=0}^{T_x-1} C(X_t, A_t)I[T_x \leq n]\right] + E_x^{\pi^{*n}}\left[\sum_{t=T_x}^{n} C(X_t, A_t)I[T_x \leq n]\right] \\ &\quad + E_x^{\pi^{*n}}\left[\sum_{t=0}^{n} C(X_t, A_t)I[T_x > n]\right] \\ &= E_x^{f^*}\left[\sum_{t=0}^{T_x-1} C(X_t, A_t)I[T_x \leq n]\right] + E_x^{f^*}[V_{n-T_x}(X_{T_x})I[T_x \leq n]] \\ &\quad + E_x^{f^*}\left[\sum_{t=0}^{n} C(X_t, A_t)I[T_x > n]\right] \\ &\leq E_x^{f^*}\left[\sum_{t=0}^{T_x-1} C(X_t, A_t)\right] + E_x^{f^*}[V_{n-T_x}(x)I[T_x \leq n]], \end{aligned}$$

where the nonnegativity of the cost function was used to obtain the second inequality. Thus,

$$V_n(x) - E_x^{f^*}[V_{n-T_x}(x)I[T_x \le n]] \le E_x^{f^*}\left[\sum_{t=0}^{T_x-1} C(X_t, A_t)\right]. \tag{5.4}$$

Now observe that

$$\begin{aligned} V_n(x) - E_x^{f^*}[V_{n-T_x}(x)I[T_x \le n]] &= V_n(x)P_x^{f^*}[T_x > n] \\ &\quad + \sum_{k=1}^{n}(V_n(x) - V_{n-k}(x))P_x^{f^*}[T_x = k]. \end{aligned} \tag{5.5}$$

On the other hand, (2.9) and the dominated convergence together imply that $E_x^{f^*}[T_x I[T_x > n]] \to 0$, whereas $\lim_{n\to\infty} V_n(x)/n = g$ (see [13]). Therefore, it follows that

$$\begin{aligned} 0 \le V_n(x)P_x^{f^*}[T_x > n] &= \frac{V_n(x)}{n} nP_x^{f^*}[T_x > n] \\ &\le \frac{V_n(x)}{n} E_x^{f^*}[T_x I[T_x > n]] \to g \cdot 0 = 0 \text{ as } n \to \infty. \end{aligned} \tag{5.6}$$

Now let $\{n(r)\}$ be a sequence increasing to ∞ such that

$$\lim_{r\to\infty} g_{n(r)}(x) = U(x), \tag{5.7}$$

and note that

$$\liminf_{r\to\infty} g_{n(r)-t}(x) \ge L(x), \quad t = 1, 2, 3, \ldots, \tag{5.8}$$

where the inequality is due to the definition of $L(x)$ as the limit inferior of the *whole* sequence $\{g_n(x)\}$. Hence, for all $s \ge 1$, (5.7) and (5.8) together yield that

$$\begin{aligned} \liminf_{r\to\infty}[V_{n(r)}(X) - V_{n(r)-s}(x)] &= \liminf_{r\to\infty} \sum_{t=0}^{s-1} g_{n(r)-t}(x) \\ &= \liminf_{r\to\infty}\left[g_{n(r)}(x) + \sum_{t=1}^{s-1} g_{n(r)-t}(x)\right] \\ &\ge U(x) + (s-1)L(x), \end{aligned}$$

so an application of Fatou's lemma [21] leads to

$$\begin{aligned} \liminf_{r\to\infty} \sum_{k=1}^{n(r)}(V_{n(r)}(x) - V_{n(r)-k}(x))P_x^{f^*}[T_x = k] &\ge \sum_{k=1}^{\infty}(U(x) + (k-1)L(x))P_x^{f^*}[T_x = k] \\ &= U(x) + E_x^{f^*}[T_x - 1]L(x). \end{aligned} \tag{5.9}$$

Replacing n with $n(r)$ in (5.4) and taking limit inferior as $r \to \infty$ in both sides of the resulting inequality it follows, via (5.5), (5.6), and (5.9), that

$$U(x) + E_x^{f^*}[T_x - 1]L(x) \le E_x^{f^*}\left[\sum_{t=0}^{T_x-1} C(X_t, A_t)\right]. \tag{5.10}$$

This immediately implies that $U(x)$ is finite, and then so is $L(x)$, since $0 \le L(\cdot) \le U(\cdot)$, by Definition 5.1. Next observe that, by Lemma 2.1(iv), $g + h(x) = C(x, f^*(x)) + \sum_y p_{xy}(f^*(x))h(y)$, $x \in S$, so that P_{f^*} has an invariant distribution, by Lemma 4.1(ii). Since f^* is AO, (5.1) follows from (5.10) after simple rearrangements using that $g = E_x^{f^*}[\sum_{t=0}^{T_x-1} C(X_t, A_t)]/E_x^{f^*}[T_x]$; see Remark 2.2, especially (2.10). □

6. Preliminaries: Third part. The objective of this section is to show that the relative value functions in Definition 3.2 are appropriately bounded and that the function L in Definition 5.1 has a minimizer.

THEOREM 6.1. (i) *There exists* $N > 0$ *and a function* $B : S \to [0, \infty)$ *such that* $-N \le R_n(x) \le B(x)$ *for all* $x \in S$ *and* $n \in \mathbb{N}$.

(ii) *There exists* $z^* \in S$ *such that* $L(z^*) \le L(x)$, $x \in S$.

The proof of this result relies on Lemmas 6.1 and 6.2 below.

LEMMA 6.1. *Let* $G \subset S$ *be a finite set. There exists a finite constant* $K(G) > 0$ *such that*

$$|V_n(x) - V_n(y)| \le K(G) \quad \text{for all } x, y \in G,\ n \in \mathbb{N}.$$

Proof. Let $x \in S$ and $n \in \mathbb{N}$ be fixed, and define the policy π^{*n} as in the proof of Theorem 5.1. Recalling that π^{*n} coincides with f^* before time T_x, it follows that for all $y \in S$, $E_y^{\pi^{*n}}[\sum_{t=0}^{n} C(X_t, A_t)I[T_x > n]] = E_y^{f^*}[\sum_{t=0}^{n} C(X_t, A_t)I[T_x > n]]$ and $E_y^{\pi^{*n}}[\sum_{t=0}^{T_x-1} C(X_t, A_t)I[T_x \le n]] = E_y^{f^*}[\sum_{t=0}^{T_x-1} C(X_t, A_t)I[T_x \le n]]$. Also, using that for each $k \le n$, π^{*n} coincides with π^{n-k} from time k onward in the event $[T_x = k]$, it follows that

$$\begin{aligned} E_y^{\pi^{*n}}\left[\sum_{t=T_x}^{n} C(X_t, A_t)I[T_x \le n]\right] &= E_y^{\pi^{*n}}[V_{n-T_x}(x)I[T_x \le n]] \\ &= E_y^{f^*}[V_{n-T_x}(x)I[T_x \le n]], \end{aligned}$$

since policy π^{*n} coincides with f^* before T_x; see the arguments in the paragraph following (5.3). Then (3.1) yields

$$\begin{aligned} V_n(y) &\le E_y^{\pi^{*n}}\left[\sum_{t=0}^{n} C(X_t, A_t)\right] \\ &= E_y^{\pi^{*n}}\left[\sum_{t=0}^{n} C(X_t, A_t)I[T_x > n]\right] + E_y^{\pi^{*n}}\left[\sum_{t=0}^{T_x-1} C(X_t, A_t)I[T_x \le n]\right] \\ &\quad + E_y^{\pi^{*n}}\left[\sum_{t=T_x}^{n} C(X_t, A_t)I[T_x \le n]\right] \\ &= E_y^{f^*}\left[\sum_{t=0}^{n} C(X_t, A_t)I[T_x > n]\right] + E_y^{f^*}\left[\sum_{t=0}^{T_x-1} C(X_t, A_t)I[T_x \le n]\right] \\ &\quad + E_y^{f^*}[V_{n-T_x}(x)I[T_x \le n]] \\ &\le E_y^{f^*}\left[\sum_{t=0}^{T_x-1} C(X_t, A_t)\right] + E_y^{f^*}[V_{n-T_x}(x)I[T_x \le n]]. \end{aligned}$$

Combining this inequality with (3.2) it follows immediately that

$$V_n(y) - V_n(x) \le E_y^{f^*}\left[\sum_{t=0}^{T_x-1} C(X_t, A_t)\right], \tag{6.1}$$

and the conclusion follows setting $K(G) := \max_{x,y\in G}\{E_y^{f^*}[\sum_{t=0}^{T_x-1} C(X_t, A_t)]\}$. □

This lemma will be used to prove part (i) of Theorem 6.1. First notice that, by Theorem 5.1, $U(z) = \limsup_{n\to\infty} g_n(z)$ is finite. Since $g_n(\cdot) \ge 0$, this yields that $\{g_n(z)\}$ is a bounded sequence. Throughout the remainder $b > 0$ is fixed and satisfies

$$b \ge g_n(z), \quad n \in \mathbb{N}. \tag{6.2}$$

In this case

$$kb \ge \sum_{s=0}^{k-1} g_{n-s} = \sum_{s=0}^{k-1}(V_{n-s} - V_{n-s-1}) = V_n(z) - V_{n-k}(z), \quad n-k \ge -1,\ n,k \in \mathbb{N}, \tag{6.3}$$

and since $V_{-1} \equiv 0$,

$$b \ge \frac{1}{n+1} V_n(z). \tag{6.4}$$

On the other hand, $G = G(b+1)$ is the *finite* set guaranteed by Assumption 2.3(i), i.e.,

$$C(x,a) \ge b+1, \quad (x,a) \in (S \setminus G) \times A. \tag{6.5}$$

Finally, set

$$T_G := \min\{n > 0 |\, X_n \in G\} = \min_{x\in G} T_x. \tag{6.6}$$

Proof of Theorem 6.1(i). Let $\pi^n \in \mathbb{M}$ be as in (3.1), and note that

$$\begin{aligned} V_n(x) &= E_x^{\pi^n}\left[\sum_{t=0}^{n} C(X_t, A_t)\right] \\ &= E_x^{\pi^n}\left[\sum_{t=0}^{n} C(X_t, A_t) I[T_G > n]\right] + E_x^{\pi^n}\left[\sum_{t=0}^{T_G-1} C(X_t, A_t) I[T_G \le n]\right] \\ &\quad + E_x^{\pi^n}\left[\sum_{t=T_G}^{n} C(X_t, A_t) I[T_G \le n]\right] \\ &= E_x^{\pi^n}\left[\sum_{t=0}^{n} C(X_t, A_t) I[T_G > n]\right] + E_x^{\pi^n}\left[\sum_{t=0}^{T_G-1} C(X_t, A_t) I[T_G \le n]\right] \\ &\quad + E_x^{\pi^n}[V_{n-T_G}(X_{T_G}) I[T_G \le n]], \end{aligned} \tag{6.7}$$

where Bellman's optimality principle was used to obtain the last equality. Next observe that $M := \max\{|V_n(z) - V_n(y)| \,|\, y \in G,\ n \in \mathbb{N}\}$ is finite, by Lemma 6.1, so (6.7) leads to

$$\begin{aligned} V_n(x) &\ge E_x^{\pi^n}\left[\sum_{t=0}^{n} C(X_t, A_t) I[T_G > n]\right] + E_x^{\pi^n}\left[\sum_{t=0}^{T_G-1} C(X_t, A_t) I[T_G \le n]\right] \\ &\quad + E_x^{\pi^n}[V_{n-T_G}(z) I[T_G \le n]] - M, \end{aligned}$$

and using (6.3) and (6.4),

$$\begin{aligned}
V_n(x) - V_n(z) &\geq E_x^{\pi^n}\left[\sum_{t=0}^{n}(C(X_t, A_t) - \frac{1}{n+1}V_n(z))I[T_G > n]\right] \\
&\quad + E_x^{\pi^n}\left[\sum_{t=0}^{T_G-1} C(X_t, A_t)I[T_G \leq n]\right] \\
&\quad + E_x^{\pi^n}[(V_{n-T_G}(z) - V_n(z))I[T_G \leq n]] - M \\
&\geq E_x^{\pi^n}\left[\sum_{t=0}^{n}(C(X_t, A_t) - b)I[T_G > n]\right] + E_x^{\pi^n}\left[\sum_{t=0}^{T_G-1} C(X_t, A_t)I[T_G \leq n]\right] \\
&\quad - E_x^{\pi^n}[bT_G I[T_G \leq n]] - M \\
&= E_x^{\pi^n}\left[\sum_{t=0}^{n}(C(X_t, A_t) - b)I[T_G > n]\right] \\
&\quad + E_x^{\pi^n}\left[\sum_{t=0}^{T_G-1}(C(X_t, A_t) - b)I[T_G \leq n]\right] - M.
\end{aligned}$$

Now observe that $X_t \notin G$ for $1 \leq t < T_G$, which yields $C(X_t, A_t) - b > 1$ (see (6.5) and (6.6)), so from the last displayed relation

$$\begin{aligned}
V_n(x) - V_n(z) &\geq E_x^{\pi^n}[(C(X_0, A_0) - b + n)I[T_G > n]] \\
&\quad + E_x^{\pi^n}[(C(X_0, A_0) - b + T_G - 1)I[T_G \leq n]] - M \\
&\geq -b + E_x^{\pi^n}[T_G \wedge n - 1] - M \quad \text{(since } C(\cdot,\cdot) \geq 0\text{)} \\
&\geq -b - M - 1,
\end{aligned} \tag{6.8}$$

and then $R_n(x) = V_n(x) - V_n(z) \geq -N$ for all state x and $n \in \mathbb{N}$, where $N := M + b + 1$. To conclude set $B(y) := E_y^{f^*}[\sum_{t=0}^{T-1} C(X_t, A_t)]$ and note that (6.1) with $x = z$ yields that $R_n(y) \leq B(y)$ for all state y; by (2.11) with f^* instead of f, $B(\cdot)$ is finite. □

The arguments used above will be used to prove the following lemma, which plays an important role in the proof of the second part of Theorem 6.1.

LEMMA 6.2. *Let $\pi^k \in \mathbb{M}$ be as in* (3.1)*, and suppose that for some sequence $\{n(r)\}$ increasing to infinity, $\pi^{n(r)} \to \pi \in \mathbb{M}$ as $r \to \infty$. Then, for all $x \in S$,*

$$P_x^{\pi}[T_G < \infty] = 1.$$

Proof. Recall that G is a finite set. A simple induction argument combining Assumption 2.1 with Proposition 18 in [21, p. 232] shows that for all $x \in S$, $n \in \mathbb{N}$, and $y \in G$, the mapping

(6.9)

$$\pi \mapsto P_x^{\pi}[X_s \notin G, 1 \leq s < n, X_n = y] = P_x^{\pi}[X_n = y, T_G = n], \quad \pi \in \mathbb{M}, \quad \text{is continuous,}$$

and since $P_x[T_G = n] = \sum_{y \in G} P_x^{\pi}[X_n = y, T_G = n]$,

$$\pi \mapsto P_x^{\pi}[T_G = n], \quad \pi \in \mathbb{M}, \quad \text{is also continuous.} \tag{6.10}$$

Now let $x \in S$ be arbitrary but fixed, and let $B(x)$ be the upper bound for $\{R_n(x)\}$ given in Theorem 6.1(i). Using (6.8) it follows that

$$M + b + 1 + B(x) \geq E_x^{\pi^n}[T_G \wedge n], \quad n \in \mathbb{N}.$$

Thus, if $q_n(\cdot)$ is the distribution of $T_G \wedge n$ with respect to $P_x^{\pi^n}$, the above relation and Markov's inequality together imply that

$$\sum_{k \geq K} q_n(k) \leq \frac{M + 1 + b + B(x)}{K}, \quad K > 0,$$

so that $\{q_n(\cdot)\}$ is a tight family of probability distributions on the subsets of $\mathbb{N} \setminus \{0\}$. Now suppose that $\pi^{n(r)} \to \pi$ as $r \to \infty$. In this case,

$$\begin{aligned} \lim_{r \to \infty} q_{n(r)}(k) &= \lim_{r \to \infty} P_x^{\pi^{n(r)}}[T_G \wedge n(r) = k] \\ &= \lim_{r \to \infty} P_x^{\pi^{n(r)}}[T_G = k] \quad (T_G \wedge n(r) = T_G \quad \text{when} \quad n(r) > k) \\ &= P_x^{\pi}[T_G = k], \end{aligned}$$

where the third equality follows from (6.10). Since $\{q_{n(r)}\}$ is a tight family, this convergence implies that $1 = \sum_k P_x^{\pi}[T_G = k] = P_x^{\pi}[T_G < \infty]$. □

Using this lemma the proof of Theorem 6.1 can be completed as follows.

Proof of Theorem 6.1(ii). Let n be a positive integer, and select π^n as in (3.1). For each $x \in S$ and $0 < k < n$,

$$\begin{aligned} V_n(x) &= E_x^{\pi^n}\left[\sum_{t=0}^{T_G \wedge k-1} C(X_t, A_t) + \sum_{t=T_G \wedge k}^{n} C(X_t, A_t)\right] \\ &= E_x^{\pi^n}\left[\sum_{t=0}^{T_G \wedge k-1} C(X_t, A_t) + V_{n-T_G \wedge k}(X_{T_G \wedge k})\right], \end{aligned} \tag{6.11}$$

where the last equality follows from Bellman's optimality principle. Now define a new policy $\pi^{n;k}$ as follows:

(a) $\pi_0^{n;k}(\cdot|x_0) = \pi_0^n(\cdot|x_0), \quad x_0 \in S.$

Suppose now that t is a positive integer, and pick $h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t) \in H_t$.

(b) If $x_s \notin G$ for all $s \in \{1, 2, \ldots, t\}$,

$$\pi_t^{n;k}(\cdot|h_t) = \begin{cases} \pi_t^n(\cdot|h_t) & \text{if } t < k, \\ \pi_{t-k}^{n-1-k}(\cdot|x_k, a_k, \ldots, x_t) & \text{if } t \geq k. \end{cases}$$

(c) Suppose that $x_s \notin G$ for $1 \leq s < r \leq t$ and $x_r \in G$. In this case,

$$\pi_t^{n;k}(\cdot|h_t) = \pi_{t-(r \wedge k)}^{n-1-(r \wedge k)}(\cdot|x_{r \wedge k}, a_{r \wedge k}, \ldots, x_t).$$

In words, a controller using $\pi^{n;k}$ operates as follows: the controller starts choosing actions according to π^n (see (a)) and keeps on using π^n until a state in G is reached or time gets value k, whatever occurs first. Then, at time $T_G \wedge k = s$, the controller switches to policy π^{n-1-s} as if the process had started again. According to this interpretation,

$$E_x^{\pi^{n;k}}\left[\sum_{t=0}^{T_G \wedge k-1} C(X_t, A_t)\right] = E_x^{\pi^n}\left[\sum_{t=0}^{T_G \wedge k-1} C(X_t, A_t)\right]$$

and

$$E_x^{\pi^{n;k}}\left[\sum_{t=T_G \wedge k}^{n-1} C(X_t, A_t)\right] = E_x^{\pi^n}[V_{n-1-T_G \wedge k}(X_{T_G \wedge k})].$$

Combining (3.1) with these equalities it follows that

$$V_{n-1}(x) \le E_x^{\pi^{n;k}}\left[\sum_{t=0}^{T_G\wedge k-1} C(X_t, A_t) + \sum_{T_G\wedge k}^{n-1} C(X_t, A_t)\right]$$
$$= E_x^{\pi^n}\left[\sum_{t=0}^{T_G\wedge k-1} C(X_t, A_t) + V_{n-1-T_G\wedge k}(X_{T_G\wedge k})\right],$$

which together with (6.11) and Definition 3.2(ii) yields

$$\begin{aligned} g_n(x) &\ge E_x^{\pi^n}[g_{n-T_G\wedge k}(X_{T_G\wedge k})] \\ &\ge E_x^{\pi^n}[g_{n-T_G}(X_{T_G})I[T_G \le k]] \quad (\text{since } g_t(\cdot) \ge 0) \\ &= \sum_{s=1}^{k}\sum_{y\in G} P_x^{\pi^n}[T_G = s, X_s = y]g_{n-s}(y). \end{aligned} \tag{6.12}$$

Now pick a subsequence $\{n(r)\}$ such that $\lim_{r\to\infty} g_{n(r)}(x) = L(x)$. After taking a subsequence, if necessary, there is no loss of generality in assuming that $\pi^{n(r)} \to \pi \in \mathbb{M}$ as $r \to \infty$, since $\mathbb{M}$ is compact metric. Replacing n with $n(r)$ in (6.12) and taking limit as $r \to \infty$ in the resulting inequality, it follows, via (6.9), that

$$\begin{aligned} L(x) &= \lim_{r\to\infty} g_{n(r)}(x) \\ &\ge \liminf_{r\to\infty} \sum_{s=1}^{k}\sum_{y\in G} P_x^{\pi^{n(r)}}[T_G = s, X_s = y]g_{n(r)-s}(y) \\ &\ge \sum_{s=1}^{k}\sum_{y\in G} \liminf_{r\to\infty} P_x^{\pi^{n(r)}}[T_G = s, X_s = y]g_{n(r)-s}(y) \\ &= \sum_{s=1}^{k}\sum_{y\in G} P_x^{\pi}[T_G = s, X_s = y] \liminf_{r\to\infty} g_{n(r)-s}(y) \\ &\ge \sum_{s=1}^{k}\sum_{y\in G} P_x^{\pi}[T_G = s, X_s = y]L(y) \\ &= E_x^{\pi}[L(X_{T_G})I[T_G \le k]], \end{aligned}$$

where (5.8) was used to obtain the third inequality, and letting k increase to ∞, this implies that

$$L(x) \ge E_x^{\pi}[L(X_{T_G})I[T_G < \infty]].$$

Finally, let $z^* \in G$ be such that $L(z^*) = \min\{L(y)|y \in G\}$; such a point exists, since G is finite. In this case the last displayed relation yields that

$$L(x) \ge E_x^{\pi}[L(z^*)I[T_G < \infty]] = L(z^*)P_x^{\pi}[T_G < \infty] = L(z^*),$$

where Lemma 6.2 was used to obtain the second equality. This completes the proof of Theorem 6.1, since $x \in S$ was arbitrary. □

7. Preliminaries: Fourth part. This is the last step before the proof of Theorem 3.1, and the main objective is to establish the following result.

THEOREM 7.1. (i) *Let z^* be as in Theorem* 6.1(ii), *and suppose that $\{n(r)\}$ is a sequence of positive integers converging to ∞ satisfying*

$$\lim_{r\to\infty} g_{n(r)}(z^*) = L(z^*). \tag{7.1}$$

Then

$$\lim_{r\to\infty} g_{n(r)-1}(z^*) = L(z^*). \tag{7.2}$$

(ii) *There exists a sequence $\{R_k^* : S \to \mathbb{R} \mid k \in \mathbb{N}\}$ and a sequence of stationary policies $\{f_k^*\} \subset \mathbb{F}$ satisfying the following: for some $N' \in (0, \infty)$ and $B' : S \to [0, \infty)$,*

(a) $R_k^* \in \Pi_{x\in S}[-N', B'(x)]$ *for all* $k \in \mathbb{N}$.

(b) $L(z^*) + R_k^*(x) \geq C(x, f_k^*(x)) + \sum_y p_{xy}(f_k^*(x))R_{k+1}^*(y)$, $x \in S$, $k \in \mathbb{N}$.

Proof. For each positive integer n select $f_n \in \mathbb{F}$ such that

$$V_n(x) = C(x, f_n(x)) + \sum_y p_{xy}(f_n(x))V_{n-1}(y), \quad x \in S. \tag{7.3}$$

(i) By (3.1), $V_{n-1}(x) \leq C(x, f_n(x)) + \sum_y p_{xy}(f_n(x))V_{n-2}(y)$, so

$$g_n(x) = V_n(x) - V_{n-1}(x) \geq \sum_y p_{xy}(f_n(x))(V_{n-1}(y) - V_{n-2}(y)) = \sum_y p_{xy}(f_n(x))g_{n-1}(y). \tag{7.4}$$

Now let $\{n(r)\}$ be as in (7.1), let L' be an *arbitrary* limit point of $\{g_{n(r)-1}(z^*)\}$, and note that (7.2) will be proven if it can be shown that

$$L' = L(z^*). \tag{7.5}$$

With this in mind, note that by taking a subsequence if necessary, it can be assumed that

$$g_{n(r)-1}(z^*) \to L' \quad \text{and} \quad f_{n(r)} \to f \in \mathbb{F} \text{ as } r \to \infty; \tag{7.6}$$

recall that $\mathbb{F}$ is compact metric. Next replace n by $n(r)$ and set $x = z^*$ in (7.4). Taking limit inferior as $r \to \infty$ in the resulting inequality it follows, via (7.1), Assumption 2.1, and Fatou's lemma, that

$$\begin{aligned} L(z^*) = \liminf_{r\to\infty} g_{n(r)}(z^*) &\geq \liminf_{r\to\infty} \sum_y p_{z^*y}(f_{n(r)}(z^*))g_{n(r)-1}(y) \\ &\geq \sum_y p_{z^*y}(f(z^*)) \liminf_{r\to\infty} g_{n(r)-1}(y) \\ &\geq \sum_y p_{z^*y}(f(z^*))L(y) \quad \text{(see (5.8))} \\ &\geq L(z^*), \end{aligned} \tag{7.7}$$

where the last inequality is due to Theorem 6.1(ii). Therefore, all inequalities in (7.7) are equalities, and then

$$\sum_y p_{z^*y}(f(z^*)) \liminf_{r\to\infty} g_{n(r)-1}(y) = \sum_y p_{z^*y}(f(z^*))L(y).$$

Combining this with (5.8) it follows that

$$\liminf_{r\to\infty} g_{n(r)-1}(y) = L(y) \quad \text{if } p_{z^*y}(f(z^*)) > 0,$$

but, by Assumption 2.4, this implies that $\liminf_{r\to\infty} g_{n(r)-1}(z^*) = L(z^*)$, which combined with the first convergence in (7.6) yields (7.5), and as already mentioned, this completes the proof of part (i).

(ii) Pick a sequence $\{n(r)\}$ such that $g_{n(r)}(z^*) \to L(z^*)$ as $r \to \infty$. By repeated application of part (i),

$$\lim_{r\to\infty} g_{n(r)-s}(z^*) = L(z^*) \quad \text{for all } s \in \mathbb{N}. \tag{7.8}$$

Replacing n with $n(r) - s$ in (7.3), simple rearrangements using Definition 3.2(ii) yield that

$$g_{n(r)-s}(z^*) + \tilde{R}_{n(r)-s}(x) = C(x, f_{n(r)-s}(x)) + \sum_y p_{xy}(f_{n(r)-s})(x)\tilde{R}_{n(r)-s-1}(x), \tag{7.9}$$
$$x \in S, \quad n(r) > s,$$

where, $\tilde{R}_{n(r)-s}(x) := V_{n(r)-s}(x) - V_{n(r)-s}(z^*) = R_{n(r)-s}(x) + [V_{n(r)-s}(z) - V_{n(r)-s}(z^*)]$. By Lemma 6.1, $\max\{|V_{n(r)-s}(z) - V_{n(r)-s}(z^*)| \mid r, s \in \mathbb{N},\ n(r) \geq s\} =: K < \infty$, and combining this with Theorem 6.1(i) it follows that

$$\tilde{R}_{n(r)-s} \in \mathbb{D} := \Pi_{x\in S}[-N - K, B(x) + K]. \tag{7.10}$$

Now set $\tilde{R}_t \equiv 0$ and $f_t \equiv f_1$ for $t < 0$, and note that

$$\rho_{n(r)} := (\tilde{R}_{n(r)-s} | s \in \mathbb{N}) \in \mathbb{D}^\infty =: \mathbb{E},$$
$$\phi_{n(r)} := (f_{n(r)-s} | s \in \mathbb{N}) \in \mathbb{F}^\infty = \mathbb{M}.$$

Since $\mathbb{E}$ and $\mathbb{M}$ are compact metric, taking a subsequence if necessary, it can be assumed that in addition to (7.8), as $r \to \infty$, $\rho_{n(r)} \to \rho^* := (R_s^* | s \in \mathbb{N}) \in \mathbb{E}$ and $\phi_{n(r)} \to \phi^* := (f_s^* | s \in \mathbb{N}) \in \mathbb{M}$. These convergences are equivalent to the following: for each $s \in \mathbb{N}$

$$\lim_{r\to\infty} \tilde{R}_{n(r)-s}(x) = R_s^*(x) \in [-N - K, B(x) + K], \quad x \in S. \tag{7.11}$$

$$\lim_{r\to\infty} f_{n(r)-s}(x) = f_s^*(x), \quad x \in S. \tag{7.12}$$

Setting $N' := N + K$ and $B'(x) = B(x) + K$, (7.11) shows that the sequence $\{R_s^*\}$ satisfies part (a). To conclude, take limit as $r \to \infty$ in both sides of (7.9). In this situation, using (7.8), (7.11), (7.12), and Assumption 2.1, it follows that for all $s \in \mathbb{N}$ and $x \in S$

$$\begin{aligned} L(z^*) + R_s^*(x) &= C(x, f_s^*(x)) + \lim_{r\to\infty} \sum_y p_{xy}(f_{n(r)-s}^*(x))\tilde{R}_{n(r)-s-1}(y) \\ &\geq C(x, f_s^*(x)) + \sum_y \liminf_{r\to\infty} p_{xy}(f_{n(r)-s}^*(x))\tilde{R}_{n(r)-s-1}(y) \\ &= C(x, f_s^*(x)) + \sum_y p_{xy}(f_s^*(x))R_{s+1}^*(y), \end{aligned}$$

where Fatou's lemma was used to obtain the inequality. Then $\{R_s^*\}$ and $\{f_s^*\}$ satisfy condition (b), and the proof is complete. □

8. Proof of the main result. The preliminaries in the previous sections will be now used to establish Theorem 3.1.

Proof of Theorem 3.1. (i) Let $\{R_k^* : S \to \mathbb{R}\}$ and $\{f_k^*\} \subset \mathbb{F}$ be as in Theorem 7.1. Set $\pi^* := \{f_s^*\} \in \mathbb{M}$. An induction argument using part (ii(b)) of Theorem 7.1 yields that for all $x \in S$ and $n \in \mathbb{N}$,

$$L(z^*) + \frac{1}{n+1}R_0^*(x) \geq \frac{1}{n+1}E_x^{\pi^*}\left[\sum_{t=0}^{n} C(X_t, A_t)\right] + \frac{1}{n+1}E_x^{\pi^*}[R_{n+1}^*(X_{n+1})]. \tag{8.1}$$

On the other hand, note that $R^*_{n+1}(\cdot) \geq -N' \in (-\infty, 0]$, so

$$\liminf_{n\to\infty} \frac{E_x^{\pi^*}[R^*_{n+1}(X_{n+1})]}{(n+1)} \geq 0.$$

Therefore, taking limit superior as $n \to \infty$ in both sides of (8.1), it follows that

$$L(z^*) \geq \limsup_{n\to\infty} \frac{1}{n+1} E_x^{\pi^*}\left[\sum_{t=0}^{n} C(X_t, A_t)\right] = J(x, \pi^*),$$

and then

$$L(z^*) \geq g,$$

since g is the optimal average cost; see Lemma 2.1(i). Combining this inequality with Theorem 5.1 and Theorem 6.1(ii) it follows that

$$g \leq L(z^*) \leq L(x) \leq \frac{U(x) - L(x)}{E_x^{f^*}[T_x]} + L(x) \leq g, \quad x \in S,$$

which immediately yields that

$$L(x) = g \quad \text{and} \quad \frac{U(x) - L(x)}{E_x^{f^*}[T_x]} = 0, \quad x \in S. \tag{8.2}$$

To conclude, recall that policy f^* is obtained as in Lemma 2.1(iv) and that, as already noted in the proof of Theorem 5.1, the corresponding transition matrix P_{f^*} has an invariant distribution. Thus, $E_x^{f^*}[T_x]$ is finite for all state x (see Remark 2.2(i)), and it follows that (8.2) is equivalent to

$$U(x) = L(x) = g, \quad x \in S,$$

i.e., $\lim_{n\to\infty} g_n(x) = g$ for all state x; see Definition 3.2(ii).

(ii) By Theorem 6.1(i),

$$R_n \in \Pi_{x\in S}[-N, B(x)] =: \mathbb{K}, \quad n \in \mathbb{N}, \tag{8.3}$$

for some positive constant N and a certain function $B : S \to [0, \infty)$. Since $\mathbb{K}$ is a compact metric space, it is sufficient to show that any limit point—say, $Q \in \mathbb{K}$—of $\{R_n\}$ coincides with the function $h(\cdot)$ in Lemma 2.1. Thus, pick a sequence $\{n(r)\}$ increasing to ∞ such that

$$\lim_{r\to\infty} R_{n(r)}(x) = Q(x) \in [-N, B(x)], \quad x \in S. \tag{8.4}$$

From Definition 3.2 it follows that $R_n(x) - R_{n-1}(x) = g_n(x) - g_n(z)$ and then $R_n(x) - R_{n-1}(x) \to 0$ as $n \to \infty$, by part (i). Therefore, (8.4) implies that

$$\lim_{r\to\infty} R_{n(r)-1}(x) = Q(x) \in [-N, B(x)], \quad x \in S. \tag{8.5}$$

Now pick a policy $f_{n(r)} \in \mathbb{F}$ such that

$$V_{n(r)}(x) = C(x, f_{n(r)}(x)) + \sum_y p_{xy}(f_{n(r)}(x)) V_{n(r)-1}(y)$$

for all state x, which after simple rearrangements using Definition 3.2 can be written as

$$g_{n(r)}(z) + R_{n(r)}(x) = C(x, f_{n(r)}(x)) + \sum_y p_{xy}(f_{n(r)}(x))R_{n(r)-1}(y), \quad x \in S. \tag{8.6}$$

Recalling that $\mathbb{F}$ is a compact metric, after taking a subsequence if necessary, it can be assumed that in addition to (8.4) and (8.5), $f_{n(r)} \to f \in \mathbb{F}$ as $r \to \infty$. In this case, letting r increase to ∞ in both sides of (8.6), it follows, via part (i), Assumption 2.1, (8.4), and (8.5), that

$$\begin{aligned} g + Q(x) &= \lim_{r\to\infty} \{g_{n(r)}(z) + R_{n(r)}(x)\} \\ &= \lim_{r\to\infty} \left\{ C(x, f_{n(r)}(x)) + \sum_y p_{xy}(f_{n(r)}(x))R_{n(r)-1}(y) \right\} \\ &= C(x, f(x)) + \lim_{r\to\infty} \sum_y p_{xy}(f_{n(r)}(x))R_{n(r)-1}(y) \\ &\geq C(x, f(x)) + \sum_y \liminf_{r\to\infty} p_{xy}(f_{n(r)}(x))R_{n(r)-1}(y) \\ &= C(x, f(x)) + \sum_y p_{xy}(f(x))Q(y), \end{aligned} \tag{8.7}$$

where Fatou's lemma was used to obtain the inequality; this is possible, since $R_n(\cdot) \geq -N$. Next observe that

$$Q(z) = 0 \quad \text{and} \quad Q(\cdot) \geq -N \in (-\infty, 0]; \tag{8.8}$$

see (8.4) and recall that $R_n(z) = 0$ for all n by Definition 3.2. Note now that (8.7) and (8.8) allow to use Lemma 4.1(iv) with W replaced by Q to obtain

$$g + Q(x) = \min_{a\in A} \left[C(x, a) + \sum_y p_{xy}(a)Q(y) \right], \quad x \in S;$$

i.e., Q satisfies the ACOE. To conclude observe that the assumptions in Theorem 4.1 are satisfied with $h_1 := Q$ and $h_2 = h$; see (8.8) and parts (ii) and (iii) of Lemma 2.1. Therefore, Theorem 4.1 implies that $Q = h$, and as already noted, this shows that $\{R_n(\cdot)\}$ converges pointwise to $h(\cdot)$.

(iii) Since $R_n(\cdot)$ is bounded below, by Theorem 6.1(i), Assumption 2.1 implies that for each $x \in S$ the mapping $a \mapsto C(x, a) + \sum_y p_{xy}(a)R_n(y)$, $a \in A$, is lower semicontinuous and then has a minimizer $f_n(x)$. In this case,

$$\begin{aligned} C(x, f_n(x)) + \sum_y p_{xy}(f_n(x))R_n(y) &= \min_{a\in A} \left[C(x, a) + \sum_y p_{xy}(a)R_n(y) \right] \\ &= \min_{a\in A} \left[C(x, a) + \sum_y p_{xy}(a)V_n(y) \right] - V_n(z) \\ &= V_{n+1}(x) - V_n(z), \end{aligned}$$

and then (see Definition 3.2)

$$C(x, f_n(x)) + \sum_y p_{xy}(f_n(x))R_n(y) = R_{n+1}(x) + g_{n+1}(z). \tag{8.9}$$

To conclude let $f \in \mathbb{F}$ be a limit point of $\{f_n\}$, and select a sequence $\{n(r)\}$ increasing to ∞ such that for all $x \in S$, $f_{n(r)}(x) \to f(x)$ as $r \to \infty$. In this case, replacing n by $n(r)$ in (8.9) and taking limit as $r \to \infty$ in both sides of the resulting equality, it follows, using Assumption 2.1, Fatou's lemma, and parts (i) and (ii), that for all $x \in S$

$$\begin{aligned} g + h(x) &= \lim_{r\to\infty} [g_{n(r)+1}(z) + R_{n(r)+1}(x)] \\ &= \lim_{r\to\infty} \left[C(x, f_{n(r)}(x)) + \sum_y p_{xy}(f_{n(r)}(x)) R_{n(r)}(y) \right] \\ &= C(x, f(x)) + \lim_{r\to\infty} \sum_y p_{xy}(f_{n(r)}(x)) R_{n(r)}(y) \\ &\geq C(x, f(x)) + \sum_y \liminf_{r\to\infty} p_{xy}(f_{n(r)}(x)) R_{n(r)}(y) \\ &= C(x, f(x)) + \sum_y p_{xy}(f(x)) h(x), \end{aligned}$$

and from the ACOE (see (2.4)), it follows that $f(x)$ is a minimizer of the mapping $a \mapsto C(x, a) + \sum_y p_{xy}(a)h(y)$, $a \in A$, so that $f \in \mathbb{F}$ is AO, by Lemma 2.1(iv). □

9. An example. This section contains an example illustrating the application of Theorem 3.1 to a single-server queueing system.

Example 9.1. Let A be a finite set endowed with the discrete topology and suppose that $\{U_{na}, D_{na} \mid a \in A, n \in \mathbb{N}\}$ is a collection of independent $\mathbb{N}$-valued random variables satisfying (i)–(iii) below.

(i) For each $a \in A$, the random variables $\{U_{na} \mid n \in \mathbb{N}\}$ are identically distributed with common distribution $\{q_a(\cdot)\}$: $P[U_{na} = k] = q_a(k)$, $n, k \in \mathbb{N}$, where

$$q_a(0) \in (0, 1). \tag{9.1}$$

(ii) For each $a \in A$ and $n \in \mathbb{N}$,

$$P[D_{na} = 1] = 1 - P[D_{na} = 0] = \delta_a \in (0, 1). \tag{9.2}$$

These random variables will be interpreted as the arriving and service streams in a single-server queueing model with state space $S = \mathbb{N}$ and action space A. For each time $n \in \mathbb{N}$, let $X_n = x \in S$ be the number of customers waiting for service at the beginning of the period $[n, n+1)$. If action $A_n = a$ is applied, then the number of customers arriving between times n and $n+1$ is U_{na}, whereas if $X_n > 0$, the server can provide a complete service in that period with probability δ_a; $D_{na} (= 0, 1)$ is interpreted as the number of customers leaving the system after service completion in $[n, n+1)$. Formally, this can be summarized in the following evolution equation:

$$X_{n+1} = \begin{cases} X_n + U_{na} - D_{na} & \text{if } X_n > 0 \text{ and } A_n = a, \\ U_{na} & \text{if } X_n = 0 \text{ and } A_n = a. \end{cases} \tag{9.3}$$

Note that this equation immediately determines the transition law $p_{xy}(a) = P[X_{n+1} = y | X_n = x, A_n = a]$.

(iii) For some $a^* \in A$, $E[U_{na^*}^2] =: \lambda_{a^*}^{(2)} < \infty$ and $E[U_{na^*}] =: \lambda_{a^*} < \delta_{a^*}$.

Finally, define the cost function by $C(x, a) := x$, $(x, a) \in S \times A$.

PROPOSITION 9.1. *Assumptions* 2.1–2.4 *are satisfied in Example* 9.1. *Therefore, the conclusions in Theorem* 3.1 *occur.*

Proof. (1) Since A is finite and $C(x,a) = x$, Assumption 2.1 clearly holds.

(2) Let f be an arbitrary stationary policy, and note that (9.1)–(9.3) together imply that for all $x, y \in S$ and $n \in \mathbb{N}$, $P_y^f[X_{n+1} \geq x+1 | X_n = x] \geq P[U_{n\,f(x)} > 0, D_{n\,f(x)} = 0] = (1-q_{f(x)}(0))(1-\delta_{f(x)}) > 0$, and that if $x > 0$, $P_x^f[X_{n+1} = x-1 | X_n = x] = q_{f(x)}(0)\delta_{f(x)} > 0$. Using these inequalities, a simple induction argument yields that

$$P_y^f[X_n \geq y+n] > 0, \quad y \in S,\ n \in \mathbb{N}, \tag{9.4}$$

and

$$P_y^f[X_n = y-n] > 0, \quad y, n \in S,\ y > n. \tag{9.5}$$

Assumption 2.2 follows immediately from these facts. Indeed, let $x, y \in S$. By (9.4), there exists a positive integer n such that $P_x^f[X_n = w] > 0$ for some $w > x \vee y$, and then (9.5) yields, for $m = w - y$, that $P_w^f[X_m = y] > 0$. Therefore, $P_x^f[X_{n+m} = y] \geq P_x^f[X_n = w] P_w^f[X_m = y] > 0$, showing that Assumption 2.2 holds, since $f \in \mathbb{F}$ was arbitrary.

(3) Note that part (i) of Assumption 2.3 clearly holds, since $C(x,a) = x$ for all $(x,a) \in S \times A$. To verify part (ii) define real constants l_i, $i = 0, 1, 2$, by

$$l_2 := \frac{1}{\delta_{a^*} - \lambda_{a^*}}, \quad l_1 := \frac{E[(U_{n,a^*} - D_{n\,a^*})^2] - (\delta_{a^*} - \lambda_{a^*})}{\delta_{a^*} - \lambda_{a^*}},$$

and

$$l_0 := \frac{1 + l_1\lambda_{a^*} + l_2\lambda_{a^*}^{(2)}}{q_{a^*}(0)}.$$

Now set $l(x) := l_0 + l_1 x + l_2 x^2$, $x \in S$, and define the stationary policy f by $f(x) := a^*$, $x \in S$. In this case, straightforward calculations using (9.3) show that

$$1 + C(x, f(x)) + \sum_{y \neq 0} p_{x\,y}(f(x)) l(y) \leq l(x), \quad x \in S,$$

so $l(\cdot)$ is a Lyapunov function of the Markov chain induced by f, and this yields that $J(\cdot, f)$ is finite [16, 28]; alternatively, the finiteness of $J(\cdot, f)$ can be obtained from Proposition 4 in [23]. This completes the verification of Assumption 2.3.

(4) Note that $p_{x\,x}(a) = P[X_{n+1} = x | X_n = x, A_n = a] \geq P[U_{n\,a} = 0] P[D_{n\,a} = 0] = q_a(0)(1 - \delta_a) > 0$ by (9.1) and (9.2), so Assumption 2.4 occurs. □

In the above example notice that, if $E[U_{n\,a}] > \delta_a$ for some $a \in A$, then it is not difficult to see that the policy $f_a \in \mathbb{F}$ given by $f_a(\cdot) \equiv a$ induces a transient Markov chain.

10. Conclusion. The value iteration procedure has been studied in the class of communicating MDPs characterized by Assumptions 2.1–2.4, and it was shown in Theorem 3.1 that, within this framework, the differential costs and relative value functions produced by the VI scheme converge to the (unique) solution of the ACOE given in Lemma 2.1. The approach used in this work—following the ideas in [7, 10]—is not based on an analysis of the differences $V_n(\cdot) - (n+1)g$ but relies on a direct study of the differential costs; see Theorems 5.1, 6.1(ii), and 7.1. This allows one to avoid other (somewhat restrictive) conditions imposed in other works, such as Assumption 5 in [24] and the boundedness of the first error function [17].

Acknowledgment. The author is grateful to the unknown reviewers for their careful reading of the original manuscript, constructive criticism and helpful suggestions.

REFERENCES

[1] V. S. BORKAR, *Controlled Markov chains and stochastic networks*, SIAM J. Control Optim., 21 (1983), pp. 652–666.

[2] ———, *On minimum cost per unit of time control of Markov chains*, SIAM J. Control Optim., 22 (1984), pp. 965–978.

[3] R. CAVAZOS-CADENA, *Weak conditions for the existence of average optimal stationary policies in average Markov decision chains with unbounded costs*, Kybernetika, 25 (1989), pp. 145–156.

[4] ———, *Solution to the optimality equation in a class of Markov decision chains with the average cost criterion*, Kybernetika, 27 (1991), pp. 26–37.

[5] ———, *Recent results on conditions for the existence of average optimal stationary policies*, Ann. Oper. Res., 28 (1991), pp. 3–28.

[6] R. CAVAZOS-CADENA AND L. I. SENNOTT, *Comparing recent assumptions for the existence of optimal stationary policies*, Oper. Res. Lett., 11 (1992), pp. 33–37.

[7] R. CAVAZOS-CADENA, *Undiscounted value iteration in stable Markov decision chains with bounded rewards*, J. Math. Systems Estim. Control, 6 (1994), pp. 243–246.

[8] ———, *Cesàro convergence of the undiscounted value iteration method in Markov decision processes under the Lyapunov stability condition*, Bol. Soc. Mat. Mexicana (2), 38 (1993), pp. 33–46.

[9] ———, *Value Iteration in Controlled Markov Chains with Penalized Costs: Cesàro Convergence Results*, Report 03–91–DEC, Universidad Autónoma Agraria Antonio Narro, Saltillo Coah, México, 1991.

[10] R. CAVAZOS-CADENA AND E. FERNÁNDEZ-GAUCHERAND, *Value iteration in stable Markov decision chains with unbounded costs: Necessary and sufficient conditions for pointwise convergence*, J. Appl. Probab., (1996), to appear.

[11] J. DUGUNDJI, *Topology*, Allyn and Bacon, Boston, 1966.

[12] A. FEDERGRUEN AND P. J. SCHWEITZER, *A survey of asymptotic value iteration for undiscounted Markovian decision processes*, in Recent Developments in Markov Decision Processes, R. Hartley, L. C. Thomas, and D. J. White, eds., Academic Press, New York, 1980.

[13] M. K. GOSH AND S. I. MARCUS, *On strong average optimality of Markov decision processes with unbounded costs*, *Oper. Res. Lett.*, 11 (1992), pp. 99–104.

[14] O. HERNÁNDEZ-LERMA, *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.

[15] K. HINDERER, *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes in Oper. Res. 33, Springer-Verlag, New York, 1970.

[16] A. HORDIJK, *Dynamic Programming and Markov Potential Theory*, Mathematical Centre Tract 51, Mathematisch Centrum, Amsterdam, 1974.

[17] A. HORDIJK, P. J. SCHWEITZER, AND H. C. TIJMS, *The asymptotic behavior of the minimal total expected cost for the denumerable state Markov decision model*, J. Appl. Probab., 12 (1975), pp. 298–305.

[18] M. LOÈVE, *Probability Theory* I, Springer-Verlag, New York, 1978.

[19] R. MONTES-DE-OCA AND O. HERNÁNDEZ-LERMA, *Value iteration in average cost Markov control processes on Borel spaces*, Acta Appl. Math., 42 (1996), pp. 203–221.

[20] M. L. PUTERMAN, *Markov Decision Processes*, Wiley, New York, 1994.

[21] H. L. ROYDEN, *Real Analysis*, 2nd ed., Macmillan, New York, 1968.

[22] S. M. ROSS, *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.

[23] L. I. SENNOTT, *Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs*, Oper. Res., 37 (1989), pp. 626–633.

[24] ———, *Value iteration in countable state Markov decision processes with unbounded costs*, Ann. Oper. Res., 28 (1991), pp. 261–272.

[25] ———, *The convergence of value iteration in average cost Markov decision chains*, (1994), submitted.

[26] P. J. SCHWEITZER AND A. FEDERGRUEN, *The asymptotic behavior of undiscounted value iteration in Markov decision proceses*, Math. Oper. Res., 2 (1977), pp. 360–381.

[27] P. J. SCHWEITZER, *Iterative solution of the functional equations for undiscounted Markov renewal programming*, J. Math. Anal. Appl., 34 (1971), pp. 495–501.

[28] L. C. THOMAS, *Connectedness conditions for denumerable state Markov decision processes*, in Recent Developments in Markov Decision Processes, R. Hartley, L. C. Thomas, and D. J. White, eds., Academic Press, New York, 1980.

[29] R. R. WEBER AND S. STIDHAM, *Optimal control of service rates in networks of queues*, Advances in Applied Probability, 19 (1987), pp. 202–218.

[30] D. J. WHITE, *Dynamic programming, Markov chains, and the method of succesive approximations*, J. Math. Anal. Appl., (1963), pp. 373–376.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 1874–1890, December 1996

003

CAUSAL FEEDBACK OPTIMAL CONTROL FOR VOLTERRA INTEGRAL EQUATIONS*

A. J. PRITCHARD† AND YUNCHENG YOU‡

Abstract. The optimal control problem for Volterra integral equations with respect to quadratic criteria is studied by a projection causality approach. The work features a synthesis result where the optimal control is implemented via a linear causal feedback in which the feedback operator is determined by solving an independent Fredholm integral operator equation.

Key words. Volterra integral equations, linear quadratic control, causal state feedback, Fredholm integral operator equation

AMS subject classifications. 93C22, 45D05

1. Introduction.

The state-space approach in modern control theory provides a mathematical framework in which many control problems for ordinary differential equations, partial differential equations, and functional differential equations can be formulated and treated in a unified manner via abstract operators on Banach spaces and semigroup theory; cf. [2], [9], [3], [12], [14], [15]. On the other hand, engineering control designs often feature direct input-output relations in either time or frequency domain. Some difficulties in this approach arise in handling control problems governed by partial differential equations, due to the fact that functions of several complex variables are involved. But it also has merit since the transfer functions are theoretically independent of whether the outputs are related to the inputs via the solution of an initial value problem. A very general class of input-output relations, which includes all initial value problems for linear evolutionary equations in Banach spaces as a proper subset, is described by Volterra integral equations. These are what we study in this paper.

We shall consider quadratic optimal control problems for general linear Volterra equations in Hilbert spaces. The open-loop control is easily obtained, but since we do not assume that there exists a realization in terms of a differential equation on the Hilbert space, we cannot apply standard theory to obtain a state feedback via the solution of a Riccati equation. Thus the principal objective of the paper will be to derive a closed-loop synthesis which provides a feedback optimal control in a causal sense. We establish the results for the single-variable case, but they are easily generalized to the bivariate and multivariate cases, which have some direct applications to multidimensional signal processing and data analysis; cf. [8].

To introduce the model equation and optimal control problem, consider the following input-output relation given by a Volterra integral operator:

$$y(t) = f(t) + \int_0^t K(t,\tau)u(\tau)\,d\tau, \qquad 0 \le t \le T, \tag{1}$$

where $T > 0$ is finite and fixed. We let $\Delta = \{(t,\tau) \in [0,T]^2;\ 0 \le \tau \le t \le T\}$, where $[0,T]^2 = [0,T] \times [0,T]$.

Let Y and U be real Hilbert spaces. We assume that

$$f \in C([0,T]; Y), \tag{A1}$$

$$K \in C(\Delta; \mathcal{L}(U;Y)), \tag{A2}$$

*Received by the editors September 24, 1994; accepted for publication (in revised form) June 24, 1995.
†Department of Mathematics, University of Warwick, Coventry CV4 7AL, UK.
‡Department of Mathematics, University of South Florida, Tampa, FL 33620.

where the continuity is in the strong sense; i.e., for each $u \in U$, the function $K(t, \tau)u : \Delta \to Y$ is strongly continuous. Thus by the Banach–Steinhaus theorem, one has $\sup_{(t,\tau)\in\Delta} \|K(t, \tau)\|_{\mathcal{L}(U;Y)} < \infty$. For convenience, we make a convention that $K(t, \tau) = 0$ whenever $(t, \tau) \in [0, T]^2 \backslash \Delta$.

We denote $\mathcal{Y} = L^2(0, T; Y)$ and $\mathcal{U} = L^2(0, T; U)$. The quadratic cost functional is

$$J(u, f) = \langle Gy(T), y(T)\rangle + \int_0^T [\langle Q(t)y(t), y(t)\rangle + \langle R(t)u(t), u(t)\rangle]\, dt, \tag{2}$$

where the inner products $\langle \cdot, \cdot \rangle$ are related to the spaces Y or U according to the context. We assume furthermore that

(A3)

$$G \in \mathcal{L}(Y),\ Q(\cdot) \in C(0, T; \mathcal{L}(Y)),\ R(\cdot) \in C(0, T; \mathcal{L}(U)), \quad \text{where } G, Q(t), R(t)$$

are self-adjoint and there exists a constant $\delta > 0$ such that $R(t) \geq \delta I_U, t \in [0, T]$.

The problem is to find an optimal control $u^* \in \mathcal{U}$ which minimizes $J(u, f)$ over $\mathcal{U}$ for any given f satisfying (A1). We shall refer to this optimal control problem as (OCP).

Before getting into the mathematical theory, let us remark that more general controlled linear Volterra integral equations can be reduced to the form (1). For instance consider the integral equation

$$y(t) = f(t) + \int_0^t \Lambda(t, \tau)y(\tau)\, d\tau + \int_0^t N(t, \tau)u(\tau)\, d\tau, \qquad 0 \leq t \leq T, \tag{3}$$

with appropriate assumptions similar to those above. Then it is easy to prove the following lemma, which reduces (3) to (1).

LEMMA 1.1. *Define an operator* $\mathcal{T} \in \mathcal{L}(C([0, T]; Y))$ *by*

$$(\mathcal{T}y)(t) = \int_0^t \Lambda(t, \tau)y(\tau)\, d\tau, \qquad t \in [0, T],$$

where $\Lambda \in C(\Delta; \mathcal{L}(Y))$. *Then the following statements hold.*

(i) $\mathcal{T}$ *is a quasi-nilpotent operator so that* $\sigma(\mathcal{T}) = \{0\}$.

(ii) *There exists a unique solution* y *of* (3), *which can be expressed by*

$$y(t) = g(t) + \int_0^t K(t, \tau)u(\tau)\, d\tau, \qquad 0 \leq t \leq T, \tag{4}$$

where

$$g(t) = f(t) + \int_0^t R(t, \tau)f(\tau)\, d\tau, \qquad 0 \leq t \leq T,$$

$$K(t, \tau)u = N(t, \tau)u + \int_\tau^t R(t, s)N(s, \tau)u\, ds, \qquad (t, \tau) \in \Delta,$$

and the resolvent kernel $R(t, s)$ *is given by*

$$R(t, s) = \sum_{j=1}^{\infty} \Lambda_j(t, s) \tag{5}$$

with

$$\Lambda_1(t,s) = \Lambda(t,s), \quad \Lambda_{j+1}(t,s)y = \int_s^t \Lambda(t,\rho)\Lambda_j(\rho,s)y\,d\rho, \quad j = 1,2,\ldots.$$

Here the series in (5) *converges in the operator norm.*

As we mentioned earlier, the difficulty in this work lies in the determination of a feedback optimal control in the causal sense. Namely, the optimal control $u^*(t)$ at any time t should not involve future information of the corresponding trajectory $y^*(\cdot)$. But it would be naive to seek an optimal control whose real-time value depends only on the past information of the function $f(\cdot)$, since the whole input $f(t)$, $t \in [0,T]$, determines the optimality conditions.

2. Existence and open-loop result. We now define various linear operators on the function spaces $\mathcal{Y}$ and $\mathcal{U}$: $\Gamma \in \mathcal{L}(\mathcal{U},\mathcal{Y})$, $\Gamma_T \in \mathcal{L}(\mathcal{U},\mathcal{Y})$, $\mathcal{Q} \in \mathcal{L}(\mathcal{Y})$, $\mathcal{R} \in \mathcal{L}(\mathcal{U})$,

$$(\Gamma u)(t) = \int_0^t K(t,\tau)u(\tau)\,d\tau, \quad t \in [0,T], \quad u \in \mathcal{U};$$

$$\Gamma_T u = \int_0^T K(T,\tau)u(\tau)\,d\tau = (\Gamma u)(T), \qquad u \in \mathcal{U};$$

$$(\mathcal{Q}y)(t) = Q(t)y(t), \quad t \in [0,T], \quad y \in \mathcal{Y};$$

$$(\mathcal{R}u)(t) = R(t)u(t), \quad t \in [0,T], \quad u \in \mathcal{U}.$$

It is easy to see that $\mathcal{Q}$ and $\mathcal{R}$ are self-adjoint and the adjoints of Γ and Γ_T are given by

$$(\Gamma^* y)(t) = \int_t^T K^*(\tau,t)y(\tau)\,d\tau, \quad t \in [0,T], \quad y \in \mathcal{Y};$$

$$(\Gamma_T^* \varphi)(t) = K^*(T,t)\varphi, \quad t \in [0,T], \quad \varphi \in Y.$$

An important role will be played by an operator $\Phi \in \mathcal{L}(\mathcal{U})$, where

$$\Phi = \mathcal{R}I + \Gamma^*\mathcal{Q}\Gamma + \Gamma_T^* G T_T. \tag{6}$$

Our final assumption is that

$$\text{there exists } \epsilon > 0 \text{ such that } \quad \Phi \geq \epsilon I_{\mathcal{U}}. \tag{A4}$$

A preliminary characterization of the optimal control is given in the following theorem.

THEOREM 2.1. *Suppose that* (A1)–(A4) *hold. Then for any given* $f \in C([0,T],Y)$ *there exists a unique optimal control* u (*we drop the* * *notation where there is no confusion*) *for (OCP). The pair* $\{u,y\}$ *is optimal if and only if the following relation is satisfied:*

$$u(t) = -R(t)^{-1}\left[K^*(T,t)Gy(T) + \int_t^T K^*(s,t)Q(s)y(s)\,ds\right], \qquad t \in [0,T]. \tag{7}$$

Proof. Substituting the expressions

$$y = f + \Gamma u \quad \text{and} \quad y(T) = f(T) + \Gamma_T u \tag{8}$$

into the cost functional (2), we obtain

$$J(u,f) = \langle \Phi u, u\rangle_{\mathcal{U}} + 2\langle \Gamma^*\mathcal{Q}f + \Gamma_T^* G f(T), u\rangle_{\mathcal{U}} + \text{ const}(f).$$

By assumption (A4), $\Phi^{-1} \in \mathcal{L}(\mathcal{U})$; therefore, the unique optimal control function $u \in \mathcal{U}$ is given by

$$u = -\Phi^{-1}(\Gamma^* \mathcal{Q} f + \Gamma_T^* G f(T)) \tag{9}$$

or equivalently

$$\begin{aligned} \mathcal{R}u &= -[\Gamma^* \mathcal{Q}(\Gamma u + f) + \Gamma_T^* G(\Gamma_T u + f(T))] \\ &= -[\Gamma^* \mathcal{Q} y + \Gamma_T^* G y(T)]. \end{aligned} \tag{10}$$

Replacing Γ^*, Γ_T^*, $\mathcal{Q}$, and $\mathcal{R}$ by their explicit forms yields (7). □

Another equivalent and useful expression for the open-loop optimal control u is given by the following corollary.

COROLLARY 2.2. *Under the same assumptions, the optimal control u is the unique solution of the following equation in $\mathcal{U}$:*

$$\begin{aligned} R(t)u(t) + \int_0^T L(t,\tau)u(\tau)\,d\tau \\ = -\left[K^*(T,t)Gf(T) + \int_t^T K^*(s,t)Q(s)f(s)\,ds\right], \qquad t \in [0,T], \end{aligned} \tag{11}$$

where

(12)

$$L(t,\tau)u = K^*(T,t)GK(T,\tau)u + \int_{\max(t,\tau)}^T K^*(\rho,t)Q(\rho)K(\rho,\tau)u\,d\rho, \qquad (t,\tau) \in \Delta.$$

Proof. This is simply a consequence of the variational equation

$$\Phi u = -(\Gamma^* \mathcal{Q} f + \Gamma_T^* G f(T))$$

and the concrete integral expression of the operator

$$[(\Gamma^* \mathcal{Q} \Gamma + \Gamma_T^* G \Gamma_T)u](t) = \int_0^T L(t,\tau)u(\tau)\,d\tau. \qquad \square$$

It follows that the minimum of $J(u, f)$ over $\mathcal{U}$ is given by

$$J_*(f) := \min_{u \in \mathcal{U}} J(u, f) = \left\langle \Psi \begin{pmatrix} f \\ f(T) \end{pmatrix}, \begin{pmatrix} f \\ f(T) \end{pmatrix} \right\rangle_{\mathcal{Y} \times Y}, \tag{13}$$

with

$$\Psi = \begin{bmatrix} \mathcal{Q} - \mathcal{Q}\Gamma\Phi^{-1}\Gamma^*\mathcal{Q} & -\mathcal{Q}\Gamma\Phi^{-1}\Gamma_T^* G \\ -G\Gamma_T\Phi^{-1}\Gamma^*\mathcal{Q} & G - G\Gamma_T\Phi^{-1}\Gamma_T^* G \end{bmatrix}.$$

The next task is to explore the possibility of producing a causal feedback optimal control.

3. Causal projections. In the open-loop relation (7) the real-time value of the optimal control $u(t)$, $t \in [0, T]$, is given in terms of the future-time values of the corresponding state trajectory $y = y(s; u, f)$, $t \leq s \leq T$. The key issue here is to convert such a noncausal dependence into a causal one. And as noted earlier, we are not able to proceed via a differential Riccati equation because the Volterra integral equation (1) may not have a state-space realization. We need, therefore, a different approach in order to treat the causality problem and a new way to accomplish the synthesis.

DEFINITION 3.1. *We define a truncation operator* π_ξ *by*

$$(\pi_\xi u)(t) = \left\{ \begin{array}{ll} u(t) & \text{for } t \in [0, \xi] \\ 0 & \text{for } t \in (\xi, T] \end{array} \right\}, \tag{14}$$

where $0 \leq \xi \leq T$ *is a parameter which can be chosen arbitrarily.*

Obviously, both π_ξ and $I - \pi_\xi$ are projection operators on the function space $\mathcal{U}$. They are idempotent and commute with the operators $\mathcal{R}$ and $\mathcal{R}^{-1}$. We shall call these projections *causal projections*. Similar truncations were introduced in the literature by Miller and Sell [11]; see also the books by Gripenberg, Londen, and Staffans [4] and Bensoussan et al. [1]. In the control literature they have been used by Ichikawa [7], Vinter and Kwong [13], and Delfour [5], [6]. We denote

$$\mathcal{U}_\xi^+ = \text{range } (I - \pi_\xi) = (I - \pi_\xi)\mathcal{U},$$

which is a closed subspace of $\mathcal{U}$, and define a parametrized operator Φ_ξ^+ by

$$\Phi_\xi^+ = (I - \pi_\xi)\Phi \,|\, \mathcal{U}_\xi^+. \tag{15}$$

It has the following properties.

LEMMA 3.2. *For any given* $\xi \in [0, T)$, *the operator* $\Phi_\xi^+ \in \mathcal{L}(\mathcal{U}_\xi^+)$ *is positive definite and self-adjoint. Moreover.*

$$\|(\Phi_\xi^+)^{-1}\|_{\mathcal{L}(\mathcal{U}_\xi^+)} \leq \text{ const } (\textit{uniform in } \xi \in [0, T)). \tag{16}$$

Proof. For any u and v in $\mathcal{U}_\xi^+$, we have

$$\begin{aligned} \langle \Phi_\xi^+ u, v\rangle_{\mathcal{U}_\xi^+} &= \langle \Phi_\xi^+ u, v\rangle_{\mathcal{U}} = \langle \Phi u, (I - \pi_\xi)v\rangle_{\mathcal{U}} = \langle \Phi u, v\rangle_{\mathcal{U}} = \langle u, \Phi v\rangle_{\mathcal{U}} \\ &= (\text{by tracing back in a similar way})\langle u, \Phi_\xi^+ v\rangle_{\mathcal{U}_\xi^+}. \end{aligned}$$

Moreover for any $u \in \mathcal{U}_\xi^+$,

$$\langle \Phi_\xi^+ u, u\rangle_{\mathcal{U}_\xi^+} = \langle \Phi_\xi^+ u, u\rangle_{\mathcal{U}} \geq \epsilon \|u\|_{\mathcal{U}}^2 = \epsilon \|u\|_{\mathcal{U}_\xi^+}^2.$$

Thus Φ_ξ^+ is boundedly invertible and $\|(\Phi_\xi^+)^{-1}\|_{\mathcal{L}(\mathcal{U}_\xi^+)} \leq 1/\epsilon$ for all $\xi \in [0, T)$. □

Note that Φ_T^+ is the zero operator, but $\mathcal{U}_T^+ = \{0\}$, and so we may also take the inverse $(\Phi_T^+)^{-1}$ as the zero operator. Hence, the above lemma can be extended to the case $\xi = T$.

DEFINITION 3.3. *For any state trajectory* $y(\cdot)$ *corresponding to an admissible control* $u(\cdot)$, *we define the* ξ*-causal trajectory* $y_\xi(\cdot)$ *by*

$$y_\xi(t) = f(t) + \int_0^t K(t, \tau)(\pi_\xi u)(\tau)\, d\tau, \qquad t \in [0, T]. \tag{17}$$

One can call a ξ-causal trajectory $y_\xi(\cdot)$ a semicausal trajectory because the control function is truncated, but not the function f. Note that (17) can be written as $y_\xi = f + \Gamma\pi_\xi u$, and it is obvious that

$$y = y_\xi + \Gamma(I - \pi_\xi)u, \qquad y(T) = y_\xi(T) + \Gamma_T(I - \pi_\xi)u. \tag{18}$$

DEFINITION 3.4. *For any $\xi \in [0, T]$, we define a ξ-evolutionary operator $N_\xi \in \mathcal{L}(\mathcal{Y} \times Y)$ by*

$$N_\xi = I_{\mathcal{Y}\times Y} - \begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(\Phi_\xi^+)^{-1}(I - \pi_\xi)(\Gamma^*\mathcal{Q}, \Gamma_T^* G). \tag{19}$$

Now we prove two important identities which will be used later in establishing the synthesis equations.

LEMMA 3.5. *The following hold:*

$$(\Phi_\xi^+)^{-1}(I - \pi_\xi)(\Gamma^*\mathcal{Q}, \Gamma_T^* G) = (I - \pi_\xi)\mathcal{R}^{-1}(\Gamma^*\mathcal{Q}, \Gamma_T^* G)N_\xi : \mathcal{Y} \times Y \to \mathcal{U}_\xi^+, \tag{20}$$

$$\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(\Phi_\xi^+)^{-1}(I - \pi_\xi) = N_\xi\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(I - \pi_\xi)\mathcal{R}^{-1} : \mathcal{U} \to \mathcal{Y} \times Y. \tag{21}$$

Proof. The verification of the two identities is straightforward, using the idempotent property of $(I - \pi_\xi)$ and the commutative property $(I - \pi_\xi)\mathcal{R}^{-1} = \mathcal{R}^{-1}(I - \pi_\xi)$. In detail, (20) follows from

$$\begin{aligned}
(I - \pi_\xi)\mathcal{R}^{-1}(\Gamma^*\mathcal{Q}, \Gamma_T^* G)N_\xi &= (I - \pi_\xi)\mathcal{R}^{-1}(\Gamma^*\mathcal{Q}, \Gamma_T^* G)\\
&\quad - (I - \pi_\xi)\mathcal{R}^{-1}(\Gamma^*\mathcal{Q}, \Gamma_T^* G)\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(\Phi_\xi^+)^{-1}(I - \pi_\xi)(\Gamma^*\mathcal{Q}, \Gamma_T^* G)\\
&= [\mathcal{R}^{-1} - \mathcal{R}^{-1}(I - \pi_\xi)(\Gamma^*\mathcal{Q}\Gamma + \Gamma_T^* G\Gamma_T)(\Phi_\xi^+)^{-1}](I - \pi_\xi)(\Gamma^*\mathcal{Q}, \Gamma_T^* G)\\
&= [\mathcal{R}^{-1} - \mathcal{R}^{-1}(I - \pi_\xi)(\Phi - \mathcal{R})(\Phi_\xi^+)^{-1}](I - \pi_\xi)(\Gamma^*\mathcal{Q}, \Gamma_T^* G)\\
&= [\mathcal{R}^{-1} - \mathcal{R}^{-1}\Phi_\xi^+(\Phi_\xi^+)^{-1} + (I - \pi_\xi)(\Phi_\xi^+)^{-1}](I - \pi_\xi)(\Gamma^*\mathcal{Q}, \Gamma_T^* G)\\
&= (\Phi_\xi^+)^{-1}(I - \pi_\xi)(\Gamma^*\mathcal{Q}, \Gamma_T^* G).
\end{aligned}$$

Equation (21) follows from

$$\begin{aligned}
N_\xi\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(I - \pi_\xi)\mathcal{R}^{-1} &= \begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(I - \pi_\xi)\mathcal{R}^{-1}\\
&\quad - \begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(\Phi_\xi^+)^{-1}(I - \pi_\xi)(\Gamma^*\mathcal{Q}, \Gamma_T^* G)\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(I - \pi_\xi)\mathcal{R}^{-1}\\
&= \begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}[\mathcal{R}^{-1} - (\Phi_\xi^+)^{-1}(I - \pi_\xi)(\Gamma^*\mathcal{Q}\Gamma + \Gamma_T^* G\Gamma_T)\mathcal{R}^{-1}](I - \pi_\xi)\\
&= \begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}[\mathcal{R}^{-1} - (\Phi_\xi^+)^{-1}(I - \pi_\xi)(\Phi - \mathcal{R})\mathcal{R}^{-1}](I - \pi_\xi)\\
&= \begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(\Phi_\xi^+)^{-1}(I - \pi_\xi). \qquad \square
\end{aligned}$$

4. Semicausal optimality principle. In the previous section, we defined several new concepts: π_ξ, Φ_ξ^+, N_ξ, and semicausal trajectories $y_\xi(\cdot)$. We will now study their interrelations, and it will turn out that the result can be regarded as a generalization of the well-known optimality principle. We call such a new result the *semicausal optimality principle*.

THEOREM 4.1. *Let $\xi \in [0, T]$ and $f \in \mathcal{Y}$ be given. The optimal state trajectory $y(\cdot)$ and the corresponding semicausal trajectory $y_\xi(\cdot)$ are related by*

$$\begin{pmatrix} y \\ y(T) \end{pmatrix} = N_\xi \begin{pmatrix} y_\xi \\ y_\xi(T) \end{pmatrix}. \tag{22}$$

Proof. Note that (22) is an equality in the space $\mathcal{Y} \times Y$. To show it, we substitute (18) into (10) to obtain

$$\mathcal{R}u + (\Gamma^* \mathcal{Q} \Gamma + \Gamma_T^* G \Gamma_T)(I - \pi_\xi)u = -(\Gamma^* \mathcal{Q} y_\xi + \Gamma_T^* G y_\xi(T)).$$

Premultiplying the above equality by $(I - \pi_\xi)$ yields

$$\Phi_\xi^+ (I - \pi_\xi)u = -(I - \pi_\xi)(\Gamma^* \mathcal{Q}, \Gamma_T^* G) \begin{pmatrix} y_\xi \\ y_\xi(T) \end{pmatrix}.$$

By Lemma 3.2, it follows that

$$(I - \pi_\xi)u = -(\Phi_\xi^+)^{-1}(I - \pi_\xi)(\Gamma^* \mathcal{Q}, \Gamma_T^* G) \begin{pmatrix} y_\xi \\ y_\xi(T) \end{pmatrix}. \tag{23}$$

Then, substituting (23) into (18), we see that (22) holds. □

This theorem shows that the entire optimal state trajectory $y(\cdot)$ is determined by the semicausal trajectory $y_\xi(\cdot)$. The latter depends *only* on the optimal control $u(t)$, $0 \le t \le \xi$, and the real-time information of the function $f(\cdot)$. Furthermore, we get an immediate synthesis result.

THEOREM 4.2. *$u(\cdot)$ is the optimal control if and only if*

$$u(t) = -R^{-1}(t)[(\Gamma^* \mathcal{Q})(t), (\Gamma_T^* G)(t)] N_t \begin{pmatrix} y_t \\ y_t(T) \end{pmatrix}, \qquad t \in [0, T], \tag{24}$$

where $y_t(\cdot)$ is the semicausal trajectory with parameter ξ equal to t and

$$(\Gamma^* \mathcal{Q} y)(t) = \int_t^T K^*(\tau, t) Q(\tau) y(\tau)\, d\tau, \quad y \in \mathcal{Y}, \quad (\Gamma_T^* G)(t)\phi = K^*(T, t) G \phi, \quad \phi \in Y. \tag{25}$$

Proof. From (10), (22), and the expressions for Γ^* and Γ_T^*, we see that

$$u(t) = -R^{-1}(t)[(\Gamma^* \mathcal{Q})(t), (\Gamma_T^* G)(t)] N_\xi \begin{pmatrix} y_\xi \\ y_\xi(T) \end{pmatrix}, \qquad t \in [0, T], \tag{26}$$

where u is the optimal control and $\xi \in [0, T]$ is arbitrary. The choice $\xi = t$ yields (24).

Conversely, in order to show that if $u(\cdot)$ satisfies (24), then it is optimal, we need only show that the pair $\{u(\cdot), y(\cdot, u, f)\}$ satisfying (24) is unique. Since both (1) and (24) are linear, it suffices to show that when $f = 0$, the unique solution of (24) is $u(t) = 0$, $t \in [0, T]$. Below we prove this. Since $f = 0$, we have $y_t = \Gamma \pi_t u$. If this pair satisfies (24), then

$$\|u(t)\|_U \le \|R^{-1}(t)\|_{\mathcal{L}(U)} \|(\Gamma^* \mathcal{Q})(t), (\Gamma_T^* G)(t)\|_{\mathcal{L}(\mathcal{Y} \times Y, U)} \|N_t\|_{\mathcal{L}(\mathcal{Y} \times Y)} \left\| \begin{pmatrix} y_t \\ y_t(T) \end{pmatrix} \right\|_{\mathcal{Y} \times Y}.$$

By (25), (19), and Lemma 3.2, we have

$$\|(\Gamma^*\mathcal{Q})(t), (\Gamma_T^*G)(t)\|_{\mathcal{L}(\mathcal{Y}\times Y,U)}\|N_t\|_{\mathcal{L}(\mathcal{Y}\times Y)} \leq \text{const (uniform in } t)$$

and

$$\left\|\begin{pmatrix} y_t \\ y_t(T)\end{pmatrix}\right\|_{\mathcal{Y}\times Y} = \left\|\begin{pmatrix}\Gamma \\ \Gamma_T\end{pmatrix}\pi_t u\right\|_{\mathcal{Y}\times Y} \leq (\text{const})\,\|\pi_t u\|_{\mathcal{U}},$$

where again the constant is uniform in t. Therefore

$$\|u(t)\|_U^2 \leq (\text{const})\,\|\pi_t u\|_{\mathcal{U}}^2 = (\text{const})\int_0^T \|\pi_t u(s)\|_U^2\,ds = (\text{const})\int_0^t \|u(s)\|_U^2\,ds.$$

Applying Gronwall's inequality it follows that $u(t) = 0$, $t \in [0, T]$, and this completes the proof. □

Although the expression (24) looks like a causal feedback, it involves the abstract operators N_t and $(\Phi_t^+)^{-1}$, and we really need more effort to reach a computable implementation. This will be the subject of the next section.

5. Feedback optimal control. Based on the result given in Theorem 4.2, we will now focus on the further manipulation of the abstract operator N_t. We want to convert it into another feedback gain operator which can be accessed in a computational manner. For this purpose we define, for $\xi \in [0, T]$, the operator $B_\xi(t, \tau) \in \mathcal{L}(U)$ by

$$B_\xi(t,\tau) = -R^{-1}(t)[(\Gamma^*\mathcal{Q})(t), (\Gamma_T^*G)(t)]N_\xi\begin{pmatrix}K(\cdot,\tau)\\ K(T,\tau)\end{pmatrix}, \qquad (t,\tau)\in\Omega. \tag{27}$$

Here $\Omega = [0, T]\times[0, T]$. In the following we will write operator equations without specifying their action on various functions. This is really an abuse of notation since the integrands will in general not be uniformly measurable. However, we feel that the proper interpretation is always clear.

LEMMA 5.1. *For any given $\xi \in [0, T]$, there exists a unique strongly continuous solution $B_\xi(t, \tau)$ of the following integral equation:*

$$B_\xi(t,\tau) + \int_\xi^T R^{-1}(t)L(t,s)B_\xi(s,\tau)\,ds = -R^{-1}(t)L(t,\tau), \qquad (t,\tau)\in\Omega, \tag{28}$$

where $L(t, \tau)$ is given by (12). *This solution is given by the expression in* (27) *and is such that*

$$\sup_{(\xi,t,\tau)\in[0,T]^3}\|B_\xi(t,\tau)\| < \infty. \tag{29}$$

Proof. First we show that the expression on the right-hand side of (27) really is a solution of (28), and then it is easy to see that the solution is strongly continuous and satisfies the equiboundedness condition (29). We have

$$\begin{aligned} B_\xi(t,\tau) &= -R^{-1}[(\Gamma^*\mathcal{Q})(t), (\Gamma_T^*G)(t)]N_\xi\begin{pmatrix}K(\cdot,\tau)\\ K(T,\tau)\end{pmatrix}\\ &= -R^{-1}(t)[(\Gamma^*\mathcal{Q})(t), (\Gamma_T^*G)(t)](I_{\mathcal{Y}\times Y}\\ &\quad -\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(\Phi_\xi^+)^{-1}(I-\pi_\xi)(\Gamma^*\mathcal{Q},\Gamma_T^*G)\begin{pmatrix}K(\cdot,\tau)\\ K(T,\tau)\end{pmatrix}\end{aligned}$$

$$= - R^{-1}(t)[(\Gamma^*\mathcal{Q})(t)K(\cdot,\tau) + (\Gamma_T^*G)(t)K(T,\tau)]$$

$$+ R^{-1}(t)[(\Gamma^*\mathcal{Q})(t)\Gamma + (\Gamma_T^*G)(t)\Gamma_T](\Phi_\xi^+)^{-1}(I-\pi_\xi)(\Gamma^*\mathcal{Q}, \Gamma_T^*G)\begin{pmatrix} K(\cdot,\tau) \\ K(T,\tau)\end{pmatrix}$$

$$= - R^{-1}(t)[(\Gamma^*\mathcal{Q})(t)K(\cdot,\tau) + (\Gamma_T^*G)(t)K(T,\tau)]$$

$$+ R^{-1}(t)[(\Gamma^*\mathcal{Q}(t)\Gamma + \Gamma_T^*G)(t)\Gamma_T](I-\pi_\xi)\mathcal{R}^{-1}(\Gamma^*\mathcal{Q}, \Gamma_T^*G)N_\xi\begin{pmatrix} K(\cdot,\tau) \\ K(T,\tau)\end{pmatrix}$$

(in this step (20) is used)

$$= - R^{-1}(t)L(t,\tau) - R^{-1}(t)[(\Gamma^*\mathcal{Q})(t)\Gamma + (\Gamma_T^*G)(t)\Gamma_T](I-\pi_\xi)B_\xi(\cdot,\tau)$$

$$= - R^{-1}(t)L(t,\tau) - \int_0^T R^{-1}(t)L(t,s)((I-\pi_\xi)B_\xi)(s,\tau)\,ds$$

$$= - R^{-1}(t)L(t,\tau) - \int_\xi^T R^{-1}(t)L(t,s)B_\xi(s,\tau)\,ds, \qquad (t,\tau)\in\Omega.$$

Next we prove the uniqueness. It is enough to show that the homogeneous operator equation

$$\tilde{B}_\xi(t,\tau) + \int_\xi^T R^{-1}(t)L(t,s)\tilde{B}_\xi(s,\tau)\,ds = 0, \qquad (t,\tau)\in\Omega, \tag{30}$$

admits only the null solution. Applying the operator $(I-\pi_\xi)\mathcal{R}$ to (30), we obtain

$$(I-\pi_\xi)(\mathcal{R}I + \Gamma^*\mathcal{Q}\Gamma + \Gamma_T^*G\Gamma_T)[(I-\pi_\xi)\tilde{B}_\xi(\cdot,\tau)] = 0 \quad \text{in} \quad \mathcal{U}_\xi^+,$$

i.e.,

$$\Phi_\xi^+(I-\pi_\xi)\tilde{B}_\xi(\cdot,\tau) = 0 \quad \text{in} \quad \mathcal{U}_\xi^+.$$

By Lemma 3.2, this reduces to

$$(I-\pi_\xi)\tilde{B}_\xi(\cdot,\tau) = 0, \qquad \tau\in[0,T].$$

So $\tilde{B}_\xi(t,\tau) = 0$ for $(t,\tau)\in(\xi,T]\times[0,T]$. Substituting this in (30) yields $\tilde{B}_\xi(t,\tau) = 0$ for $(t,\tau)\in\Omega$, and this completes the proof. □

Now we present the main results on feedback optimal control.

THEOREM 5.2. *Suppose that* (A1)–(A4) *hold. Then* $u(\cdot)$ *is the optimal control if and only if*

$$\begin{aligned} u(t) = &-R^{-1}(t)[(\Gamma^*\mathcal{Q})(t)y_t + (\Gamma_T^*G)(t)y_t(T)] \\ &- \int_t^T B_t(t,s)R^{-1}(s)[(\Gamma^*\mathcal{Q})(s)y_t + (\Gamma_T^*G)(s)y_t(T)]\,ds, \qquad t\in[0,T], \end{aligned} \tag{31}$$

where $B_\xi(t,\tau)$ *is the unique solution of the integral operator equation* (28) *of Fredholm type and* $y_t(\cdot)$ *is the corresponding real-time semicausal trajectory which depends only on the past information* $\{u(\tau): 0\le\tau<t\}$.

Proof. Suppose that $u(\cdot)$ is the optimal control. Then by (24), the definition (19) of N_t, and the identity (21), we have

$$
\begin{aligned}
u(t) &= -R^{-1}(t)[(\Gamma^*\mathcal{Q})(t), (\Gamma_T^*G)(t)]N_t\begin{pmatrix} y_t \\ y_t(T)\end{pmatrix} \\
&= -R^{-1}(t)[(\Gamma^*\mathcal{Q})(t), (\Gamma_T^*G)(t)]\begin{pmatrix} y_t \\ y_t(T)\end{pmatrix} \\
&\quad + R^{-1}(t)[(\Gamma^*\mathcal{Q})(t), (\Gamma_T^*G)(t)]\begin{pmatrix}\Gamma \\ \Gamma_T\end{pmatrix}(\Phi_t^+)^{-1}(I-\pi_t)(\Gamma^*\mathcal{Q}, \Gamma_T^*G)\begin{pmatrix} y_t \\ y_t(T)\end{pmatrix} \\
&= -R^{-1}(t)[(\Gamma^*\mathcal{Q})(t), (\Gamma_T^*G)(t)]\begin{pmatrix} y_t \\ y_t(T)\end{pmatrix} \\
&\quad + R^{-1}(t)[(\Gamma^*\mathcal{Q})(t), (\Gamma_T^*G)(t)]N_t\begin{pmatrix}\Gamma \\ \Gamma_T\end{pmatrix}(I-\pi_t)\mathcal{R}^{-1}(\Gamma^*\mathcal{Q}, \Gamma_T^*G)\begin{pmatrix} y_t \\ y_t(T)\end{pmatrix} \\
&= -R^{-1}(t)[(\Gamma^*\mathcal{Q})(t)y_t + (\Gamma_T^*G)(t)y_t(T)] \\
&\quad + R^{-1}(t)[(\Gamma^*\mathcal{Q})(t), (\Gamma_T^*G)(t)]N_t\int_t^T \begin{pmatrix} K(\cdot, s) \\ K(T, s)\end{pmatrix} \\
&\quad \times R(s)^{-1}[(\Gamma^*\mathcal{Q})(s)y_t + (\Gamma_T^*G)(s)y_t(T)]\,ds \\
&= -R^{-1}(t)[(\Gamma^*\mathcal{Q})(t)y_t + (\Gamma_T^*G)(t)y_t(T)] \\
&\quad - \int_t^T B_t(t, s)R(s)^{-1}[(\Gamma^*\mathcal{Q})(s)y_t + (\Gamma_T^*G)(s)y_t(T)]\,ds,
\end{aligned}
$$

where in the last step we used Lemma 5.1 and (27). Therefore, (31) is satisfied by the optimal control $u(\cdot)$ and its corresponding semicausal trajectory $y_{(\cdot)}$.

Conversely, for any given f, there is only one pair $\{u, y\}$ which satisfies the relation (31). This can be proven in a way similar to the proof of Theorem 4.2 by using Gronwall's inequality. The proof is omitted. □

Finally we replace the operators Γ^* and Γ_T^* with their explicit expressions and obtain the following synthesis result.

THEOREM 5.3. *Suppose that* (A1)–(A4) *hold. Then* $u(\cdot)$ *is the optimal control if and only if*

$$
u(t) = -H(T, t)Gy_t(T) - \int_0^T H(\tau, t)Q(\tau)y_t(\tau)\,d\tau, \qquad t \in [0, T], \tag{32}
$$

where y_t *is the corresponding semicausal trajectory, the feedback operator* $H(\cdot, \cdot)$ *is* independently *determined by*

$$
H(\tau, t) = R(t)^{-1}K^*(\tau, t) + \int_t^T B_t(t, s)R(s)^{-1}K^*(\tau, s)\,ds, \tag{33}
$$

and $B_t(t, s)$ *is the unique solution of the Fredholm integral equation* (28).

Proof. Substituting (25) into (31), we obtain

$$
\begin{aligned}
u(t) &= -R^{-1}(t)\left[K^*(T, t)Gy_t(T) + \int_t^T K^*(\tau, t)Q(\tau)y_t(\tau)\,d\tau\right] \\
&\quad - \int_t^T B_t(t, s)R^{-1}(s)\left[K^*(T, s)Gy_t(T) + \int_s^T K^*(\tau, s)Q(\tau)y_t(\tau)\,d\tau\right]ds.
\end{aligned} \tag{34}
$$

Note that we have a convention that $K(\tau, s) = K^*(\tau, s)$ for $(\tau, s) \in \Omega\backslash\Delta$. Interchanging the order of integration in the last term of (34) gives

$$\begin{aligned} u(t) &= -H(T,t)Gy_t(T) - \int_0^T R^{-1}(t)K^*(\tau,t)Q(\tau)y_t(\tau)\,d\tau \\ &\quad - \int_t^T B_t(t,s)R^{-1}(s)\int_0^T K^*(\tau,s)Q(\tau)y_t(\tau)\,d\tau\,ds \\ &= -H(T,t)Gy_t(T) - \int_0^T H(\tau,t)Q(\tau)y_t(\tau)\,d\tau. \end{aligned}$$

Conversely, since we can deduce (31) from (32), the control process satisfying (32) must be optimal by Theorem 5.2. □

COROLLARY 5.4. *$u(\cdot)$ is the optimal control if and only if it satisfies the following linear integral equation:*

(35)
$$u(t) = -H(T,t)Gf(T) - \int_0^T H(\tau,t)Q(\tau)f(\tau)\,d\tau - \int_0^t \Pi(t,s)u(s)\,ds, \qquad t \in [0,T],$$

where $H(\cdot\,,\cdot)$ is defined by (33) *and the operator $\Pi(\cdot\,,\cdot)$ is given by*

$$\Pi(t,\tau) = H(T,t)GK(T,\tau) + \int_0^T H(s,t)Q(s)K(s,\tau)\,ds. \tag{36}$$

Proof. By the definition (17) and the convention on K, we have

$$y_t(\tau) = f(\tau) + \int_0^t K(\tau,s)u(s)\,ds, \qquad \tau \in [0,T]. \tag{37}$$

Substituting (37) in (32) gives (35). □

Remark 5.5. We have achieved a feedback optimal control given by (32). In this closed-loop formula, there are two ingredients. One is the semicausal trajectory $y_t(\cdot)$ and its terminal value at time T but with the parameter at the real time t. The other is the feedback operator $H(\tau, t)$, given by (33), and essentially relies on another operator $B_\xi(\tau, t)$, which is obtained by solving the linear integral equation (28). This latter equation is totally independent of the function f. For this reason we call the Fredholm integral equation (28) the *synthesis equation*. It plays a role similar to that of the operator Riccati equation.

Since the general Volterra equation does not have a semigroup evolutionary property, the direct feedback implementation of the optimal control in terms of the actual trajectory $\{y(\tau) : 0 \le \tau \le t\}$ is not possible, because the future information of the function f is not counted. In view of this, the semicausal trajectory feedback is the best that can be hoped for. However, when the Volterra integral actually represents a mild solution of a linear evolution equation, our results incorporate both the regulator and the tracking problems. Indeed we are able to obtain

- the usual feedback via Riccati equations for the regulator problem,
- the additional input which solves the tracking problem.

The approach is similar to that carried out in [10], and the first is illustrated in the next section.

6. The reduction for controlled evolution equations. In this section we show that when the Volterra integral equation is given by the mild solution of a time-invariant evolution equation, our results reduce to the well-known ones in terms of Riccati equations. Specifically we assume that

$$\dot{y}(t) = Ay(t) + Bu(t), \quad 0 \le t \le T, \quad y(0) = y_0 \in Y,$$

where A generates a strongly continuous semigroup $S(t)$, $t \ge 0$, on Y and $B \in \mathcal{L}(U, Y)$ so that

$$f(t) = S(t)y_0, \qquad K(t, \tau) = S(t - \tau)B. \tag{38}$$

Consider the differential Riccati equation

$$\langle \dot{\mathbb{P}}(t)x, y\rangle + \langle \mathbb{P}(t)x, Ay\rangle + \langle Ax, \mathbb{P}(t)y\rangle + \langle Qx, y\rangle - \langle \mathbb{P}(t)BR^{-1}B^*\mathbb{P}(t)x, y\rangle = 0, \quad \mathbb{P}(T) = G,$$

where $t \in [0, T]$, $x, y \in D(A)$, and we have assumed for simplicity that Q and R are time invariant. We will prove that the optimal control as given in Theorem 5.3 is $u(t) = -R^{-1}B^*\mathbb{P}(t)y(t)$, $t \in [0, T]$. The proof of this reduction will be given in detail so that one can see the correspondence between the standard approach and the one presented here. In this way more insight is obtained into the operator method and causality argument that we have used for Volterra integral equations.

LEMMA 6.1. *If* (38) *holds for* (1), *then the optimal control* $u(\cdot)$ *is given by*

$$u(t) = -E(t, t)y(t), \qquad t \in [0, T], \tag{39}$$

where $y(\cdot)$ *is the corresponding trajectory*

$$E(t, \tau) = H(T, t)GS(T - \tau) + \int_t^T H(s, t)QS(s - \tau)\, ds, \qquad (t, \tau) \in [0, T]^2, \tag{40}$$

$H(\cdot, \cdot)$ *is defined by* (33), *and we have the convention* $S(t) = 0$ *whenever* $t < 0$.

Proof. By (17),

$$\begin{aligned} y_\xi(t) &= S(t)y_0 + \int_0^t S(t - \tau)B\pi_\xi u(\tau)\, d\tau, \qquad t \ge 0, \\ &= y(t), \qquad t \le \xi, \\ &= S(t - \xi)S(\xi)y_0 + S(t - \xi)\int_0^\xi S(\xi - \tau)Bu(\tau)\, d\tau, \qquad t > \xi, \\ &= S(t - \xi)y(\xi), \qquad t > \xi. \end{aligned}$$

Hence

$$y_\xi(t) = \begin{cases} y(t), & t \le \xi, \\ S(t - \xi)y(\xi), & t > \xi, \end{cases} \tag{41}$$

and the result follows by substitution of (41) in (32). □

Using (38), the expressions (12) and (33) for $L(\cdot, \cdot)$ and $H(\cdot, \cdot)$ take the form

$$\begin{aligned} L(t, \tau) &= B^*S^*(T - t)GS(T - \tau)B + B^*\int_{\max(t,\tau)}^T S^*(s - t)QS(s - \tau)B\, ds \\ &= B^*S^*(T - t)GS(T - \tau)B + B^*\int_t^T S^*(s - t)QS(s - \tau)B\, ds \end{aligned}$$

and

$$\begin{aligned} H(\tau, t) &= R^{-1}B^*S^*(\tau - t) + \int_t^T B_t(t, s)R^{-1}B^*S^*(\tau - s)\, ds \\ &= R^{-1}B^*S^*(\tau - t) + \int_t^\tau B_t(t, s)R^{-1}B^*S^*(\tau - s)\, ds. \end{aligned} \tag{42}$$

We define an operator $W(\cdot, \cdot)$ by

(43)
$$W(t,\xi) = S(t-\xi) - \Gamma(t)(\Phi_\xi^+)^{-1}(I-\pi_\xi)(\Gamma^*Q, \Gamma_T^*G)\begin{pmatrix} S(\cdot-\xi) \\ S(T-\xi)\end{pmatrix}, \qquad 0 \le \xi \le t \le T,$$

where $W(T,T) = I$. We have the convention that $W(t,\xi) = 0$ for $t < \xi$ and with slight abuse $\Gamma(t)u = (\Gamma u)(t)$, $u \in \mathcal{U}$.

LEMMA 6.2. *The operator W has the following properties:*

(i) *for any initial data, the optimal trajectory $y(\cdot)$ satisfies*

(44)
$$y(t) = W(t,\xi)y(\xi), \qquad 0 < \xi \le t \le T.$$

(ii) $\sup_{0\le\xi\le t\le T} \|W(t,\xi)\|_{\mathcal{L}(Y)} < \infty$.

(iii) $W(t,\eta)W(\eta,\xi) = W(t,\xi), \qquad 0 < \xi \le \eta \le t \le T.$

(iv) *$W(t,\xi)$ is strongly continuous in $t \in [\xi, T]$ and in $\xi \in (0,t]$, respectively.*

Proof. To prove (i) we note the following facts:

$$\begin{aligned} y &= y_\xi + \Gamma(I-\pi_\xi)u, \\ (I-\pi_\xi)u &= -(\Phi_\xi^+)^{-1}(I-\pi_\xi)(\Gamma^*Qy_\xi + \Gamma_T^*Gy_\xi(T)), \\ y_\xi(t) &= S(t-\xi)y(\xi), \qquad 0 < \xi \le t \le T. \end{aligned}$$

Hence

$$\begin{aligned} y(t) &= S(t-\xi)y(\xi) - \Gamma(t)(\Phi_\xi^+)^{-1}(I-\pi_\xi)(\Gamma^*Q, \Gamma_T^*G)\begin{pmatrix} S(\cdot-\xi) \\ S(T-\xi)\end{pmatrix}y(\xi) \\ &= W(t,\xi)y(\xi), \qquad (t,\xi) \in \{0 < \xi \le t \le T : \xi \ne T\}. \end{aligned}$$

For $\xi = t = T$, (44) holds since $W(T,T) = I$.

The proof of (ii) follows from Lemma 3.2.

For (iii), let $0 < \xi \le \eta \le t \le T$. By using (43) and (19), we find

$$\begin{aligned} W(t,\eta)W(\eta,\xi) &= S(t-\xi) - S(t-\eta)\Gamma(\eta)(\Phi_\xi^+)^{-1}(I-\pi_\xi)(\Gamma^*Q, \Gamma_T^*G)\begin{pmatrix} S(\cdot-\xi) \\ S(T-\xi)\end{pmatrix} \\ &\quad - \Gamma(t)(\Phi_\eta^+)^{-1}(I-\pi_\eta)(\Gamma^*Q, \Gamma_T^*G)\begin{pmatrix} S(\cdot-\eta) \\ S(T-\eta)\end{pmatrix}S(\eta-\xi) \\ &\quad + \Gamma(t)(\Phi_\eta^+)^{-1}(I-\pi_\eta)(\Gamma^*Q, \Gamma_T^*G)\begin{pmatrix} S(\cdot-\eta) \\ S(T-\eta)\end{pmatrix} \\ &\quad \times \Gamma(\eta)(\Phi_\xi^+)^{-1}(I-\pi_\xi)(\Gamma^*Q, \Gamma_T^*G)\begin{pmatrix} S(\cdot-\xi) \\ S(T-\xi)\end{pmatrix} \\ &= S(t-\xi) - S(t-\eta)\Gamma(\eta)(I-\pi_\xi)R^{-1}(\Gamma^*Q, \Gamma_T^*G)N_\xi\begin{pmatrix} S(\cdot-\xi) \\ S(T-\xi)\end{pmatrix} \\ &\quad - \Gamma(t)(I-\pi_\eta)R^{-1}(\Gamma^*Q, \Gamma_T^*G)N_\eta \\ &\quad \times \left\{I - \begin{pmatrix} S(\cdot-\eta) \\ S(T-\eta)\end{pmatrix}\Gamma(\eta)(I-\pi_\xi)R^{-1}(\Gamma^*Q, \Gamma_T^*G)N_\xi\right\}\begin{pmatrix} S(\cdot-\xi) \\ S(T-\xi)\end{pmatrix}. \end{aligned}$$

By (19), (20), and (21), we have

$$\begin{aligned}
N_\eta - N_\xi &= N_\eta(I - N_\xi) - (I - N_\eta)N_\xi \\
&= N_\eta\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(I - \pi_\xi)R^{-1}(\Gamma^*Q, \Gamma_T^*G)N_\xi - N_\eta\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(I - \pi_\eta)R^{-1}(\Gamma^*Q, \Gamma_T^*G)N_\xi \\
&= N_\eta\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(\pi_\eta - \pi_\xi)R^{-1}(\Gamma^*Q, \Gamma_T^*G)N_\xi \\
&= N_\eta\begin{pmatrix}S(\cdot - \xi)\\ S(T - \xi)\end{pmatrix}\Gamma(\eta)(I - \pi_\xi)R^{-1}(\Gamma^*Q, \Gamma_T^*G)N_\xi.
\end{aligned}$$

Substituting this into the parenthesis $\{\ldots\}$ of the last term in the above expression, we have

$$\begin{aligned}
W(t,\eta)W(\eta,\xi) &= S(t - \xi) - S(t - \eta)\Gamma(\eta)(I - \pi_\xi)R^{-1}(\Gamma^*Q, \Gamma_T^*G)N_\xi\begin{pmatrix}S(\cdot - \xi)\\ S(T - \xi)\end{pmatrix} \\
&\quad - \Gamma(t)(I - \pi_\eta)R^{-1}(\Gamma^*Q, \Gamma_T^*G)N_\xi\begin{pmatrix}S(\cdot - \xi)\\ S(T - \xi)\end{pmatrix} \\
&= S(t - \xi) - \int_\xi^\eta S(t - \sigma)BR^{-1}((\Gamma^*Q)(\sigma), (\Gamma_T^*G)(\sigma))N_\xi\begin{pmatrix}S(\cdot - \xi)\\ S(T - \xi)\end{pmatrix}d\sigma \\
&\quad - \int_\eta^t S(t - \sigma)BR^{-1}((\Gamma^*Q)(\sigma), (\Gamma_T^*G)(\sigma))N_\xi\begin{pmatrix}S(\cdot - \xi)\\ S(T - \xi)\end{pmatrix}d\sigma \\
&= S(t - \xi) - \Gamma(t)(I - \pi_\xi)R^{-1}(\Gamma^*Q, \Gamma_T^*G)N_\xi\begin{pmatrix}S(\cdot - \xi)\\ S(T - \xi)\end{pmatrix} \\
&= S(t - \xi) - \Gamma(t)(\Phi_\xi^+)^{-1}(I - \pi_\xi)(\Gamma^*Q, \Gamma_T^*G)\begin{pmatrix}S(\cdot - \xi)\\ S(T - \xi)\end{pmatrix} \\
&= W(t,\xi)
\end{aligned}$$

for any $0 < \xi \le \eta \le t \le T$ with $\xi < T, \eta < T$, but if $\xi < T$ and $\eta = t = T$, then the equality remains valid, viz. $W(T,T)W(T,\xi) = W(T,\xi)$. Clearly it is true for $\xi = \eta = t = T$, and hence (iii) is proven.

For (iv), by the definition of W it is clear that $W(t,\xi)$ is strongly continuous in $t \in [\xi, T]$. Using this continuity and properties (ii) and (iii), it is easy to prove the strong continuity in $\xi \in (0, t]$. The detail is omitted. □

LEMMA 6.3. *Let $P : [0,T] \to \mathcal{L}(Y)$ be defined by*

$$\begin{aligned}
P(t) &= S^*(T - t)GS(T - t) + \int_t^T S^*(s - t)QS(s - t)\,ds \\
&= -((\tilde{\Gamma}^*Q)(t), (\tilde{\Gamma}_T G)(t))N_t\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(I - \pi_t)R^{-1}(\Gamma^*Q, \Gamma_T^*G)\begin{pmatrix}S(\cdot - t)\\ S(T - t)\end{pmatrix},
\end{aligned} \tag{45}$$

where the new operators $\tilde{\Gamma}^* \in \mathcal{L}(\mathcal{Y})$ and $\tilde{\Gamma}_T^* \in \mathcal{L}(Y, \mathcal{Y})$ are defined by

$$(\tilde{\Gamma}^*\phi)(t) = \int_t^T S^*(\sigma - t)\phi(\sigma)d\sigma, \quad (\tilde{\Gamma}_T^*)(t)\phi_0 = S^*(T - t)\phi_0, \quad t \in [0,T].$$

Then if E is given by (40), *we have*

$$E(t,t) = R^{-1}B^*P(t), \qquad t \in [0,T]. \tag{46}$$

Proof. From (27), (40), and (42), we have

$$
\begin{aligned}
E(t,t) &= H(T,t)GS(T-t) + \int_t^T H(s,t)QS(s-t)\,ds \\
&= R^{-1}B^*\left[S^*(T-t)GS(T-t) + \int_t^T S^*(s-t)QS(s-t)\,ds\right] \\
&\quad - R^{-1}B^*\int_t^T ((\tilde{\Gamma}^*\mathcal{Q})(t), (\tilde{\Gamma}_T^*G)(t))N_t\begin{pmatrix} S(\cdot-\sigma) \\ S(T-\sigma)\end{pmatrix} \\
&\quad \times BR^{-1}B^*S^*(T-\sigma)GS(T-t)\,d\sigma \\
&\quad - R^{-1}B^*\int_t^T\int_t^s ((\tilde{\Gamma}^*\mathcal{Q})(t), (\tilde{\Gamma}_T^*G)(t))N_t\begin{pmatrix} S(\cdot-\sigma) \\ S(T-\sigma)\end{pmatrix} \\
&\quad \times BR^{-1}B^*S^*(s-\sigma)QS(s-t)d\sigma\,ds \\
&= R^{-1}B^*\left[S^*(T-t)GS(T-t) + \int_t^T S^*(s-t)QS(s-t)\,ds\right] \\
&\quad - R^{-1}B^*((\tilde{\Gamma}\mathcal{Q})(t), (\tilde{\Gamma}_T^*G)(t))N_t\int_t^T \begin{pmatrix} S(\cdot-\sigma) \\ S(T-\sigma)\end{pmatrix} \\
&\quad \times BR^{-1}B^*\left[S^*(T-\sigma)GS(T-t) + \int_\sigma^T S^*(s-\sigma)QS(s-t)\,ds\right]d\sigma \\
&= R^{-1}B^*\left[S^*(T-t)GS(T-t) + \int_t^T S^*(s-t)QS(s-t)\,ds\right] \\
&\quad - R^{-1}B^*((\tilde{\Gamma}^*Q)(t), (\tilde{\Gamma}_T^*G)(t))N_t\begin{pmatrix}\Gamma \\ \Gamma_T\end{pmatrix}(I-\pi_t)R^{-1}(\Gamma^*Q, \Gamma_T^*G)\begin{pmatrix} s(\cdot-t) \\ S(T-t)\end{pmatrix},
\end{aligned}
$$

which completes the proof. □

LEMMA 6.4. *Let P and W be defined by* (45) *and* (43), *respectively. Then*

$$
P(t)y = S^*(T-t)GW(T,t) + \int_t^T S^*(s-t)QW(s,t)\,ds, \qquad t\in[0,T]. \tag{47}
$$

Proof. From (43) and (20), it follows that

$$
\begin{aligned}
&S^*(T-t)GW(T,t) + \int_t^T S^*(s-t)QW(s,t)\,ds \\
&\quad = S^*(T-t)GS(T-t) + \int_t^T S^*(s-t)QS(s-t)\,ds \\
&\qquad - S^*(T-t)G\Gamma(T)(\Phi_t^+)^{-1}(I-\pi_t)(\Gamma^*Q, \Gamma_T^*G)\begin{pmatrix} S(\cdot-t) \\ S(T-t)\end{pmatrix} \\
&\qquad - \int_t^T S^*(s-t)Q\Gamma(s)(\Phi_t^+)^{-1}(I-\pi_t)(\Gamma^*Q, \Gamma_T^*G)\begin{pmatrix} S(\cdot-t) \\ S(T-t)\end{pmatrix}ds \\
&\quad = S^*(T-t)GS(T-t) + \int_t^T S^*(s-t)QS(s-t)\,ds
\end{aligned}
$$

$$
\begin{aligned}
&- ((\tilde{\Gamma}^* Q)(t), (\tilde{\Gamma}_T^* G)(t))\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(\Phi_t^+)^{-1}(I-\pi_t)(\Gamma^* Q, \Gamma_T^* G)\begin{pmatrix}S(\cdot - t)\\ S(T-t)\end{pmatrix}\\
&= S^*(T-t)GS(T-t) + \int_t^T S^*(s-t)QS(s-t)\,ds\\
&\quad - ((\tilde{\Gamma}^* Q)(t), (\tilde{\Gamma}_T^* G)(t))N_t\begin{pmatrix}\Gamma\\ \Gamma_T\end{pmatrix}(I-\pi_t)R^{-1}(\Gamma^* Q, \Gamma_T^* G)\begin{pmatrix}S(\cdot - t)\\ S(T-t)\end{pmatrix}\\
&= P(t), \qquad t \in [0, T]. \qquad \square
\end{aligned}
$$

LEMMA 6.5. *The relation*

$$
W(t,\xi) = S(t-\xi) - \int_\xi^t W(t,\sigma)BR^{-1}B^*P(\sigma)S(\sigma-\xi)\,d\sigma \tag{48}
$$

holds, and P is the unique strongly continuous and self-adjoint solution of the integral Riccati equation

(49)

$$
P(t) = S^*(T-t)GS(T-t) + \int_t^T S^*(s-t)[Q - P(s)BR^{-1}B^*P(s)]S(s-t)\,ds, \qquad t \in [0, T].
$$

Proof. First we prove (48). Let the function on the right of (48) be denoted by $\theta(t,\xi)$. Then it can be verified that both

$$
g_1(t) = W(t,\xi) \quad \text{and} \quad g_2(t) = \theta(t,\xi), \quad t \in (\xi, T],
$$

are continuous solutions of the Volterra integral equation

$$
g(t) = S(t-\xi) - \int_\xi^t S(t-\sigma)BR^{-1}B^*P(\sigma)g(\sigma)\,d\sigma, \qquad t \in [\xi, t].
$$

Then by the uniqueness of the solution, (48) is valid.

Now we prove (49). By (47) and (48), for $t \in [0, T]$, we have

$$
\begin{aligned}
P(t) &= S^*(T-t)GW(T,t) + \int_t^T S^*(s-t)QW(s,t)\,ds\\
&= S^*(T-t)GS(T-t) + \int_t^T S^*(s-t)QS(s-t)\,ds\\
&\quad - S^*(T-t)G\int_t^T W(T,s)BR^{-1}B^*P(s)S(s-t)\,ds\\
&\quad - \int_t^T S^*(\eta-t)Q\int_t^\eta W(\eta,s)P(s)BR^{-1}B^*P(s)S(s-t)\,ds\,d\eta\\
&= S^*(T-t)GS(T-t) + \int_t^T S^*(s-t)QS(s-t)\,ds\\
&\quad - \int_t^T S^*(s-t)BR^{-1}B^*P(s)S(s-t)\,ds\\
&= S^*(T-t)GS(T-t) + \int_t^T S^*(s-t)[Q - P(s)BR^{-1}B^*P(s)]S(s-t)\,ds.
\end{aligned}
$$

By (47) and Lemma 6.2(iv), we see that $P(\cdot)$ is strongly continuous. In [2], it is proven that the strongly continuous solution of (49) is unique. By transposition, it can be seen that $P^*(\cdot)$ is also a strongly continuous solution of (49). Hence, by uniqueness, $P^*(t) = P(t)$, $t \in [0, T]$, and the proof is complete. □

We have finally reached our goal in the reduction. This is stated in the following theorem.

THEOREM 6.6. *Suppose that* (A3) *and* (A4) *hold for the data given by* (38). *Then the optimal control is given by*

$$u(t) = -R^{-1}B^*\mathbb{P}(t)y(t), \qquad t \in [0, T], \tag{50}$$

where $\mathbb{P}$ *is the unique strongly continuous and self-adjoint solution of the differential Riccati equation and* $y(\cdot)$ *is the corresponding trajectory.*

Proof. This follows from Lemmas 6.1, 6.3, and 6.5 and the known fact that the integral Riccati equation (49) is equivalent to the aforementioned differential Riccati equation and hence $P = \mathbb{P}$. □

REFERENCES

[1] A. BENSOUSSAN, G. DA PRATO, M. DELFOUR, AND S. MITTER, *Representation and Control of Infinite Dimensional Systems*, Birkhäuser, Boston, 1992.

[2] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Inform. Sci. 8, Springer-Verlag, New York, 1978.

[3] R. F. CURTAIN AND H. ZWART, *An Introduction to Infinite Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.

[4] G. GRIPENBERG, S. O. LONDEN, AND O. J. STAFFANS, *Volterra Integral and Functional Equations*, Cambridge University Press, Cambridge, UK, 1990.

[5] M. DELFOUR, *Linear optimal control with state and control variable delays*, Automatica, 20 (1984), pp. 69–77.

[6] ———, *The linear quadratic optimal control problem with delays in state and control variables: A state space approach*, SIAM J. Control Optim., 24 (1986), pp. 835–883.

[7] A. ICHIKAWA, *Optimal Control and Filtering of Evolution Equations with Delay in Control and Observation*, Control Theory Centre Report 53, University of Warwick, 1977.

[8] T. KACZOREK, *Two Dimensional Linear Systems*, Lecture Notes in Control and Inform. Sci. 68, Springer-Verlag, New York, 1985.

[9] I. LASIECKA AND R. TRIGGIANI, *Differential and Algebraic Riccati Equations with Application to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*, Lecture Notes in Control and Inform. Sci. 164, Springer-Verlag, New York, 1991.

[10] E. B. LEE AND Y. YOU, *Quadratic optimisation for infinite dimensional linear differential difference type systems: Synthesis via the Fredholm equations*, SIAM J. Control Optim., 28 (1990), pp. 265–293.

[11] R. MILLER AND G. SELL, *Volterra integral equations and topological dynamics*, Mem. Amer. Math. Soc., 102 (1970).

[12] A. J. PRITCHARD AND D. SALAMON, *The linear quadratic problem for infinite dimensional systems with unbounded input and output operators*, SIAM J. Control Optim., 25 (1987), pp. 121–144.

[13] R. VINTER AND R. KWONG, *The infinite time quadratic control problem for linear systems with state and control delays, an evolution equation approach*, SIAM J. Control Optim., 19 (1981), pp. 139–153.

[14] Y. YOU, *Quadratic differential games of general linear systems*, J. Math. Pures Appl., 69 (1990), pp. 261–283.

[15] ———, *Optimal control for linear systems with quadratic indefinite criterion on Hilbert spaces*, Chinese Ann. Math. Ser. B, 4 (1983), pp. 21–31.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 1891–1913, December 1996

004

ON THE USE OF CONSISTENT APPROXIMATIONS FOR THE OPTIMAL DESIGN OF BEAMS*

C. KIRJNER NETO† AND E. POLAK†

Abstract. This paper presents a discretization strategy, based on the concept of consistent approximations, for certain optimal beam design problems, where the beam is modeled using Euler–Bernoulli beam theory. It is shown that any accumulation point of the sequence of the stationary points of the family of resulting approximating problems is a stationary point of the original, infinite-dimensional problem. The construction of approximating problems requires the development of a relaxation of constraints to ensure existence of solutions. The numerical solution of the approximating problems, by means of nonlinear programming algorithms that are not scale invariant, must be preceded by a change of variables to guard against deterioration of performance. The use of such approximating problems, in conjunction with a diagonalization strategy, is illustrated by a numerical example.

Key words. optimal design, discretization theory, epiconvergence, consistent approximations, algorithm convergence theory

AMS subject classifications. Primary, 49Q10; Secondary, 65K10

1. Introduction.

Over the past 15 years, there has been a great deal of activity in the development of numerical methods for the solution of optimal design problems (see, e.g., [4, 5, 9, 10, 12, 16] and references therein). A major class of these methods is based on the construction of an infinite sequence of finite-dimensional approximating problems by means of numerical integration techniques. To be of any value, such approximating problems must be consistent in the sense that their solutions converge to those of the original problem. Now, in the context of optimization problems, the concept of a solution is not unique; one may mean a *global* minimizer, a *local* minimizer, or, quite commonly, simply a *stationary point*.

A scan through the optimal design literature (see, e.g., [6, 9, 13, 16]) shows that only the question of convergence of global minimizers of approximating problems to a global minimizer of the original problem is usually addressed. However, it has been observed empirically (see [7]) that some numerical methods used to solve boundary value problems result in approximating problems with local minimizers that converge to nonstationary points of the original problem. A simple example of how such a pathology can occur is given in §2.

The impact of the multiplicity of solution concepts in optimization on the issue of consistency of approximation is only now beginning to be recognized. In [18] we find an abstract theory of consistent approximations for optimization problems. This theory provides guidance to the construction of approximating finite-dimensional optimization problems and to the use of efficient diagonalization techniques for constructing sequences of finite-dimensional approximating problems in the approximate solution of an infinite-dimensional optimization problem.[1] Within this theory, optimization problems are not endowed with a specific structure, and hence optimality conditions are expressed in terms of zeros of *optimality functions*.[2] Approximating problems are judged to be consistent if the constrained epigraphs of their cost functions and the hypographs of their optimality functions converge, in the Kuratowski sense, to those of the original problem. Consistency in the sense of [18] ensures the convergence of the global

*Received by the editors April 18, 1994; accepted for publication (in revised form) July 13, 1995. This research was supported by Air Force Office of Scientific Research contract AFOSR-90-0068 and National Science Foundation grant ECS-8916168.

†Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720.

[1]Referring to [11], we see that using a diagonalization strategy in getting an approximation at specified precision to a solution of an infinite-dimensional optimization problem is always more efficient than simply solving an approximating problem of the given precision.

[2]Optimality functions are usually the value functions of quadratic optimization problems that arise when one attempts to verify whether a classical optimality condition is satisfied or not.

minimizers, local minimizers, and stationary points of approximating problems to corresponding points of the original problem.

To apply the results in [18] to optimal design, one must carry out three tasks. The first, and often the easiest, is the selection of bases for finite-dimensional subspaces of the original infinite-dimensional design space. Next, one must show that the selected numerical integration method converges with the same rate *on all* the finite-dimensional subspaces used, uniformly in the decision variables. This fact shows that there is an important interaction between the selection of the subspaces and the selection of the method of integration (see, e.g. [22]). The literature on numerical integration usually does not include such results, and their development can be a serious source of difficulty. Finally, to ensure that the approximating problems have solutions, one often has to invent a satisfactory relaxation of the constraints.

To date, the applicability of the theory of consistent approximations in [18] has been established for discrete optimal control problems [19, 22] resulting from the use of Euler and Runge–Kutta methods of integration on ordinary differential equations.

This paper uses the theory of consistent approximations in developing an approach to the solution of a class of optimal Euler–Bernoulli beam design problems with continuum constraints, such as constraints on vertical deflection, shear stress, and normal stress at the extreme fiber. The continuum constraints make these problems nondifferentiable and hence quite difficult to solve.

In the process of developing consistent approximations for the beam problems, we have addressed two important issues that are usually ignored. The first is that of finding a rational method for relaxing constraints so as to ensure that the approximating problems have solutions. The second issue is that of preserving problem conditioning. It is well known in nonlinear programming that changing norms can have a profoundly adverse effect on the behavior of algorithms, such as the phase I–phase II method of centers in [20], that are not scale invariant. When the basis functions used for the finite-dimensional subspaces on which the approximating problems are defined are not orthonormal, the corresponding Euclidean coefficient spaces are not isometric with the function subspaces. Hence, if one compares the behavior of a nonscale invariant algorithm in solving an approximating problem formulated in a finite-dimensional subspace of the original function space using the original norm, with its behavior in solving the same approximating problem formulated in the space of coefficients using the Euclidean norm, one often finds that in the Euclidean space this algorithm performs much worse. To correct for this possibility, one must develop appropriate transformations in Euclidean space, as we have done in this paper.

For ease of exposition we will restrict ourselves to beams with rectangular cross section, fixed width, and distributed loads. Although beams with nonuniform cross sections are more difficult to manufacture when weight is at a premium, as in aerospace applications, the construction of minimum weight beams may be quite realistic. Moreover, the problem of determining the optimal dimensions of a uniform beam subject to continuum constraints is a particular case of the problems with which we will deal. It is straightforward to generalize our results to beams whose cross sections are not necessarily rectangular, provided that the cross sections have horizontal and vertical axes of symmetry and the plane containing the vertical axis of symmetry also contains the loads. For instance, one can extend our results to the design of rectangular beams with varying depth and width or the design of a cylindrical beam with varying radius.

The paper is organized as follows. In §2 we summarize the basic definitions and results on consistent approximations in [18]. In §3 we develop a mathematical formulation of the optimal design problem together with an optimality function. Then we construct a sequence of

approximating problems, together with their optimality functions, which we show to be consistent in the sense of [18]. In §4 we develop transformations of variables that compensate for the fact that our coefficient spaces are not isometric with the corresponding finite-dimensional function subspaces. Also, we present a diagonalization strategy for the numerical solution of the optimal design problems under consideration. In §5 we present the results of a numerical experiment. Finally, in §6 we present our conclusions.

2. Consistent approximations. We begin by presenting a summary of the main definitions and results related to the concept of consistent approximations introduced in [18].

Let B be a topological vector space, and consider the problem

$$\mathbf{P} \qquad \min_{z \in Z} f(z) \tag{2.1a}$$

where $f : \mathsf{B} \to \mathbb{R}$ is continuous and $Z \subset \mathsf{B}$ is the constraint set. Let $\{\mathsf{B}_N\}_{N=1}^{\infty}$ be a family of finite-dimensional subspaces of B such that $\mathsf{B}_N = \mathsf{B}$ if B is finite dimensional ($\mathbb{R}^n$) and $\mathsf{B}_N \subset \mathsf{B}_{N+1}$ for all N otherwise. Consider the family of approximating problems

$$\mathbf{P}_N \qquad \min_{z \in Z_N} f_N(z), \qquad N \in \mathbb{N}, \tag{2.1b}$$

where $f_N : \mathsf{B}_N \to \mathbb{R}$ is continuous and $Z_N \subset \mathsf{B}_N$. To be of any use to us at all, the problems $\mathbf{P}_N$ must, at least, converge epigraphically to $\mathbf{P}$; i.e., the epigraphs $E_N \triangleq \{(z^0, z) \in \mathbb{R} \times Z_N | z^0 \geq f_N(z)\}$ of the problems $\mathbf{P}_N$ must converge in the sense of Kuratowski to the epigraph $E \triangleq \{(z^0, z) \in \mathbb{R} \times Z | z^0 \geq f(z)\}$ of the problem $\mathbf{P}$. Equivalently, we have the following definition.

DEFINITION 2.1 (see [1, 8]). *The problems in the family* $\{\mathbf{P}_N\}_{N=1}^{\infty}$ converge epigraphically *to* $\mathbf{P}$, ($\mathbf{P}_N \to^{\text{Epi}} \mathbf{P}$) *if* (a) *for every* $z \in Z$ *there exists a sequence* $\{z_N\}_{N=1}^{\infty}$ *with* $z_N \in Z_N$ *such that* $z_N \to z$ *and* $\lim f_N(z_N) \leq f(z)$ *and* (b) *for every sequence* $\{z_{N_k}\}_{k=1}^{\infty}$ *with* $z_{N_k} \in Z_{N_k}$ *such that* $z_{N_k} \to z$ *as* $k \to \infty$, $z \in Z$, *and* $\lim f_{N_k}(z_{N_k}) \geq f(z)$.

Epigraphic convergence, or epiconvergence for short, can be viewed as a "zeroth-order" consistency property. In particular, it ensures the following result.

THEOREM 2.2. *Suppose that* $\mathbf{P}_N \to^{\text{Epi}} \mathbf{P}$ *and that* $\{\hat{z}_N\}_{N=1}^{\infty}$ *is a sequence such that* $\hat{z}_N \in Z_N$ *for all* N *and* $\hat{z}_N \to \hat{z}$.

(a) *If the* $\hat{z}_N$ *are global minimizers for the* $\mathbf{P}_N$, *then* $\hat{z}$ *is a global minimizer of* $\mathbf{P}$.

(b) *If* $\hat{z}_N$ *are strict local minimizers for the* $\mathbf{P}_N$ *whose radii of attraction do not converge to zero as* $N \to \infty$, *then* $\hat{z}$ *is a local minimizer of* $\mathbf{P}$.

The reader is referred to [1, 8] for the proof of Theorem 2.2(a) and to [18] for the proof of Theorem 2.2(b).

Optimization algorithms, applied to the finite-dimensional problems $\mathbf{P}_N$, are known only to compute stationary points. As the following example shows, epiconvergence alone does not rule out the possibility that stationary points of the $\mathbf{P}_N$ converge to a nonstationary point of $\mathbf{P}$. Let $\mathsf{B} = \mathbb{R}^2$ so that $z = (x, y)$, and let $f(z) = f_N(z) = (x-2)^2$, $N \in \mathbb{N}$. Let

$$Z \triangleq \{(x, y) \in \mathbb{R}^2 | x^2 + y^2 \leq 2\}, \tag{2.2a}$$

and, for all $N \in \mathbb{N}$ let

$$Z_N \triangleq \left\{(x, y) \in \mathbb{R}^2 | (x-y)^2(x^2+y^2-2) \leq 0, x^2 + y^2 \leq 2 + \frac{1}{N}\right\}. \tag{2.2b}$$

Then we see that $\mathbf{P}_N \to^{\text{Epi}} \mathbf{P}$. Nevertheless, the point (1, 1) is feasible and satisfies the F. John

optimality condition for all $\mathbf{P}_N$, but it is not a stationary point for the problem $\mathbf{P}$. The reason for this is an incompatibility of the constraint sets Z_N with the constraint set Z that shows up only at the level of optimality conditions.

To eliminate the possibility of pathologies such as in the above example, as well as some others (e.g., failure of derivatives to converge), [18] imposed a second condition on the approximating problems in terms of optimality conditions that can be viewed as a "first-order" consistency requirement. For the purpose of this condition, it is convenient to characterize stationary points as the zeros of optimality functions $\theta : \mathsf{D} \to \mathbb{R}$ for $\mathbf{P}$ and $\theta_N : \mathsf{D}_N \to \mathbb{R}$ for $\mathbf{P}_N$, $N \in \mathbb{N}$, where $\mathsf{D} \subset \mathsf{B}$ and $\mathsf{D}_N \subset \mathsf{B}_N$; i.e., the optimality functions may not be defined on the entire space. We will assume that $\mathsf{D}_N \subset \mathsf{D} \cap \mathsf{B}_N$ for all $N \in \mathbb{N}$.

DEFINITION 2.3. *A function* $\theta : \mathsf{D} \to \mathbb{R}$ *is an* optimality function *for* $\mathbf{P}$ *if* (i) $Z \subset \mathsf{D}$, (ii) $\theta(\cdot)$ *is sequentially upper semicontinuous,* (iii) $\theta(z) \leq 0$ *for all* $z \in \mathsf{D}$, *and* (iv) $\theta(\hat{z}) = 0$ *for any* $\hat{z} \in Z$ *that is a local minimizer for* $\mathbf{P}$. *Similarly, a function* $\theta_N : \mathsf{D}_N \to \mathbb{R}$ *is an* optimality function *for* $\mathbf{P}_N$ *if* (i) $Z_N \subset \mathsf{D}_N$, (ii) $\theta_N(\cdot)$ *is sequentially upper semicontinuous,* (iii) $\theta_N(z) \leq 0$ *for all* $z \in \mathsf{D}_N$, *and* (iv) $\theta_N(\hat{z}_N) = 0$ *for any* $\hat{z}_N \in Z_N$ *that is a local minimizer for* $\mathbf{P}_N$.

DEFINITION 2.4. *Let* $\theta(\cdot), \theta_N(\cdot), N \in \mathbb{N}$, *be optimality functions for* $\mathbf{P}$, $\mathbf{P}_N$, *respectively. The pairs* $(\mathbf{P}_N, \theta_N)$, *in the sequence* $\{(\mathbf{P}_N, \theta_N)\}_{N=1}^{\infty}$ *are* weakly consistent approximations *to the pair* $(\mathbf{P}, \theta)$ *if* (i) $\mathbf{P}_N \xrightarrow{\text{Epi}} \mathbf{P}$ *and* (ii) *for any sequence* $\{z_N\}_{N \in K}$, $K \subset \mathbb{N}$, *with* $z_N \in \mathsf{D}_N$ *for all* $N \in K$ *such that* $z_N \to z$, $\limsup \theta_N(z_N) \leq \theta(z)$.

As a result of this definition, we immediately get the following result, which subsumes Theorem 2.2.

THEOREM 2.5. *Suppose that the pairs* $(\mathbf{P}_N, \theta_N)$ *in the sequence* $\{(\mathbf{P}_N, \theta_N)\}_{N=1}^{\infty}$ *are weakly consistent approximations to the pair* $(\mathbf{P}, \theta)$ *and that* $\{\hat{z}_N\}_{N=1}^{\infty}$ *is a sequence such that* $\hat{z}_N \in Z_N$ *for all* N *and* $\hat{z}_N \to \hat{z}$.

(a) *If the* $\hat{z}_N$ *are global minimizers for the* $\mathbf{P}_N$, *then* $\hat{z}$ *is a global minimizer of* $\mathbf{P}$.

(b) *If* $\hat{z}_N$ *are strict local minimizers whose radii of attraction do not converge to zero, as* $N \to \infty$, *then* $\hat{z}$ *is a local minimizer of* $\mathbf{P}$.

(c) *If* $\overline{\lim} \theta_N(\hat{z}_N) = 0$, *then* $\theta(\hat{z}) = 0$.

If we define a point $\hat{z}$ to be stationary for $\mathbf{P}$ if $\theta(\hat{z}) = 0$, then we see that Definition 2.3 permits nonfeasible points to be stationary (e.g., they can be stationary points for a problem with relaxed or modified constraints). This phenomenon can be removed by imposing an additional condition, as is done in the following definition.

DEFINITION 2.6. *Let* $\theta(\cdot), \theta_N(\cdot), N \in \mathbb{N}$, *be optimality functions for* $\mathbf{P}$, $\mathbf{P}_N$, *respectively. The pairs* $(\mathbf{P}_N, \theta_N)$, *in the sequence* $\{(\mathbf{P}_N, \theta_N)\}_{N=1}^{\infty}$ *are* consistent approximations *to* $(\mathbf{P}, \theta)$, *if they are weakly consistent approximations, and in addition* $\theta(z) < 0$ *for all* $z \notin Z$ *and* $\theta_N(z) < 0$ *for all* $z \notin Z_N$, $N \in \mathbb{N}$.

3. Consistent approximations for the optimal design of a fixed beam. In this section we formulate the optimal fixed beam design problem and decompose it into a family of consistent approximations. First, we present the equations for modeling a fixed beam from which the bending moment, the shear, and the displacement of the beam can be computed, as can the corresponding sensitivity equations. Second, we formulate the optimal design problem and define an optimality function for it. Third, we choose a dense family of finite-dimensional subsets of the design space and obtain discrete counterparts of the equations modeling the beam as well as the corresponding sensitivity equations. Fourth, we formulate the finite-dimensional, approximating optimal design problems and define optimality functions for them. Fifth, we prove that the approximating problems, with their respective optimality functions, constitute a family of consistent approximations for the original problem and its optimality function.

3.1. Mathematical model of the beam. Consider a fixed beam of length $L > 0$ and rectangular cross section with constant width $b > 0$ and variable depth defined by a positive, Lipschitz continuous function $h : [0, L] \to \mathbb{R}$. The material of the beam has modulus of elasticity $E > 0$, and the beam is subjected to a vertical load with density $l(h, \cdot)$ of the form

$$l(h, x) = m(x) - Kh(x), \qquad x \in [0, L], \tag{3.1}$$

where $K \geq 0$ is a given constant and $m(\cdot)$ is the density of an external load applied to the beam. We assume that $m(\cdot)$ is piecewise Lipschitz continuous with finitely many points of discontinuity in $[0, L]$. We model the beam using Euler–Bernoulli beam theory.

We will obtain an expression for the bending moment in a fixed beam by using the bending moment in a similarly loaded cantilever beam and duality theory. For a cantilever beam of length L, depth determined by the function $h(\cdot)$, and subject to the load density $l(h, \cdot)$, the bending moment $M_c(h, \cdot)$ is the unique solution of the final value problem

$$M_c''(h, x) = l(h, x), \quad x \in [0, L]; \quad M_c(h, L) = 0; \quad M_c'(h, L) = 0, \tag{3.2a}$$

where the prime denotes differentiation with respect to x. It is convenient to rewrite (3.2a) as follows:

$$\frac{d}{dx}\begin{bmatrix} M_c(h, x) \\ M_c'(h, x) \end{bmatrix} = \begin{bmatrix} M_c'(h, x) \\ l(h, x) \end{bmatrix}, \quad x \in [0, L]; \quad M_c(h, L) = M_c'(h, L) = 0. \tag{3.2b}$$

The bending moment $M(h, \cdot)$ in a fixed beam of length L, depth determined by the function $h(\cdot)$, and subject to the load density $l(h, \cdot)$ differs from $M_c(h, \cdot)$ only by an affine term in x that is due to the reactions at the support points. Hence we have that

$$M(h, x) = M_c(h, x) + g_1(h)x + g_2(h), \qquad x \in [0, L], \tag{3.2c}$$

where $g_1(h)$ and $g_2(h)$ are reals, depending on h, that can be computed by using a variational formulation. Let

$$S \triangleq \{p \in L_2[0, L] | p(x) = M_c(h, x) + ax + b, a, b \in \mathbb{R}, x \in [0, L]\}.$$

It follows from the dual formulation of the variational problem associated with the bending of the fixed beam (see [13, 21]) that $M(h, \cdot)$ is the minimizer of the convex functional $W(h, \cdot) : S \to \mathbb{R}$ defined by

$$W(h, p) \triangleq \frac{12}{Eb} \int_0^L \frac{p(x)^2}{h(x)^3}\, dx.$$

From (3.2c) and the first-order optimality condition for $W(h, \cdot)$ we conclude that the vector $g(h) \triangleq [g_1(h) g_2(h)]^T$ satisfies

$$\begin{bmatrix} \int_0^L \frac{x^2}{h(x)^3}\, dx & \int_0^L \frac{x}{h(x)^3}\, dx \\ \int_0^L \frac{x}{h(x)^3}\, dx & \int_0^L \frac{1}{h(x)^3}\, dx \end{bmatrix} \begin{bmatrix} g_1(h) \\ g_2(h) \end{bmatrix} = \begin{pmatrix} -\int_0^L \frac{M_c(h, x)x}{h(x)^3}\, dx \\ -\int_0^L \frac{M_c(h, x)}{h(x)^3}\, dx \end{pmatrix}. \tag{3.2d}$$

The shear force at $x \in [0, L]$ is given by

$$V(h, x) = -M'(h, x) = -M_c'(h, x) - g_1(h), \qquad x \in [0, L], \tag{3.2e}$$

and the deflection of the beam $y(h, \cdot)$ is the solution of the initial value problem

$$\frac{d}{dx}\begin{bmatrix} y(h,x) \\ y'(h,x) \end{bmatrix} = \begin{bmatrix} y'(h,x) \\ 12M(h,x)/Ebh(x)^3 \end{bmatrix}, x \in [0, L], y(h,0) = y'(h,0) = 0. \tag{3.2f}$$

For design purposes we assume that the depth function is an element of the set

$$H_{\text{ad}} \triangleq \{h \in C[0, L] | 0 < \alpha \leq h(x) \leq \beta, |dh(x)/dx| \leq \gamma, \text{ for a.e. } x \in [0, L]\}, \tag{3.3}$$

where $0 < \alpha < \beta < \infty$ and $\gamma \geq 0$ are given constants and $C[0, L]$ is the space of continuous real-valued functions defined on $[0, L]$.

The "natural" norm on $C[0, L]$ for establishing continuity and differentiability of solutions of (3.2b)–(3.2f) with respect to depth functions $h(\cdot)$ is the sup-norm $\|\cdot\|_\infty$. However, when we define optimality functions for our design problems, by extension of optimality functions for problems defined on $\mathbb{R}^n$, which is a Hilbert space, it is much more natural to use the $L_2[0, L]$ norm $\|\cdot\|_2$. Hence, we will work in the inner product space $(C[0, L], \|\cdot\|_2, \langle\cdot, \cdot\rangle_2)$, where $\langle\cdot, \cdot\rangle_2$ denotes the usual inner product on $L_2[0, L]$.

DEFINITION 3.1. *Let $(V, \|\cdot\|_V)$ be a normed space, and let $\zeta : H_{\text{ad}} \subset (C[0, L], \|\cdot\|_2, \langle\cdot, \cdot\rangle_2) \to (V, \|\cdot\|_V)$. We will say that*

(a) *$\zeta(\cdot)$ is Lipschitz continuous relative to H_{ad} if there exists a $C \in (0, \infty)$ such that for all $h, h' \in H_{\text{ad}}$*

$$\|\zeta(h) - \zeta(h')\|_V \leq C\|h - h'\|_2; \tag{3.4a}$$

(b) *$\zeta(\cdot)$ is differentiable relative to H_{ad} if for any $h \in H_{\text{ad}}$ there exists a map $D\zeta(h; \cdot) \in \mathsf{L}(C[0, L], V)$ (the space of continuous linear maps from $(C[0, L], \|\cdot\|_2, \langle\cdot, \cdot\rangle_2)$ into $(V, \|\cdot\|_V)$), called the H_{ad}-derivative of $\zeta(\cdot)$ at h, such that*

$$\lim_{\substack{h' \in H_{\text{ad}} \\ \|h'-h\|_2 \to 0}} \frac{\|\zeta(h') - \zeta(h) - D\zeta(h; h' - h)\|_V}{\|h' - h\|_2}; \tag{3.4b}$$

(c) *$\zeta(\cdot)$ is Lipschitz continuously differentiable relative to H_{ad} if it is differentiable relative to H_{ad}, and the mapping $h \in H_{\text{ad}} \mapsto D\zeta(h, \cdot) \in \mathsf{L}(C[0, L], V)$ is Lipschitz continuous relative to H_{ad}.*

It can be deduced from the results in [3] that the mappings $h \mapsto M_c(h, \cdot)$, and $h \mapsto M_c'(h, \cdot)$ from $H_{\text{ad}} \subset (C[0, L], \|\cdot\|_2, \langle\cdot, \cdot\rangle_2)$ into $(C[0, L], \|\cdot\|_\infty)$ are Lipschitz continuously differentiable relative to H_{ad}. Their H_{ad}-derivatives are denoted by $D_1M_c'(h, \cdot; \cdot)$ and $D_1M_c(h, \cdot; \cdot)$, respectively.

Let $A(h)$ denote the matrix on the left-hand side of (3.2d), and let $b(h)$ denote the vector on the right-hand side of (3.2d) so that, abstractly, (3.2d) becomes $A(h)g(h) = b(h)$. To establish differentiability of the mapping $h \mapsto g(h)$ we need the following result.

LEMMA 3.2. *There exist constants $c, C \in (0, \infty)$ such that for all $h \in H_{\text{ad}}$ and $w \in \mathbb{R}^2$ we have*

$$c\|w\|^2 \leq w^T A(h)w \leq C\|w\|^2. \tag{3.5a}$$

The proof of Lemma 3.2 can be found in the appendix.

It can be easily seen that each entry of $A(h)$ and $b(h)$ is Lipschitz continuously differentiable relative to H_{ad}. We denote their H_{ad}-derivatives by $DA(h; \cdot)$ and $Db(h; \cdot)$. It follows from (3.2d) and Lemma 3.2 that the mapping $h \in H_{\text{ad}} \mapsto g(h) \in \mathbb{R}^2$ is also Lipschitz continuously differentiable relative to H_{ad}. Hence, (3.2c), (3.2e), and (3.2f) imply

$h \mapsto M(h, \cdot)$, $h \mapsto V(h, \cdot)$, and $h \mapsto y(h, \cdot)$, all mapping $H_{\text{ad}} \subset (C[0, L], \|\cdot\|_2, \langle \cdot, \cdot \rangle_2)$ into $(C[0, L], \|\cdot\|_\infty)$, are Lipschitz continuously differentiable relative to H_{ad}. We denote their H_{ad}-derivatives at h by $Dg(h; \cdot)$, $D_1 M(h, \cdot; \cdot)$, $D_1 V(h, \cdot; \cdot)$, and $D_1 y(h, \cdot; \cdot)$, respectively.

Given $h, h' \in H_{\text{ad}}$, let $\delta h \triangleq h' - h$. It can be shown, using (3.1) and (3.2b)–(3.2f), that the following relations hold:

$$\text{(3.5b)} \qquad \frac{d}{dx}\begin{bmatrix} D_1 M_c(h, x; \delta h) \\ D_1 M_c'(h, x; \delta h) \end{bmatrix} = \begin{bmatrix} D_1 M_c'(h, x; \delta h) \\ -K \delta h(x) \end{bmatrix}, \quad x \in [0, L],$$

$$\begin{bmatrix} D_1 M_c(h, L; \delta h) \\ D_1 M_c'(h, L; \delta h) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix};$$

$$\text{(3.5c)} \qquad DA(h; \delta h) = -3 \begin{bmatrix} \int_0^L \frac{x^2 \delta h(x)}{h(x)^4}\, dx & \int_0^L \frac{x \delta h(x)}{h(x)^4}\, dx \\ \int_0^L \frac{x \delta h(x)}{h(x)^4}\, dx & \int_0^L \frac{\delta h(x)}{h(x)^4}\, dx \end{bmatrix};$$

$$\text{(3.5d)} \qquad Db(h; \delta h) = \begin{pmatrix} \int_0^L \frac{3M_c(h, x)\delta h(x) - DM_c(h, x; \delta h)h(x)}{h(x)^4} x\, dx \\ \int_0^L \frac{3M_c(h, x)\delta h(x) - DM_c(h, x; \delta h)h(x)}{h(x)^4}\, dx \end{pmatrix};$$

$$\text{(3.5e)} \qquad A(h) Dg(h; \delta h) = Db(h; \delta h) - DA(h; \delta h) g(h);$$

$$\text{(3.5f)} \qquad D_1 M(h, x; \delta h) = D_1 M_c(h, x; \delta h) + Dg_1(h; \delta h)x + Dg_2(h; \delta h), \qquad x \in [0, L];$$

$$\text{(3.5g)} \qquad D_1 V(h, x; \delta h) = -D_1 M_c'(h, x; \delta h) - Dg_1(h; \delta h), \qquad x \in [0, L];$$

$$\frac{d}{dx}\begin{bmatrix} D_1 y(h, x; \delta h) \\ D_1 y'(h, x; \delta h) \end{bmatrix} = \begin{bmatrix} D_1 y'(h, x; \delta h) \\ \frac{12}{Ebh(x)^3}\left(D_1 M(h, x; \delta h) - 3M(h, x)\frac{\delta h(x)}{h(x)} \right) \end{bmatrix}, x \in [0, L];$$

$$\text{(3.5h)} \qquad D_1 y(h, 0; \delta h) = D_1 y'(h, 0; \delta h) = 0.$$

3.2. Formulation of the optimal design problem. We will consider optimal beam design problems with continuum constraints on the shear force, the bending moment, and the deflection. For example, suppose that we wish to minimize the weight or volume of a fixed beam of constant width subject to constraints on the maximum normal stress at the extreme fiber $\sigma_{\max}(h, \cdot)$, on the maximum shear stress $\tau_{\max}(h, \cdot)$, and on the deflection $y(h, \cdot)$ of the form

$$\text{(3.6a)} \qquad |\sigma_{\max}(h, x)| \le a^1, \quad |\tau_{\max}(h, x)| \le a^2, \quad |\sigma_{\max}(h, x)| \le a^3 \quad \forall x \in [0, L],$$

where the a^j's are given positive constants. To obtain a convenient mathematical formulation for this problem we define the cost function as

$$\text{(3.6b)} \qquad f(h) \triangleq \int_0^L h(\tau)\, d\tau;$$

the bounds as $r^1 = r^2 = a^1$, $r^3 = r^4 = a^2$, $r^5 = r^6 = a^3$; and the constraint functions as

$$\text{(3.6c)} \qquad \phi^1(h, x) \triangleq \sigma_{\max}(h, x) = \frac{6}{b}\frac{M(h, x)}{h(x)^2}, \quad \phi^2(h, x) \triangleq -\phi^1(h, x), \qquad x \in [0, L],$$

$$\text{(3.6d)} \qquad \phi^3(h, x) \triangleq \tau_{\max}(h, x) = \frac{3}{2}\frac{V(h, x)}{bh(x)}, \quad \phi^4(h, x) \triangleq -\phi^3(h, x), \qquad x \in [0, L],$$

(3.6e) $$\phi^5(h,x) \triangleq y(h,x), \phi^6(h,x) \triangleq -\phi^5(h,x), \qquad x \in [0,L].$$

In terms of these functions and bounds, the above-described problem becomes

(3.6f) $$\min_{h\in H_{ad}} \left\{ f(h) | \max_{1\le j\le 6} \max_{x\in[0,L]} \phi^j(h,x) - r^j \le 0 \right\}.$$

The above example is a particular case of optimal design problems of the form

(3.7a) $$\mathbf{P} \qquad \min_{h\in H_{ad}} \left\{ f(h) | \max_{j\in\mathbf{q}} \max_{x\in[0,L]} \phi^j(h,x) - r^j(x) \le 0 \right\},$$

where for any integer $q > 0$, $\mathbf{q} \triangleq \{1, 2, \ldots, q\}$ and the functions $f(\cdot)$ and $\phi^j(\cdot,\cdot)$, $j \in \mathbf{q}$, are of the form

(3.7b) $$f(h) = \int_0^L \phi^0(h,x)\,dx,$$

(3.7c) $$\phi^j(h,x) = \tilde{\phi}^j(h(x), M(h,x), V(h,x), y(h,x), x), \qquad j \in \bar{\mathbf{q}},$$

with $M(h,\cdot)$, $V(h,\cdot)$, and $y(h,\cdot)$ determined by (3.2b)–(3.2f) and, for $j \in \bar{\mathbf{q}} \triangleq \{0, 1, \ldots, q\}$, $\tilde{\phi}^j : [\alpha,\beta] \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times [0,L] \to \mathbb{R}$.

Assumption 3.3. (a) The functions $r^j(\cdot)$, $j \in \mathbf{q}$, are Lipschitz continuously differentiable on $[0, L]$ and satisfy

(3.8) $$\min_{j\in\mathbf{q}} \min_{x\in[0,L]} r^j(x) = \hat{r} > 0.$$

(b) The functions $\tilde{\phi}^j(\cdot,\cdot,\cdot,\cdot,\cdot)$, $j \in \bar{\mathbf{q}}$, are Lipschitz continuously differentiable.

(c) The feasible set for **P** is nonempty.

Existence of a solution to **P** follows from the Ascoli–Arzelà theorem, which implies that the set H_{ad} is compact in $(C[0,L], \|\cdot\|_2, \langle\cdot,\cdot\rangle_2)$. The proofs of existence of solutions for similar problems can be found in [6, 9, 15].

It follows from (3.7b), (3.7c), and the Lipschitz continuous differentiability of $\tilde{\phi}^j(\cdot,\cdot,\cdot,\cdot,\cdot)$, $j \in \bar{\mathbf{q}}$, with respect to all their arguments and of $M(h,\cdot)$, $V(h,\cdot)$, and $y(h,\cdot)$ with respect to h that $h \mapsto \phi^j(h,\cdot)$, $j \in \bar{\mathbf{q}}$, and $h \mapsto f(h)$ are Lipschitz continuously differentiable functions relative to H_{ad}. We will denote by $D_1\phi^j(h,\cdot;\cdot)$, $j \in \bar{\mathbf{q}}$, and $Df(h;\cdot)$ their H_{ad}-derivatives.

LEMMA 3.4. *There exists a constant* $C \in (0,\infty)$ *such that for any* $h, \tilde{h}, h', h'' \in H_{ad}$,

(3.9a) $$|Df(h;h'-h) - Df(\tilde{h};h''-\tilde{h})| \le C[\|h-\tilde{h}\|_2 + \|h'-h''\|_2]$$

and for all $j \in \bar{\mathbf{q}}$,

(3.9b) $$\|D_1\phi^j(h,\cdot;h'-h) - D_1\phi^j(\tilde{h},\cdot;h''-\tilde{h})\|_\infty \le C[\|h-\tilde{h}\|_2 + \|h'-h''\|_2].$$

Proof. Both inequalities are direct consequences of the Lipschitz continuity relative to H_{ad} of the H_{ad}-derivatives of $h \mapsto M(h,\cdot)$, $h \mapsto V(h,\cdot)$, and $h \mapsto y(h,\cdot)$ and the Lipschitz continuous differentiability of the functions $\tilde{\phi}^j(\cdot,\cdot,\cdot,\cdot,\cdot)$, $j \in \bar{\mathbf{q}}$, in (3.7c). □

We define the constraint violation function $\psi : H_{ad} \to \mathbb{R}$ for **P** by

(3.10a) $$\psi(h) \triangleq \max_{j\in\mathbf{q}} \max_{x\in[0,L]} \phi^j(h,x) - r^j(x)$$

and the surrogate cost function $F : H_{ad} \times H_{ad} \to \mathbb{R}$, suggested by method of centers-type algorithms, by

$$F(h, h') \triangleq \max\{f(h') - f(h) - \omega\psi(h)_+, \max_{j\in\mathbf{q}} \max_{x\in[0,1]} \phi^j(h', x) - r^j(x) - \psi(h)_+\}, \tag{3.10b}$$

where $\psi(h)_+ \triangleq \max\{\psi(h), 0\}$ and $\omega > 0$ is a parameter. Note that (i) for all $h \in H_{ad}$, $F(h, h) = 0$, and (ii) if $\hat{h} \in H_{ad}$ is a local minimizer for **P**, then since $\psi(h) > 0$ when h is infeasible and $f(h) \geq f(\hat{h})$ for all feasible h in a ball about $\hat{h}$, $\hat{h}$ must also be a local minimizer for the surrogate problem

$$\min_{h\in H_{ad}} F(\hat{h}, h). \tag{3.10c}$$

This fact is used in [2] to obtain the following first-order optimality condition for **P**.

PROPOSITION 3.5. *If $\hat{h}$ is a local minimizer for* **P**, *then*

$$\hat{h} \in H_{ad} \text{ and } d_2F(\hat{h}, \hat{h}; h' - \hat{h}) \geq 0 \quad \forall h' \in H_{ad} \tag{3.10d}$$

where $d_2F(\hat{h}, \hat{h}; h' - \hat{h})$ denotes the (one-sided) directional derivative of $F(\cdot, \cdot)$ at $(\hat{h}, \hat{h})$ with respect to the second argument in the direction $h' - \hat{h}$.

Referring to [17], we see that one way to verify whether (3.10d) holds at a given $h \in H_{ad}$ is to define an optimality function $\theta : H_{ad} \to \mathbb{R}$ for **P** based on a *convex*, first-order approximation $\tilde{F}(h, h')$ to $F(h, h')$. Thus we define

$$\begin{aligned}\tilde{F}(h, h') \triangleq \max \Big\{ & Df(h; h' - h) - \omega\psi(h)_+, \\ & \max_{j\in\mathbf{q}} \max_{x\in[0,L]} \phi^j(h, x) + D_1\phi^j(h, x; h' - h) - \psi(h)_+ \Big\} \\ & + \frac{1}{2}\|h' - h\|_2^2\end{aligned} \tag{3.10e}$$

and

$$\theta(h) \triangleq \min_{h'\in H_{ad}} \tilde{F}(h, h'). \tag{3.11}$$

THEOREM 3.6. (a) *The function $\theta(\cdot)$ is well defined and takes values in $(-\infty, 0]$.*

(b) *$\theta : H_{ad} \to \mathbb{R}$ is upper semicontinuous.*

(c) *For any $\hat{h} \in H_{ad}$, $\theta(\hat{h}) = 0$ if and only if* either *$\psi(\hat{h}) \leq 0$ and* (3.10d) *holds* or *$\psi(\hat{h}) > 0$ and $0 \in \partial\psi(\hat{h})$, where $\psi(\hat{h})$ denotes the Clarke generalized gradient* [2] *of $\psi(\cdot)$ at $\hat{h}$ (i.e., $\hat{h}$ satisfies the first-order optimality condition for the problem $\min_{h\in H_{ad}} \psi(h)$).*

Proof. We start by showing that $\theta(\cdot)$ is well defined. In view of (3.7b), (3.7c), and Assumption 3.3(b), it should be clear that for all $j \in \bar{\mathbf{q}}$ and $x \in [0, L]$ the mappings $h \mapsto \phi^j(h, x)$ and $h \mapsto \psi(h)$ are continuous on H_{ad}. Hence, as a consequence of the definition of $\tilde{F}(\cdot, \cdot)$, (3.9a), and (3.9b) we have that $\tilde{F} : H_{ad} \times H_{ad} \to \mathbb{R}$ is continuous. Since $H_{ad} \subset (C[0, L], \|\cdot\|_2, \langle\cdot, \cdot\rangle)$ is compact (by the Ascoli–Arzelà theorem), it follows from (3.11) that $\theta(\cdot)$ is well defined.

The fact that $\theta(\cdot)$ takes values in $(-\infty, 0]$, and part (c) can be deduced from Propositions 5.4 and 5.5 in [17]. We now prove part (b).

Suppose $\{h_j\}_{j=0}^{\infty} \subset H_{ad}$ is such that $h_j \to h \in H_{ad}$ as $j \to \infty$. Let $h' \in H_{ad}$ be such that $\theta(h) = \tilde{F}(h, h')$. Then

$$\theta(h_j) \leq \tilde{F}(h_j, h') \quad \forall j \in \mathbb{N}. \tag{3.12a}$$

Hence, taking $\overline{\lim}$ on both sides and using the continuity of $\tilde{F}(\cdot, \cdot)$, we get

$$\overline{\lim_{j\to\infty}}\ \theta(h_j) \leq \overline{\lim_{j\to\infty}}\ \tilde{F}(h_j, h') = \tilde{F}(h, h') = \theta(h). \tag{3.12b}$$

COROLLARY 3.7. *$\theta(\cdot)$ is an optimality function for* $\mathbf{P}$.

3.3. Choice of finite-dimensional subsets of H_{ad} and discretization of the beam equations. According to the theory in §2, to define approximating problems $\mathbf{P}_N$ we must begin by selecting a family of finite-dimensional subspaces of $C[0, L]$. We specify these subspaces by constructing basis sets for them.

For every integer $N > 0$ we let $\Delta_N \triangleq L/N$ and define the mesh $T_N \triangleq \{0, \Delta_N, 2\Delta_N, \ldots, L\}$ with nodes $x_{N,k} = (k-1)\Delta_N$, $k \in \mathbf{N+1}$.

Let

$$P_{N,k}(x) \triangleq \begin{cases} \dfrac{x - x_{N,k-1}}{\Delta_N} & \forall x \in [x_{N,k-1}, x_{N,k}],\ k \in \{2, \cdots, N+1\}; \\ \dfrac{x_{N,k+1} - x}{\Delta_N} & \forall x \in [x_{N,k}, x_{N,k+1}],\ k \in \mathbf{N}; \\ 0 & \text{otherwise.} \end{cases} \tag{3.13a}$$

We denote by H_N the span of the basis set $\{P_{N,k}\}_{k=1}^{N+1}$. Clearly, H_N is a subspace of $C[0, L]$, and for each $h \in H_{\text{ad},N}$ there exists a unique $(\eta_1, \eta_2, \ldots, \eta_{N+1})^T \in \mathbb{R}^{N+1}$ such that

$$h(x) = \sum_{k=1}^{N+1} \eta_k P_{N,k}(x), \qquad x \in [0, L]. \tag{3.13b}$$

We let $H_{\text{ad},N} \triangleq H_N \cap H_{\text{ad}}$.

Next, we discretize the equations describing the behavior of the beam. We will use Euler's method to discretize the initial value problem (3.2f) and the final value problem (3.2b) and the rectangle rule to approximate the integrals in (3.2d). The choice of this simple discretization scheme is motivated by the fact that, in the context of optimal design problems, it is not clear whether using higher order integration schemes is more efficient. To see this consider an unconstrained optimal design problem $\mathbf{P}_u$ defined over H_{ad} and a sequence of approximating problems $\mathbf{P}_{u,N}$ defined over $H_{\text{ad},N}$ and obtained by discretizing $\mathbf{P}_u$. We assume that to define the $\mathbf{P}_{u,N}$ one uses an integration scheme of order $r \geq 1$ to discretize the differential equations and the integrals defining $\mathbf{P}_u$. Let $f^0(\cdot)$ and $f_N^0(\cdot)$ be the cost functions for $\mathbf{P}_u$ and $\mathbf{P}_{u,N}$, respectively. Then one can show, under appropriate assumptions, that there exists a $K \in (0, \infty)$ such that for all N

$$-K[\Delta_N + \Delta_N^r] \leq f^0(\hat{h}) - f_N^0(\hat{h}_N) \leq K\Delta_N, \tag{3.13c}$$

where $\hat{h}$ is the minimizer of $\mathbf{P}_u$ and $\hat{h}_N$ is the minimizer of $\mathbf{P}_{u,N}$. The term $-K\Delta_N$ in the left-hand side of (3.13c) is due to the replacement of H_{ad} by $H_{\text{ad},N}$, while the term $-K\Delta_N^r$ is due to the replacement of the differential equations and integrals by discrete counterparts arising from the use of an rth-order integration method. Hence, if we use a first-order method (i.e., $r = 1$) according to (3.13c) the uncertainty interval in the computation of the optimal value is of length $3K\Delta_N$. As $r \to \infty$, the length of the interval of uncertainty decreases to $K\Delta_N$. Hence we see that the difference between the optimal cost of the approximating problem and that of the original problem is of order $O(\Delta_N)$ regardless of the accuracy of the integration method used. Therefore, it is not clear that much is gained by using the

more computationally intensive higher order integration methods to solve optimal design methods.

Let $M_N(h, x_{N,k})$, $V_N(h, x_{N,k})$, and $y_N(h, x_{N,k})$, $k \in \mathbf{N+1}$, denote the discrete bending moment, the discrete shear force, and the discrete deflection, respectively, defined by the recursions

$$\begin{aligned} \begin{bmatrix} M_{c,N}(h, x_{N,k}) \\ M'_{c,N}(h, x_{N,k}) \end{bmatrix} &= \begin{bmatrix} M_{c,N}(h, x_{N,k+1}) - \Delta_N M'_{c,N}(h, x_{N,k+1}) \\ M_{c,N}(h, x_{N,k+1}) - \Delta_N l(h, x_{N,k+1}) \end{bmatrix}, \qquad k \in \mathbf{N}, \\ \begin{bmatrix} M_{c,N}(h, x_{N,N+1}) \\ M'_{c,N}(h, x_{N,N+1}) \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}; \end{aligned} \tag{3.14a}$$

$$\begin{bmatrix} \sum_{j=1}^{N} \frac{(j-1)^2 \Delta_N^3}{h(x_{N,j})^3} & \sum_{j=1}^{N} \frac{(j-1)\Delta_N^2}{h(x_{N,j})^3} \\ \sum_{j=1}^{N} \frac{(j-1)\Delta_N^2}{h(x_{N,j})^3} & \sum_{j=1}^{N} \frac{\Delta_N}{h(x_{N,j})^3} \end{bmatrix} \begin{bmatrix} g_{1,N}(h) \\ g_{2,N}(h) \end{bmatrix} = \begin{pmatrix} -\sum_{j=1}^{N} \frac{(j-1)\Delta_N^2}{h(x_{N,j})^3} M_{c,N}(h, x_{N,j}) \\ -\sum_{j=1}^{N} \frac{\Delta_N}{h(x_{N,j})^3} M_{c,N}(h, x_{N,j}) \end{pmatrix}; \tag{3.14b}$$

$$M_N(h, x_{N,k}) = M_{c,N}(h, x_{N,k}) + g_{1,N}(h) x_{N,k} + g_{2,N}(h), \qquad k \in \mathbf{N+1}; \tag{3.14c}$$

$$V_N(h, x_{N,k}) = -M'_{c,N}(h, x_{N,k}) - g_{1,N}(h), \qquad k \in \mathbf{N+1}; \tag{3.14d}$$

and, for $k \in \mathbf{N}$,

$$\begin{aligned} \begin{bmatrix} y_N(h, x_{N,k+1}) \\ y'_N(h, x_{N,k+1}) \end{bmatrix} &= \begin{bmatrix} y_N(h, x_{N,k}) + \Delta_N y'_N(h, x_{N,k}) \\ y'_N(h, x_{N,k}) + \Delta_N \frac{12 M_N(h, x_{N,k})}{E b h(x_{N,k})^3} \end{bmatrix}, \\ \begin{bmatrix} y_N(h, x_{N,1}) \\ y'_N(h, x_{N,1}) \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{aligned} \tag{3.14e}$$

Let $A_N(h)$ denote the matrix on the left-hand side of (3.14b), and let $b_N(h)$ denote the vector on the right-hand side of (3.14b) so that, abstractly, (3.14b) becomes $A_N(h) g_N(h) = b_N(h)$.

LEMMA 3.8. *There exist constants* $c, C \in (0, \infty)$ *such that for all* $N \geq 1$, $h \in H_{ad,N}$, *and* $w \in \mathbb{R}^2$, *we have*

$$c\|w\|^2 \leq w^T A_N(h) w \leq C\|w\|^2. \tag{3.15}$$

The proof of Lemma 3.8 is similar to that of Lemma 3.2 and hence we omit it.

It follows from the implicit function theorem (see, e.g., [14]) that the maps $h \mapsto M_{c,N}(h, \cdot)$, $h \mapsto M'_{c,N}(h, \cdot)$ from $H_{ad,N}$ into $\mathbb{R}^{N+1}$ are Lipschitz continuously differentiable. Lipschitz continuous differentiablity of the map $h \mapsto g_N(h)$ from $H_{ad,N}$ into $\mathbb{R}^2$ and the maps $h \mapsto M_N(h, \cdot)$, $h \mapsto V_N(h, \cdot)$, and $h \mapsto y_N(h, \cdot)$ from $H_{ad,N}$ into $\mathbb{R}^{N+1}$ follows from the Implicit Function Theorem, (3.14b)–(3.14e), and from Lemma 3.8.

Given $h, h' \in H_{ad,N}$, let $\delta h \triangleq h' - h$. One can show, by differentiating (3.14a)–(3.14e), that

$$\begin{bmatrix} D_1 M_{c,N}(h, x_{N,k}; \delta h) \\ D_1 M'_{c,N}(h, x_{N,k}; \delta h) \end{bmatrix} = \begin{bmatrix} D_1 M_{c,N}(h, x_{N,k+1}; \delta h) - \Delta_N D_1 M'_{c,N}(h, x_{N,k+1}; \delta h) \\ D_1 M_{c,N}(h, x_{N,k+1}; \delta h) - \Delta_N K \delta h(x_{N,k+1}) \end{bmatrix}, \qquad k \in \mathbf{N},$$

$$D_1 M'_{c,N}(h, x_{N,N+1}; \delta h) = D_1 M_{c,N}(h, x_{N,N+1}; \delta h) = 0; \tag{3.16a}$$

(3.16b) $$DA_N(h_N;\delta h) = -3\begin{bmatrix} \sum_{j=1}^{N}\frac{(j-1)^2\Delta_N^3\delta h(x_{N,j})}{h_N(x_{N,j})^4} & \sum_{j=1}^{N}\frac{(j-1)\Delta_N^2\delta h(x_{N,j})}{h_N(x_{N,j})^4} \\ \sum_{j=1}^{N}\frac{(j-1)\Delta_N^2\delta h(x_{N,j})}{h_N(x_{N,j})^4} & \sum_{j=1}^{N}\frac{\Delta_N\delta h(x_{N,j})}{h_N(x_{N,j})^4} \end{bmatrix};$$

(3.16c)
$$Db_N(h_N;\delta h) = \begin{pmatrix} -\sum_{j=1}^{N}(j-1)\Delta_N^2\left[\frac{DM_{c,N}(h_N,x_{N,j};\delta h)}{h_N(x_{N,j})^3} - \frac{3M_{c,N}(h_N,x_{N,j})}{h_N(x_{N,j})^4}\delta h(x_{N,j})\right] \\ -\sum_{j=1}^{N}\Delta_N\left[\frac{DM_{c,N}(h_N,x_{N,j};\delta h)}{h_N(x_{N,j})^3} - \frac{3M_{c,N}(h_N,x_{N,j})}{h_N(x_{N,j})^4}\delta h(x_{N,j})\right] \end{pmatrix};$$

(3.16d) $$A_N(h_N)Dg_N(h_N;\delta h) = Db_N(h_N;\delta h) - DA_N(h_N;\delta h)g_N(h_N);$$

(3.16e)
$$D_1M_N(h,x_{N,k};\delta h) = D_1M_{c,N}(h,x_{N,k};\delta h)+Dg_{1,N}(h;\delta h)x_{N,k}+Dg_{2,N}(h;\delta h), \quad k\in\mathbf{N+1};$$

(3.16f) $$D_1V_N(h,x_{N,k};\delta h) = -D_1M'_N(h,x_{N,k};\delta h) - Dg_{1,N}(h;\delta h), \qquad k\in\mathbf{N+1};$$

and $D_1y_N(h,x_{N,k};\delta h) = \delta y_N(h,x_{N,k})$, $k\in\mathbf{N+1}$, where $\delta y_N(h,x_{N,k})$ is the solution of

$$\begin{bmatrix}\delta y_{k+1}\\ \delta y'_{k+1}\end{bmatrix} = \begin{bmatrix} \delta y_k + \Delta_N\delta y'_k \\ \delta y'_k + \frac{12\Delta_N}{Ebh(x_{N,k})^3}\left(D_1M_N(h,x_{N,k};\delta h) - \frac{3M_N(h,x_{N,k})}{h(x_{N,k})}\delta h(x_{N,k})\right)\end{bmatrix}, \quad k\in\mathbf{N},$$

(3.16g) $$\delta y_1 = \delta y'_1 = 0.$$

3.4. Formulation of the approximating problems. We now define the family of approximating problems $\mathbf{P}_N$, $N = 1, 2, \ldots$, as

(3.17a) $\mathbf{P}_N$
$$\min_{h\in H_{ad,N}}\left\{f_N(h)\,|\,\max_{j\in\mathbf{q}}\max_{k\in\mathbf{N+1}}\phi_N^j(h,x_{N,k}) - (1+\Delta_N^{1/2})r^j(x_{N,k})\le 0\right\},$$

where

(3.17b) $$f_N(h)\triangleq\sum_{k=1}^{N}\phi_N^0(h,x_{N,k})\Delta_N,$$

(3.17c) $$\phi_N^j(h,x_{N,k}) = \tilde{\phi}^j(h(x_{N,k}),M_N(h,x_{N,k}),V_N(h,x_{N,k}),y_N(h,x_{N,k}),x_{N,k}), \quad j\in\bar{\mathbf{q}}.$$

Equation (3.17c) defines the functions $\phi_N^j(h,\cdot)$, $j\in\bar{\mathbf{q}}$, only on the mesh points $x_{N,k}$, $k\in\mathbf{N+1}$. We define the functions $\phi_N^j(h,\cdot):[0,L]\to\mathbb{R}$ as the piecewise affine interpolation of the values $\phi_N^j(h,x_{N,k})$, $k\in\mathbf{N+1}$.

The term $\Delta_N^{1/2}$ in (3.17a) is added to guarantee that for N large enough the feasible set for $\mathbf{P}_N$ is nonempty. This relaxation of the constraints will be needed in the proof of Theorem 3.10(a).

It follows from (3.17c) and Assumption 3.3(b) that the functions $\phi_N^j(\cdot,\cdot)$, $j\in\bar{\mathbf{q}}$, are Lipschitz continuously differentiable on $H_{ad,N}$. The derivatives of the mappings $h\mapsto\phi_N^j(h,\cdot)$, $j\in\bar{\mathbf{q}}$, and $h\mapsto f(h)$, which we denote by $D_1\phi_N^j(h,\cdot;\cdot)$ and $Df_N(h;\cdot)$, are easily obtained from (3.17b), (3.17c) and (3.16a)–(3.16g) by applying the chain rule.

Next, we define the constraint violation function $\psi_N : H_{ad,N}\to\mathbb{R}$ for $\mathbf{P}_N$ and the surrogate cost $F_N : H_{ad,N}\times H_{ad,N}\to\mathbb{R}$ by

$$\psi_N(h) \triangleq \max_{j\in\mathbf{q}} \max_{k\in\mathbf{N+1}} \phi_N^j(h, x_{N,k}) - (1 + \Delta_N^{1/2}) r^j(x_{N,k}), \tag{3.18a}$$

(3.18b)

$$F_N(h, h') \triangleq \max \left\{ f_N(h') - f_N(h) - \omega\psi_N(h)_+, \max_{j\in\mathbf{q}} \max_{k\in\mathbf{N+1}} \phi_N^j(h', x_{N,k}) - \psi_N(h)_+ \right\}.$$

Any local minimizer $\hat{h} \in H_{\mathrm{ad},N}$ of $\mathbf{P}_N$ is also a local minimizer of $F_N(\hat{h}, \cdot)$ on $H_{\mathrm{ad},N}$, and hence it satisfies the optimality condition

$$\hat{h} \in H_{\mathrm{ad},N} \text{ and } d_2 F_N'(\hat{h}, \hat{h}; h' - \hat{h}) \geq 0 \quad \forall h' \in H_{\mathrm{ad},N}. \tag{3.19}$$

Following the pattern set in §3.2 we define an optimality function $\theta_N : H_{\mathrm{ad},N} \to \mathbb{R}$ based on a *convex*, first-order approximation $\tilde{F}_N(h, h')$ of $F_N(h, h')$ defined by

$$\tilde{F}_N(h, h') \triangleq \max \left\{ Df_N(h; h' - h) - \omega\psi_N(h)_+, \right.$$

$$\left. \max_{j\in\mathbf{q}} \max_{k\in\mathbf{N+1}} \phi_N^j(h, x_{N,k}) - \psi_N(h)_+ + D_1\phi_N^j(h, x_{N,k}; h' - h) \right\} \tag{3.20a}$$

$$+ \frac{1}{2}\|h' - h\|_2^2,$$

$$\theta_N(h) \triangleq \min_{h'\in H_{\mathrm{ad},N}} \tilde{F}_N(h, h'). \tag{3.20b}$$

It is not difficult to show, using (3.20a), that $\theta_N(h)$ can be computed by solving a positive definite quadratic program. An evaluation of $\theta_N(h)$ provides a computational method for checking whether $h \in H_{\mathrm{ad},N}$ satisfies the basic optimality condition (3.19).

Results analogous to Theorem 3.6 and Corollary 3.7 hold for $\theta_N(\cdot)$.

3.5. Epiconvergence and consistency of approximations. We start by establishing some results relating the functions defining $\mathbf{P}$ with those defining $\mathbf{P}_N$. The proof of the following lemma is given in the appendix.

LEMMA 3.9. (a) *For every $h \in H_{\mathrm{ad}}$, and every integer $N \geq 1$, there exists an $h_N \in H_{\mathrm{ad},N}$ such that*

$$\max_{x\in[0,L]} |h(x) - h_N(x)| \leq \gamma\Delta_N. \tag{3.21a}$$

(b) *There exists a constant $C \in (0, \infty)$ such that for all $j \in \bar{\mathbf{q}}$, $N \geq 1$, $h \in H_{\mathrm{ad}}$, and $h_N \in H_{\mathrm{ad},N}$,*

$$\max_{x\in[0,L]} |\phi^j(h, x) - \phi_N^j(h_N, x)| \leq C[\Delta_N + \|h - h_N\|_2], \tag{3.21b}$$

$$|\psi(h) - \psi_N(h_N)| \leq C[\Delta_N^{1/2} + \|h - h_N\|_2], \tag{3.21c}$$

$$|f(h) - f_N(h_N)| \leq C[\Delta_N + \|h - h_N\|_2]. \tag{3.21d}$$

In view of the definitions of $\psi(h)$ and $\psi_N(h)$, it is clear that the feasible sets $\mathbf{Z}$ and $\mathbf{Z}_N$ for $\mathbf{P}$ and $\mathbf{P}_N$ are given by

$$\mathbf{Z} = \{h \in H_{\mathrm{ad}} | \psi(h) \leq 0\}, \qquad \mathbf{Z}_N = \{h \in H_{\mathrm{ad},N} | \psi_N(h) \leq 0\}. \tag{3.21e}$$

THEOREM 3.10 (epiconvergence). (a) *For every $h \in \mathbf{Z}$, there exists a sequence $\{h_N\}_{N=N_0}^{\infty}$, with $h_N \in \mathbf{Z}_N$, such that $f_N(h_N) \to f(h)$ as $N \to \infty$.* (b) *Let $\{h_N\}_{N=N_0}^{\infty}$ be a sequence such that $h_N \in \mathbf{Z}_N$ and $h_N \to \hat{h}$ as $N \to \infty$; then $\hat{h} \in \mathbf{Z}$, and $f_N(h_N) \to f(\hat{h})$.*

Proof. Suppose $h \in \mathbf{Z}$ is given. Then, by Lemma 3.9(a), for each integer N there exists an $h_N \in H_{\text{ad},N}$ such that (3.21a) holds. Clearly, $h_N \to h$ as $N \to \infty$. It follows from (3.21d) that $f_N(h_N) \to f(h)$ as $N \to \infty$. To complete the proof of part (a) it remains to show that there exists an N_0 such that $h_N \in \mathbf{Z}_N$ for all $N \geq N_0$. Indeed, since $h \in \mathbf{Z}$ by assumption we have $\psi(h) \leq 0$. Hence, using (3.21b) and (3.21a) we obtain

$$
\begin{aligned}
\psi_N(h_N) &\leq \psi_N(h_N) - \psi(h) \\
&= \max_{j \in \mathbf{q}} \max_{x \in [0,L]} [\phi_N^j(h_N, x) - r^j(x)(1 + \Delta_N^{1/2})] - \max_{j \in \mathbf{q}} \max_{x \in [0,L]} [\phi^j(h, x) - r^j(x)] \\
&\leq \max_{j \in \mathbf{q}} \max_{x \in [0,L]} [|\phi_N^j(h_N, x) - \phi^j(h, x)| - r^j(x)\Delta_N^{1/2}] \qquad (3.22) \\
&\leq C[\Delta_N + \|h - h_N\|_2] - \hat{r}\Delta_N^{1/2} \leq C[\Delta_N + \|h - h_N\|_\infty] - \hat{r}\Delta_N^{1/2} \\
&\leq C(1 + \gamma)\Delta_N - \hat{r}\Delta_N^{1/2},
\end{aligned}
$$

where $\hat{r} > 0$ is as in (3.8). It follows from (3.22) that there exists an N_0 such that for all $N \geq N_0$, $\psi_N(h_N) \leq 0$, which proves (a).

Let $\{h_N\}_{N=N_0}^\infty$ be a sequence as in (b). Since H_{ad} is closed and $\mathbf{Z}_N \subset H_{\text{ad},N} \subset H_{\text{ad}}$ for all $N \in \mathbb{N}$, it follows that $\hat{h} \in H_{\text{ad}}$. The facts that $\hat{h} \in \mathbf{Z}$, that is, that $\psi(\hat{h}) \leq 0$, and that $f_N(h_N) \to f(\hat{h})$ follow directly from (3.21c) and (3.21d), respectively. □

Next, we establish approximation results relating the derivatives of the functions defining $\mathbf{P}$ and $\mathbf{P}_N$.

LEMMA 3.11. *There exists a constant $C < \infty$ such that for all positive integers N and h, $h' \in H_{\text{ad},N}$,*

$$
|Df(h; h' - h) - Df_N(h; h' - h)| \leq C\Delta_N \|h' - h\|_2, \qquad (3.23a)
$$

$$
\max_{k \in \mathbf{N+1}} |D_1\phi^j(h, x_{N,k}; h' - h) - D_1\phi_N^j(h, x_{N,k}; h' - h)| \leq C\Delta_N \|h' - h\|_2. \qquad (3.23b)
$$

Lemma 3.11 is proved in the appendix.

LEMMA 3.12. *There exists a constant $C \in (0, \infty)$ such that for all positive integers N and h, $h' \in H_{\text{ad},N}$,*

$$
|\tilde{F}(h, h') - \tilde{F}_N(h, h')| \leq C\Delta_N^{1/2}. \qquad (3.24)
$$

Lemma 3.12 follows from the boundedness of $H_{\text{ad},N}$, (3.21c), and (3.21d), the definitions of $\tilde{F}(\cdot, \cdot)$ and $\tilde{F}_N(\cdot, \cdot)$ in (3.10e) and (3.20a), respectively, and Lemma 3.11.

THEOREM 3.13. *Suppose that $\{h_N\}_{N=N_0}^\infty$, with $h_N \in H_{\text{ad},N}$, is such that $h_N \to h$ as $N \to \infty$. Then $h \in H_{\text{ad}}$ and $\overline{\lim}_{N \to \infty} \theta_N(h_N) \leq \theta(h)$.*

Proof. Let $h' \in H_{\text{ad}}$ be such that $\theta(h) = \tilde{F}(h, h')$. Let $\{h'_N\}_{N=N_0}^\infty$ be such that $h'_N \in H_{\text{ad},N}$ and $h'_N \to h'$ as $N \to \infty$. Then we have

$$
\theta_N(h_N) \leq \tilde{F}_N(h_N, h'_N) \leq \tilde{F}(h_N, h'_N) + C\Delta_N^{1/2}, \qquad (3.25a)
$$

where we made use of Lemma 3.12 to obtain the last inequality. Hence, taking $\overline{\lim}$ on both sides and using the fact that $\tilde{F}(\cdot, \cdot)$ is continuous, we obtain

$$
\overline{\lim_{N \to \infty}}\, \theta_N(h_N) \leq \overline{\lim_{N \to \infty}}\, \tilde{F}(h_N, h'_N) = \tilde{F}(h, h') = \theta(h). \qquad (3.25b)
$$

COROLLARY 3.14. *The sequence $\{(\mathbf{P}_N, \theta_N)\}_{N=1}^\infty$ is a family of weakly consistent approximations to the pair $(\mathbf{P}, \theta)$. Furthermore, if for all $h \in H_{\text{ad}}$ such that $\psi(h) > 0$, $0 \notin \partial\psi(h)$, then $\{(\mathbf{P}_N, \theta_N)\}_{N=1}^\infty$ is a family of consistent approximations to $(\mathbf{P}, \theta)$.*

4. Transcription of $\mathbf{P}_N$ into a nonlinear programming problem. We will now establish a transcription of the problem $\mathbf{P}_N$, which is defined on the finite-dimensional subspace $H_N \subset C[0, L]$, into an equivalent nonlinear programming problem, $\bar{\mathbf{P}}_N$, defined on $\mathbb{R}^{N+1}$. Given any $h \in H_N$ there exists a unique vector $\eta = (\eta_1, \ldots, \eta_{N+1})^T \in \mathbb{R}^{N+1}$ satisfying (3.13b). In fact, in view of (3.13a), we have that $\eta_k = h(x_{N,k})$, $k \in \mathbf{N+1}$. Let the mapping $W_N : H_N \to \mathbb{R}^{N+1}$ be defined by

$$W_N(h) \triangleq (\eta_1, \eta_2, \ldots, \eta_{N+1})^T. \tag{4.1a}$$

Clearly, W_N is a bijection and the components of $\eta \in \mathbb{R}^{N+1}$ are the coordinates of $h \in H_N$ with respect to the basis set $\{P_{N,k}(\cdot)\}_{k=1}^{N+1}$. Since the basis set $\{P_{N,i}(\cdot)\}_{i=1}^{N+1}$ is not orthonormal, given a vector $h(\cdot) = \sum_{i=1}^{N+1} \eta_i P_{N,i}(\cdot) \in H_N$, it is *not true* that $\|h\|_2^2 = \sum_{i=1}^{N+1} \eta_i^2 \triangleq \|W_N(h)\|^2$; i.e., $W_N(\cdot)$ is not an isometry. Now suppose that we define the problem $\bar{\mathbf{P}}_N$ using the map $W_N(\cdot)$ and try to solve P_N by solving $\bar{\mathbf{P}}_N$, using a nonlinear programming algorithm, without taking into account the fact that W_N is not an isometry. Then it turns out that we are trying to solve the problem $\mathbf{P}_N$ using a nonlinear programming algorithm with a modified metric. As is well known, changing the metric can cause the performance of nonscale invariant algorithms to deteriorate considerably.

To compute in $\mathbb{R}^{N+1}$, using a metric that is equivalent to the one on H_N one can either modify existing nonlinear programming software, something not easily undertaken when using a standard library of programs, or define the $\bar{\mathbf{P}}_N$ in terms of coordinates corresponding to an orthonormal basis. Problems $\bar{\mathbf{P}}_N$ thus defined can be solved by standard nonlinear programming algorithms, in the space of coordinates corresponding to this basis, without incurring penalties due to induced ill-conditioning.

We will now show how to define $\bar{\mathbf{P}}_N$ in terms of coordinates corresponding to an orthonormal basis for H_N.

Let $Q_N \in \mathbb{R}^{(N+1)\times(N+1)}$ be such that

$$(Q_N)_{ij} \triangleq \int_0^L P_{N,i}(x) P_{N,j}(x)\, dx, \tag{4.1b}$$

where for any matrix $A \in \mathbb{R}^{(N+1)\times(N+1)}$, A_{ij} denotes the i,jth entry of A. One can verify that

$$Q_N = \frac{\Delta_N}{6} \begin{bmatrix} 2 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 4 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & 4 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 1 & 4 & 1 \\ 0 & \cdot & \cdot & \cdot & \cdot & 1 & 2 \end{bmatrix}. \tag{4.1c}$$

Let $T_N : H_N \to \mathbb{R}^{N+1}$ be defined by

$$T_N(h) \triangleq Q_N^{1/2} W_N(h) \tag{4.2a}$$

so that, for any $h \in H_N$ and $\bar{h} = T_N(h)$, we have

$$\begin{aligned} \|h\|_2^2 &= \int_0^L \sum_{i,j=1}^{N+1} (W_N(h))_j P_{N,j}(x) (W_N(h))_i P_{N,i}(x)\, dx \\ &= (Q_N^{1/2} W_N(h))^T (Q_N^{1/2} W_N(h)) = \|\bar{h}\|^2. \end{aligned} \tag{4.2b}$$

Equation (4.2b) implies that the distance between two elements h and h' of $H_N \subset (C[0, L], \|\cdot\|_2, \langle\cdot,\cdot\rangle)$ is equal to the Euclidean distance between $T_N(h)$ and $T_N(h')$. It is not difficult to see that for each $h \in H_N$, $T_N(h)$ is the $(N+1)$-tuple made up of the coordinates of h with

respect to the basis set $\{B_{N,k}(\cdot)\}_{k=1}^{N+1}$ for H_N, defined by

$$B_{N,k}(\cdot) \triangleq W_N^{-1}(Q_N^{-1/2}e_k) = \sum_{i=1}^{N+1}(Q_N^{-1/2})_{ik}P_{N,i}(\cdot), \qquad k \in \mathbf{N+1}, \tag{4.2c}$$

where e_k denotes the kth canonical basis vector in $\mathbb{R}^{N+1}$. $\{B_{N,k}(\cdot)\}_{k=1}^{N+1}$ is an orthonormal basis set for H_N. Indeed,

$$\begin{aligned}\langle B_{N,k}(\cdot), B_{N,l}(\cdot)\rangle_2 &= \sum_{i,j=1}^{N+1}(Q_N^{-1/2})_{ik}(Q_N^{-1/2})_{jl}\langle P_{N,i}, P_{N,i}\rangle_2 \\ &= \sum_{i,j=1}^{N+1}(Q_N^{-1/2})_{ik}(Q_N^{-1/2})_{jl}(Q_N)_{ij} = e_k e_l^T.\end{aligned} \tag{4.2d}$$

Next, let $\bar{H}_N \triangleq T_N(H_N)$ and $\bar{H}_{\mathrm{ad},N} \triangleq T_N(H_{\mathrm{ad},N}) \subset \mathbb{R}^{N+1}$ so that

$$\begin{aligned}\bar{H}_{\mathrm{ad},N} = \{\bar{h} \in \mathbb{R}^{N+1} | &\alpha \le Q_N^{-1/2}\bar{h}_k \le \beta, k \in \mathbf{N+1}, |(Q_N^{-1/2}\bar{h})_{k+1} - (Q_N^{-1/2}\bar{h})_k| \\ &\le \gamma\Delta_N, k \in \mathbf{N}\}.\end{aligned} \tag{4.2e}$$

We define the mappings $\bar{\phi}_N^j(\cdot, x_{N,k}) : \bar{H}_{\mathrm{ad},N} \to \mathbb{R}$, $j \in \bar{\mathbf{q}}$, $k \in \mathbf{N+1}$, and $\bar{f}_N : \bar{H}_{\mathrm{ad},N} \to \mathbb{R}$ by

$$\bar{\phi}_N^j(\bar{h}, x_{N,k}) \triangleq \phi_N^j(T_N^{-1}(\bar{h}), x_{N,k}), \tag{4.3a}$$

$$\bar{f}_N(\bar{h}) \triangleq \sum_{k=1}^{N}\bar{\phi}_N^0(\bar{h}, x_{N,k})\Delta_N, \tag{4.3b}$$

which can be computed using (3.14a)–(3.14e), (3.17b), and (3.17c).

Finally, we define a family the problems $\bar{\mathbf{P}}_N$, $N = 1, 2, \ldots$, by

$$\bar{\mathbf{P}}_N \qquad \min_{\bar{h}\in\bar{H}_{\mathrm{ad},N}}\left\{\bar{f}_N(\bar{h}) | \max_{j\in\mathbf{q}}\max_{k\in\mathbf{N+1}}\bar{\phi}_N^j(\bar{h}, x_{N,k}) - (1+\Delta_N^{1/2})r^j(x_{N,k}) \le 0\right\}. \tag{4.4}$$

The following result, which establishes a correspondence between $\mathbf{P}_N$ and $\bar{\mathbf{P}}_N$, $N = 1, 2, \ldots$, follows directly from (3.17a)–(3.17c), (4.3a), (4.3b), and (4.4).

PROPOSITION 4.1. *Problems $\mathbf{P}_N$ and $\bar{\mathbf{P}}_N$ are equivalent in the following sense.* (a) *$h \in H_N$ is feasible for $\mathbf{P}_N$ if and only if $\bar{h} = T_N(h) \in \mathbb{R}^{N+1}$ is feasible for $\bar{\mathbf{P}}_N$; and* (b) *$h \in H_N$ is a global/local minimizer for $\mathbf{P}_N$ if and only if $\bar{h} = T_N(h) \in \mathbb{R}^{N+1}$ is a global/local minimizer for $\bar{\mathbf{P}}_N$.*

We define the constraint violation function $\bar{\psi}_N : \bar{H}_{\mathrm{ad},N} \to \mathbb{R}$; the surrogate cost $\bar{F}_N : \bar{H}_{\mathrm{ad},N} \times \bar{H}_{\mathrm{ad},N} \to \mathbb{R}$; the convex first-order approximation to $\bar{F}_N(\bar{h}, \bar{h}')$, denoted by $\tilde{\bar{F}}_N : \bar{H}_{\mathrm{ad},N} \times \bar{H}_{\mathrm{ad},N} \to \mathbb{R}$; and the optimality function $\bar{\theta}_N : \bar{H}_{\mathrm{ad},N} \to \mathbb{R}$ by

$$\bar{\psi}_N(\bar{h}) \triangleq \max_{j\in\mathbf{q}}\max_{k\in\mathbf{N+1}}\bar{\phi}_N^j(\bar{h}, x_{N,k}) - (1+\Delta_N^{1/2})r^j(x_{N,k}), \tag{4.5a}$$

$$\begin{aligned}\bar{F}_N(\bar{h}, \bar{h}') \triangleq \max\Big\{&\bar{f}_N(\bar{h}') - \bar{f}_N(\bar{h}) - \omega\bar{\psi}_N(\bar{h})_+, \\ &\max_{j\in\mathbf{q}}\max_{k\in\mathbf{N+1}}\bar{\phi}_N^j(\bar{h}', x_{N,k}) - \bar{\psi}_N(\bar{h})_+\Big\},\end{aligned} \tag{4.5b}$$

$$\begin{aligned}\tilde{\bar{F}}_N(\bar{h}, \bar{h}') \triangleq \max\Big\{&\langle\nabla\bar{f}_N(\bar{h}), \bar{h}' - \bar{h}\rangle - \omega\bar{\psi}_N(\bar{h})_+, \\ &\max_{j\in\mathbf{q}}\max_{k\in\mathbf{N+1}}\bar{\phi}_N^j(\bar{h}, x_k) - \bar{\psi}_N(\bar{h})_+ + \langle\nabla_1\bar{\phi}_N^j(\bar{h}, x_{N,k}), \bar{h}' - \bar{h}\rangle\Big\} \\ &+ \frac{1}{2}\|\bar{h}' - \bar{h}\|^2,\end{aligned} \tag{4.5c}$$

$$\bar{\theta}_N(\bar{h}) \triangleq \min_{\bar{h}'\in\bar{H}_{\mathrm{ad},N}}\tilde{\bar{F}}_N(\bar{h}, \bar{h}'). \tag{4.5d}$$

PROPOSITION 4.2. (a) $\bar{\theta}_N : \bar{H}_{\mathrm{ad},N} \to \mathbb{R}$ *is an optimality function for* $\bar{\mathbf{P}}_N$.

(b) *For any* $h, h' \in H_{\mathrm{ad},N}$, $\bar{h} = T_N(h)$, *and* $\bar{h}' = T_N(h')$ *we have*

$$F_N(h, h') = \bar{F}_N(\bar{h}, \bar{h}'), \tag{4.6a}$$

$$\theta_N(h) = \bar{\theta}_N(\bar{h}). \tag{4.6b}$$

Proof. A proof of part (a) can be found in [17]. Next we prove (b). First, in view of (3.17b), (3.18a), (4.3a), (4.3b), and (4.5a) it is clear that $\psi(h) = \bar{\psi}(\bar{h})$ and $f(h) = \bar{f}(\bar{h})$. Hence, it follows from (3.18b) and (4.5b) that $F_N(h, h') = \bar{F}_N(\bar{h}, \bar{h}')$.

Second, given $\delta h \in H_N$ let $\overline{\delta h} \triangleq T_N(\delta h)$. Then for all $j \in \bar{\mathbf{q}}$ and $k \in \mathbf{N+1}$ we have, using (4.3a) and the chain rule, that

$$\begin{aligned} \langle \nabla_1 \bar{\phi}_N^j(\bar{h}, x_{N,k}), \overline{\delta h} \rangle &= D_1 \bar{\phi}_N^j(\bar{h}, x_{N,k}; \overline{\delta h}) = D_1 \phi_N^j(T_N^{-1}(\bar{h}), x_{N,k}; T_N^{-1}(\overline{\delta h})) \\ &= D_1 \phi_N^j(h, x_{N,k}; \delta h), \end{aligned} \tag{4.7a}$$

which, together with (3.17b) and (4.4b), implies

$$\langle \nabla \bar{f}_N(\bar{h}), \bar{h}' - \bar{h} \rangle = Df_N(h; h' - h). \tag{4.7b}$$

It follows from (3.20a), (4.2b), (4.3a), (4.3b), (4.5c), (4.7a), and (4.7b) that for all $h, h' \in H_{\mathrm{ad},N}$,

$$\tilde{F}_N(h, h') = \bar{\tilde{F}}_N(\bar{h}, \bar{h}'), \tag{4.7c}$$

where $\bar{h} = T_N(h)$ and $\bar{h}' = T_N(h')$. Equality (4.6b) follows directly from the definition of $\bar{H}_{\mathrm{ad},N}$, (3.20b), (4.5d), and (4.7c). □

We now show how the gradients $\nabla_1 \bar{\phi}_N^j(\bar{h}, x_{N,k})$, $j \in \bar{\mathbf{q}}$, $k \in \mathbf{N+1}$, can be computed. Let e_i denote the ith canonical basis vector in $\mathbb{R}^{N+1}$. Then, for $\bar{h} \in \bar{H}_{\mathrm{ad},N}$, $j \in \bar{\mathbf{q}}$, and $k \in \mathbf{N+1}$, we have from (4.7a) that

$$\langle \nabla_1 \bar{\phi}_N^j(\bar{h}, x_{N,k}), e_i \rangle = D_1 \phi_N^j(T_N^{-1}(\bar{h}), x_{N,k}; T_N^{-1}(e_i)), \qquad i \in \mathbf{N+1}, \tag{4.8a}$$

which can be computed using (3.17c), (3.16a)–(3.16g), and the chain rule.

We will apply the algorithm described in [20] to solve problem **P** using the framework of consistent approximations, as suggested in [18].

ALGORITHM 4.3.

Parameters: $a, b, s \in (0, 1)$, $w, \epsilon > 0$ and $N_0 \in \mathbb{N}$.

Data. $h_0 \in H_{\mathrm{ad},N_0}$.

Step 0. Set $i = 0$.

Step 1.

Inner-Step 0. Set $N = N_i$, $\bar{h}_i = T_N(h_i)$.

Inner-Step 1. Compute

$$\bar{\theta}_N(\bar{h}_i) = \min_{\bar{h}' \in H_{\mathrm{ad},N}} \bar{\tilde{F}}_N(\bar{h}_i, \bar{h}'), \tag{4.9a}$$

$$d_i = \arg \min_{\bar{h}' \in \bar{H}_{\mathrm{ad},N}} \bar{\tilde{F}}_N(\bar{h}_i, \bar{h}'). \tag{4.9b}$$

Inner-Step 2. If $\bar{\theta}_N(\bar{h}_i) = 0$, set $\bar{h}_* = \bar{h}_i$ and go to Step 3. Else, compute the *step size*

$$\lambda_i \triangleq \arg \max_{k \in \mathbb{N}} \{b^k \mid \bar{F}_N(\bar{h}_i + b^k d_i, \bar{h}_i) \le b^k a \bar{\theta}_N(\bar{h}_i)\}. \tag{4.9c}$$

Inner-Step 3. Set

$$\bar{h}_* = \bar{h}_i + \lambda_i d_i. \tag{4.9d}$$

Step 2. If

$$\bar{F}_N(\bar{h}_i, \bar{h}_*) \le -\epsilon \Delta_N^s, \tag{4.9e}$$

go to Step 3. Else, set $h_i = T_N^{-1}(\bar{h}_i)$, replace N_i by $2N_i$ and go to Inner-Step 0.

Step 3. Set $h_{i+1} = T_{N_i}^{-1}(\bar{h}_*)$, $N_{i+1} = N_i$, replace i by $i + 1$, and go to Step 1.

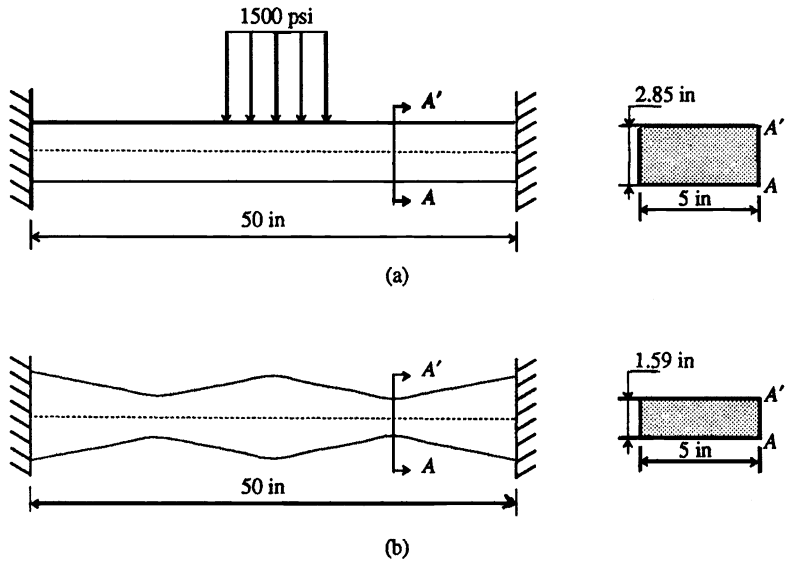


FIG. 5.1. (a) *Initial design*; (b) *final design.*

The following theorem on the convergence properties of Algorithm 4.3 can be deduced from Theorem 5.15 in [18].

THEOREM 4.4. *Suppose that Algorithm* 4.3 *has constructed an infinite sequence* $\{h_i\}_{i=0}^{\infty}$ *that has an accumulation point* $\hat{h}$. *Then* $\theta(\hat{h}) = 0$.

5. Numerical results. We will illustrate the use of consistent approximations and Algorithm 4.3 in solving a particular problem of the kind **P**. In our example, we assumed that $E = 10^7$ psi, $L = 50$ in, $b = 5$ in, $K = 0$ (we neglected the weight of the beam), $\alpha = 1.0$ in, $\beta = 5.0$ in, and $\gamma = 0.15$. We imposed continuum constraints on the maximum normal stress, on the maximum shear, and on the deflection, as follows:

$$|\sigma_{\max}(h, x)| \leq 30{,}000 \text{ psi} \quad \forall x \in [0, L], \tag{5.1a}$$

$$|\tau_{\max}(h, x)| \leq 15{,}000 \text{ psi} \quad \forall x \in [0, L], \tag{5.1b}$$

$$|y(h, x)| \leq 0.1 \text{ in} \quad \forall x \in [0, L]. \tag{5.1c}$$

The cost function was proportional to the total mass of the beam,

$$f(h) = \int_0^L h(x)\, dx. \tag{5.1d}$$

The load applied to the beam was

$$l(x) = \begin{cases} -1500 \text{ psi} & \text{if } x \in [20, 30], \\ 0 & \text{otherwise.} \end{cases} \tag{5.1e}$$

The initial discretization was set to $N = 8$ points; and the initial $h(\cdot)$ was constant, with value 2.85 in (see Fig. 5.1(a)). This initial design, whose cost is 142.5, corresponds to the uniform beam of least mass that satisfies the constraints (for this $h(\cdot)$ the constraint on the displacement is active and the other two are slack).

In Fig. 5.1(b), we find the beam obtained after 16 inner-steps of Algorithm 4.1. The discretization level at the end of the 16th inner-step was $N = 128$. The corresponding cost

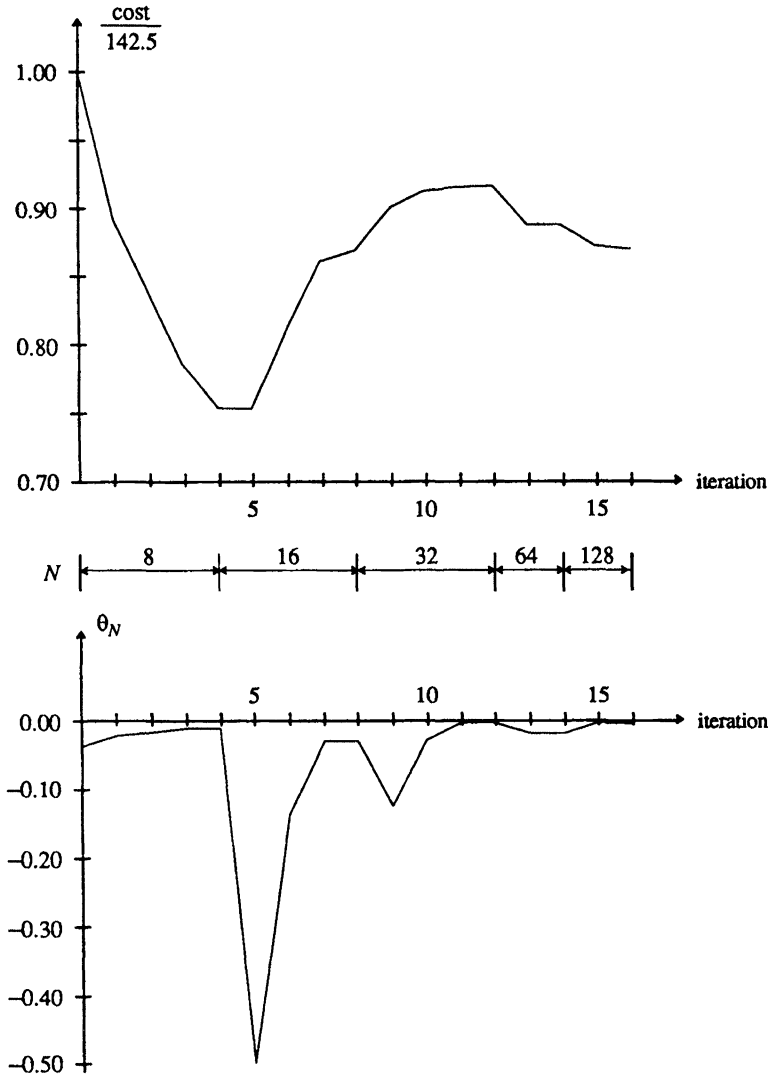


FIG. 5.2. *Computed cost and computed optimality function.*

was 124.05, about 87% of the initial cost. For the final design, the constraint on the deflection of the beam was active and the constraints on the maximum normal stress and on the maximum shear stress were slack.

In Fig. 5.2 we present the computed cost at each iteration as a percentage of the initial cost, 142.5, and the computed value of the optimality function θ_N, at each iteration. The number of discretization points used at each iteration is also shown in Fig. 5.2. As our analysis indicates, for each given discretization the optimality function is driven to zero, but when the discretization is refined (at iterations 4, 8, 12, and 14) the value of the optimality function may decrease. However, as the algorithm progresses, the optimality function is eventually driven to zero and, therefore, the computed depth functions $h_i(\cdot)$ approach a stationary point.

6. Conclusion. We have shown that one can obtain consistent approximations, satisfying the axioms formulated in [18], for certain classes of optimal beam design problems involving Euler–Bernoulli beams subject to continuum constraints, which include displacement,

maximum shear stress, and maximum normal stress constraints. We have also demonstrated numerically how an algorithm first described in [20] and proposed for use with consistent approximations in [18] can be used to obtain an arbitrarily good approximation to a stationary point of these design problems.

Appendix A.

Proof of Lemma 3.2. First we show that for all $h \in H_{\text{ad}}$ the determinants of the two principal minors of $A(h) \in \mathbb{R}^{2\times 2}$ are positive, which implies that $A(h)$ is positive definite. Clearly, $\int_0^L x^2/h(x)^3\,dx \geq L^3/3\beta^3 > 0$. Next, because all $h \in H_{\text{ad}}$ take values in $[\alpha, \beta]$, we have that for all $p_1, p_2, \epsilon \in (0, \infty)$

$$\int_0^1 \frac{(p_1x - p_2)^2}{h(x)^3}\,dx \geq \int_0^1 \frac{(p_1x - p_2)^2}{\beta^3}\,dx = \frac{p_1^2}{3\beta^3} + \frac{p_2^2}{\beta^3} - \frac{p_1p_2}{\beta^3} \geq \frac{p_1^2}{3\beta^3} + \frac{p_2^2}{\beta^3} - \frac{\epsilon p_1^2}{\beta^3} - \frac{p_2^2}{4\epsilon\beta^3}. \tag{A.1a}$$

If we set $\epsilon = 1/4$ and choose

$$p_1^2 = \int_0^1 \frac{1}{h(x)^3}\,dx \quad \text{and} \quad p_2 = \frac{1}{p_1}\int_0^1 \frac{x}{h(x)^3}\,dx, \tag{A.1b}$$

we get from (A.1a) that

$$\begin{aligned} \det A(h) &= \int_0^1 \frac{x^2}{h(x)^3}\,dx \int_0^1 \frac{1}{h(x)^3}\,dx - \left(\int_0^1 \frac{x}{h(x)^3}\,dx\right)^2 \\ &= \int_0^1 \frac{(p_1x - p_2)^2}{h(x)^3}\,dx \geq \frac{p_1^2}{12\beta^3} \geq \frac{1}{12\beta^6}, \end{aligned} \tag{A.1c}$$

which implies that $A(h) \in \mathbb{R}^{2\times 2}$ is positive definite.

Next, we observe that all entries of $A(h)$ are bounded by $(L^3 + 1)/\alpha^3$. Hence there exists an $M \in (0, \infty)$ such that for all $w \in \mathbb{R}^2$

$$w^T A(h) w \leq M\|w\|^2. \tag{A.1d}$$

Since the strictly positive lower bounds on the principal determinants are independent of $h \in H_{\text{ad}}$, it follows that the smallest eigenvalue of $A(h)$ is bounded away from 0 for all $h \in H_{\text{ad}}$, which implies the desired result. □

Proof of Lemma 3.8(a). Given $h \in H_{\text{ad}}$, let h_N be the linear interpolate of h on the mesh T_N. Clearly, $h_N \in H_{\text{ad},N}$. From (3.3), we have that h is Lipschitz continuous with Lipschitz constant γ, and hence $\|h - h_N\|_\infty \leq \gamma\Delta_N$, which proves (3.21a). □

PROPOSITION A1. *There exists a constant $C \in (0, \infty)$ such that for all $N \geq 1$, $h \in H_{\text{ad}}$, and $h_N \in H_{\text{ad},N}$,*

$$\max_{k\in\mathbf{N+1}} |M_c'(h, x_{N,k}) - M_{c,N}'(h_N, x_{N,k})| \leq C[\Delta_N + \|h - h_N\|_2], \tag{A.2a}$$

$$\max_{k\in\mathbf{N+1}} |M_c(h, x_{N,k}) - M_{c,N}(h_N, x_{N,k})| \leq C[\Delta_N + \|h - h_N\|_2], \tag{A.2b}$$

$$\|A(h) - A_N(h_N)\| \leq C[\|h - h_N\|_2 + \Delta_N], \tag{A.2c}$$

$$\|b(h) - b_N(h_N)\| \leq C[\|h - h_N\|_2 + \Delta_N], \tag{A.2d}$$

$$\|g(h) - g_N(h_N)\| \leq C[\|h - h_N\|_2 + \Delta_N], \tag{A.2e}$$

$$\max_{k\in\mathbf{N+1}} |M(h, x_{N,k}) - M_N(h_N, x_{N,k})| \leq C[\Delta_N + \|h - h_N\|_2], \tag{A.2f}$$

$$\max_{k\in\mathbf{N+1}} |V(h, x_{N,k}) - V_N(h_N, x_{N,k})| \leq C[\Delta_N + \|h - h_N\|_2], \tag{A.2g}$$

$$\max_{k\in\mathbf{N+1}} |y(h, x_{N,k}) - y_N(h_N, x_{N,k})| \leq C[\Delta_N + \|h - h_N\|_2]. \tag{A.2h}$$

Proof. Let $h \in H_{\text{ad}}$ and $h_N \in H_{\text{ad},N}$ be given. First, from (3.2a) it follows that

$$M_{\text{c}}''(h, x) - M_{\text{c}}''(h_N, x) = -K(h(x) - h_N(x)), \qquad x \in [0, L];$$

$$\text{(A.3a)} \qquad M_{\text{c}}'(h, L) - M_{\text{c}}'(h_N, L) = 0, \qquad M_{\text{c}}(h, L) - M_{\text{c}}(h_N, L) = 0.$$

Hence, integrating both sides of (A.3a) and using Holder's inequality, we get that

$$\text{(A.3b)} \qquad |M_{\text{c}}'(h, x) - M_{\text{c}}'(h_N, x)| \le K\sqrt{L}\|h - h_N\|_2, \qquad x \in [0, L].$$

If we integrate (A.3a) twice and use (A.3b) we get

$$\text{(A.3c)} \qquad |M_{\text{c}}(h, x) - M_{\text{c}}(h_N, x)| \le KL^{3/2}\|h - h_N\|_2, \qquad x \in [0, L].$$

Next, we show that there exists a $C \ge \max\{K\sqrt{L}, KL^{3/2}\}$ such that for all $N \in \mathbb{N}$ and $h_N \in H_{\text{ad},N}$,

$$\text{(A.3d)} \qquad \max_{k \in \mathbf{N+1}} |M_{\text{c},N}'(h_N, x_{N,k}) - M_{\text{c}}'(h_N, x_{N,k})| \le C\Delta_N,$$

$$\text{(A.3e)} \qquad \max_{k \in \mathbf{N+1}} |M_{\text{c},N}(h_N, x_{N,k}) - M_{\text{c}}(h_N, x_{N,k})| \le C\Delta_N,$$

where $M_{\text{c},N}'(h_N, x_{N,k})$ and $M_{\text{c},N}(h_N, x_{N,k})$ are determined by (3.14a). Equations (A.3d), (A.3e), (A.3b), and (A.3c) and the triangle inequality imply (A.2a) and (A.2b).

First we prove (A.3d). Recall that $m(\cdot)$ is assumed to be piecewise Lipschitz continuous (see (3.1)). From (3.1) and (3.3), it follows that for any $h_N \in H_{\text{ad},N} \subset H_{\text{ad}}$, $l(h_N, \cdot)$ is also piecewise Lipschitz continuous and has finitely many points of discontinuity in $[0, L]$. Hence, there exists a constant C', independent of $N \in \mathbb{N}$ and of $h_N \in H_{\text{ad},N}$, such that C' is a Lipschitz constant for $l(h_N, \cdot)$ on any subinterval of $[0, L]$ in which $l(h_N, \cdot)$ is continuous. Consider the mesh T_N. In each mesh interval $[x_{N,k}, x_{N,k+1}]$, $k \in \mathbf{N}$, $l(h_N, \cdot)$ is either Lipschitz continuous or has at least one point of discontinuity. There are at most finitely many mesh intervals, say $p \ge 0$, in which $l(h_N, \cdot)$ is discontinuous. Clearly, p is no larger than the number of discontinuities of $m(\cdot)$ and, hence, is independent of $N \in \mathbb{N}$. If we apply Euler's method to integrate the second equation in (3.2d), obtaining the second equation in (3.14a), the local truncation error, on each mesh interval where $l(h_N, \cdot)$ has at least one discontinuity, is bounded by $2\Delta_N \max_{x \in [0,L]} |l(h_N, x)|$. In the intervals where $l(h_N, \cdot)$ is Lipschitz continuous, and there are at most $N - p$ of these, the local truncation error of Euler's method is bounded by $C'\Delta_N^2$ such that

$$\text{(A.3f)} \qquad |M_{\text{c},N}'(h_N, x_{N,k}) - M_{\text{c}}'(h_N, x_{N,k})| \le C'\Delta_N^2(N - p) + 2p \max_{x \in [0,L]} |l(h_N, x)|\Delta_N,$$

which implies (A.3d). To prove (A.3e) we first note that if we set

$$\text{(A.3g)} \qquad K' \triangleq \max_{x \in [0,L]} |m(x)| + K\beta,$$

where K is as in (3.1), then K' is Lipschitz constant for $M_{\text{c}}'(h, x)$ for all $h \in H_{\text{ad}}$; that is,

$$\text{(A.3h)} \qquad |M_{\text{c}}'(h, x) - M_{\text{c}}'(h, y)| \le K'|x - y| \quad \forall\, h \in H_{\text{ad}}\ \forall\, x, y \in [0, L].$$

Now consider the first differential equation in (3.2b) and its discrete counterpart in (3.14a). Their solutions satisfy

$$\text{(A.3i)} \qquad M_{\text{c}}(h, x_{N,k}) = M_{\text{c}}(h, x_{N,k+1}) + \int_{x_{N,k+1}}^{x_{N,k}} M_{\text{c}}'(h, x)\, dx, \qquad k \in \mathbf{N},$$

$$\text{(A.3j)} \qquad M_{\text{c},N}(h, x_{N,k}) = M_{\text{c},N}(h, x_{N,k+1}) + \int_{x_{N,k+1}}^{x_{N,k}} M_{\text{c},N}'(h, x_{N,k+1})\, dx, \qquad k \in \mathbf{N}.$$

Defining $e_k \triangleq |M_c(h, x_{N,k}) - M_{c,N}(h, x_{N,k})|$, $k \in \mathbf{N+1}$, and subtracting (A.3j) from (A.3i), we obtain, after taking absolute values on both sides,

$$e_k \leq e_{k+1} + \int_{x_{N,k+1}}^{x_{N,k}} |M_c'(h, x) - M_{c,N}'(h, x_{N,k+1})|\, dx, \qquad k \in \mathbf{N}, e_{N+1} = 0. \tag{A.3k}$$

Adding and subtracting $M_c'(h, x_{N,k+1})$ to the integrand in (A.3k) and using the triangle inequality we get

$$\begin{aligned} e_k \leq e_{k+1} &+ \int_{x_{N,k+1}}^{x_{N,k}} |M_c'(h, x) - M_c'(h, x_{N,k+1})|\, dx + \int_{x_{N,k+1}}^{x_{N,k}} |M_c'(h, x_{N,k+1}) \\ &- M_{c,N}'(h, x_{N,k+1})|\, dx, \end{aligned} \tag{A.3l}$$

which, in view of (A.3d) and (A.3h) implies that there exists a constant $C < \infty$ such that

$$e_k \leq e_{k+1} + C\Delta_N^2, \qquad k \in \mathbf{N}, e_{N+1} = 0, \tag{A.3m}$$

which in turn implies $|e_k| \leq C\Delta_N$ for all $k \in \mathbf{N+1}$ and hence proves (A.3e).

Inequalities (A.2c) and (A.2d) follow from the fact that $h(x)$ and $h_N(x)$ take values in $[\alpha, \beta]$, Holder's inequality, the fact that the rectangle rule is $O(\Delta_N)$, and from the definitions of $A(h)$ and $b(h)$ and of $A_N(h_N)$ and $b_N(h_N)$ (see (3.2d) and (3.14d)). To prove (A.2e), we first note that $A(h)g(h) = b(h)$ and $A_N(h_N)g_N(h_N) = b_N(h_N)$ imply

$$g(h) - g_N(h_N) = A(h)^{-1}[(A_N(h_N) - A(h))g_N(h_N) + (b(h) - b_N(h_N))], \tag{A.3n}$$

which in view of Lemmas 3.2 and 3.8 and (A.2c,d) implies (A.2e).

Inequalities (A.2f)–(A.2h) follow directly from (3.2c), (3.2e), and (3.2f) and (A.2a), (A.2b), (A.2e), and (A.2f). □

Proof of Lemma 3.8(b). In view of Assumption 3.3(b), (3.21b) is a direct consequence of (A.2f)–(A.2h), and the definitions of $\phi^j(\cdot, \cdot)$ and $\phi_N^j(\cdot, \cdot)$ in (3.7c) and (3.6b), respectively. Next, if we let $R \triangleq \max_{j \in \mathbf{q}} \max_{x \in [0,L]} r^j(x)$ and make use of (3.21b), it follows from (3.10a) and (3.18a) that

$$\begin{aligned} \psi(h) - \psi_N(h_N) &\leq \max_{j \in \mathbf{q}} \max_{x \in [0,L]} \{\phi^j(h, x) - r^j(x) - \phi_N^j(h_N, x) + (1 - \Delta_N^{1/2}) r^j(x)\} \\ &\leq C[\Delta_N + \|h - h_N\|_2] + R\Delta_N^{1/2} \leq C[\Delta_N^{1/2} + \|h - h_N\|_2]. \end{aligned} \tag{A.4a}$$

In a similar way, an upper bound for $\psi_N(h_N) - \psi(h)$ can be obtained, namely,

$$\psi_N(h_N) - \psi(h) \leq C[\Delta_N^{1/2} + \|h - h_N\|_2], \tag{A.4b}$$

which together with (A.4a) implies (3.21c).

Finally, we prove (3.21d). First we note that because all $h \in H_{ad}$ take values between $[\alpha, \beta]$ and $l(\cdot, \cdot)$ is bounded, $V(h, \cdot)$, $M(h, \cdot)$, and $y(h, \cdot)$ are Lipschitz continuous on $[0, L]$, with a common Lipschitz constant independent of $h \in H_{ad}$. In view of Assumption 3.3(b) we get that there exists a constant C such that for all $x, y \in [0, L]$

$$|\phi^0(h, x) - \phi^0(h, y)| \leq C|x - y|. \tag{A.4c}$$

Hence,

$$|f(h) - f_N(h_N)| \leq \sum_{j=1}^{N} \int_{x_{N,j}}^{x_{N,j+1}} \{|\phi^0(h, x) - \phi^0(h, x_{N,j})| + |\phi^0(h, x_{N,j}) - \phi_N^0(h_N, x_{N,j})|\}\, dx, \tag{A.4d}$$

which in view of (A.4c) and (3.21c) implies that there exists a constant C such that

$$|f(h) - f_N(h_N)| \leq C[\Delta_N + \|h - h_N\|_2]. \tag{A.4e}$$

□

The proof of the following result is similar to that of Proposition A.1 and hence is omitted.

PROPOSITION A.2. *There exists a constant* $C \in (0, \infty)$ *such that for all integers* $N > 0$, $h, h' \in H_{\mathrm{ad},N}$, *we have*

$$\max_{k\in\mathbf{N+1}} |D_1 M_{\mathrm{c}}(h, x_{N,k}; h' - h) - D_1 M_{\mathrm{c},N}(h, x_{N,k}; h' - h)| \le C\|h' - h\|_2 \Delta_N, \tag{A.5a}$$

$$\max_{k\in\mathbf{N+1}} |D_1 M'_{\mathrm{c}}(h, x_{N,k}; h' - h) - D_1 M'_{\mathrm{c},N}(h, x_{N,k}; h' - h)| \le C\|h' - h\|_2 \Delta_N, \tag{A.5b}$$

$$\|DA(h; h' - h) - DA_N(h; h' - h)\| \le C\Delta_N \|h' - h\|_2, \tag{A.5c}$$

$$\|Db(h; h' - h) - Db_N(h; h' - h)\| \le C\Delta_N \|h' - h\|_2, \tag{A.5d}$$

$$\|Dg(h; h' - h) - Dg_N(h; h' - h)\| \le C\Delta_N \|h' - h\|_2. \tag{A.5e}$$

$$\max_{k\in\mathbf{N+1}} |D_1 M(h, x_{N,k}; h' - h) - D_1 M_N(h, x_{N,k}; h' - h)| \le C\|h' - h\|_2 \Delta_N, \tag{A.5f}$$

$$\max_{k\in\mathbf{N+1}} |D_1 V(h, x_{N,k}; h' - h) - D_1 V_N(h, x_{N,k}; h' - h)| \le C\|h' - h\|_2 \Delta_N, \tag{A.5g}$$

$$\max_{k\in\mathbf{N+1}} |D_1 y(h, x_{N,k}; h' - h) - D_1 y_N(h, x_{N,k}; h' - h)| \le C\|h' - h\|_2 \Delta_N. \tag{A.5h}$$

Proof of Lemma 3.11. The result follows directly from (A.5f)–(A.5h) and Assumption 3.3. □

REFERENCES

[1] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Pitman, London, 1984.

[2] F. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[3] T. E. BAKER AND E. POLAK, *On the optimal control of systems described by evolution equations*, SIAM J. Control Optim., 32 (1994), pp. 224–260.

[4] N. V. BANICHUK, *Problems and Methods of Optimal Structural Design*, Plenum Press, New York, 1983.

[5] ———, *Introduction to Optimization of Structures*, Springer-Verlag, New York, 1990.

[6] D. BEGIS AND R. GLOWINSKI, *Application de la methode des elements finis a l'approximation d'un probleme de domaine optimal. Methodes de resolution des problemes approaches*, Appl. Math. Optim., 2 (1975), pp. 130–169.

[7] J. BORGGAARD, On the presence of shocks in domain optimization of Euler flows, personal communication, 1995.

[8] S. DOLECKI, G. SALINETI, AND R. J.-B. WETS, *Convergence of functions: Equisemicontinuity*, Trans. Amer. Math. Soc., 276 (1983), pp. 409–429.

[9] J. HASLINGER AND P. NEITTAANMAKI, *Finite Element Approximation for Optimal Shape Design: Theory and Applications*, Wiley, New York, 1988.

[10] E. J. HAUG AND J. S. ARORA, *Applied Optimal Design: Mechanical and Structural Systems*, Wiley, New York, 1979.

[11] L. HE AND E. POLAK, *Effective diagonalization strategies for the solution of a class of optimal design problems*, IEEE Trans. Automat. Control, 35 (1990), pp. 258–267.

[12] E. J. HAUG AND J. CEA, eds., *Optimization of Distributed Parameter Structures*, Sijthoff and Noordhoff, Rockville, MD, 1981.

[13] I. HLAVACEK, I. BOCK, AND J. LOVISEK, *Optimal control of a variational inequality with applications to structural analysis. Part* I. *Optimal design of a beam with unilateral supports*, Appl. Math. Optim., 11 (1984), pp. 111–142.

[14] S. LANG, *Real Analysis*, 2nd ed., Addison-Wesley, Reading, MA, 1983.

[15] R. MAKINEN, *On Numerical Methods for State Constrained Optimal Shape Design Problems*, Internat. Ser. Numer. Math. 91, Birkhäuser, Basel, 1989, pp. 283–299.

[16] O. PIRONNEAU, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, New York, 1984.

[17] E. POLAK, *On the mathematical foundations of nondifferentiable optimization in engineering design*, SIAM Rev., 29 (1987), pp. 21–91.

[18] ———, *On the use of consistent approximations in the solution of semi-infinite optimization and optimal control problems*, Math. Programming B, 62 (1993), pp. 385–414.

[19] ———, *Computational Methods in Optimization: A Unified Approach*, 2nd ed., Academic Press, New York, to appear.

[20] E. POLAK AND L. HE, *A unified steerable phase* I*-phase* II *method of feasible directions for semi-infinite optimization*, J. Optim. Theory Appl., 69 (1991), pp. 83–107.

[21] J. N. REEDY, *Energy and Variational Methods in Applied Mechanics*, Wiley, New York, 1984.

[22] A. SCHWARTZ AND E. POLAK, *Consistent approximations for optimal control problems based on Runge-Kutta integration*, SIAM J. Control Optim., 34 (1996), pp. 1235–1269.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 1914–1946, December 1996

005

CLASSIFICATION OF GENERIC SINGULARITIES FOR THE PLANAR TIME-OPTIMAL SYNTHESIS*

B. PICCOLI†

Abstract. This paper is concerned with control systems on the plane with control appearing linearly. It is known that under generic conditions the problem of reaching points from the origin in minimum time admits a regular synthesis. The minimum time function is piecewise smooth, possibly nondifferentiable on a set that is a finite union of embedded submanifolds of dimension 1 or 0, called singularities. The purpose of the present paper is to provide a classification of all types of singularities arising under generic conditions.

Key words. time-optimal control, two-dimensional system, regular synthesis

AMS subject classifications. 93C10, 93B20

1. Introduction.

This paper is concerned with the control system on the plane

$$\dot{x} = F(x) + u\,G(x), \qquad |u| \le 1, \tag{1.1}$$

with $F(0) = 0$.

Let $R(\tau)$ be the reachable set for (1.1) at a given time τ. Under generic conditions on the vector fields F and G, it is known that the problem of reaching points $x \in R(\tau)$ from the origin in minimum time admits a regular synthesis [8]–[11]. Indeed, one can partition the set $R(\tau)$ into finitely many embedded submanifolds $\mathcal{M}_i$ such that, on each $\mathcal{M}_i$, the corresponding optimal control is either $u \equiv \pm 1$ or singular, i.e., $u = \varphi_S(x) \notin \{-1, +1\}$, where

$$\varphi_S(x) = -\frac{\nabla \Delta_B(x) \cdot F(x)}{\nabla \Delta_B(x) \cdot G(x)}.$$

See (2.11) for the definition of Δ_B.

In particular, the minimum time function is piecewise smooth on $R(\tau)$, possibly nondifferentiable on the boundaries of the submanifolds $\mathcal{M}_i$. These boundaries, of dimension 1 or 0, were called, in [8], frame curves (FCs) and frame points, respectively.

The purpose of the present paper is to provide a classification of all types of FCs and frame points, arising under generic conditions on F and G. In §§3 and 4 we classify all generic types of singularities, up to a relation of topological equivalence, and construct explicit examples of each one of them. We refer to the examples as representations of the various equivalence classes.

This paper is a first part of a research aimed at the generic classification of feedbacks, analogous to the work of Peixoto; for ordinary differential equations see [7]. The research will be completed in a subsequent paper.

2. Basic definitions.

We consider the Banach space $\mathcal{V}$ of $\mathcal{C}^3$ vector fields on $\mathbb{R}^2$, having all derivatives, up to order three, continuous and bounded on the plane. We endow the space $\mathcal{V}$ with the norm

$$F = (F_1, F_2), \qquad \| F \|_{\mathcal{C}^3} = \sup \{ |D^\alpha F_i(x)| : x \in \mathbb{R}^2;\ |\alpha| \le 3;\ i = 1, 2 \}.$$

We denote by $\Xi \subset \mathcal{V} \times \mathcal{V}$ the Banach subspace of couples (F, G) such that $F(0) = 0$, endowed with the restriction of the product norm. With every couple $(F, G) \in \Xi$ we associate the control system (1.1), called Σ, and write $\Sigma \in \Xi$.

*Received by the editors September 24, 1993; accepted for publication (in revised form) August 3, 1995.
†Scuola Internazionale Superiore di Studi Avanzati-ISAS, via Beirut 2–4, 34014 Trieste, Italy.

We recall that the *Lie bracket* of two vector fields F and G is the vector field

$$[F, G] \doteq \nabla G \cdot F - \nabla F \cdot G. \tag{2.1}$$

A *control* is a measurable function $u : [a, b] \mapsto [-1, 1]$, where $-\infty < a \leq b < +\infty$. A *trajectory of* Σ corresponding to u is an absolutely continuous curve $\gamma : [a, b] \mapsto \mathbb{R}^2$ which satisfies the equation

$$\dot{\gamma}(t) = F(\gamma(t)) + u(t)G(\gamma(t)) \tag{2.2}$$

for almost every t in the domain of u. We write $\mathrm{Dom}(u)$ (resp., $\mathrm{Dom}(\gamma)$) to indicate the domain of u (resp., γ) and $\gamma \restriction [c, d]$ to denote the restriction of γ to $[c, d] \subset \mathrm{Dom}(\gamma)$. The initial point of γ is denoted by $\mathrm{In}(\gamma) = \gamma(a)$, and its terminal point by $\mathrm{Term}(\gamma) = \gamma(b)$. The time along γ is defined as

$$T(\gamma) = b - a. \tag{2.3}$$

A trajectory $\gamma \in \mathrm{Traj}(\Sigma)$ (the set of trajectories of Σ) is *time optimal* if, for every trajectory γ' having the same initial and terminal points, one has $T(\gamma') \geq T(\gamma)$.

If $u_1 : [a, b] \mapsto [-1, 1]$ and $u_2 : [b, c] \mapsto [-1, 1]$ are controls, their *concatenation* $u_2 * u_1$ is the control

$$(u_2 * u_1)(t) \doteq \begin{cases} u_1(t) & \text{for } t \in [a, b], \\ u_2(t) & \text{for } t \in (b, c]. \end{cases}$$

If $\gamma_1 : [a, b] \mapsto \mathbb{R}^2$, $\gamma_2 : [b, c] \mapsto \mathbb{R}^2$ are trajectories of Σ for u_1 and u_2 such that $\gamma_1(b) = \gamma_2(b)$, then the *concatenation* $\gamma_2 * \gamma_1$ is the trajectory

$$(\gamma_2 * \gamma_1)(t) = \begin{cases} \gamma_1(t) & \text{for } t \in [a, b], \\ \gamma_2(t) & \text{for } t \in [b, c]. \end{cases}$$

For convenience, we also define the vector fields

$$X = F - G, \qquad Y = F + G.$$

We say that γ is an X-trajectory and write $\gamma \in \mathrm{Traj}(X)$ if it corresponds to the constant control -1. Similarly we define Y-trajectories.

If a trajectory γ is concatenation of an X-trajectory and a Y-trajectory, then we say that γ is a $Y * X$-trajectory. (The X-trajectory comes first.) Similarly we define trajectories of type $X * Y$, $X * Y * X$, and so on. For a complete description of notations see [9].

If $\tau \geq 0$, we denote by $R(\tau)$ the *reachable set* within time τ:

$$R(\tau) = \{x : \exists \gamma \in \mathrm{Traj}(\Sigma) \quad \text{s.t. } \gamma(0) = 0 \in \mathbb{R}^2, \ \gamma(t) = x, \ \text{for some } t \leq \tau\}. \tag{2.4}$$

The *minimum time function*, $T : \mathbb{R}^2 \mapsto [0, +\infty]$, is defined by

$$T(x) \doteq \inf\{\tau : x \in R(\tau)\}. \tag{2.5}$$

The control system Σ is *locally controllable* if, for each $\tau > 0$, the set $R(\tau)$ contains a neighborhood of the origin. The following lemma is well known (see [6]).

LEMMA 2.1. *If $F(0) = 0$ and the vector fields G, $[F, G]$ are linearly independent at the origin, then the system $\Sigma = (F, G)$ in (1.1) is locally controllable.*

A *synthesis* for the control system Σ at time τ is a family $\Gamma = \{\gamma_x : [0, b_x] \mapsto \mathbb{R}^2 | x \in R(\tau)\}$ of trajectories satisfying the following conditions:

(a) For each $x \in R(\tau)$ one has $\gamma_x(0) = 0$, $\gamma_x(b_x) = x$.

(b) If $y = \gamma_x(t)$, where $t \in \text{Dom}(\gamma)$, then $\gamma_y = \gamma_x \restriction [0, t]$.

A synthesis for the system Σ is *time optimal* if, for each $x \in R(\tau)$, one has $\gamma_x(T(x)) = x$, where T is the minimum time function defined at (2.5).

An *admissible pair* for the system Σ is a couple (u, γ) such that u is a control and γ is a trajectory corresponding to u. We use the symbol $\text{Adm}(\Sigma)$ to denote the set of admissible pairs and say that $(u, \gamma) \in \text{Adm}(\Sigma)$ is optimal if γ is optimal.

A *variational vector field along* $(u, \gamma) \in \text{Adm}(\Sigma)$ is the vector-valued absolutely continuous function $v : \text{Dom}(\gamma) \mapsto \mathbb{R}^2$ that satisfies the equation

$$\dot{v}(t) = ((\nabla F)(\gamma(t)) + u(t)(\nabla G)(\gamma(t))) \cdot v(t) \tag{2.6}$$

for almost all $t \in \text{Dom}(\gamma)$.

A *variational covector field along* $(u, \gamma) \in \text{Adm}(\Sigma)$ is an absolutely continuous function $\lambda : \text{Dom}(\gamma) \mapsto \mathbb{R}^2_*$ that satisfies the equation

$$\dot{\lambda}(t) = -\lambda(t) \cdot ((\nabla F)(\gamma(t)) + u(t)(\nabla G)(\gamma(t))) \tag{2.7}$$

for almost all $t \in \text{Dom}(\gamma)$. Here $\mathbb{R}^2_*$ denotes a space of row vectors.

The Hamiltonian $\mathcal{H} : \mathbb{R}^2_* \times \mathbb{R}^2 \times \mathbb{R} \mapsto \mathbb{R}$ is defined as

$$\mathcal{H}(\lambda, x, u) = \lambda \cdot (F(x) + uG(x)). \tag{2.8}$$

If λ is a variational covector field along $(u, \gamma) \in \text{Adm}(\Sigma)$, we say that λ is *maximizing* if

$$\mathcal{H}(\lambda(t), \gamma(t), u(t)) = \max \{\mathcal{H}(\lambda(t), \gamma(t), w) : |w| \leq 1\} \tag{2.9}$$

for almost all $t \in \text{Dom}(\gamma)$.

The *Pontryagin maximum principle* (PMP) states that if $(u, \gamma) \in \text{Adm}(\Sigma)$ is time optimal, then there exists

(PMP1) a nontrivial maximizing variational covector field λ along (u, γ);

(PMP2) a constant $\lambda_0 \leq 0$ such that $\mathcal{H}(\lambda(t), \gamma(t), u(t)) + \lambda_0 = 0$ for almost all $t \in \text{Dom}(\gamma)$.

In this case λ is called an *adjoint covector field along* (u, γ) or simply an *adjoint variable*, and we say that (γ, λ) satisfies the PMP or that γ is an *extremal trajectory*. Moreover, the function $\phi_\lambda(t) \doteq \lambda(t) \cdot G(\gamma(t))$ is called the switching function along (u, γ, λ).

Consider $(u, \gamma) \in \text{Adm}(\Sigma)$, $t_0 \in \text{Dom}(\gamma)$, and $v_0 \in \mathbb{R}^2$. We write $v(v_0, t_0; t)$ to denote the value at time t of the variational vector field along (u, γ) satisfying (2.6) together with the boundary condition $v(t_0) = v_0$. If $t_0, t_1 \in \text{Dom}(\gamma)$, we say that t_0 and t_1 are *conjugate* along γ if the vectors $v(G(\gamma(t_1)), t_1; t_0)$ and $G(\gamma(t_0))$ are linearly dependent. Let D and D' be two $\mathcal{C}^3$ connected one-dimensional embedded submanifolds of $\mathbb{R}^2$. We say that D' is a *conjugate curve* to D along the X-trajectories if there is a bijective function $\psi : D \mapsto D'$ with the following properties. If γ_x is the X-trajectory satisfying $\gamma_x(0) = x$, then $\psi(x) = \gamma_x(t(x))$ for some time t depending continuously on x, and the times 0, $t(x)$ are conjugate along γ_x. Conjugate curves along the Y-trajectories are defined similarly.

For each $x \in \mathbb{R}^2$, one can form the 2×2 matrices whose columns are the vectors F, G, or $[F, G]$. As in [9], we shall use the following scalar functions on $\mathbb{R}^2$:

$$\Delta_A(x) \doteq \det\big(F(x), G(x)\big), \tag{2.10}$$

$$\Delta_B(x) \doteq \det\big(G(x), [F, G](x)\big), \tag{2.11}$$

where det stands for determinant. A point $x \in \mathbb{R}^2$ is called an *ordinary point* if

$$\Delta_A(x) \cdot \Delta_B(x) \neq 0. \tag{2.12}$$

On the set of ordinary points we define the scalar functions f and g as the coefficients of the linear combination

$$[F, G](x) = f(x)F(x) + g(x)G(x). \tag{2.13}$$

In [9, p. 447] it was shown that

$$f(x) = -\frac{\Delta_B(x)}{\Delta_A(x)}. \tag{2.14}$$

These functions play a key role in the study of the structure of time-optimal trajectories. Of particular interest are the curves formed by zeros of the function Δ_B called turnpikes. These are the only curves that can be run by optimal trajectories corresponding to controls different from ± 1.

A point x at which $\Delta_A(x)\Delta_B(x) = 0$ is called a *nonordinary point*. A C^2 one-dimensional connected embedded submanifold S of $\mathbb{R}^2$, with the property that every $x \in S$ is a nonordinary point, is a *nonordinary arc* and it is *isolated* (or it is an INOA) if there exists a set U satisfying the following conditions:

(C1) U is an open connected subset of $\mathbb{R}^2$.

(C2) S is a relatively closed subset of U.

(C3) If $x \in U \setminus S$, then x is an ordinary point.

(C4) The set $U \setminus S$ has exactly two connected components.

A *turnpike* is an isolated nonordinary arc that satisfies the following conditions:

(S1) For each $x \in S$ the vectors $X(x)$ and $Y(x)$ are not tangent to S and point to opposite sides of S.

(S2) For each $x \in S$ one has $\Delta_B(x) = 0$ and $\Delta_A(x) \neq 0$.

(S3) Let U be an open set which satisfies (C1)–(C4) above. If U_X and U_Y are the connected components of $U \setminus S$ labelled in such a way that $X(x)$ points into U_X and $Y(x)$ points into U_Y, then the function f in (2.14) satisfies

$$f(x) > 0 \quad \text{on} \quad U_Y, \qquad\qquad f(x) < 0 \quad \text{on} \quad U_X.$$

Next, consider a turnpike S and a point $x_0 \in S$. We wish to construct a trajectory $\gamma \in \text{Traj}(\Sigma)$ such that $\gamma(t_0) = x_0$ and $\gamma(t) \in S$ for each $t \in \text{Dom}(\gamma) \doteq [t_0, t_1]$. After straightforward calculations we have that γ must correspond to the feedback control

$$\varphi_S(x) \doteq -\frac{\nabla\Delta_B \cdot F(x)}{\nabla\Delta_B \cdot G(x)}. \tag{2.15}$$

The turnpike S is said to be *regular* if the function φ_S in (2.15) satisfies

$$|\varphi_S(x)| < 1, \qquad x \in S. \tag{2.16}$$

A trajectory γ is said to be a *Z- or S-trajectory* or *singular trajectory* if there exists a regular turnpike S such that $\{\gamma(t) : t \in \text{Dom}(\gamma)\} \subset S$; in this case we write $\gamma \in \text{Traj}(Z)$.

Given a trajectory γ we denote by $n(\gamma)$ the smallest integer such that there exist $\gamma_i \in \text{Traj}(X) \cup \text{Traj}(Y) \cup \text{Traj}(Z)$, $i = 1, \ldots, n(\gamma)$, verifying

$$\gamma = \gamma_{n(\gamma)} * \cdots * \gamma_1.$$

We call $n(\gamma)$ the number of arcs of γ.

Given $\tau > 0$, define Π_τ to be the class of systems having an a priori bound on the number of arcs of optimal trajectories:

$$\Pi_\tau = \{\Sigma \in \Xi : \exists N(\Sigma) \text{ s.t. } \forall \gamma : [0, \tau] \to \mathbb{R}^2, \text{ optimal, } n(\gamma) \leq N(\Sigma)\}.$$

A subset of Ξ is said to be *generic* if it contains an open and dense subset of Ξ. In [8] the following theorem was proven.

THEOREM 2.2. *For every $\tau > 0$ the set Π_τ is a generic subset of Ξ.*

Given a system $\Sigma \in \Pi_\tau$ it is possible to construct a synthesis Γ for Σ. We can follow the classical idea of constructing extremal trajectories and deleting those trajectories which are not globally optimal. At the end we obtain a set of trajectories from which we can extract a synthesis. By construction, this synthesis is optimal. For synthesis theory see [2]–[4] and [14].

We describe an algorithm $\mathcal{A}$ by induction. At step N, we construct precisely those trajectories which are concatenation of N bang- or singular arcs and satisfy the PMP. The endpoints of the arcs forming these trajectories, corresponding to the switching times of the control, are determined by certain nonlinear equations. Under generic conditions, such equations can be solved by the implicit function theorem, thus determining a smooth switching locus. Eventually the algorithm will partition the reachable set $R(\tau)$ into finitely many open regions (where the optimal feedback control is either $u = 1$ or $u = -1$), separated by boundary curves and points, called *frame curves* and *frame points*, respectively.

At each step, it may happen that distinct extremal trajectories reach the same point x_0 at different times. It is therefore necessary to delete from the synthesis those trajectories which are not globally optimal. This procedure will usually produce new "overlap curves," consisting of points reached in minimum time by two distinct trajectories, one ending with the control value $u = 1$, the other with $u = -1$.

If at step N the algorithm $\mathcal{A}$ does not construct any new trajectory, then we say that $\mathcal{A}$ *stops at step N (for Σ at time τ)* or that $\mathcal{A}$ *succeeds for* Σ. From Theorem 2.2, it is clear that under generic assumptions there exists $N(\Sigma)$ such that $\mathcal{A}$ stops before step $N(\Sigma)$ and every γ constructed by $\mathcal{A}$ is optimal. By definition, $\mathrm{Fr}(R(\tau))$ is a frame curve, and its intersections with other frame curves are frame points.

If $\mathcal{A}$ stops, then for each $x \in R(\tau)$ there exists a set of constructed trajectories that reach x. Define $\Gamma_x \doteq \{\gamma : \gamma \text{ is constructed by } \mathcal{A}, \mathrm{Term}(\gamma) = x\}$.

We want to select, for each $x \in R(\tau)$, a trajectory from Γ_x to form a synthesis. Define K_k to be the set of points $x \in R(\tau)$ reached by at least one constructed trajectory γ satisfying $n(\gamma) \leq k$. Note that K_k is compact for each k and $K_{N(\Sigma)} = R(\tau)$. We proceed by induction on k. Given $x \in K_k \setminus K_{k-1}$, we consider the optimal trajectories $\gamma \in \Gamma_x$ formed by k arcs, for which the following holds. If $y = \gamma(t)$ is the initial point of the last arc of γ, then $\gamma \restriction [0, t]$ has been selected from Γ_y by induction. Finally, if there is more than one such trajectory, then we select one, say, according to the preference order X, Y, Z on the type of the last arc.

In this way, at step $N(\Sigma)$, we have constructed a synthesis for Σ at time τ. We use the symbol $\Gamma_{\mathcal{A}}(\Sigma, \tau)$ to denote this synthesis and call it *the synthesis generated by the algorithm $\mathcal{A}$.*

THEOREM 2.3. *Consider $\Sigma \in \Xi$ and $\tau > 0$. If $\mathcal{A}$ stops for Σ at time τ, then $\Gamma_{\mathcal{A}}(\Sigma, \tau)$ is an optimal synthesis.*

Remark 2.1. The points of overlap curves are reached by two different optimal trajectories. Moreover, if a trajectory $\gamma \neq \gamma^{\pm}$ (see (F1) and (F2) below) goes through an endpoint of an overlap curve, say, at time t_0, then all points $\gamma(t)$, $t \geq t_0$, are reached in two different optimal way. See Example 13 in §4. These are the only cases in which a point can be reached by two different optimal trajectories. Now, if a trajectory γ as above enters a turnpike, then

there is an open region on which the synthesis is not unique. But this situation is not generic. Therefore, for systems in a generic set the trajectories γ of the above type do not reach turnpikes, and the synthesis is unique, excluding a finite set of one-dimensional manifolds. Hence we have that the synthesis $\Gamma_{\mathcal{A}}(\Sigma, \tau)$ depends on the algorithm $\mathcal{A}$, but generically the optimal synthesis is *essentially* unique.

Given $x \in R(\tau)$ we denote by γ_x, u_x the trajectory of $\Gamma_{\mathcal{A}}(\Sigma, \tau)$ and the corresponding control such that $\gamma_x(t_x) = x$. If x does not belong to any frame curve, then we denote by $u_{\mathcal{A}}(x)$ the control $u_x(t_x)$. We have $|u_{\mathcal{A}}| = 1$.

The algorithm constructs only six types of frame curves:

(F1) the trajectory γ^-, starting from 0 and corresponding to the control $u^- \equiv -1$.

(F2) the trajectory γ^+, starting from 0 and corresponding to the control $u^+ \equiv 1$.

(F3) the topological frontier of the reachable set: $\mathrm{Fr}(R(\tau))$.

(F4) conjugate curves to other frame curves, also called switching curves.

(F5) regular turnpikes.

(F6) overlap curves.

To denote these types of curves we use, respectively, the symbols: X, Y, F, C, S, and K. Therefore we say that an FC D is an X-curve if $D \subset \gamma^-(\mathrm{Dom}(\gamma^-))$ and similarly for the other types of curves. We write $D \in \Gamma_{\mathcal{A}}(\Sigma, \tau)$ to denote the fact that D is an FC constructed by $\mathcal{A}$.

Now consider two systems Σ_1 and Σ_2, a time $\tau \geq 0$, and two open sets $U_1 \subset R_1(\tau)$ and $U_2 \subset R_2(\tau)$; here R_1 and R_2 denote the reachable sets of Σ_1 and Σ_2, respectively. Assume that $\mathcal{A}$ succeeds for Σ_1 and Σ_2 at time τ. We will say that $\Gamma_1 = \Gamma_{\mathcal{A}}(\Sigma_1, \tau) \restriction U_1$ and $\Gamma_2 = \Gamma_{\mathcal{A}}(\Sigma_2, \tau) \restriction U_2$ are *equivalent* if there exists an homeomorphism $\varphi : U_1 \mapsto U_2$ such that

(EC1) φ induces a bijection on Γ_i: $\{\varphi(\gamma_x(t)) : t \in \mathrm{Dom}(\gamma_x)\} \cap U_1 = \{\gamma_{\varphi(x)}(t) : t \in \mathrm{Dom}(\gamma_{\varphi(x)})\} \cap U_2$ for each $x \in U_1$; if the two sets are oriented for increasing t, then φ preserves the orientation.

(EC2) φ induces a bijection on FCs; i.e., for each FC D_1 of Γ_1 we have that $\varphi(D_1)$ is an FC of Γ_2 of the same type and vice versa, assuming that the types X- and Y- are equivalent. In this case we write $\Gamma_1 \restriction U_1 \equiv \Gamma_2 \restriction U_2$.

Remark 2.2. Note that in the definition of equivalence there are no requests about the time along γ_x; in fact there are no conditions of the type $\varphi(\gamma_x(t)) = \gamma_{\varphi(x)}(t)$. It is necessary to give a not-too-strict definition of equivalence to have a discrete set of equivalence classes. The same problem occurs in the definition of equivalence for a singular point of a dynamical system. In this case the orbital equivalence was introduced (see [1]).

Given x_1 and x_2 we say that $\Gamma_1 \restriction x_1$ and $\Gamma_2 \restriction x_2$ are equivalent, or we write $\Gamma_1 \restriction x_1 \equiv \Gamma_2 \restriction x_2$, if there exist U_1 and U_2 neighborhoods of x_1 and x_2, respectively, such that $\Gamma_1 \restriction U_1 \equiv \Gamma_2 \restriction U_2$. We say that two FCs D_i of Γ_i, $i = 1, 2$, are *equivalent* if for each $y_1 \in D_1 \setminus \partial D_1$, $y_2 \in D_2 \setminus \partial D_2$ we have that $\Gamma_1 \restriction y_1 \equiv \Gamma_2 \restriction y_2$. Similarly two frame points x_i of Γ_i, $i = 1, 2$, are *equivalent* if $\Gamma_1 \restriction x_1 \equiv \Gamma_2 \restriction x_2$.

If $\mathcal{A}$ succeeds for Σ at time τ, then under generic conditions, all frame points are intersections of two FCs. Using the same notation used for FCs, we will say that the origin is an (X, Y)-point; in fact, $0 \in \gamma^+ \cap \gamma^-$. Similarly if a frame point x is the intersection of two FCs D_1 and D_2 of respective type V_1 and V_2, then we say that x is a (V_1, V_2)-point. As for FCs we write $x \in \Gamma_{\mathcal{A}}(\Sigma, \tau)$ to denote the fact that x is a frame point constructed by $\mathcal{A}$.

Given $\varepsilon > 0$ we say that two systems $\Sigma_1 = (F_1, G_1)$, $\Sigma_2 = (F_2, G_2)$ are *ε-near* if

$$\max \{\| F_1 - F_2 \|_{C^3}, \| G_1 - G_2 \|_{C^3}\} \leq \varepsilon. \tag{2.17}$$

Consider a system Σ for which $\mathcal{A}$ succeeds at a time τ and a frame point x of $\Gamma_{\mathcal{A}}(\Sigma, \tau)$. We say that x is *structurally stable* if there exist $\varepsilon > 0$, $\delta > 0$ such that for each system Σ', ε-near

to Σ, there exists a unique frame point x' of the same type verifying

$$\| x - x' \| \leq \delta, \tag{2.18}$$

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma', \tau) \restriction x'. \tag{2.19}$$

We are interested only in structurally stable frame points: from conditions (2.18) and (2.19) we know that these are the only points that are *observable*; i.e., a small perturbation of the system does not change the structure of the synthesis near these points.

3. Frame curves. In this section we give a complete description of the FCs generated by the algorithm $\mathcal{A}$. We use the notation introduced in §2 for the six types (F1)–(F6) of FCs. From now on we consider a fixed $\tau \geq 0$ and a fixed system Σ for which $\mathcal{A}$ succeeds at time τ.

An FC D is *simple* if $D \setminus \partial D$ does not contain any frame point. Every FC can be divided into a finite number of simple FCs. The classification of simple FCs in connection with the classification of frame points, described in the following section, gives a complete classification of FCs. In fact two FCs D_1 and D_2 are equivalent if we can divide them into two families of simple FCs $D_1^1, \ldots, D_1^n$ and $D_2^1, \ldots, D_2^n$ such that

$$D_1^i \equiv D_2^i, \qquad D_1^i \cap D_1^j \equiv D_2^i \cap D_2^j \quad \forall\, i, j \in \{1, \ldots, n\},$$

where we assume, by definition, that $\emptyset \equiv \emptyset$. Therefore, throughout this section we consider only simple FCs.

***X*-curve.** Consider an X-curve D and $x \in D \setminus \partial D$. There exists a neighborhood U of x such that the control $u_{\mathcal{A}}$ is constant in each one of the two connected components U_1 and U_2 of $U \setminus D$. If, for example, $u_{\mathcal{A}} = 1$ on U_1, then Y-trajectories leave from D entering U_1. It is clear that there are only two possibilities:

(X1) $u_{\mathcal{A}} = 1$ on U_1 and $u_{\mathcal{A}} = -1$ on U_2, or vice versa;

(X2) $u_{\mathcal{A}} = -1$ on $U_1 \cup U_2$.

Example 1. Let $\tau > 2$, and consider the control system

$$\begin{cases} \dot{x}_1 = u, \\ \dot{x}_2 = x_1 + \dfrac{1}{2} x_1^2. \end{cases} \tag{3.1}$$

The X- and Y-trajectories can be described, giving x_2 as a function of x_1, and are, respectively, cubic polynomials of the following type:

$$x_2 = -\frac{x_1^3}{6} - \frac{x_1^2}{2} + \alpha, \qquad \alpha \in \mathbb{R}, \tag{3.2}$$

$$x_2 = \frac{x_1^3}{6} + \frac{x_1^2}{2} + \alpha, \qquad \alpha \in \mathbb{R}. \tag{3.3}$$

With a straightforward computation we obtain

$$[F, G] = \begin{pmatrix} 0 \\ -1 - x_1 \end{pmatrix}.$$

Then

$$\Delta_B(x) = \det \begin{pmatrix} 0 & 1 \\ -1 - x_1 & 0 \end{pmatrix} = 1 + x_1. \tag{3.4}$$

From (3.4) it follows that every turnpike is a subset of $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 = -1\}$. Indeed, at the second step the algorithm $\mathcal{A}$ constructs the turnpike

$$S = \left\{(x_1, x_2) : x_1 = -1, x_2 \leq -\frac{1}{3}\right\}. \tag{3.5}$$

Given b, consider the trajectories $\gamma_1 : [0, b] \mapsto \mathbb{R}^2$ for which there exists $t_0 \in [0, b]$ such that $\gamma_1 \restriction [0, t_0]$ is a Y-trajectory and $\gamma_1 \restriction [t_0, b]$ is an X-trajectory, and the trajectories $\gamma_2 : [0, b] \mapsto \mathbb{R}^2$, $b > 2$, for which there exists $t_1 \in [2, b]$ such that $\gamma_2 \restriction [0, t_1]$ is an X-trajectory and $\gamma_2 \restriction [t_1, b]$ is a Y-trajectory. For every $b > 2$, these trajectories cross each other in the region of the plane above the cubic (3.3) with $\alpha = 0$ and determine an overlap curve K that originates from the point $(-2, -\frac{2}{3})$. We use the symbols $x^{+-}(b, t_0)$ and $x^{-+}(b, t_1)$ to indicate, respectively, the terminal points of γ_1 and γ_2 above. Explicitly we have

$$x_1^{+-} = 2t_0 - b, \qquad x_2^{+-} = -\frac{(2t_0 - b)^3}{6} - \frac{(2t_0 - b)^2}{2} + t_0^2 + \frac{t_0^3}{3}, \tag{3.6}$$

$$x_1^{-+} = b - 2t_1, \qquad x_2^{-+} = \frac{(b - 2t_1)^3}{6} + \frac{(b - 2t_1)^2}{2} - t_1^2 + \frac{t_1^3}{3}. \tag{3.7}$$

Now the equation

$$x^{+-}(b, t_0) = x^{-+}(b, t_1), \tag{3.8}$$

as b varies in $[2, +\infty[$, describes the set K. From (3.6), (3.7), and (3.8) it follows that

$$t_0 = b - t_1, \qquad t_1\Big(-2t_1^2 + (2 + 3b)t_1 + (-b^2 - 2b)\Big) = 0.$$

Solving for t_1 we obtain three solutions:

$$t_1 = 0, \qquad t_1 = b, \qquad t_1 = 1 + \frac{b}{2}. \tag{3.9}$$

The first two equations of (3.9) are trivial, while the third determines a point of K so that

$$K = \left\{(x_1, x_2) : x_1 = -2, x_2 \geq -\frac{2}{3}\right\}.$$

The set $R(\tau)$ is portrayed in Fig. 1.

Consider the system Σ_1 of Example 1 at time $\tau_1 > 2$. If (X1) holds true, then

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_1, \tau_1) \restriction \gamma^-(t_0), \tag{EC1a}$$

where

$$0 < t_0 < 1. \tag{EC1b}$$

Hence D is equivalent to $\gamma^- \restriction [0, 1]$. In this case we say that D is of type X_1 or that D is an X_1-curve.

If (X2) holds true, then

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_1, \tau_1) \restriction \gamma^-(t_0), \tag{EC2a}$$

where

$$1 < t_0 < 2. \tag{EC2b}$$

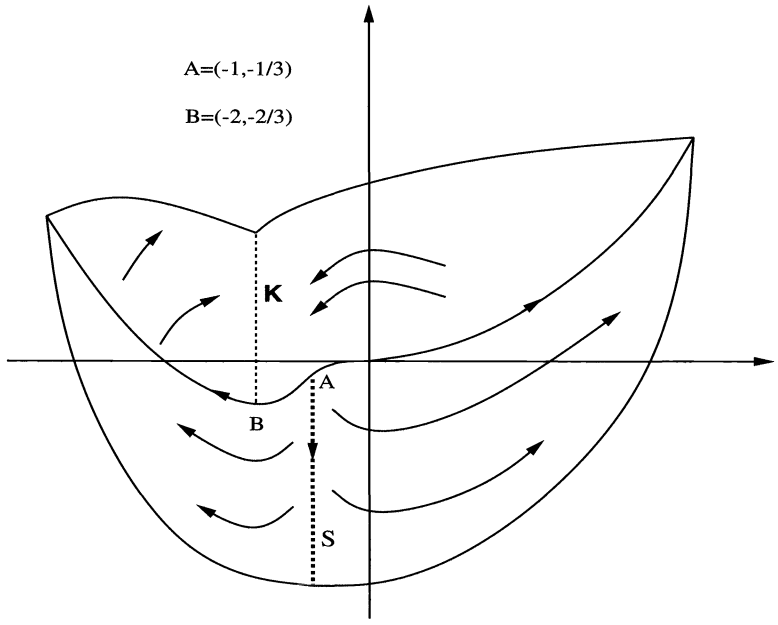


FIG. 1.

Then D is equivalent to $\gamma^- \restriction [1, 2]$. As before, we say that D is of type X_2 or that D is an X_2-curve.

***Y*-curve.** This case can be treated as the previous one, and we have the same equivalences. Now, the only difference is the sign of $u_{\mathcal{A}}$.

***F*-curve.** Consider an F-curve D and $x \in D \setminus \partial D$. There exists a neighborhood U of x in $R(\tau)$ such that $u_{\mathcal{A}}$ is constant on $U \setminus Fr(R(\tau))$. Consider the system Σ_1 at time τ_1 of Example 1. Choose $0 < \varepsilon < 1$, and let x_1 be the point reached by the trajectory corresponding to the control

$$u_1 = -1 \text{ on } [0, \varepsilon], \qquad u_1 = 1 \text{ on } [\varepsilon, \tau].$$

We have

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_1, \tau_1) \restriction x_1. \tag{EC3}$$

***C*-curve.** Let D be a C-curve, and consider a point x of $D \setminus \partial D$. There exists a neighborhood U of x such that the control $u_{\mathcal{A}}$ is constant in each one of the two connected components of $U \setminus D$. From the description of the switching curves it is clear that $u_{\mathcal{A}}$ is equal to 1 on one component and equal to -1 on the other.

Example 2. Consider $\tau > \pi$ and the control system

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -x_1 + u. \end{cases} \tag{3.10}$$

This example is accurately described in [6, pp. 11–14] and [5, p. 80].

The X- and Y-trajectories are circles centered at $(-1, 0)$ and at $(1, 0)$, respectively. The algorithm $\mathcal{A}$ constructs $\gamma^{\pm}$ only up to time π; indeed after this time they are not extremal.

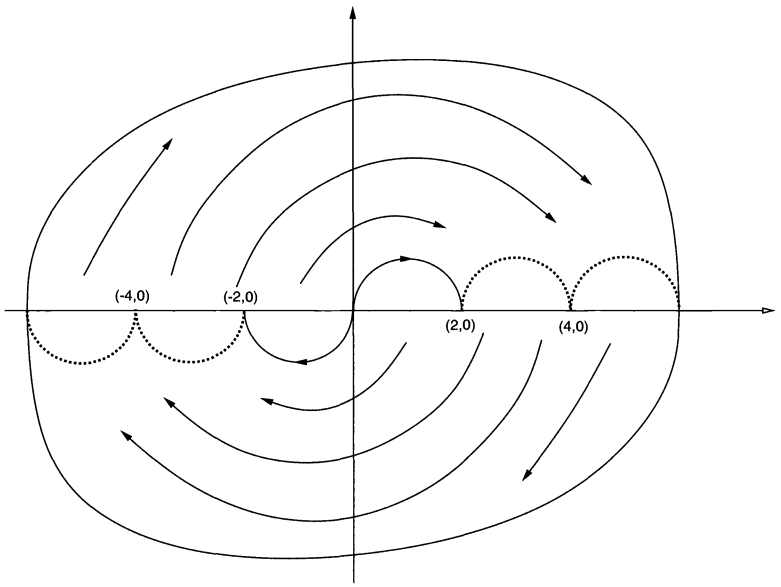


FIG. 2.

At step $n+1$ the following switching curves originate:

(a) all the semicircles of radius 1 centered at $(2n+1, 0)$ and contained in the half plane $\{(x_1, x_2) : x_2 \geq 0\}$.

(b) all the semicircles of radius 1 centered at $(-2n-1, 0)$ and contained in the half plane $\{(x_1, x_2) : x_2 \leq 0\}$.

Along the switching curves described in (a) the constructed trajectories arrive as Y-trajectories and leave as X-trajectories; i.e., the controls switch from $+1$ to -1. The opposite happens along the switching curves described in (b).

The set $R(\tau)$ is represented in Fig. 2.

Consider the system Σ_2 of Example 2 at time $\tau_2 > \frac{3}{2}\pi$. We have

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_2, \tau_2) \restriction (-3, -1). \tag{EC4}$$

***S*-curve.** Let x be a point of the relative interior of an S-curve D. As for the previous case, there exists a neighborhood U of x such that $u_{\mathcal{A}}$ is constant in each connected component of $U \setminus D$. From the definition of turnpike we have that $u_{\mathcal{A}}$ has different signs on the two components.

Consider the system Σ_1 at time τ_1 of Example 1. The following equivalence holds:

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_1, \tau_1) \restriction \left(-1, -\frac{1}{2}\right). \tag{EC5}$$

***K*-curve.** Consider a K-curve D and $x \in D \setminus \partial D$. If U is a suitably small neighborhood of x, then the control $u_{\mathcal{A}}$ is constant in each connected component of $U \setminus D$. As before, $u_{\mathcal{A}}$ has different signs on the two components. Consider the system Σ_1 of Example 1 at time $\tau_1 > 4$. We have the equivalence

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_1, \tau_1) \restriction (-2, 0). \tag{EC6}$$

Thanks to this analysis we have the following theorem.

THEOREM 3.1. *Consider $\Sigma \in \Xi$ and $\tau > 0$. If $\mathcal{A}$ succeeds for Σ at time τ and D is a simple FC of $\Gamma_{\mathcal{A}}(\Sigma, \tau)$, then D is of one of the following six types:*

$$X_1, \quad X_2, \quad F, \quad C, \quad S, \quad K,$$

and we have, respectively, one of the equivalences (EC1)–(EC6).

4. Frame points. In this section we give a complete description of the local structure of $\Gamma_{\mathcal{A}}$ in a neighborhood of a frame point. More precisely, only structurally stable frame points are considered. Therefore, all frame points will be intersections of no more than two FCs. Indeed, an intersection of three or more FCs can be destroyed by an arbitrary small perturbation (see (2.17)) of the system.

From now on we consider a fixed $\tau > 0$ and a fixed system Σ for which $\mathcal{A}$ succeeds at time τ. In particular Σ is locally controllable. For each type of frame point there are only a finite number of equivalence classes.

Before starting to examine frame points, case by case, we make a general observation. Consider a frame point x and two FCs D_1 and D_2 such that $\{x\} = D_1 \cap D_2$. There are four possible cases:

(FP0) $x \in D_1 \setminus \partial D_1, x \in D_2 \setminus \partial D_2$,
(FP1) $x \in D_1 \setminus \partial D_1, x \in \partial D_2$,
(FP2) $x \in \partial D_1, x \in D_2 \setminus \partial D_2$,
(FP3) $x \in \partial D_1, x \in \partial D_2$.

It is easy to check that, by construction, (FP0) can never occur. The case in which one frame curve is of type X, Y, F, S, or K is immediate. The case in which D_1 and D_2 are both C FCs is a consequence of the following observation. If we assume that D_1 and D_2 are not tangent and this is a generic situation, then there are some curves of zeros of either Δ_A or Δ_B to which x belongs. Indeed, near x, there are trajectories switching from control $+1$ to -1 and vice versa. Moreover, from Theorem 3.9 of [9] it follows that the possibility of switching from control $+1$ to -1 and vice versa depends on the sign of the function f of (2.14). In all possible cases, we obtain the existence of trajectories having two switchings near such curves. But this is prohibited by Lemma 7.1 of [9].

However, for each point we have to examine the other three possibilities.

The classification of frame points will be based on the types of the two intersecting curves D_1 and D_2. We will use the notation, introduced in §2, for frame points and the symbols $\gamma^{\pm}$, F, C, S, and K to indicate the curve of types (F1)–(F6), respectively.

(X, Y)-point. Consider an (X, Y)-point x of $\Gamma_{\mathcal{A}}(\Sigma, \tau)$. If $x = (0, 0)$, then it is a structurally stable (X, Y)-frame point. Indeed if Σ' is ε-near to Σ and ε is sufficiently small, then Σ' is locally controllable and $\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction (0, 0) \equiv \Gamma_{\mathcal{A}}(\Sigma', \tau) \restriction (0, 0)$. Let Σ_1 be the system of Example 1 at time $\tau_1 > 0$, then

$$\text{(EP1)} \qquad \Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction (0, 0) \equiv \Gamma_{\mathcal{A}}(\Sigma_1, \tau_1) \restriction (0, 0).$$

Now suppose that $x \neq (0, 0)$. It follows that $x = \gamma^-(t^-) = \gamma^+(t^+)$, $t^- > 0$, $t^+ > 0$. We have that $t^- = t^+$; otherwise one of the two trajectories would have been deleted from the synthesis. Since the condition $t^- = t^+$ can be destroyed by a small perturbation, x is not structurally stable. In fact in this case x belongs to an overlap curve; hence it is the intersection of at least three FCs.

(X, F)-point. Let x be an (X, F)-frame point. The cases (FP1) and (FP3) cannot occur because $\partial\big(\mathrm{Fr}(R(\tau))\big) = \emptyset$. Therefore we are in the case (FP2). There exists a neighborhood U of x (in $R(\tau)$) such that $u_{\mathcal{A}}$ is constant in each one of the two connected components U_1 and U_2 of $U \setminus (\gamma^- \cup F)$. One of the two following cases holds:

(XF1) $u_{\mathcal{A}} = -1$ on $U_1 \cup U_2$,

(XF2) $u_{\mathcal{A}} = 1$ on U_1 and $u_{\mathcal{A}} = -1$ on U_2, or vice versa.

Consider the system Σ_1 of Example 1 (§3) at time τ_1, and let $x_1 = \gamma^-(\tau_1)$. If (XF1) holds true and

$$1 < \tau_1 < 2, \tag{EP2a}$$

then

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_1, \tau_1) \restriction x_1. \tag{EP2b}$$

In this case we say that x is a frame point of type $(X, F)_1$.

If (XF2) holds true, then some Y-trajectories arise from γ^- and reach F. Let

$$\tau_1 > 2. \tag{EP3a}$$

Then

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_1, \tau_1) \restriction x_1, \tag{EP3b}$$

and x is a frame point of type $(X, F)_2$.

(X, C)-point. Assume that (FP1) holds true. There exists a neighborhood U of the (X, C)-frame point x such that $u_{\mathcal{A}}$ is constant in each one of the three connected components U_1, U_2, and U_3 of $U \setminus (\gamma^- \cup C)$. We label U_1, U_2, and U_3 in such a way that U_3 is the connected component of $U \setminus \gamma^-$ that does not contain $C \cap U$; U_1 comes before U_2 along γ^- for the orientation of increasing time. Because of the definition of C-curve we have one of the following:

(XC1) $u_{\mathcal{A}} = 1$ on U_1,

(XC2) $u_{\mathcal{A}} = 1$ on U_2.

Example 3. Consider $\tau > \frac{1}{3}\ln(4)$ and the system Σ:

$$\begin{cases} \dot{x}_1 = 3x_1 + u, \\ \dot{x}_2 = x_1^2 + x_1. \end{cases} \tag{4.1}$$

Since $\Sigma \in \Xi$ and

$$[F, G] = \begin{pmatrix} -3 \\ -2x_1 - 1 \end{pmatrix},$$

the system is locally controllable. The X-trajectory passing through the point (x_1^0, x_2^0) at time 0 is

$$x_1(t) = \left(x_1^0 - \frac{1}{3}\right) e^{3t} + \frac{1}{3}, \tag{4.2}$$

$$\begin{aligned} x_2(t) = {} & \frac{1}{6}\left(x_1^0 - \frac{1}{3}\right)^2 e^{6t} + \frac{5}{9}\left(x_1^0 - \frac{1}{3}\right) e^{3t} \\ & + \frac{4}{9}t + x_2^0 - \frac{1}{6}\left(x_1^0 - \frac{1}{3}\right)^2 - \frac{5}{9}\left(x_1^0 - \frac{1}{3}\right). \end{aligned} \tag{4.3}$$

While the Y-trajectory passing through the point (x_1^0, x_2^0) at time 0 is

$$x_1(t) = \left(x_1^0 + \frac{1}{3}\right) e^{3t} - \frac{1}{3}, \tag{4.4}$$

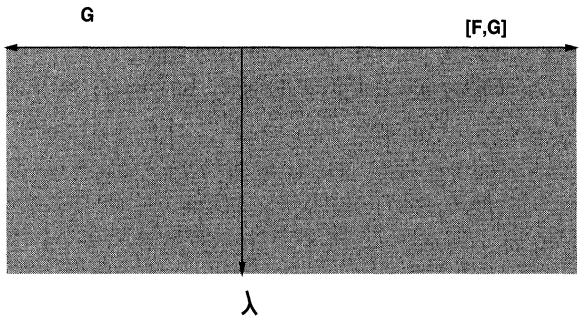


FIG. 3.

$$x_2(t) = \frac{1}{6}\left(x_1^0 + \frac{1}{3}\right)^2 e^{6t} + \frac{1}{9}\left(x_1^0 + \frac{1}{3}\right) e^{3t} - \frac{2}{9}t + x_2^0 - \frac{1}{6}\left(x_1^0 + \frac{1}{3}\right)^2 - \frac{1}{9}\left(x_1^0 + \frac{1}{3}\right). \tag{4.5}$$

The equation for turnpikes is

$$0 = \Delta_B(x_1, x_2) = -(2x_1 + 1).$$

Hence every turnpike is a subset of $S = \{(x_1, x_2) : x_1 = -\frac{1}{2}\}$. We have that the control φ_S to stay on S (cf. (2.15)) is

$$\varphi_S(x_1, x_2) = \frac{3}{2}.$$

Then there is no regular turnpike.

Now consider the pairs $(\gamma_s, u_s) \in \mathrm{Adm}(\Sigma)$, $\mathrm{In}(\gamma_s) = 0$, $\mathrm{Dom}(\gamma_s) = [0, s + \varepsilon_s]$ $(\varepsilon_s \geq 0)$ such that γ_s originates as an X-trajectory and switches at time s, going on as a Y-trajectory up to the time $s + \varepsilon_s$. Let $(\gamma^*, u^*) = (\gamma_{s^*}, u_{s^*})$ be the pair that verifies

$$\gamma^*(s^*) = \left(-\frac{1}{2}, -\frac{13}{72} + \frac{4}{9}\ln\left(\sqrt[3]{\frac{5}{2}}\right)\right). \tag{4.6}$$

Define $\varepsilon^* = \varepsilon_{s^*}$, and assume that γ^* satisfies the PMP with adjoint variable λ^*. We know that

$$\lambda^*(s^*) \cdot G(\gamma^*(s^*)) = 0, \qquad \Delta_B(\gamma^*(s^*)) = 0. \tag{4.7a}$$

Then

$$\frac{d}{dt}(\lambda^*(t) \cdot G(\gamma^*(t)))\Big|_{t=s^*} = \lambda^*(s^*) \cdot [F, G](\gamma^*(s^*)) = 0. \tag{4.7b}$$

From (4.7a), (4.7b), and straightforward calculations we have at $\gamma^*(s^*)$ the situation of Fig. 3.

Now it is easy to verify that

$$\forall t \in [s^*, s^* + \varepsilon^*] \qquad \Delta_B\big(\gamma^*(t)\big) > 0. \tag{4.8}$$

Thus, for each $t \in [s^*, s^* + \varepsilon^*]$, the pair of vectors $(G(t), [F, G](t))$ forms a positive-oriented base of $\mathbb{R}^2$. Since $u^*(t) \restriction [s^*, s^* + \varepsilon^*] \equiv 1$, it follows that the two functions $\lambda^*(t) \cdot G(\gamma^*(t))$ and $\lambda^*(t) \cdot [F, G](\gamma^*(t))$ are positive in a right neighborhood of s^*. Therefore G and $[F, G]$

lie in the darkened region of Fig. 3. But this is prohibited by (4.8), as observed above; hence it follows that $\varepsilon^* = 0$.

Similarly, if the trajectories γ_s, $s \in [\ln(\sqrt[3]{2}), s^*]$, satisfy the PMP, then ε_s is small. More precisely the algorithm $\mathcal{A}$ constructs trajectories that have a second switching point, and these switching points form a switching curve C_1 originating at (4.6).

The above geometric reasoning is very general; however, in this case we can compute explicit calculations. Suppose that γ_s satisfies the PMP with adjoint variable λ^s. The equation for λ^s is

$$\dot{\lambda}^s(t) = -\lambda^s(t) \cdot \big(\nabla F\big(\gamma_s(t)\big) + u_s(t)\, \nabla G\big(\gamma_s(t)\big)\big) = -\lambda^s(t) \cdot \big(\nabla F\big(\gamma_s(t)\big)\big). \tag{4.9}$$

Let x_1^s be the first component of γ_s. For time $t \geq s$, the explicit form of (4.9) is

$$(\dot{\lambda}_1^s, \dot{\lambda}_2^s)(t) = \left(-3\lambda_1^s(t) - \lambda_2^s(t)\left[2\left(x_1^s(s) + \frac{1}{3}\right) e^{3(t-s)} + \frac{1}{3}\right], 0\right). \tag{4.10}$$

Denote by ϕ_s the switching function along $(\gamma_s, u_s, \lambda^s)$. The solution to (4.10) with initial condition

$$\lambda^s(s) \cdot G\big(\gamma_s(s)\big) = 0$$

is

$$\lambda_2^s(t) \equiv \lambda_2^s(0),$$

$$\phi_s(t) = \lambda_1^s(t) = \lambda_2^s(t)\left[\frac{1}{3}\left(x_1^s(s) + \frac{2}{3}\right) e^{-3(t-s)} - \frac{1}{3}\left(x_1^s(s) + \frac{1}{3}\right) e^{3(t-s)} - \frac{1}{9}\right].$$

Now, the equation $\phi_s(t) = 0$ has two solutions:

$$t_1^s = s, \qquad t_2^s = s + \ln\left(\sqrt[3]{\frac{-3x_1^s(s) - 2}{3x_1^s(s) + 1}}\right).$$

Thus

$$x_1^s(t_2^s) = -x_1^s(s) - 1$$

gives the first component of switching points belonging to C_1.

Let γ^- be, as before, the X-trajectory verifying $\mathrm{In}(\gamma^-) = \gamma^-(0) = 0$. The point $\gamma^-(\ln \sqrt[3]{4})$ is conjugate to the origin along γ^-. Consider the trajectories γ_r, $\mathrm{In}(\gamma_r) = 0$, $\mathrm{Dom}(\gamma_r) = [0, b_r]$ $(b_r \geq r)$ that originate as Y-trajectories and have a switching at time r, going on as X-trajectories. Again we can make direct calculations and obtain the existence of a second switching time (if $r < \ln \sqrt[3]{2}$):

$$t_r = r + \ln\left(\sqrt[3]{-1 - \frac{10}{6\, a_r}}\right), \tag{4.11}$$

where $a_r = \big(\gamma_r(r) - \frac{1}{3}\big)$. These switching points form another switching curve C_2 that intersects γ^- at $\gamma^-(\ln \sqrt[3]{4})$. If we denote by x_1^r the first component of γ_r, then from (4.11) it follows that

$$x_1^r(t_r) = -x_1^r(r) - 1.$$

In Fig. 4, the reachable set $R(\tau)$ is represented.

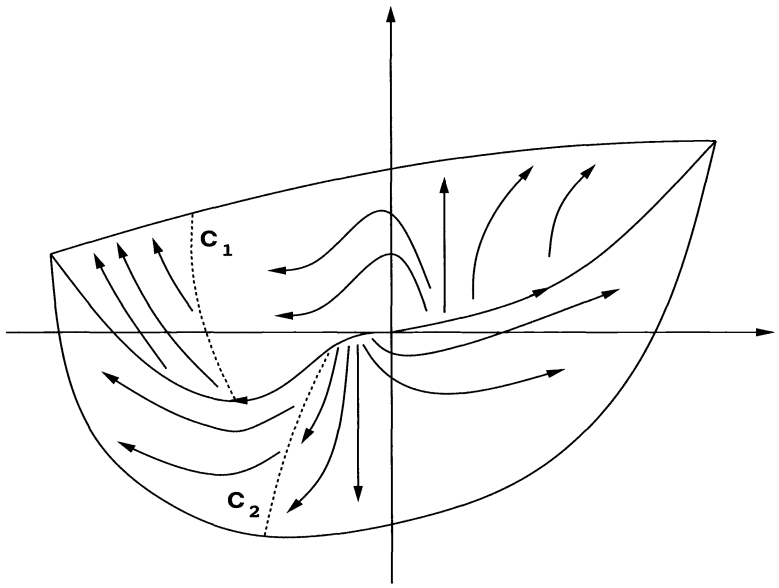


FIG. 4.

Consider the system Σ_3 at time τ_3 of Example 3. If (XC1) holds true, then $u_{\mathcal{A}} = -1$ on $U_2 \cup U_3$ and the Y-trajectories leaving γ^- reach C. We have

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_3, \tau_3) \restriction \gamma^-\left(\frac{1}{3} \ln \frac{5}{2}\right). \tag{EP4}$$

In this case we say that x is of type $(X, C)_1$.

If (XC2) holds true, then $u_{\mathcal{A}} = -1$ on $U_1 \cup U_3$. Hence

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_3, \tau_3) \restriction \gamma^-\left(\frac{1}{3} \ln 4\right), \tag{EP5}$$

and we say that x is of type $(X, C)_2$.

Remark 4.1. Consider the equivalence (EP4). In Example 3, $\gamma^-(\frac{1}{3} \ln \frac{5}{2})$ belongs to a nonordinary arc that is not a turnpike. This happens for every frame point x of type $(X, C)_1$. Indeed, assume $x = \gamma^-(t_x)$ and let (γ_r, u_r), $r \in [t_x - \varepsilon, t_x + \varepsilon]$ $(\varepsilon > 0)$, be the pair such that $\gamma_r(0) = \gamma^-(r)$ and $u_r \equiv 1$. Let λ_r be the covector field along (γ_r, u_r) satisfying

$$\lambda_r(0) \cdot G\big(\gamma_r(0)\big) = 0, \qquad \det\big[\lambda_r(0), G\big(\gamma_r(0)\big)\big] > 0, \qquad \| \lambda_r(0) \| = 1,$$

and consider the function

$$\psi(r, s) = \lambda_r(s) \cdot G\big(\gamma_r(s)\big).$$

The equation $\psi(r, s) = 0$ has two branches of solutions in $(t_x, 0)$. Then we have

$$0 = \left.\frac{\partial \psi}{\partial s}\right|_{(t_x, 0)} = \lambda_{t_x}(0) \cdot [F, G](\gamma^-(t_x)).$$

Now $0 = \lambda_{t_x}(0) \cdot G(x) = \lambda_{t_x}(0) \cdot [F, G](x)$ and $\lambda_{t_x}(0) \neq 0$. Then

$$\Delta_B(x) = \det\Big(G(x), [F, G](x)\Big) = 0.$$

It follows that if $\nabla(\Delta_B(x)) \neq 0$, then x belongs to an INOA (see §2 and [9]). This INOA cannot be a regular turnpike; otherwise it would have been constructed by the algorithm $\mathcal{A}$.

The case (FP2) is not generic. Indeed if (FP2) holds, then there exists a neighborhood U of x in C such that for each $y \in U$ there exists a trajectory γ_y that switches at $y = \gamma_y(t_y)$. One side of C with respect to x is reached by trajectories γ_y that arise from an FC D_1. The other side is reached by trajectories that originate from a different FC, say, D_2. Then at x, two different switching curves meet each other and x is not stable.

Suppose that (FP3) holds true. If C lies on the left (right) of γ^-, then $u_{\mathcal{A}} \equiv -1$ to the left (right) of γ^-. Consider the system Σ_2 of Example 2 at time $\tau_2 > \pi$. Then we have that

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_2, \tau_2) \restriction (2, 0). \tag{EP6}$$

In this case x is of type $(X, C)_3$.

(X, S)-point. Let x be an (X, S)-point x, and assume that (FP1) holds true. There exists a neighborhood U of x such that $u_{\mathcal{A}}$ is constant on each one of the three connected components U_1, U_2, and U_3 of $U \setminus (\gamma^- \cup S)$. We suppose that U_1, U_2, and U_3 are labelled in such a way that U_3 is the connected component of $U \setminus \gamma^-$ that does not contain $S \cap U$; U_1 comes before U_2 along γ^- for the orientation of increasing time. From the definition of a turnpike it follows that $u_{\mathcal{A}} = 1$ on U_1 and $u_{\mathcal{A}} = -1$ on $U_2 \cup U_3$. Consider the system Σ_1 at time τ_1 of the first example (§3). The following equivalence holds:

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_1, \tau_1) \restriction \left(-1, -\frac{1}{3}\right). \tag{EP7}$$

The cases (FP2) and (FP3) cannot occur because, from the description of turnpikes, it follows that γ^- cannot terminate at x.

(X, K)-point. Assume that (FP1) holds true. As before, there exists a neighborhood U of x such that $u_{\mathcal{A}}$ is constant in each one of the three connected components U_1, U_2, U_3 of $U \setminus (\gamma^- \cup K)$. We label U_1, U_2, and U_3 in such a way that U_3 is the connected component of $U \setminus \gamma^-$ that does not contain $K \cap U$; U_1 comes before U_2 along γ^- for increasing time. We have that $u_{\mathcal{A}} = 1$ on U_2 and $u_{\mathcal{A}} = -1$ on $U_1 \cup U_3$. Under generic assumptions, the Y-trajectories arising from γ^- reach K. In fact, if the opposite happens, then $X(x)$ and $Y(x)$ are parallel and have the same versus, but this is not generic. Consider again the system Σ_1 at time τ_1 of the first example (§3). In this case we have

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_1, \tau_1) \restriction \left(-2, -\frac{2}{3}\right), \tag{EP8}$$

and we say that x is of type $(X, K)_1$.

Now let (FP2) hold. For every sufficiently small neighborhood U of x, we have that $u_{\mathcal{A}}$ is constant in each one of the two connected components U_1 and U_2 of $U \setminus K$. If, for example, U_1 contains $\gamma^- \cap U$, then $u_{\mathcal{A}} = -1$ on U_1 and $u_{\mathcal{A}} = 1$ on U_2.

Example 4. Consider ε, $0 < \varepsilon < 1$, $\tau > \frac{\pi}{\sqrt{1-\varepsilon}}$, and the system Σ:

$$\begin{cases} \dot{x}_1 = \varepsilon x_2 + u x_2, \\ \dot{x}_2 = u(1 - x_1). \end{cases} \tag{4.12}$$

It is easy to check that

$$[F, G] = \begin{pmatrix} -\varepsilon(1 - x_1) \\ -\varepsilon x_2 \end{pmatrix}. \tag{4.13}$$

From (4.13) and Lemma 2.1 we have that the system is locally controllable.

The X-trajectory passing through the point (x_1^0, x_2^0) at time 0 is

$$x_1(t) = (x_1^0 - 1)\cos(\sqrt{1-\varepsilon}\,t) + x_2^0\sqrt{1-\varepsilon}\,\sin(\sqrt{1-\varepsilon}\,t) + 1, \tag{4.14}$$

$$x_2(t) = \frac{(x_1^0 - 1)}{\sqrt{1-\varepsilon}}\sin(\sqrt{1-\varepsilon}\,t) - x_2^0\cos(\sqrt{1-\varepsilon}\,t). \tag{4.15}$$

The Y-trajectory passing through the point (x_1^0, x_2^0) at time 0 is

$$x_1(t) = (x_1^0 - 1)\cos(\sqrt{1+\varepsilon}\,t) + x_2^0\sqrt{1+\varepsilon}\,\sin(\sqrt{1+\varepsilon}\,t) + 1, \tag{4.16}$$

$$x_2(t) = -\frac{(x_1^0 - 1)}{\sqrt{1+\varepsilon}}\sin(\sqrt{1+\varepsilon}\,t) + x_2^0\cos(\sqrt{1+\varepsilon}\,t). \tag{4.17}$$

The equation for turnpikes is

$$0 = \Delta_B(x_1, x_2) = -\varepsilon x_2^2 + \varepsilon(1 - x_1)^2. \tag{4.18}$$

Hence every turnpike is a subset of $S = \{(x_1, x_2) : x_2 = \pm(1 - x_1)\}$. Using (4.16)–(4.18) it is easy to verify that the trajectory γ^+ intersects the set S in a point (x_1^+, x_2^+) of the first quadrant. The algorithm $\mathcal{A}$ constructs the turnpike $S_1 = \{(x_1, x_2) : x_2 = 1 - x_1, x_1^+ \le x_1 < 1\}$. The singular control φ_S^1 along the turnpike S_1 (cf. (2.15)) is

$$\varphi_S^1(x_1, x_2) = -\frac{\varepsilon x_2}{1 - x_1 + x_2} > -1. \tag{4.19}$$

From (4.19) we have

$$\dot{x}_1(\varphi_S^1) = \frac{\varepsilon}{2}(1 - x_1). \tag{4.20}$$

Hence the point $(1, 0)$ is not reached in finite time by a singular trajectory.

Similarly, using (4.14), (4.15) it is easy to verify that the trajectory γ^- intersects the set S in a point of the fourth quadrant:

$$(x_1^-, x_2^-) \doteq \gamma^-\left(\frac{1}{\sqrt{1-\varepsilon}}\arccos\left(\sqrt{\frac{1}{2-\varepsilon}}\right)\right). \tag{4.21}$$

Hence the algorithm $\mathcal{A}$ constructs the turnpike $S_2 = \{(x_1, x_2) : x_2 = x_1 - 1, x_1 \le x_1^-\}$. Indeed, the control φ_S^2 (cf. (2.15)) is

$$\varphi_S^2(x_1, x_2) = \frac{\varepsilon x_2}{1 - x_1 - x_2}.$$

The trajectories $\gamma^\pm$ are very close to the circle A of center $(1, 0)$ and radius 1; γ^+ runs clockwise, and γ^- counterclockwise. From (4.14)–(4.17) we have that γ^+ lies inside A, γ^- outside, and

$$\gamma^+ \cap \gamma^- \cap A = \{(0, 0), (2, 0)\}.$$

However, the two trajectories $\gamma^\pm$ do not meet each other at $(2, 0)$; indeed,

$$(2, 0) = \gamma^+\left(\frac{\pi}{\sqrt{1+\varepsilon}}\right) = \gamma^-\left(\frac{\pi}{\sqrt{1-\varepsilon}}\right).$$

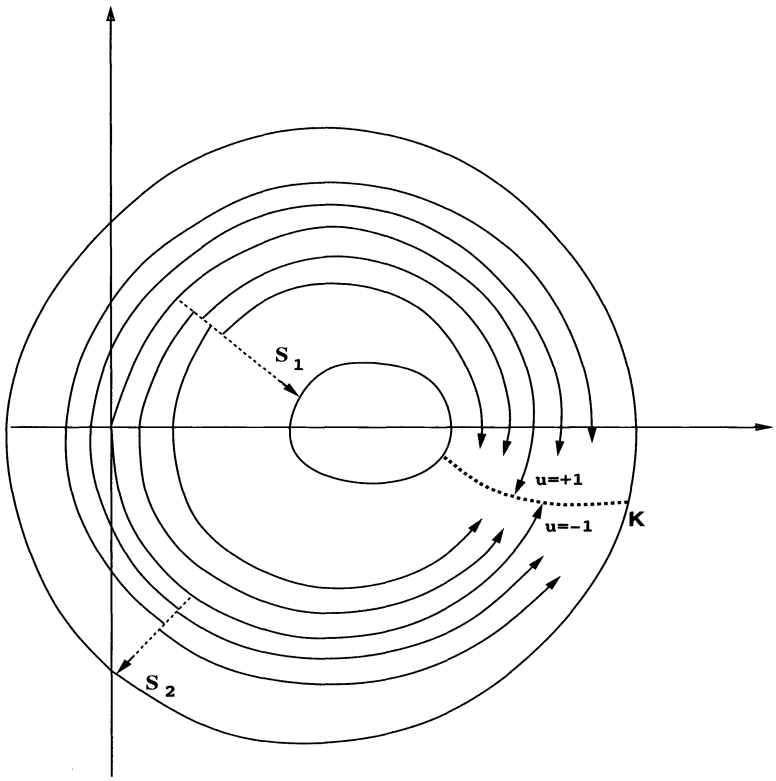


FIG. 5.

But the $X * Y$ and $Y * X$ trajectories constructed by the algorithm give rise to an overlap curve K, and $\gamma^{\pm}$ end on it. In Fig. 5, $R(\tau)$ is represented.

Consider the system Σ_4 at time τ_4 of Example 4, and let $\bar{x}$ be the point in which γ^- intersects the overlap curve. We have

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_4, \tau_4) \restriction \bar{x}, \tag{EP9}$$

and we say that x is of type $(X, K)_2$.

Assume that (FP3) holds true and that X and K are not tangent. There exists a neighborhood U of x such that $u_{\mathcal{A}}$ is constant in each one of the two connected components U_1 and U_2 of $U \setminus (\gamma^- \cup K)$. Suppose that U_1 and U_2 are labelled in such a way that the vector $X(x)$ points into U_2. It is clear that $u_{\mathcal{A}} = -1$ on U_1 and $u_{\mathcal{A}} = 1$ on U_2. The Y-trajectories leaving from γ^- do not reach K.

Example 5. Consider the system (cf. Example 1)

$$\begin{cases} \dot{x}_1 = u, \\ \dot{x}_2 = \dfrac{1}{2}x_1^2 + x_1 \end{cases} \tag{4.22}$$

and the two embedded submanifolds

$$M_1 = \left\{(x_1, x_2) : x_1 = 0, -\frac{2}{3} \le x_2 \le 0\right\},$$

$$M_2 = \left\{(x_1, x_2) : x_1 = \sqrt{2}, -\frac{2}{3} \le x_2 \le \frac{2}{3}\right\}.$$

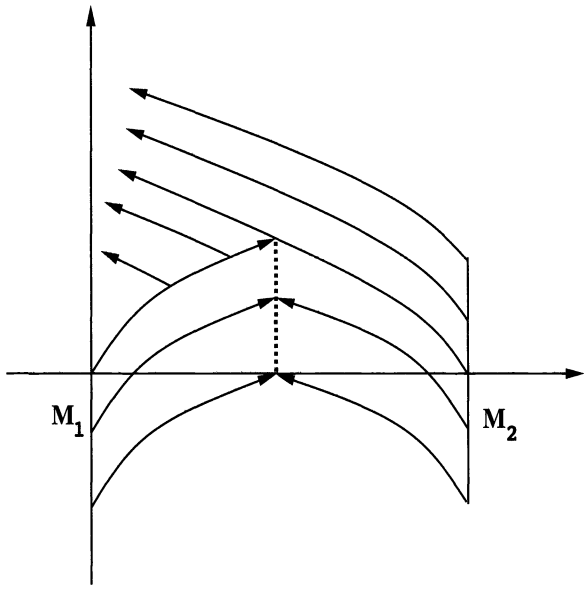


FIG. 6.

We assume that for each $y \in M_1$ there exists a Y-trajectory $\gamma_1(y)$ that verifies $\gamma_1(y)(0) = y$. Moreover, for each $x \in M_2$ there exists an X-trajectory $\gamma_2(x)$ that arises from x at time 0. Finally, an X-trajectory originates from each point of the Y-trajectory $\gamma_1(0)$, i.e., $\gamma_1(0)$ is the trajectory γ^+ of a given system.

At the point $(1, \frac{2}{3})$ the trajectories $\gamma_1(0)$, $\gamma_2\big((\sqrt{2}, 0)\big)$ meet each other. After this point $\gamma_1(0)$ is not constructed by $\mathcal{A}$ because the trajectories $\gamma_2\big((\sqrt{2}, c)\big)$, $c \geq 0$, achieve a better performance. The trajectories $\gamma_1\big((0, -c)\big)$ and $\gamma_2\big((\sqrt{2}, -c)\big)$, meeting each other, give rise to an overlap curve:

$$K = \left\{(x_1, x_2) : x_1 = -1, 0 \leq x_2 \leq \frac{2}{3}\right\}.$$

In Fig. 6, this local example is portrayed.

Consider the synthesis Γ_5 of Example 5. We have that

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_5 \restriction \left(1, \frac{2}{3}\right), \tag{EP10}$$

and we say that x is of type $(X, K)_3$.

(Y, F)-, (Y, C)-, (Y, S)-, (Y, K)-points. These points can be treated as the corresponding points with Y replaced by X, and we have the same equivalences. In this case the only difference is the sign of $u_{\mathcal{A}}$.

(X, X)-, (Y, Y)-, (F, F)-points. It is easy to verify that points of these types cannot exist.

(F, C)-point. Consider an (F, C)-point x. Since $\partial\big(\mathrm{Fr}(R(\tau))\big) = \emptyset$, the cases (FP2) and (FP3) cannot occur. Then (FP1) holds true. There exists a neighborhood U of x in $R(\tau)$ such that $u_{\mathcal{A}}$ is constant in each one of the two connected components of $U \setminus (F \cup C)$. It is clear that $u_{\mathcal{A}} = 1$ on one connected component and $u_{\mathcal{A}} = -1$ on the other. The trajectories leaving from C reach F. Consider the system Σ_2 of Example 2 at time π. We have

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_2, \pi) \restriction (-3, -1). \tag{EP11}$$

(F, S)-point. As for the previous type, only the case (FP1) can hold. There exists a neighborhood U of x in $R(\tau)$ such that $u_{\mathcal{A}}$ is constant in each one of the two connected components of $U \setminus (F \cup S)$. Again $u_{\mathcal{A}} = 1$ on one connected component and $u_{\mathcal{A}} = -1$ on the other. Under generic assumptions the trajectories leaving from S reach F. Consider the system Σ_1 at time τ_1 of Example 1. Let x_1 be the point in which the turnpike intersects the frontier of the reachable set, namely,

$$x_1 = \left(-1, -\frac{1}{3} - \frac{1}{2}(\tau_1 - 1)\right).$$

It follows that

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_1, \tau_1) \restriction x_1. \tag{EP12}$$

(F, K)-point. Consider an (F, K)-point x. The case (FP1) is the only possible one. There exists a neighborhood U of x in $R(\tau)$ such that $u_{\mathcal{A}}$ is constant in each one of the two connected components of $U \setminus (F \cup K)$. It is clear that $u_{\mathcal{A}} = 1$ on one connected component and $u_{\mathcal{A}} = -1$ on the other. Consider again the system Σ_1 at time τ_1 of Example 1. Let x_1 be the point in which the overlap curve intersects the frontier of the reachable set, namely,

$$x_1 = \left(-2, \frac{2}{3} + \frac{1}{3}\left(1 + \frac{\tau}{2}\right)^3 - \left(1 + \frac{\tau}{2}\right)^2\right).$$

We have the following:

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_1, \tau_1) \restriction x_1. \tag{EP13}$$

(C, C)-point. Let x be a (C, C)-point. From the definition of switching curve we have that the cases (FP1) and (FP2) cannot occur. Therefore (FP3) holds.

There exist two switching curves C_1 and C_2 verifying $x = C_1 \cap C_2$ and a neighborhood U of x such that $u_{\mathcal{A}}$ is constant in each connected component of $U \setminus (C_1 \cup C_2)$. We have that $u_{\mathcal{A}}$ has different signs on the two connected components. Consider the following cases:

(CCa) The curves leaving from C_1 reach C_2.

(CCb) The curves leaving from C_2 reach C_1.

It is easy to show that (CCa) and (CCb) cannot hold at the same time; otherwise there is no trajectory reaching x. Hence we have two cases:

(CC1) (CCa) holds and (CCb) does not, or vice versa.

(CC2) (CCa) and (CCb) do not hold.

Example 6. Consider the system (4.1) of Example 3 and the manifold

$$M = \{(x_1, x_2) : x_1 = 0, |x_2| \leq 1\}.$$

We assume that from every point $(0, x_2) \in M$ an X-trajectory $\gamma(x_2)$ arises, with initial time 0 and with adjoint variable satisfying

$$[\lambda_1(x_2)](0) = -\frac{9}{36} - \frac{1}{36}\operatorname{sgn}(x_2)\, x_2, \qquad [\lambda_2(x_2)](0) = -1.$$

With simple calculations we obtain the solutions to the equation $[\phi(x_2)](t) = 0$, where $\phi(x_2)$ is the switching function along $(\gamma(x_2), -1, \lambda(x_2))$:

$$t^{\pm}(x_2) = \ln\left(\sqrt[3]{\frac{5}{2} \pm \frac{1}{2}\sqrt{9 + 36\,[\lambda_1(x_2)](0)}}\right).$$

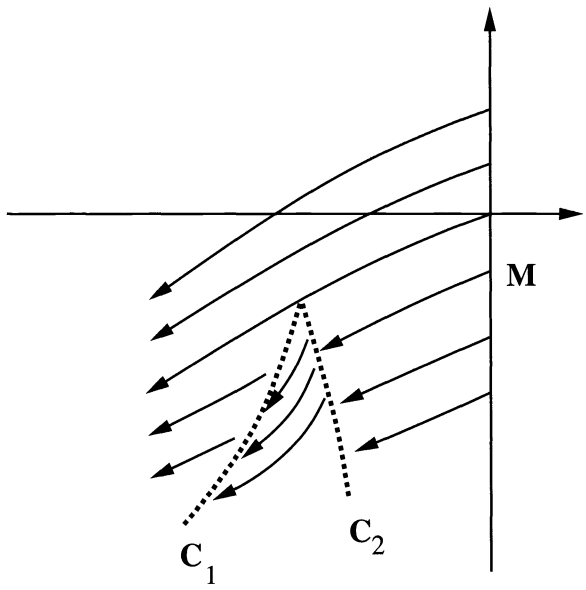


FIG. 7.

Hence, the trajectories $\gamma(x_2)$, $x_2 \leq 0$, has a switching at time $t^-(x_2)$, while the trajectories $\gamma(x_2)$, $x_2 > 0$, do not switch. These switching points form a switching curve C_1 having the point (4.6) as endpoint.

Now the equation $\phi(x_2) = 0$ has another solution after the time $t^-(x_2)$, namely,

$$t'(x_2) = t^-(x_2) + \ln\left(\sqrt[3]{\frac{-3\ p_1(x_2) - 2}{3\ p_1(x_2) + 1}}\right),$$

where $p_1(x_2)$ is the first coordinate of the first switching point of $\gamma(x_2)$. These switching points form another switching curve C_2 that meets C_1 at the point (4.6). This local example is portrayed in Fig. 7.

Consider the synthesis Γ_6 of Example 6. If (CC1) holds true, then

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_6 \restriction \left(-\frac{1}{3}, -\frac{13}{72} + \frac{4}{9}\ln\left(\sqrt[3]{\frac{5}{2}}\right)\right), \tag{EP14}$$

and we say that x is of type $(C, C)_1$.

Consider the system Σ_2 at time $\tau_2 > 3\pi$ of Example 2. If (CC2) holds true, then

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_2, \tau_2) \restriction (4, 0), \tag{EP15}$$

and we say that x is of type $(C, C)_2$.

Remark 4.2. Reasoning as in Remark 4.1 one can prove that if x is a frame point of type $(C, C)_1$, then $\Delta_B(x) = 0$.

The frame points of type $(C, C)_2$ are not *effective* singular points. Indeed, the optimal synthesis near these points is equivalent to the synthesis near a point x of a simple FC of type C, verifying $x \in C \setminus \partial C$.

(C, S)-point. Consider a (C, S)-point x. There exists a neighborhood U of x such that $u_{\mathcal{A}}$ is constant in each connected component of $U \setminus (C \cup S)$.

The cases (FP1) and (FP2) cannot occur because the control $u_{\mathcal{A}}$ changes sign crossing S (or C); moreover it has to be constant along each side of C (or S).

Therefore (FP3) holds true. There exists a C^1 diffeomorphism $\alpha : [0, \varepsilon] \mapsto \mathbb{R}^2$, $\varepsilon > 0$, such that $\alpha(t) \in C$, $\alpha(0) = x$. Consider the vectors

$$C(x) = \lim_{t \to 0} \dot{\alpha}(t), \qquad S(x) = F(x) + \varphi_S(x)G(x),$$

where φ_S is the control to stay on S (cf. (2.15)). Assume that $C(x)$ and $S(x)$ are not parallel. Let U_X and U_Y be the connected component of $U \setminus \{x + tS(x) : t \in \mathbb{R}\}$ labelled in such a way that $X(x)$ and $Y(x)$ point into U_X and U_Y, respectively. Moreover, let U_1 and U_2 be the connected component of $U \setminus (C \cup S)$ labelled in such a way that the angle with vertex x and sides $C(x)$ and $S(x)$ contained in U_1 is smaller than that one contained in U_2. Now, if U_1 is contained in U_Y, then $u_{\mathcal{A}} = 1$ on U_1; otherwise $u_{\mathcal{A}} = -1$ on U_1.

There exists $\gamma_S \in \mathrm{Traj}(\Sigma)$ such that $\gamma_S(\mathrm{Dom}(\gamma_S)) = S \cap U$. We have two cases:

(CS1) $\mathrm{In}(\gamma_S) = x$.

(CS2) $\mathrm{Term}(\gamma_S) = x$.

Assume that (CS1) holds. There are two subcases:

(CSa) Some constructed trajectories reach C from U_2.

(CSb) Some constructed trajectories reach C from U_1.

If (CSb) holds, then no nontrivial trajectory reaches x, but this is not possible. Hence (CSa) holds true. For the same reason the trajectories originating from S and entering U_2 cannot reach C.

Example 7. Consider the system (4.22) of Example 5 and the manifold

$$M = \left\{(x_1, x_2) : x_1 = 0, -\frac{1}{3} \le x_2 \le \frac{1}{3}\right\}.$$

We assume that from each $(x_1, x_2) \in M$ there arises, with initial time 0, an X-trajectory $\gamma(x_1, x_2) = \gamma(x_2)$ with adjoint variable $\lambda(x_2)$ that satisfies

$$[\lambda_1(x_2)](0) = \frac{-1 - 4\,\mathrm{sgn}(x_2)\,x_2^2}{2}, \qquad [\lambda_2(x_2)](0) = -1,$$

where $\mathrm{sgn}(x) = x \mid x \mid^{-1}$ if $x \neq 0$ and $\mathrm{sgn}(0) = 0$. Now, the switching function along $(\gamma(x_2), -1, \lambda(x_2))$ is

$$[\phi(x_2)](t) = \lambda_1(t) = -\frac{t^2}{2} + t - \frac{1 + 4\,\mathrm{sgn}(x_2)\,x_2^2}{2}.$$

If $x_2 \le 0$, the equation $[\phi(x_2)](t) = 0$ has the following solutions:

$$t_1(x_2) = 1 + 2|x_2|, \qquad t_2(x_2) = 1 - 2|x_2|;$$

otherwise there is no solution. Then every trajectory $\gamma(x_2)$, $x_2 \le 0$, switches at the point

$$[\gamma(x_2)](t_2) = \left(2|x_2| - 1, -\frac{(2|x_2| - 1)^3}{6} - \frac{(2|x_2| - 1)^2}{2} + x_2\right).$$

These switching points form a switching curve C.

The trajectory $\gamma(0)$ crosses the set $\{(x_1, x_2) : \Delta_B(x_1, x_2) = 0\} = \{(x_1, x_2) : x_1 = -1\}$ at a switching point; hence the algorithm $\mathcal{A}$ constructs the turnpike

$$S = \left\{(x_1, x_2) : x_1 = -1, x_2 \le -\frac{1}{3}\right\}.$$

This local example is represented in Fig. 8.

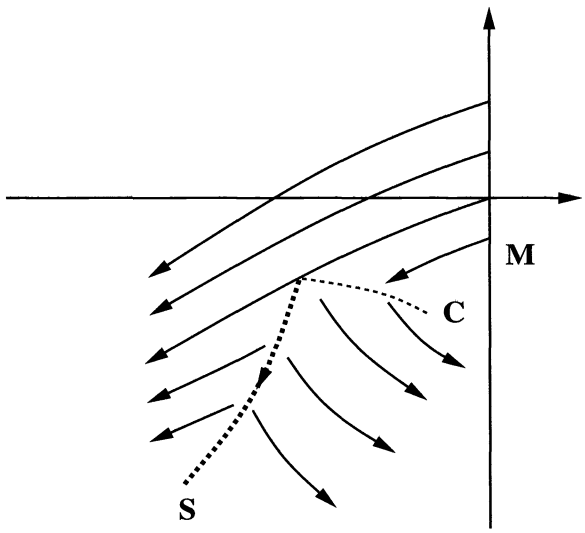


FIG. 8.

Consider the synthesis Γ_7 of Example 7. We have

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_7 \restriction \left(-1, -\frac{1}{3}\right), \tag{EP16}$$

and we say that x is of type $(C, C)_1$.

Suppose that (CS2) holds. We again have the subcases (CSa) and (CSb). The case (CSa) cannot hold. Indeed, the trajectories arising from S and entering U_2 cannot reach C, and then, from the direction of $X(x)$, $Y(x)$, we have that $\mathrm{In}(\gamma_S) = x$, contradicting (CS2).

Suppose that (CSb) holds. We have that the trajectories leaving from S and entering U_1 reach C. From Theorem 3.9 of [9] it follows that Δ_B cannot have constant sign on $V \cap U_1$ for any neighborhood V of x. Hence we have the nongeneric condition $\nabla\Delta_B(x) = 0$.

Consider again the case (CS2), and assume that $C(x)$ and $S(x)$ are parallel. The trajectories arriving onto C come from S.

Example 8. Let $\tau > \frac{7}{3} + \sqrt[3]{4}$, and consider the system

$$\begin{cases} \dot{x}_1 = u, \\ \dot{x}_2 = (x_1 + \psi(x_2)) + \dfrac{1}{2}(x_1 + \psi(x_2))^2, \end{cases} \tag{4.23}$$

where

$$\psi(x_2) = \begin{cases} 0, & x_2 > -1, \\ (x_2 + 1)^4, & x_2 \leq -1. \end{cases} \tag{4.24}$$

Observe that for $x_2 > -1$ the system is the same as in Example 1. There is a turnpike S that lies on the line $x_1 = -1$ between the points $(-1, -\frac{1}{3})$ and $(-1, -1)$. Moreover, for $x_2 \leq -1$, S is represented by the equations

$$x_1 + (x_2 + 1)^4 + 1 = 0, \qquad x_2 \leq -1. \tag{4.25}$$

Recalling (2.15), from (4.23)–(4.25) we have that the control φ_S is

$$\varphi_S(x_1, x_2) = 0 \quad \text{if } x_2 \geq -1, \qquad \varphi_S(x_1, x_2) = 2(x_2 + 1)^3 \quad \text{if } x_2 \leq -1. \tag{4.26}$$

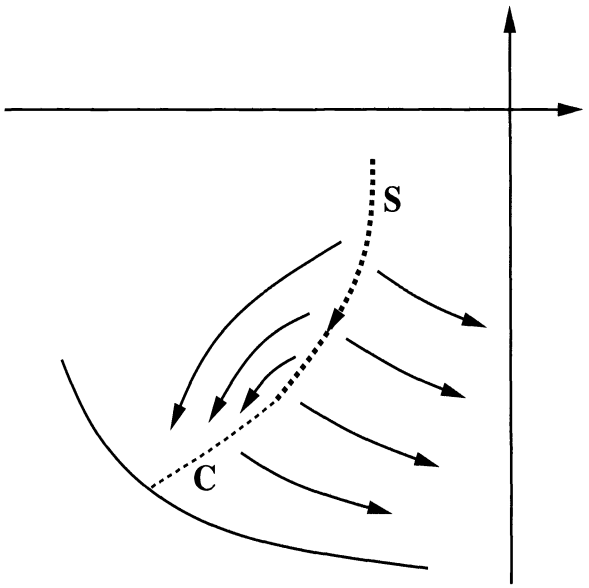


FIG. 9.

By (4.26), the turnpike S is regular up to the point

$$(\bar{x}_1, \bar{x}_2) = \left(-1 - \frac{1}{2\sqrt[3]{2}}, -1 - \frac{1}{\sqrt[3]{2}}\right). \tag{4.27}$$

Indeed

$$\varphi_S(\bar{x}_1, \bar{x}_2) = -1.$$

Hence, the algorithm $\mathcal{A}$ constructs a turnpike that ends at the the point (4.27). The set $R(\tau)$ near the point (4.27) is represented in Fig. 9.

Consider the system Σ_8 at time τ of Example 8. We have

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_8, \tau) \restriction \left(-1 - \frac{1}{2\sqrt[3]{2}}, -1 - \frac{1}{\sqrt[3]{2}}\right), \tag{EP17}$$

and we say that x is of type $(C, S)_2$.

(C, K)-point. There exists a neighborhood U of the (C, K)-point x such that $u_{\mathcal{A}}$ is constant in each connected component of $U \setminus (C \cup K)$.

The cases (FP1) and (FP2) cannot occur because the control $u_{\mathcal{A}}$ changes sign when we cross K (or C), but it also has to be constant along each side of C (or K).

Therefore (FP3) holds true. There exist two C^1 diffeomorphisms $\alpha_{1,2} : [0, \varepsilon] \mapsto \mathbb{R}^2$, $\varepsilon > 0$, such that $\alpha_1(t) \in C$, $\alpha_2(t) \in K$, $\alpha_{1,2}(0) = x$. Consider the vectors

$$C(x) = \lim_{t \to 0} \dot{\alpha}_1(t), \qquad K(x) = \lim_{t \to 0} \dot{\alpha}_2(t).$$

Suppose that $C(x)$ and $K(x)$ are not parallel. Let U_X and U_Y be the connected components of $U \setminus \{x + tK(x) : t \in \mathbb{R}\}$ labelled in such a way that $X(x)$, $Y(x)$ point into U_X and U_Y, respectively. Let U_1 and U_2 be the connected components of $U \setminus (C \cup K)$ labelled in such a way that the angle with vertex x and sides $C(x)$ and $K(x)$ contained in U_1 is smaller than that one contained in U_2. If U_1 is contained in U_X, then $u_{\mathcal{A}} = 1$ on U_1; otherwise $u_{\mathcal{A}} = -1$ on U_1.

We have two cases:

(CK1) Some constructed trajectories reach C from U_1.

(CK2) Some constructed trajectories reach C from U_2.

Assume that (CK1) holds. The trajectories originating from C cannot reach K; otherwise we have one of the not generic conditions $Y(x) = 0$, $X(x) = 0$.

Example 9. Consider the system (cf. Example 2)

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -x_1 + u \end{cases} \tag{4.28}$$

and the two manifolds

$$M_1 = \{(x_1, x_2) : x_1 = 0,\ 1 \le x_2 \le 2\},$$

$$M_2 = \{(x_1, x_2) : 1 \le x_1 \le 2,\ x_2 = 3\}.$$

The algorithm $\mathcal{A}$ succeeds for the system (4.28) at time 4. A Y-trajectory $\gamma'(x) \in \Gamma_{\mathcal{A}}(\Sigma, 4)$, with an associate adjoint variable $\lambda'(x)$, passes through each point $x \in M_1$. We suppose that from each $x \in M_1$ a Y-trajectory $\gamma(x)$ with adjoint variable $\lambda(x)$ arises at time 0. Moreover, $(\gamma(x), \lambda(x))$ is obtained from $(\gamma'(x), \lambda'(x))$ shifting the time.

Consider the line given by the equation

$$x_2 = (\sqrt{2} - 3)\, x_1 + 8 - \sqrt{8}. \tag{4.29}$$

For each $s \in [1, \frac{2}{3}\sqrt{6}]$ the trajectory $\gamma^s = \gamma\big((0, s)\big)$ intersects the line (4.29) in a point, say, (x_1^s, x_2^s).

Let $r(s) \in [1, 2]$ be such that the X-trajectory passing through $(r(s), 3)$ intersects the line (4.29) in (x_1^s, x_2^s). We assume that an X-trajectory γ_r with initial time $t(r)$ originates from each $(r, 3) \in M_2$. If $r = r(s)$ for some s, define, denoting by d the Euclidean distance,

$$t_s \doteq t\big(r(s)\big) = 2 \arcsin\left(\frac{d\big((0, s), (x_1^s, x_2^s)\big)}{2\sqrt{1 + s^2}}\right) - 2 \arcsin\left(\frac{d\big((r(s), 3), (x_1^s, x_2^s)\big)}{2\sqrt{(r(s) + 1)^2 + 9}}\right). \tag{4.30}$$

Otherwise

$$t(r) \doteq \max\{t_s : s \in [1, 2]\}.$$

We associate with every γ_r an adjoint variable λ_r verifying

$$\lambda_1^r\big(t(r)\big) = -1, \qquad \lambda_2^r\big(t(r)\big) = 0. \tag{4.31}$$

The equation (4.31) implies that $\lambda^r\big(t(r)\big) \cdot G\big((r, 3)\big) = 0$. Then to satisfy the PMP the trajectory γ_r is an X-trajectory for a time interval of length π.

The trajectories γ^s, $s \in [1, \frac{2}{3}\sqrt{6}]$, form a switching curve:

$$C = \left\{(x_1, x_2) : x_2 = \sqrt{1 - (x_1 - 3)^2},\ 2 \le x_1 \le \frac{8}{3}\right\}.$$

By direct calculations, one can verify that (4.30) ensures that the trajectories γ^s and $\gamma_{r(s)}$, $s \in [\frac{2}{3}\sqrt{6}, 2]$, meet each other, giving rise to an overlap curve:

$$K = \left\{(x_1, x_2) : (x_1, x_2) \text{ satisfies (4.29)},\ 2 \le x_1 \le \frac{8}{3}\right\}.$$

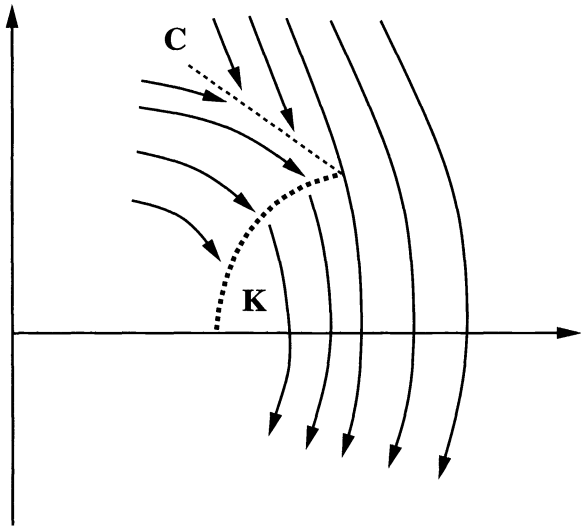


FIG. 10.

The curves K and C meet each other at the point

$$\left(\frac{8}{3}, \frac{\sqrt{8}}{3}\right).$$

In Fig. 10, this local example is portrayed.

Consider the synthesis Γ_9 of Example 9. We have

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_9 \restriction \left(\frac{8}{3}, \frac{\sqrt{8}}{3}\right), \tag{EP18}$$

and we say that x is of type $(C, K)_1$.

Example 10. Consider the system

$$\begin{cases} \dot{x}_1 = 3\, x_1 - u, \\ \dot{x}_2 = x_1^2 + x_1 \end{cases} \tag{4.32}$$

that is obtained from the system (4.1), replacing G with $-G$, and the manifold

$$M = \{(x_1, x_2) : x_1 = 0, |x_2| < \varepsilon\}.$$

We assume that from every point $(0, x_2) \in M$ a Y-trajectory $\gamma(x_2)$ with initial time $t_0(x_2)$ arises. Moreover, $\gamma(x_2)$ admits an adjoint variable satisfying

$$[\lambda_1(x_2)](0) = -\frac{9}{36} + 25\ \mathrm{sgn}(x_2)\ x_2, \qquad [\lambda_2(x_2)](0) = -1.$$

With simple calculations we obtain the solutions to the equation $[\phi(x_2)](t) = 0$, where $\phi(x_2)$ is the switching function along $(\gamma(x_2), +1, \lambda(x_2))$:

$$t^{\pm}(x_2) = t_0(x_2) + \ln\left(\sqrt[3]{\frac{5}{2} \pm \frac{1}{2}\sqrt{9 + 36\ [\lambda_1(x_2)](0)}}\right). \tag{4.33}$$

Hence the trajectories $\gamma(x_2)$, $x_2 \geq 0$, have a switching at time $t^-(x_2)$, while the trajectories $\gamma(x_2)$, $x_2 < 0$, do not switch. Let $x^-(x_2) \doteq [\gamma(x_2)](t^-(x_2))$, $x_2 \geq 0$. From (4.33) we have

$$x_1^- = -\frac{1}{2} + 5\, x_2^2, \tag{4.34}$$

and these switching points form a switching curve C_1 originating from (4.6).

Now the equation $\phi(x_2) = 0$, where $\phi(x_2)$ denotes again the switching function along $\gamma(x_2)$, after the time $t^-(x_2)$ has another solution:

$$t'(x_2) = t^-(x_2) + \ln\left(\sqrt[3]{\frac{-3\, x_1^-(x_2) - 2}{3\, x_1^-(x_2) + 1}}\right). \tag{4.35}$$

These switching points give rise to another switching curve C_2 that meets C_1 at the point (4.6).

It is easy to verify that the X-trajectories leaving from C_1 cross the trajectory $\gamma_0 \doteq \gamma(0)$ before reaching the switching curve C_2. Hence, we can define $P(x_2)$, $x_2 \geq 0$, to be the point at which $\gamma(x_2)$ meets γ_0.

Let $r(x_2)$, $x_2 \geq 0$, be such that the trajectory $\gamma(r(x_2))$ meets $\gamma(x_2)$ at the point $Q(x_2) \doteq [\gamma(x_2)](t'(x_2))$; i.e., they meet each other on C_2.

Now let $t_1(x_2)$, $t_2(x_2)$ be the time in which, respectively, $\gamma(x_2)$, γ_0 reaches $P(x_2)$ and $t_3(x_2)$, $t_4(x_2)$ be the time in which, respectively, $\gamma(x_2)$, $\gamma(r(x_2))$ reaches $Q(x_2)$. If x_2 is sufficiently small, from (4.34), (4.35) we have that

$$t_4 - t_3 < t_2 - t_1.$$

Then, taking ε sufficiently small, we can define

$$t_0(x_2) = 0 \qquad \text{if } x_2 < 0,$$

$$t_0(x_2) = \frac{(t_2 - t_1) + (t_4 - t_3)}{2} \qquad \text{if } x_2 \geq 0.$$

With this choice, the trajectories $\gamma(x_2)$, $x_2 \geq 0$, and $\gamma(x_2)$, $x_2 < 0$, meet each other, forming an overlap curve K that meets C_1 at (4.6). The curve C_2 is deleted by the algorithm. In Fig. 11, this local example is portrayed.

Assume that (CK2) holds. Consider the synthesis Γ_{10} of Example 10. We have

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{10} \restriction \left(-\frac{1}{3}, -\frac{13}{72} + \frac{4}{9}\, \ln\left(\sqrt[3]{\frac{5}{2}}\right)\right), \tag{EP19}$$

and we say that x is of type $(C, K)_2$.

Now suppose that $C(x)$ and $K(x)$ are parallel. If the trajectories leaving from C do not reach K, then the equivalence (EP18) holds and an unstable tangency between C and K is verified. If the opposite happens, then we have a stable tangency between C and K.

Example 11. Consider the system

$$\begin{cases} \dot{x}_1 = \dfrac{x_1 + 1}{2} + u\, \dfrac{x_1 - 1}{2}, \\ \dot{x}_2 = \dfrac{x_2}{2} + u\, \dfrac{x_2}{2} \end{cases} \tag{4.36}$$

and the manifold

$$M = \{(x_1, x_2) : x_1 = 0, 0 < x_2 < 1\}.$$

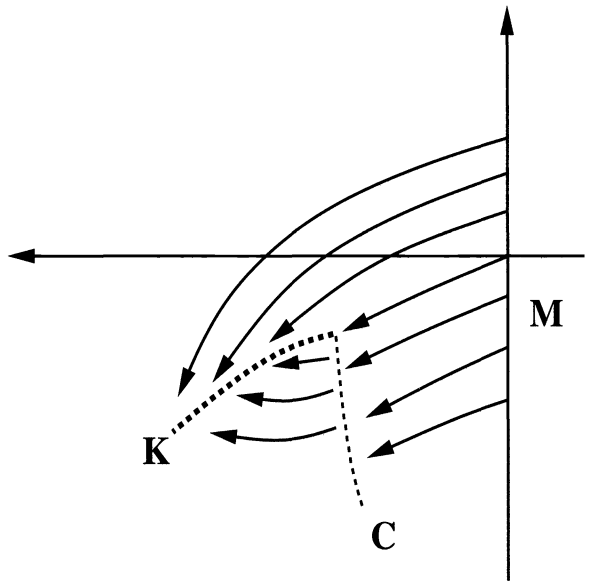


FIG. 11.

Assume that from every point $(0, x_2) \in M$, with initial time 0, an X-trajectory $\gamma(x_2)$ originates with adjoint variable satisfying

$$[\lambda_1(x_2)](0) = \frac{x_2}{1 - \sqrt{1 - x_2^2}}, \qquad [\lambda_2(x_2)](0) = -1. \tag{4.37}$$

It is easy to verify, from (4.36), (4.37), that every $\gamma(x_2)$ switches at time

$$t(x_2) = 2 - \sqrt{1 - x_2^2}$$

and the corresponding switching points form a switching curve

$$C = \{(x_1, \psi(x_1)) : 1 < x_1 < 2\}, \qquad \psi(x_1) \doteq \sqrt{1 - (2 - x_1)^2},$$

that is an arc of circle.

Observe that for ε small, $Y(\varepsilon, \psi(\varepsilon))$ points to the right of C and $Y(2 - \varepsilon, \psi(2 - \varepsilon))$ points to the left of C. Then there exists $(\bar{x}_1, \bar{x}_2) \in C$ such that $Y(\bar{x}_1, \bar{x}_2)$ is tangent to C. Define $C' \doteq \{(x_1, \psi(x_1)) \in C : x_1 \geq \bar{x}_1\}$. The trajectories $\gamma(x_2)$ that reach C' meet other trajectories $\gamma(x_2)$, giving rise to an overlap curve K. It is possible to move along C' with a trajectory of the system. Hence we can construct an envelope for the curves $\gamma(x_2) \restriction [0, t(x_2)]$ that reach C'; see [12], [13] for envelope theory. Hence the subset C' of C is removed by the algorithm. In Fig. 12 this local example is represented.

Consider the synthesis Γ_{11} of Example 11, and define $(\bar{x}_1, \bar{x}_2)$ in the same way. The following equivalences hold:

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{11} \restriction (\bar{x}_1, \bar{x}_2)$$

$$\equiv \Gamma_{10} \restriction \left(-\frac{1}{3}, -\frac{13}{72} + \frac{4}{9} \ln\left(\sqrt[3]{\frac{5}{2}} \right) \right).$$

Remark 4.3. The point $(\bar{x}_1, \bar{x}_2)$ of Example 11 and (4.6) of Example 10 are equivalent, but they are in some sense different. In fact, proceeding as in Remark 4.1 one can prove that

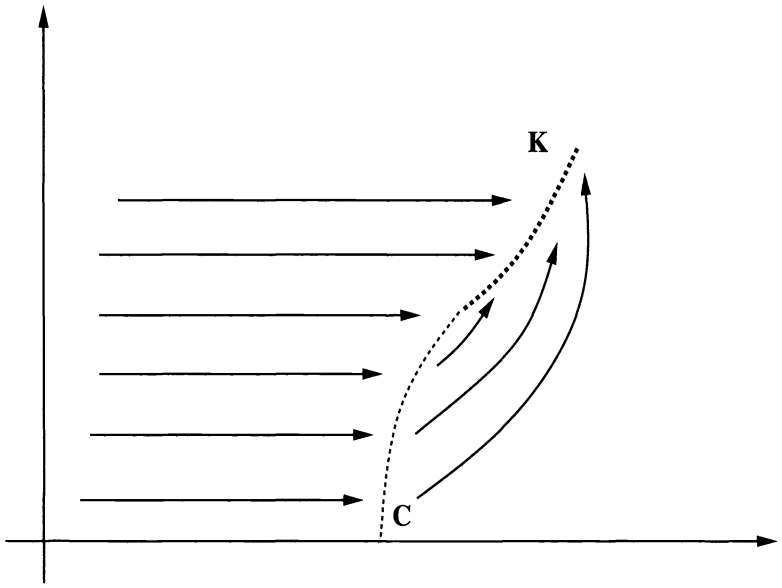


FIG. 12.

if $K(x)$ and $C(x)$ are linearly independent and (CK2) holds, then $\Delta_B(x) = 0$. If, instead, $K(x)$ and $C(x)$ are parallel, we can have that $\Delta_B(x) \neq 0$ as in Example 10.

(S, S)-point. It is easy to verify that these points cannot exist.

(S, K)-point. Consider an (S, K)-point x. The control $u_{\mathcal{A}}$ is constant in each connected component of $U \setminus (S \cup K)$ for every sufficiently small neighborhood U of x.

The cases (FP1) and (FP2) cannot occur because the control $u_{\mathcal{A}}$ changes sign when we cross K (or S) and is constant along each side of S (or K).

Therefore (FP3) holds true. The cases in which every trajectory arising from S reaches K or no trajectory leaving from S reaches K are not generic. Indeed, we have that $X(x)$ and $Y(x)$ are linearly dependent. Therefore the trajectories leaving from one side of S reach K, and those leaving from the other side do not. There exists $\gamma_S \in \mathrm{Traj}(\Sigma)$ such that $\gamma_S(\mathrm{Dom}(\gamma_S)) = S \cap U$. There are two cases:

(SK1) $\mathrm{In}(\gamma_S) = x$.

(SK2) $\mathrm{Term}(\gamma_S) = x$.

Example 12. Consider the same system and the same manifold of Example 7, and define S in the same way. We assume that from each $(0, x_2) \in M$ an X-trajectory $\gamma(x_2)$ arises with initial time

$$t_0(x_2) = -\frac{2}{3}\, x_2$$

and with adjoint variable satisfying

$$[\lambda_1(x_2)](0) = \frac{-1 - \alpha\ \mathrm{sgn}(x_2)\ x_2^2}{2}, \qquad [\lambda_2(x_2)](0) = -1,$$

where $\alpha > 0$ and sgn is defined as in Example 7. There is again a switching curve C. The X-trajectories starting from $(0, x_2)$, $x_2 \leq 0$, reach $(-1, -\frac{1}{3} - x_2) \in S$ at time

$$1 + \frac{2}{3}\, |x_2|.$$

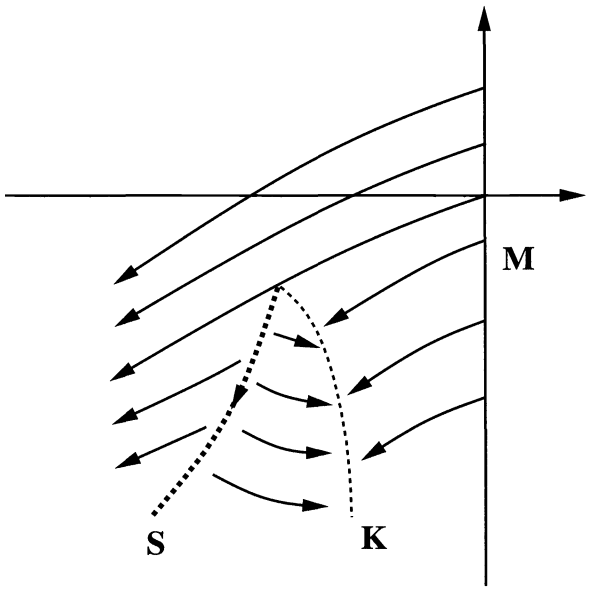


FIG. 13.

On the other hand the $Z * X$-trajectories, concatenations of an X- and a Z- trajectory, starting from 0 reach the same point at time

$$1 + \frac{|x_2|}{2}.$$

Therefore the $Y * Z * X$-trajectories, concatenations of an X-, a Z-, and a Y-trajectory, starting from the origin, and the X-trajectories starting from $(0, x_2)$, $x_2 \leq 0$, give rise to an overlap curve K having $(-1, -\frac{1}{3})$ as its endpoint.

Let $(s, k(s))$ be a parametrization of K in a neighborhood of $(-1, -\frac{1}{3})$, and define

$$\beta \doteq \left.\frac{dk(s)}{ds}\right|_{s=-1^+}.$$

Now let $(c_1(x_2), c_2(x_2))$ be a parametrization of the switching curve C. After straightforward calculations we have

$$\left.\frac{dc_2}{dc_1}\right|_{c_1=-1^+} = -\frac{3}{2} - \frac{1}{\alpha}.$$

Thus if α is sufficiently small,

$$\beta \geq -\frac{3}{2} - \frac{1}{\alpha},$$

and the overlap curve K arises. Therefore the curve C is deleted by the algorithm. This local example is portrayed in Fig. 13.

If (SK1) holds true, then consider the synthesis Γ_{12} of Example 12. We have

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \upharpoonright x \equiv \Gamma_{12} \upharpoonright \left(-1, -\frac{1}{3}\right), \tag{EP20}$$

and we say that x is of type $(S, K)_1$.

Example 13. Consider $0 < \varepsilon << 1$, $\tau > 1$, and the system Σ:

$$\begin{cases} \dot{x}_1 = u, \\ \dot{x}_2 = x_1^3 - \varepsilon x_1. \end{cases} \tag{4.38}$$

We have that

$$[F, G] = \begin{pmatrix} 0 \\ \varepsilon - 3x_1^2 \end{pmatrix};$$

hence the system is locally controllable. The X- and Y-trajectories are quartic polynomials of the following types, respectively:

$$x_2 = -\frac{x_1^4}{4} + \varepsilon \frac{x_1^2}{2} + \alpha, \qquad \alpha \in \mathbb{R},$$

$$x_2 = \frac{x_1^4}{4} - \varepsilon \frac{x_1^2}{2} + \alpha, \qquad \alpha \in \mathbb{R}.$$

The equation for turnpikes is

$$0 = \Delta_B(x_1, x_2) = \varepsilon - 3x_1^2,$$

whose set of solutions is

$$\left\{(x_1, x_2) : x_1 = \pm\sqrt{\frac{\varepsilon}{3}}\right\}. \tag{4.39}$$

Every turnpike is a subset of (4.39). The algorithm constructs the turnpikes

$$S_1' = \left\{(x_1, x_2) : x_1 = \sqrt{\frac{\varepsilon}{3}}, x_2 \leq -\frac{5}{36}\varepsilon^2\right\} \cap R(\tau),$$

$$S_2' = \left\{(x_1, x_2) : x_1 = -\sqrt{\frac{\varepsilon}{3}}, x_2 \geq \frac{5}{36}\varepsilon^2\right\} \cap R(\tau).$$

The points $(\pm\sqrt{\varepsilon}, \mp\frac{\varepsilon^2}{4})$ are conjugate to the origin along $\gamma^{\pm}$. Two overlap curves K_2 and K_1, respectively, originate at these points. The algorithm partially deletes the turnpikes S_1' and S_2', determining two new turnpikes $S_1 \subset S_1'$ and $S_2 \subset S_2'$. The new turnpikes S_1 and S_2 end on K_1 and K_2, respectively. In Fig. 14, $R(\tau)$ is represented.

Remark 4.4. Let Σ'' be the system $(\dot{x}_1, \dot{x}_2) = (u, x_1^3 - \psi(x_1)\varepsilon x_1)$, where ψ is a smooth function, $\{x : \psi(x) \neq 0\} \subset B(0, 1)$, and $\psi \restriction B(0, \frac{1}{2}) \equiv 1$. Note that this system is obtained by a small perturbation of the system Σ': $(\dot{x}_1, \dot{x}_2) = (u, x_1^3)$. The synthesis $\Gamma_{\mathcal{A}}(\Sigma', \tau)$ is formed by bang-bang trajectories with at most one switching. It is clear that Σ' is not structurally stable (in a sense that will be stated more precisely in a following paper). In fact we have that for ε small the system Σ'' is ε-near (cf. (2.17)) to Σ', but $\Gamma_{\mathcal{A}}(\Sigma'', \tau)$ has a structure completely different from $\Gamma_{\mathcal{A}}(\Sigma', \tau)$.

If (SK2) holds, consider the system Σ_{13} at time τ and the curves S_1 and K_1 of Example 13. Let $x_1 = S_1 \cap K_1$. We have

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_{13}, \tau) \restriction x_1, \tag{EP21}$$

and we say that x is of type $(S, K)_2$.

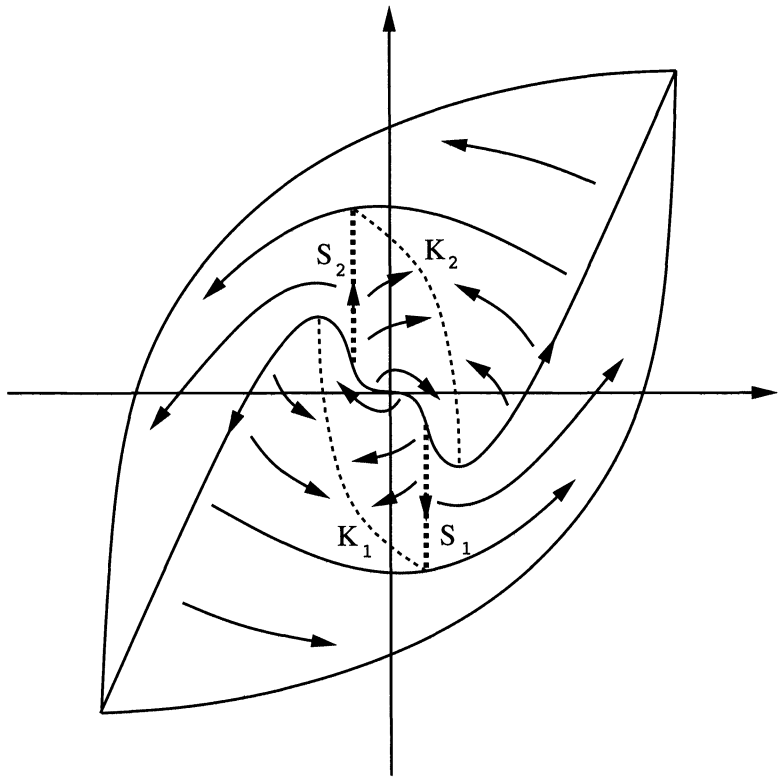


FIG. 14.

(K, K)-point. Consider a (K, K)-point x. From the definition of overlap curve we have that the cases (FP1) and (FP2) cannot occur; then (FP3) holds. Consider the system Σ_{13} at time τ of Example 13. The overlap curve K_1 is union of two overlap curves K_1' and K_1''. The set K_1' is formed by intersections of $Y * X$- and $X * Y$-trajectories, while K_1'' is formed by intersections of $Y * X$- and $X * S * Y$-trajectories. Let $x_1 = K_1' \cap K_1''$. We have

$$\Gamma_{\mathcal{A}}(\Sigma, \tau) \restriction x \equiv \Gamma_{\mathcal{A}}(\Sigma_{13}, \tau) \restriction x_1. \tag{EP22}$$

Remark 4.5. As observed for $(C, C)_2$ points (see Remark 4.2), the frame points of type (K, K) are not *effective* singular points.

From the present analysis we immediately have the following theorem.

THEOREM 5.2. *Consider $\Sigma \in \Xi$ and $\tau > 0$. If $\mathcal{A}$ succeeds at time τ for Σ and x is a frame point, then x is of one of the following 22 types:*

$$(X, Y),\ (X, F)_{1,2},\ (X, C)_{1,2,3},\ (X, S),\ (X, K)_{1,2,3},\ (F, C),$$

$$(F, S),\ (F, K),\ (C, C)_{1,2},\ (C, S)_{1,2},\ (C, K)_{1,2},\ (S, K)_{1,2},\ (K, K),$$

and we have, respectively, one of the equivalences (EP1)–(EP22).

Acknowledgments. We wish to thank Prof. A. Bressan for suggesting the problem and for much useful advice.

REFERENCES

[1] V. I. ARNOLD, *Geometrical Method in the Theory of ODE*, Springer-Verlag, New York, 1983.

[2] V. G. BOLTIANSKII, *Sufficient conditions for optimality and the justification of the dynamic programming principle*, SIAM J. Control Optim., 4 (1966), pp. 326–361.

[3] P. BRUNOVSKY, *Every normal linear system has a regular time–optimal synthesis*, Math. Slovaca, 28 (1978), pp. 81–100.

[4] ———, *Existence of regular syntheses for general problems*, J. Differential Equations, 38 (1980), pp. 317–343.

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

[6] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, Wiley, New York, 1967.

[7] M. M. PEIXOTO, *Generic properties of ordinary differential equations*, in Studies in Ordinary Differentials Equations, J. Hale, ed., Studies in Mathematics 14, Mathematical Association of America, Washington, 1977, pp. 52–92.

[8] B. PICCOLI, *Regular time–optimal syntheses for smooth planar systems*, Rend. Sem. Mat. Univ. Padova, 95 (1996), to appear.

[9] H. J. SUSSMANN, *The structure of time–optimal trajectories for single–input systems in the plane: The C^∞ nonsingular case*, SIAM J. Control Optim., 25 (1987), pp. 433–465.

[10] ———, *The structure of time–optimal trajectories for single–input systems in the plane: The general real analytic case*, SIAM J. Control Optim., 25 (1987), pp. 868–904.

[11] ———, *Regular synthesis for time–optimal control of single–input real–analytic systems in the plane*, SIAM J. Control Optim., 25 (1987), pp. 1145–1162.

[12] ———, *Envelopes, conjugate points, and optimal bang–bang extremals*, in Algebraic and Geometric Methods in Nonlinear Control Theory, M. Fliess and M. Hazewinkel, eds., D. Reidel Publishing Company, Dordrecht, 1986, pp. 325–346.

[13] ———, *Envelopes, higher–order optimality conditions, and lie brackets*, in Proc. 1989 I.E.E.E. Conf. Decision and Control.

[14] ———, *Synthesis, presynthesis, sufficient conditions for optimality and subanalytic sets*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 1–18.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 1947–1971, November 1996

006

BIFURCATION PROBLEMS FOR SOME PARAMETRIC NONLINEAR PROGRAMS IN BANACH SPACES*

AUBREY B. POORE†

Abstract. Singularities in a class of parametric nonlinear programming problems in Banach spaces are investigated using bifurcation theory. Motivated by the Fritz John first-order necessary conditions and a nonstandard normalization of the multipliers, this problem is first formulated as a system of nonlinear equations. Conditions for this system to be Fredholm are then derived, and singularities are shown to arise from a violation of one or more of the following conditions: strict complementarity, surjectivity of the Fréchet derivative of the active constraints, and a second-order condition. A branching analysis is developed for each of these singularities under a second-order nondegeneracy assumption. Examples from the calculus of variations are then used to illustrate these singularities.

Key words. bifurcation, singularity, parametric nonlinear programming, Banach spaces

AMS subject classifications. 90C31, 90C48

1. Introduction. The parametric constrained optimization problem considered in this work is that of determining the behavior of solution(s) as a parameter or vector of parameters $\alpha \in \mathbf{R}^r$ varies over a region of interest for the problem

$$\begin{aligned} &\text{minimize} && f_0(x,\alpha) \\ &\text{subject to} && F(x,\alpha) = 0, \\ &&& f_i(x,\alpha) \le 0 \text{ for } i = 1,\dots,m, \end{aligned} \tag{1.1}$$

where $f_i \in C^2(U \times V; \mathbf{R})$, $F \in C^2(U \times V; Y)$, U and V are open subsets of the Banach space X and $\mathbf{R}^r$, respectively, and Y is a second Banach space that contains the range of $F(x,\alpha)$. This formulation is sufficiently general to cover classes of parametric problem in the calculus of variations and nonlinear optimal control [1, 9, 16, 23, 56]. Physical problems leading to (1.1) contain parameters. Some are fixed and are known either precisely or imprecisely; others (sometimes called control parameters) may be varied to enhance the system. For imprecisely known parameters, the sensitivity and persistence of the solution to variations in these parameters can be of paramount importance. For control parameters that are varied over a wide range to enhance the system, persistence of minima, exchanges in the type of solutions of (1.1), differentiability, and sensitivity of the solution with respect to variations in the natural parameters are equally important. Indeed, varying system parameters can often assist in obtaining a "global" view of the solution set of (1.1).

For finite-dimensional versions of problem (1.1), local behavior of regular and singular points has been investigated extensively (see, e.g., [4, 6, 20, 25–27, 32, 41, 42, 55] and the references therein). Reviews with extensive bibliographies can be found in the works of Jongen and Weber [28]; Bonnans, Ioffe, and Shapiro [7]; and Fiacco and Ishizuka [15] and the books of Fiacco [13, 14] and Levitin [36]. Numerical continuation procedures for investigating the global behavior of the finite-dimensional versions of (1.1) have also been developed [19, 37]. Following the work of Robinson [46–48], the sensitivity analysis of generalized equations was investigated by several authors [11, 30, 35, 44, 48]. The case of infinite-dimensional space and general cone and set constraints are discussed in a number of recent papers [2, 12, 24, 38, 53, 54]. The questions of parametric dependence for these problems are indeed more difficult than for (1.1). Thus, and as reviewed by Ioffe [22], these investigations have generally made three basic assumptions: (a) local uniqueness of the unperturbed solution, (b) constraint

*Received by the editors November 16, 1994; accepted for publication (in revised form) August 3, 1995.

†Department of Mathematics, Colorado State University, Fort Collins, CO 80523. This research was supported in part by Air Force Office of Scientific Research grants AFOSR F49620-93-1-0133 and F49620-95-1-0136.

qualifications, and (c) a second-order sufficient condition. In a different direction, Ioffe [22] has relaxed the first two assumptions and to an extent the third assumption by reducing a problem of the form (1.1) to a composite unconstrained problem and developing a sensitivity theory for the latter. In this work, we also investigate the problem (1.1) in those cases where (a), (b), and (c) are relaxed, but our methods are those of bifurcation (and singularity) theory. In particular, this work extends some of our previous bifurcation work on the finite-dimensional problem [41, 55] to the infinite-dimensional problem (1.1).

Since the general setting for bifurcation theory is that of Fredholm operators, the first task is to convert the optimization problem to a system of nonlinear equations and to determine conditions under which the system is Fredholm. The formulated system is motivated by the Fritz John first-order necessary conditions, uses a nonstandard normalization of the multipliers similar to that used in our earlier work [41], and contains minimizers, maximizers, and saddle points as well as nonfeasible solutions of the problem (1.1). Given a solution of this system (see (2.4)), the classical implicit function theorem is generally applicable when the following three conditions are valid: (1) strict complementarity, (2) surjectivity of the Fréchet derivative of the active constraints (a constraint qualification), and (3) the bijectivity of the Hessian of the Lagrangian on the null space (kernel) of the Fréchet derivative of the active constraints to an appropriate subspace of the (topological) dual of X (a second-order condition). (Theorem 2.2 gives the precise conditions.) When these conditions are valid, the implicit function theorem guarantees the existence of a locally unique and smooth solution, and one can compute derivatives of the solution and multipliers with respect to the system parameters and thus perform a first-order sensitivity analysis. As a system parameter(s) is varied, one invariably encounters situations where one or more of these three conditions fail. Many of the interesting phenomena, e.g., changes in the active set, loss of the minima, exchanges in critical point type, multiple solutions, and loss of differentiability of the solution in the natural parameters of the system, occur at these singularities. Thus, neighborhoods of these singularities are regions of "extreme sensitivity."

Bifurcation theory is now quite a large subject as evidenced by books on the subject (e.g., [10, 17, 18]), and the various singularities are far too numerous to delineate in this work. In particular, we do not exploit symmetry in the problem. Instead, the objectives are to derive sufficient and almost necessary conditions for the applicability of this theory to the parametric problem (1.1) and to demonstrate a branching analysis for some of the simplest but generic singularities. Associated with the bifurcation or branching of multiple solutions is the question of the persistence of minima or changes in critical point type as the parameter or parameters are varied about the singularity. Although we treat this stability topic to some extent in the examples, a systematic study under more restrictive assumptions will be undertaken in future work.

The paper is organized as follows. Section 2 contains a summary of much of the notation used in the paper, the formulation of a system of nonlinear equations that must be satisfied by a minimizer of (1.1), the development of necessary and sufficient conditions for the solution of this system to be regular, and conditions under which this system is Fredholm. The bifurcation theory framework for the investigation of the singularities is presented in §3. A bifurcation analysis for some of the more commonly occurring singularities is presented in §§4–6 with some illustrative examples. Further research problems are discussed in §7.

2. Systems formulation and the bifurcation problems. The first objective is to convert the parametric constrained optimization problem (1.1) to a closed system of nonlinear equations; this is accomplished through the use of the Fritz John first-order necessary conditions and a nonstandard normalization of the Lagrange multipliers. Having developed this system, we next establish sufficient and almost necessary conditions (Theorem 2.2) for the Fréchet

derivative of this system to be bijective so that with sufficient smoothness the hypotheses of the implicit function theorem are satisfied. (If one adds a strong monotonicity condition (Theorem 2.3), then a minimizer exists and persists locally.) The primary interest in this work is the investigation of the local solution structure when this Fréchet derivative is no longer bijective, but is Fredholm in which case bifurcation theory is applicable. Thus the final Theorem 2.4 in this section gives conditions sufficient to ensure that the Fréchet derivative of this system is Fredholm. We first explain the notation used in subsequent discussion and review some of the needed facts from functional analysis [50].

2.1. Notation. X and Y with appropriate norms will be assumed to be Banach spaces throughout without further comment. If $F : X \times \mathbf{R}^r \to Y$, the Fréchet derivative with respect to $x \in X$ will be denoted by $D_x F(x, \alpha)$, with respect to $\alpha \in \mathbf{R}^r$ by $D_\alpha F(x, \alpha)$, and similarly for higher derivatives. X^* denotes the topological (normed) dual of the normed linear space X, i.e., the space of bounded linear functionals on X, and $\langle x, x^* \rangle_X$ represents the action of a bounded linear functional $x^* \in X^*$ on an element $x \in X$. The subscript X on $\langle x, x^* \rangle_X$ will be omitted if the context is clear. If $L : X \to Y$ is a linear transformation, $\mathcal{N}(L)$ will denote the null space (kernel) of L and $\mathcal{R}(L)$, the range (image) of L. The adjoint of a bounded linear transformation $L : X \to Y$ is that unique bounded linear transformation $L^* : Y^* \to X^*$ satisfying $\langle Lx, y^* \rangle = \langle x, L^* y^* \rangle$ for all $x \in X$ and $y^* \in Y^*$. The following closed-range theorem will be used: assuming X and Y are Banach spaces, the range of L is norm closed if and only if the range of L^* is norm (or wk*-) closed.

Let M be a subspace of the Banach space X, and N, a subspace of X^*. Then define the annihilators $M^\perp$ and ${}^\perp N$ by $M^\perp = \{x^* \in X^* : \langle x, x^* \rangle = 0 \text{ for all } x \in M\}$ and ${}^\perp N = \{x \in X : \langle x, x^* \rangle = 0 \text{ for all } x^* \in N\}$, respectively. If $L : X \to Y$ is a bounded linear transformation, then $\mathcal{N}(L) = {}^\perp \mathcal{R}(L^*)$ and $\mathcal{N}(L^*) = \mathcal{R}(L)^\perp$; and if, in addition, the range of L is closed, $\mathcal{R}(L) = {}^\perp \mathcal{N}(L^*)$ and $\mathcal{R}(L^*) = \mathcal{N}(L)^\perp$.

A bounded linear transformation $L : X \to Y$ is called a Fredholm operator if the range is closed, the codimension of the range is finite (i.e., the dimension of the quotient space $Y/\mathcal{R}(L)$ is finite), and the dimension of the null space is finite. In this case, the index of L is defined by $\text{index}\,(L) = \dim \mathcal{N}(L) - \dim Y/\mathcal{R}(L)$. (To compute this index, we make use of $\dim Y/\mathcal{R}(L) = \dim \mathcal{N}(L^*)$ for a closed $\mathcal{R}(L)$.) A bounded linear transformation $L : X \to Y$ is called a semi-Fredholm operator if it has closed range and either the codimension of the range is finite or the dimension of the null space is finite. Properties of Fredholm and semi-Fredholm operators can be found in the books by Kato [29] and Schechter [51].

Two basic assumptions in the theorems to follow are that the Fréchet derivative $D_x F(\hat{x}, \hat{\alpha})$ has closed range and $\mathcal{N}(D_x F(\hat{x}, \hat{\alpha}))$ is complemented in X (or splits the space X). A closed subspace X_1 of a Banach space is complemented in X if there exists a closed subspace X_2 of X such that $X_1 + X_2 = X$ and $X_1 \cap X_2 = \{0\}$. In this case we write $X = X_1 \oplus X_2$ and say that X is the topological (internal) direct sum of X_1 and X_2 and that X_1 splits the space. This assumption is needed since infinite-dimensional closed subspaces of Banach spaces need not be complemented. Here are some valid examples. If X is a Hilbert space, then closed subspaces are complemented. If M is a closed subspace of a Banach space X and either M or X/M is finite-dimensional, then M is complemented in X [50, p. 106]. A fact that will be used in some of the examples from the calculus of variations is that if $L : X \to Y$ is a bounded linear transformation with a finite-dimensional range, i.e., L is compact [50, p. 104], then the null space of L has finite codimension and is thus complemented in X. In nonlinear optimal control, the space $X = X_1 \oplus X_2$ is split into state and control variables. One can often use this along with the above facts to show that $\mathcal{N}(L)$ is complemented even when it and $X/\mathcal{N}(L)$ are infinite dimensional.

The Cartesian product of two Banach spaces Y and Z, denoted by $Y \times Z$, is a Banach space under the norm $\|(y, z)\|_{Y\times Z} = (\|y\|_Y^2 + \|z\|_Z^2)^{1/2}$. (There are several such equivalent norms.) Next, if $L : X \to Y$ and $K : X \to Z$ are bounded linear transformations, then $J : X \to Y \times Z$ by $Jx = (Lx, Kx)$ is also a bounded linear transformation. We shall make use of the following lemma on the closed image (range) [1, p. 80]: If the subspace $\mathcal{R}(L)$ is closed in Y and $\mathcal{R}(K|_{\mathcal{N}(L)})$ is closed in Z, then the subspace $\mathcal{R}(J)$ is closed in $Y \times Z$.

Next, define two index sets $\mathcal{A}$ and $\mathcal{A}_s$ by

$$\begin{aligned} \mathcal{A}(\hat{x}, \hat{\alpha}) &= \{i \geq 1 : f_i(\hat{x}, \hat{\alpha}) = 0\}, \\ \mathcal{A}_s(\hat{x}, \hat{\alpha}) &= \{i \in \mathcal{A} : \lambda_i \neq 0\}, \end{aligned} \tag{2.1}$$

where the subscript s on $\mathcal{A}$ is for "strongly active" and the λ_i is the multiplier corresponding to the ith constraint f_i used in the Lagrangian $\mathcal{L}$ in (2.3). By permuting the inequality constraints, we may assume $\mathcal{A}_s(\hat{x}, \hat{\alpha}) = \{1, 2, \ldots, l\}$ and $\mathcal{A}(\hat{x}, \hat{\alpha}) = \{1, 2, \ldots, k\}$ with $l \leq k$ and define

$$\begin{aligned} D_x f_A &= \begin{bmatrix} D_x f_1(\hat{x}, \hat{\alpha}) \\ \vdots \\ D_x f_l(\hat{x}, \hat{\alpha}) \end{bmatrix}, \quad D_x f_B = \begin{bmatrix} D_x f_{l+1}(\hat{x}, \hat{\alpha}) \\ \vdots \\ D_x f_k(\hat{x}, \hat{\alpha}) \end{bmatrix}, \quad \text{and} \\ D_x f_C &= \begin{bmatrix} D_x f_{k+1}(\hat{x}, \hat{\alpha}) \\ \vdots \\ D_x f_m(\hat{x}, \hat{\alpha}) \end{bmatrix}. \end{aligned} \tag{2.2}$$

Since the codimension of $\mathcal{N}(D_x f_A)$ is finite, which follows from $\mathcal{R}(D_x f_A)$ being finite dimensional, $\mathcal{N}(D_x f_A)$ is complemented in X. Let $L = D_x F(\hat{x}, \hat{\alpha})$, and suppose that the null space of L is complemented in X and the range is closed in Y. If $\tilde{L} = \begin{bmatrix} L \\ D_x f_A \end{bmatrix}$, the lemma on the closed image [1, p. 80] implies that the range of $\tilde{L}$ is closed. Also, one can show that $\mathcal{N}(\tilde{L})$ is complemented in X so that $X = X_1 \oplus X_2$, where $X_1 = \mathcal{N}(\tilde{L})$. This decomposition defines two continuous projections P_1 and P_2 ($P_i : X \to X_i$). The adjoints P_1^* and P_2^* are also continuous projections and induce a topological direct sum of the dual space $X^* = X_1^* \oplus X_2^*$, where $X_1^* = P_1^* X^* = X_2^\perp$ and $X_2^* = P_2^* X^* = X_1^\perp = \mathcal{N}(\tilde{L})^\perp = \mathcal{R}(\tilde{L}^*)$. The notation is consistent with the fact that X_i^* is the topological dual of X_i. Similarly, the adjoint P_i^{**} of P_i^* is a continuous projection of the second dual X^{**} onto $P_i^{**} X^{**} \equiv X_i^{**}$ ($P_i^{**} : X^{**} \to X_i^{**}$) with the corresponding decomposition $X^{**} = X_1^{**} \oplus X_2^{**}$, where $X_1^{**} = \mathcal{R}(\tilde{L}^*)^\perp = \mathcal{N}(\tilde{L}^{**})$.

The (weak) Lagrangian for the problem (1.1) is

$$\mathcal{L} = \langle f_0(x, \alpha), \lambda_0 \rangle_{\mathbf{R}} + \langle f(x, \alpha), \lambda \rangle_{\mathbf{R}^m} + \langle F(x, \alpha), y^* \rangle_Y, \tag{2.3}$$

where $\lambda_0 \in \mathbf{R}^*$ ($= \mathbf{R}$), $\lambda \in \mathbf{R}^{m*}$ ($= \mathbf{R}^m$), and $y^* \in Y^*$ are Lagrange multipliers. Here we have used the notation λ without a subscript to refer to the vector $(\lambda_1, \lambda_2, \ldots, \lambda_m)$ that does not include, in particular, the component λ_0. The same applies to the vector f with a range in $\mathbf{R}^m$. Finally, define $D_x^2 \mathcal{L}_{ij} = P_i^* D_x^2 \mathcal{L} P_j$, and note that $D_x^2 \mathcal{L}_{ij}^* = P_j^* D_x^2 \mathcal{L}^* P_i^{**}$.

2.2. The nonlinear system. Having completed a discussion of some of the notation, the next task is to convert the parametric optimization problem (1.1) to a system of nonlinear equations using the Fritz John first-order necessary conditions. This is the content of the following theorem.

THEOREM 2.1 (see [1]). *Let $f_i \in C^2(U \times V; \mathbf{R})$, $F \in C^2(U \times V; Y)$, U and V be open subsets of the Banach space X and $\mathbf{R}^r$, respectively, and Y be a second Banach space containing the range of $F(x, \alpha)$. Let $\hat{x} \in U$ be a local minimizer of the optimization problem*

(1.1) *at* $\hat{\alpha} \in V$, *and assume that the Fréchet derivative* $D_x F(\hat{x}, \hat{\alpha})$ *has closed range. Then there exist Lagrange multipliers* $\hat{\lambda}_i$ $(i = 0, 1, \ldots, m)$ *and* $\hat{y}^* \in Y^*$, *not all zero, such that* $\chi \equiv (x, y^*, \lambda, \lambda_0) = (\hat{x}, \hat{y}^*, \hat{\lambda}, \hat{\lambda}_0)$ *solves the system of equations*

$$G(x, y^*, \lambda, \lambda_0; \alpha) = \begin{bmatrix} D_x \mathcal{L}(x, y^*, \lambda, \lambda_0; \alpha) \\ F(x, \alpha) \\ \Lambda f(x, \alpha) \\ N(\lambda_0, \lambda, y^*) \end{bmatrix} = 0 \tag{2.4}$$

and inequalities $\lambda_0 \geq 0$ *and* $\lambda_i \geq 0$ *and* $f_i(x, \alpha) \leq 0$ *for* $i = 1, \ldots, m$. *Here,* $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ *is a diagonal matrix and the equation* $N(\lambda_0, \lambda, y^*) = 0$ *represents a normalization of the multipliers that ensures that not all multipliers are zero.*

Choices for the normalization N include

$$N_1(\lambda_0, \lambda, y^*) = \lambda_0 - 1, \tag{2.5a}$$

$$N_2(\lambda_0, \lambda, y^*) = \sum_{i=0}^{m} \lambda_i^2 + \|y^*\|^2 - 1, \tag{2.5b}$$

$$N_3(\lambda_0, \lambda, y^*) = \sum_{i=0}^{m} \lambda_i^2 + |\langle y_0, y^* \rangle|^2 - 1, \tag{2.5c}$$

where $y_0 \in Y$ is chosen so that $|\langle y_0, y^* \rangle| \geq \frac{1}{2}\|y^*\|$. The usual choice is $N_1 = 0$, which is guaranteed by a variety of constraint qualifications such as the following. If the active inequality constraints at $(\hat{x}, \hat{\alpha})$ are $\{f_i\}_{i=1}^k$, then $\lambda_0 \neq 0$ if either of the following two constraint qualifications is satisfied:

LICQ: The map $(D_x F(\hat{x}, \hat{\alpha}), D_x f_1(\hat{x}, \hat{\alpha}), \ldots, D_x f_k(\hat{x}, \hat{\alpha})) : X \to Y \times \mathbf{R}^k$ is surjective.

MFCQ: The map $D_x F(\hat{x}, \hat{\alpha}) : X \to Y$ is surjective and $\{h \in X : D_x f_i(\hat{x}, \hat{\alpha})h < 0$ for $i = 1, \ldots, k;\ D_x F(\hat{x}, \hat{\alpha})h = 0\}$ is nonempty.

LICQ is the infinite-dimensional analogue of the linear independence constraint qualification and MFCQ, the Mangasarian–Fromovitz constraint qualification. The latter is weaker than the first in that it is implied by the first. Thus, not LICQ is implied by not MFCQ, so a violation of LICQ is a weaker condition than a violation of MFCQ. Since a violation of the LICQ may lead to bifurcations, we do not use the normalization $N_1 = 0$ when $\lambda_0 = 0$. Instead, we use either the normalization N_2 or N_3, which allows for the smooth transition of λ_0 away from zero. (Such normalizations were introduced in our earlier work [41] on finite-dimensional problems.) With the exception of a Hilbert space in which case $\|y^*\|^2 = \langle y^*, y^* \rangle$, N_2 is generally not differentiable with respect to y^*. Thus the normalization N_3 will be used in those cases where $\lambda_0 = 0$.

2.3. The regular case. The next objective is to establish sufficient and almost sufficient conditions for the Fréchet derivative of the system (2.4) to be bijective at a solution of the system, in which case the conclusion of the implicit function theorem is valid.

THEOREM 2.2. *Let* $f_i \in C^2(U \times V; \mathbf{R})$ *for* $i = 0, 1, \ldots, m$, $F \in C^2(U \times V; Y)$, U *and* V *be open sets in the Banach space* X *and* $\mathbf{R}^r$, *respectively, and* Y *be a second Banach space containing the range of* $F(x, \alpha)$. *Let* $\mathcal{X} \equiv X \times Y^* \times \mathbf{R}^m \times \mathbf{R}$, *and suppose that* $(\hat{\chi}; \hat{\alpha}) = (\hat{x}, \hat{y}^*, \hat{\lambda}, \hat{\lambda}_0; \hat{\alpha}) \in \mathcal{X} \times \mathbf{R}^r$ *is a solution of* $G(\chi; \alpha) = 0$. *Suppose that the Fréchet derivative* $D_x F(\hat{x}, \hat{\alpha})$ *has closed range and its null space is complemented in* X. *Under these assumptions, a set of necessary and sufficient conditions for* $D_\chi G(\hat{\chi}, \hat{\alpha})$ *to be bijective is that*

(a) *strict complementarity holds, i.e.,* $\mathcal{A} = \mathcal{A}_s$ *and* $k = l$ *in* (2.2);

(b) *the bounded linear transformation* $D_x^2 \mathcal{L}_{11} : X_1 \to X_1^*$ *is bijective;*

(c) *the bounded linear transformation* $\tilde{L} \equiv (D_x F(\hat{x}, \hat{\alpha}), D_x f_1(\hat{x}, \hat{\alpha}), \ldots, D_x f_l(\hat{x}, \hat{\alpha})) : X \to Y \times \mathbf{R}^l$ *is surjective.*

If the bounded linear transformation $D_\chi G(\hat{\chi}, \hat{\alpha})$ *is bijective, then there exist neighborhoods* $\mathcal{B}_1$ *of* $\alpha = \hat{\alpha}$ *and* $\mathcal{B}_2$ *of* $\hat{\chi} = (\hat{x}, \hat{y}^*, \hat{\lambda}, \hat{\lambda}_0)$ *such that the following is true: there exists a function* $\phi \in C^1(\mathcal{B}_1, \mathcal{X})$ *such that* $\phi(\hat{\alpha}) = \hat{\chi}$, $G(\phi(\alpha), \alpha) = 0$ *for all* $\alpha \in \mathcal{B}_1$, *and this solution is locally unique in that if* $(\chi, \alpha) \in \mathcal{B}_1 \times \mathcal{B}_2$ *and* $G(\chi, \alpha) = 0$, *then* (χ, α) *belongs to the manifold defined by* ϕ, *i.e.,* $\chi = \phi(\alpha)$ *for some* $\alpha \in \mathcal{B}_1$. *Finally, if* f_i *and* F *are* $C^k (k \geq 2)$ (C^∞ *or real analytic*), *then* ϕ *is* C^{k-1} (C^∞ *or real analytic, respectively*) *on* $\mathcal{B}_1$.

Several remarks are in order. The term *critical point* will refer to any solution of system (2.4), *regular point* will describe any solution of (2.4) for which conditions (a), (b), and (c) of Theorem 2.2 are valid, and *singular point* is reserved for any solution of (2.4) at which $D_\chi G$ is not bijective, i.e., at which one or more of (a), (b), and (c) is violated. The importance of the three conditions (a), (b), and (c) in this theorem is that they provide a set of necessary and sufficient conditions for a violation of the bijectivity of $D_\chi G$ and thus an initial classification for the singularities and bifurcation problems. Next, note that condition (c) is just the surjectivity (linear independence) constraint qualification. Certainly, weaker constraint qualifications such as those of the MFCQ guarantee $\lambda_0 \neq 0$. In the bifurcation analysis to follow we treat the case $\lambda_0 \neq 0$ or $\lambda_0 = 0$ by the *weakest possible set of conditions*, i.e., whether or not $D_x f_0(\hat{x}, \hat{\alpha}) \in \mathcal{R}(\tilde{L})$.

In the finite-dimensional case, we [41, 55] have characterized the type (i.e., maximum, minimum, or saddle point) of a regular point of (2.4) by three sets of numbers: (1) the sign of $f_i(x, \alpha)$ for $i \notin \{1, \ldots, k\}$, (2) the sign of λ_i for $i \in \{0, 1, \ldots, k\}$, and (3) signs of the eigenvalues of $D_x^2 \mathcal{L}_{11}$. The index, i.e., dimension of the largest space on which $D_x^2 \mathcal{L}_{11}$ is negative definite, and the nullity, i.e., the dimension of the null space of $D_x^2 \mathcal{L}_{11}$, are sufficient to characterize the numbers in (3). Definitions for the index and nullity of a bounded bilinear form $B : X \times X \to \mathbf{R}$ on a Banach space X can be similarly defined [57, pp. 86–87].

Finally and for completeness, we remark that a modification of this theorem guarantees the local persistence of a minimizer as stated in the following theorem.

THEOREM 2.3. *Suppose that in addition to the hypotheses of Theorem* 2.2 *that*

$$D_x^2 \mathcal{L}(\hat{\chi}, \hat{\alpha})[h, h] \geq C \|h\|_X^2 \tag{2.6}$$

for all $h \in \mathcal{N}(\tilde{L}(\hat{x}, \hat{\alpha}))$. *Then* $\hat{x}$ *is a local minimizer of* (1.1) *at* $\alpha = \hat{\alpha}$, *and there exists a constant* $\bar{C}$ *and neighborhood* $\tilde{\mathcal{B}}_1$ *of* $\alpha = \hat{\alpha}$ *such that for each* $\alpha \in \tilde{\mathcal{B}}_1$, $D_x^2 \mathcal{L}(\chi(\alpha), \alpha)[h, h] \geq \bar{C} \|h\|_X^2$ *for all* $h \in \mathcal{N}(\tilde{L}(x(\alpha), \alpha))$ *and* $\lambda_i(\alpha) > 0$ *for each* $i = 1, \ldots, l = k$. *In particular, the local minimizer persists locally about* $\alpha = \hat{\alpha}$.

This theorem is established under weaker assumptions by Alt [3] and Shapiro [52], and thus we omit the proof. As noted by Alekseev, Tikhomirov, and Fomin [1, p. 157], the strong monotonicity condition (2.6) places severe restrictions on the space $\mathcal{N}(\tilde{L}(\hat{x}, \hat{\alpha}))$. To explain, first note that the closed subspace $\mathcal{N}(\tilde{L}(\hat{x}, \hat{\alpha}))$ of X along with the norm of X is a Banach space. Next, $\langle u, v \rangle = D_x^2 \mathcal{L}(\hat{\chi}, \hat{\alpha})[u, v]$ defines an inner product on $\mathcal{N}(\tilde{L}(\hat{x}, \hat{\alpha}))$, and due to the assumptions of Theorem 2.2 we have

$$C \|h\|_X^2 \leq \langle h, h \rangle \leq \|D_x^2 \mathcal{L}(\hat{\chi}, \hat{\alpha})\| \, \|h\|_X^2.$$

Thus the norm induced on $\mathcal{N}(\tilde{L}(\hat{x}, \hat{\alpha}))$ by the inner product $\langle \cdot, \cdot \rangle$ is equivalent to that of $\| \cdot \|_X$, so the Banach space $(\mathcal{N}(\tilde{L}(\hat{x}, \hat{\alpha})), \| \cdot \|_X)$ is linearly homeomorphic to the Hilbert space $(\mathcal{N}(\tilde{L}(\hat{x}, \hat{\alpha})), \langle \cdot, \cdot \rangle)$. Interestingly, we now have that $X_1^* = \mathcal{N}(\tilde{L}(\hat{x}, \hat{\alpha}))^*$ is identified with this Hilbert space via the Riesz representation theorem. Given these remarks, note that the strong monotonicity condition (2.6) then implies condition (b) in Theorem 2.2.

To circumvent this restrictive strong monotonicity condition (2.6), Ioffe [21] and Maurer [39] replace it with a coercivity condition in a weaker norm and use this *two-norm discrepancy* along with some compatibility conditions to establish a locally isolated minimum [21, 39]. This technique is now commonly used in the literature, particularly for the integral-type objective functions. Indeed, it is used in §4.2 to establish a minimum in an example from the calculus of variations. From the implicit function point of view, this same problem manifests itself in condition (b) in Theorem 2.2 in that the topological duals X^* and X_1^* are too large to obtain surjectivity of the bounded linear transformation $D_x^2\mathcal{L}_{11} : X_1 \to X_1^*$ for integral-type objective functions.

2.4. The Fredholm operator and bifurcation problems. Since the general framework for bifurcation problems is that of Fredholm operators, conditions must be imposed to ensure that the Fréchet derivative $D_\chi G$ is Fredholm. This is the content of the next theorem.

THEOREM 2.4. *Let $f_i \in C^2(U \times V; \mathbf{R})$ for $i = 0, 1, \ldots, m$, $F \in C^2(U \times V; Y)$, U and V be open sets in the Banach space X and $\mathbf{R}^r$, respectively, and Y be a second Banach space containing the range of $F(x, \alpha)$. Let $(\hat{\chi}; \hat{\alpha}) = (\hat{x}, \hat{y}^*, \hat{\lambda}, \hat{\lambda}_0; \hat{\alpha}) \in U \times V$ be a solution of $G(\chi; \alpha) = 0$. Suppose that $\mathcal{N}(D_x F(\hat{x}, \hat{\alpha}))$ is complemented in X. Then the Fréchet derivative $D_\chi G(\hat{\chi}, \hat{\alpha})$ is Fredholm if*

(a) *$D_x^2\mathcal{L}_{11} : X_1 \to X_1^*$ is Fredholm;*

(b) *the range of $D_x F(\hat{x}, \hat{\alpha})$ is closed and has finite codimension in Y.*

Also, if $D_x^2\mathcal{L}_{11}$ and $D_x F(\hat{x}, \hat{\alpha})$ have closed ranges and $D_\chi G(\hat{\chi}, \hat{\alpha})$ is Fredholm, then (a) *and* (b) *are valid.*

The proof of this theorem is included in the appendix. Just as we have discussed the surjectivity of $D_x^2\mathcal{L}_{11} : X_1 \to X_1^*$ in Theorem 2.2 as being too strong for integral-type objective functions, we also note that condition $D_x^2\mathcal{L}_{11} : X_1 \to X_1^*$ being Fredholm is too strong for certain applications. Such problems arise for unconstrained problem in the calculus of variations and have been resolved by two methods. The first approach is described by Zeidler [59, §29.18], wherein one introduces a generalized inner product on a Banach space. The second method is to embed the function spaces continuously and densely in a Hilbert space and then use the structure of the Hilbert space to perform the bifurcation analysis as in the work of Bobylev and Krasnosel'skii [5]. These same approaches will be investigated in future work for the constrained problem with applications to the calculus of variations and control systems.

3. A brief review of the bifurcation setting. The objective in this section is to give two bifurcation theorems (3.2 and 3.4) that will be used in the subsequent three sections. The singularities in these theorems are generic [17]. To explain these theorems within the context of a general problem $G(x, y) = 0$, we briefly describe the Lyapunov–Schmidt procedure for a reduction to a finite number of equations in a finite number of unknowns and then delineate some of the cases to which Theorems 3.2 and 3.4 apply. The setting described here is based on the books by Nirenberg [40] and Rabier [45] as well as our earlier work [55]. We start with the implicit function theorem, which is also used in §6.

THEOREM 3.1 (implicit function theorem [40]). *Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be Banach spaces, let $U_0 \times V_0 \subset \mathcal{X} \times \mathcal{Y}$ be an open neighborhood of a point (x_0, y_0) that solves $G(x, y) = 0$, and suppose that*

(a) *$G \in C^p(U_0 \times V_0, \mathcal{Z})$ for some $p \geq 1$;*

(b) *$\mathcal{R}(D_x G(x_0, y_0)) = \mathcal{Z}$;*

(c) *$\mathcal{N}(D_x G(x_0, y_0)) = \mathcal{X}_1$ has a closed complementing subspace $\mathcal{X}_2$ in $\mathcal{X}$; i.e., $\mathcal{X}$ is a topological direct sum $\mathcal{X} = \mathcal{X}_1 \oplus \mathcal{X}_2$.*

Then there exist open neighborhoods U of x_0 and V of y_0 and a solution $x_2 = u(x_1, y) \in C^p(PU \times V, \mathcal{X}_2)$ of $G(x_1 + x_2, y) = 0$, where $P : \mathcal{X} \to \mathcal{X}_1$ is a projection of $\mathcal{X}$ onto $\mathcal{X}_1$,

$u(x_{01}, y_0) = x_{02}$, $x_{01} = Px_0$, and $x_{02} = (I - P)(x_0)$. Furthermore, U and V can be chosen so that this solution is locally unique in that if $(x, y) \in U \times V$ is a solution of $G(x, y) = 0$, then $x = x_1 + u(x_1, y)$ for some $x_1 \in PU$.

The problem of the previous section was that of determining the local structure of the solutions of the equation $G(\chi, \alpha) = 0$, where $\chi \in \mathcal{X} \equiv X \times Y^* \times \mathbf{R}^m \times \mathbf{R}$, the range of G is in $\mathcal{Z} \equiv X^* \times Y \times \mathbf{R}^m \times \mathbf{R}$, and $\alpha \in \mathbf{R}^r$. The Lyapunov–Schmidt method for reducing this problem to one of a finite number of equations in a finite number of unknowns is perhaps best described by putting $\tilde{\chi} = (\chi, \alpha)$ and explaining the procedure for the generic problem $G(\tilde{\chi}) = 0$, where $G : \tilde{\mathcal{X}} \to \mathcal{Z}$. For smoothness, we require $G \in C^p(U_0, \mathcal{Z})$ for some fixed $p \geq 1$, where U_0 is an open neighborhood of a solution $\tilde{\chi}_0$ of $G = 0$. Next, assume that $D_{\tilde{\chi}} G(\tilde{\chi}_0)$ is Fredholm operator in that $\mathcal{R}(D_{\tilde{\chi}} G(\tilde{\chi}_0)) = \mathcal{Z}_1$ is a closed linear subspace of $\mathcal{Z}$ of finite codimension and $\mathcal{N}(D_{\tilde{\chi}} G(\tilde{\chi}_0)) = \tilde{\mathcal{X}}_1$ is finite dimensional. By this assumption, $\tilde{\mathcal{X}}$ and $\mathcal{Z}$ can be decomposed into topological direct sums $\tilde{\mathcal{X}} = \tilde{\mathcal{X}}_1 \oplus \tilde{\mathcal{X}}_2$ and $\mathcal{Z} = \mathcal{Z}_1 \oplus \mathcal{Z}_2$, where $\dim \mathcal{Z}_2$ and $\dim \tilde{\mathcal{X}}_1$ are finite. Let Q be an associated (continuous) projection of $\mathcal{Z}$ onto $\mathcal{Z}_1$. Applying Q and $(I - Q)$ to the equation $G(\tilde{\chi}) = 0$ yields an equivalent system

$$\begin{aligned} QG(\tilde{\chi}) &= 0, \\ (I - Q)G(\tilde{\chi}) &= 0. \end{aligned} \tag{3.1}$$

Theorem 3.1 can be applied to the problem

$$0 = g(\tilde{\chi}_2, \tilde{\chi}_1) \equiv QG(\tilde{\chi}_1 + \tilde{\chi}_2) \; : \; \tilde{\mathcal{X}}_2 \times \tilde{\mathcal{X}}_1 \to \mathcal{Z}_1 \tag{3.2}$$

to obtain the existence of a C^p locally unique solution $\tilde{\chi}_2 = u(\tilde{\chi}_1)$ (locally about χ_{01}). Hence $\tilde{\chi}_1 + u(\tilde{\chi}_1)$ is a solution of $G(\tilde{\chi}) = 0$ if and only if

$$(I - Q)G(\tilde{\chi}_1 + u(\tilde{\chi}_1)) = 0. \tag{3.3}$$

Since the range of $I - Q$ is finite dimensional, the problem $G(x, y) = 0$ has been reduced to one of a finite number of equations in a finite number of unknowns. We now consider the simplest but generic cases of the bifurcation equations (3.3).

The first case is that in which $\mathcal{Z}_1 = \mathcal{Z}$ and the index of the Fredholm operator $D_{\tilde{\chi}} G(\tilde{\chi}_0)$ is one. Then the dimension of the null space of $D_{\tilde{\chi}} G(\tilde{\chi}_0)$ is one. For the parametric problem $G(\chi, \alpha) = 0$, where $G(\chi, \alpha) : \mathcal{X} \times \mathbf{R} \to \mathcal{Z}$ (α is now a single parameter), the problem breaks into two cases:

(i) $D_\chi G(\chi_0, \alpha_0)$ is surjective.

(ii) Codimension of $\mathcal{R}(D_\chi G(\chi_0, \alpha_0))$ is one and $D_\alpha G(\chi_0, \alpha_0) \notin \mathcal{R}(D_\chi G(\chi_0, \alpha_0))$.

In case (i), $D_\chi G(\chi_0, \alpha_0)$ is bijective and the implicit function theorem (Theorem 3.1) is applicable with $y = \alpha$ and $\mathcal{X}_1 = \{0\}$ so that $\mathcal{X} = \mathcal{X}_2$. This is the content of Theorem 2.2 for a one-dimensional parameter $\alpha \in \mathbf{R}$. The second case (ii) is addressed in the following theorem.

THEOREM 3.2. *Let $\mathcal{X}$ and $\mathcal{Z}$ be Banach spaces and $\mathcal{B}_1$ and $\mathcal{B}_2$ be open neighborhoods of $\chi_0 \in \mathcal{X}$ and $\alpha_0 \in \mathbf{R}$, respectively, and assume that $G(\chi_0, \alpha_0) = 0$, $G \in C^p(\mathcal{B}_1 \times \mathcal{B}_2; \mathcal{Z})$ for some fixed $p \geq 1$, $D_{(\chi,\alpha)} G(\chi_0, \alpha_0)$ is a Fredholm operator of index one, the codimension of $\mathcal{R}(D_\chi G(\chi_0, \alpha_0))$ is one, and $D_\alpha G(\chi_0, \alpha_0) \notin \mathcal{R}(D_\chi G(\chi_0, \alpha_0))$ so that the dimension $\mathcal{N}(D_\chi G(\chi_0, \alpha_0))$ is one. Let $\phi \in \mathcal{X}$, $\phi^* \in \mathcal{X}^*$, and $\psi^* \in \mathcal{Z}^*$ span the null space of $D_\chi G(\chi_0, \alpha_0)$, the one-dimensional space complement of the range of $D_\chi G(\chi_0, \alpha_0)^*$, and the null space of $D_\chi G(\chi_0, \alpha_0)^*$, respectively. Then there exist open neighborhoods $U \subset \mathcal{B}_1 \times \mathcal{B}_2$ of (χ_0, α_0) and $I \subset \mathbf{R}$ of 0 and a function $(\chi(\epsilon), \alpha(\epsilon)) \in C^p(I; U)$ such that $(\chi(0), \alpha(0)) = (\chi_0, \alpha_0)$, $\frac{d\alpha(0)}{d\epsilon} = 0$, and $G(\chi(\epsilon), \alpha(\epsilon)) = 0$ for all $\epsilon \in I$. Furthermore, any solution $(\chi, \alpha) \in U$ of $G = 0$ is given by $(\chi, \alpha) = (\chi(\epsilon), \alpha(\epsilon))$ for some $\epsilon \in I$; and the parameterization can be chosen so that $P(\chi(\epsilon) - \chi_0) = \epsilon\phi$ and $(I - P)(\chi(\epsilon) - \chi_0) = v(\epsilon)$,*

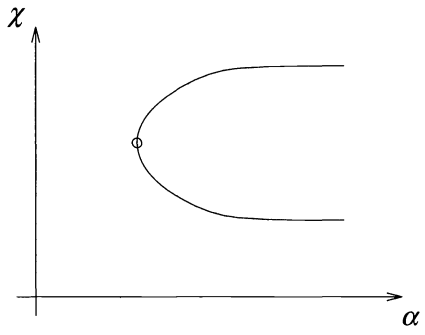


FIG. 1. *Quadratic fold point.*

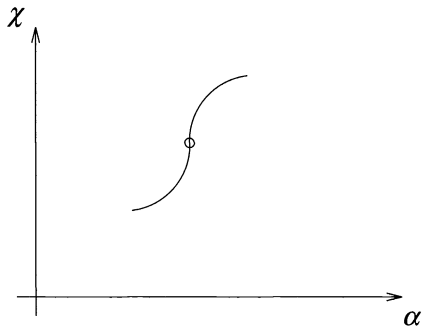


FIG. 2. *Hysteresis point.*

where $v(0) = 0$, $\frac{dv(0)}{d\epsilon} = 0$, *and the projection* P *is defined by* $Pu = \langle u, \phi^* \rangle \phi / \langle \phi, \phi^* \rangle$. *Finally, if* G *is an analytic function of* (χ, α) *near* (χ_0, α_0), *then* $(\chi(\epsilon), \alpha(\epsilon))$ *is analytic in* ϵ *near zero.*

Proof of Theorem 3.2. Let $[\phi] = \mathcal{N}(D_\chi G(\chi_0, \alpha_0))$ and $[\psi^*] = \mathcal{N}(D_\chi G(\chi_0, \alpha_0)^*)$. Make the transformation $\chi = \chi_0 + \epsilon\phi + v$, where $v \in \mathcal{N}(P) = \mathcal{R}(I - P)$, and consider the problem

$$\hat{G}(v, \alpha, \epsilon) = G(\chi_0 + \epsilon\phi + v, \alpha).$$

Then $\hat{G} : (I - P)\mathcal{X} \times \mathbf{R} \times \mathbf{R} \to \mathcal{Z}$. Now $\hat{G}(0, \alpha_0, 0) = 0$, $\hat{G}(v, \alpha, \epsilon) \in C^p(\hat{U} \times \hat{V} \times \hat{I}; \mathcal{Y})$ for some open sets $\hat{U} \subset (I - P)\mathcal{X}$ of $(I - P)\chi_0$ and $\hat{V} \subset \mathbf{R}$ and $\hat{I} \subset \mathbf{R}$ about zero. The Fréchet derivative $D_{(v,\alpha)}\hat{G}(0, \alpha, 0)$ is a bijective, bounded operator from $(I - P)\mathcal{X} \times \mathbf{R}$ onto $\mathcal{Z}$. Thus the implicit function theorem (Theorem 3.1) yields the result. □

If $p \geq 2$ and

$$\frac{d^2\alpha(0)}{d\epsilon^2} \equiv -\frac{\langle D_\chi^2 G(\chi_0, \alpha_0)[\phi, \phi], \psi^* \rangle}{\langle D_\alpha G(\chi_0, \alpha_0)[1], \psi^* \rangle} \neq 0, \tag{3.4}$$

the singularity in this theorem is called a *quadratic fold point*. Figure 1 illustrates the case $\frac{d^2\alpha(0)}{d\epsilon^2} > 0$. Likewise, if $\frac{d^2\alpha(0)}{d\epsilon^2} = 0$ but $\frac{d^3\alpha(0)}{d\epsilon^3} \neq 0$, then the singularity is called a hysteresis point, which is illustrated in Figure 2.

The second case is that in which $D_{\tilde{\chi}} G$ *is a Fredholm operator of index one and* $\mathcal{Z}_1 \equiv \mathcal{R}(D_{\tilde{\chi}} G)$ *has codimension one in* $\mathcal{Z}$, so the dimension of $\tilde{\mathcal{X}}_1 \equiv \mathcal{N}(D_{\tilde{\chi}} G(\tilde{\chi}_0))$ is two and there exists a continuous linear functional $\psi^* \in \mathcal{N}(D_\chi G^*)$ so that $\mathcal{Z}_1 = \{z \in \mathcal{Z} | \langle z, \psi^* \rangle = 0\}$. The study of the local solutions of $G(\tilde{\chi}) = 0$ about $\tilde{\chi}_0$ is thereby reduced to the investigation of the single equation

$$g(\tilde{\chi}_1) \equiv \langle G(\tilde{\chi}_1 + u(\tilde{\chi}_1)), \psi^* \rangle = 0 \text{ for } \tilde{\chi}_1 \in \tilde{\mathcal{X}}_1,$$

where $\tilde{\mathcal{X}} = \tilde{\mathcal{X}}_1 \oplus \tilde{\mathcal{X}}_2$ and $u(\tilde{\chi}_1) \in C^p((I - P)U, \tilde{\mathcal{X}}_2)$ for some open neighborhood $U \subset \tilde{\mathcal{X}}$ of $\tilde{\chi}_0$. Here $P : \tilde{\mathcal{X}} \to \tilde{\mathcal{X}}_1$ is a continuous projection. This situation breaks into three cases:

(i) $g(\tilde{\chi}_{01}) = 0$ and $D_{\tilde{\chi}_1} g(\tilde{\chi}_{01}) \neq 0$;

(ii) $g(\tilde{\chi}_{01}) = 0$, $D_{\tilde{\chi}_1} g(\tilde{\chi}_{01}) = 0$, and $D_{\tilde{\chi}_1}^2 g(\tilde{\chi}_{01})$ is nonsingular;

(iii) $g(\tilde{\chi}_{01}) = 0$, $D_{\tilde{\chi}_1} g(\tilde{\chi}_{01}) = 0$, and $D_{\tilde{\chi}_1}^2 g(\tilde{\chi}_{01})$ is singular, including the case when it vanishes identically.

For case (i) we obtain a C^p curve locally about $\tilde{\chi}_0$. We do not consider case (iii), which leads to higher-order singularities, since such singularities can be treated efficiently via singularity theory [17, 18]. For case (ii) we use the following theorem.

THEOREM 3.3 (see [40, 45]). *Let $\tilde{\mathcal{X}}$ and $\mathcal{Z}$ be Banach spaces, let $U_0 \subset \tilde{\mathcal{X}}$ be an open neighborhood of a point $\tilde{\chi}_0$ that solves $G(\tilde{\chi}) = 0$, and suppose that*

(a) *$G \in C^p(U_0, \mathcal{Z})$ for $p \geq 2$;*

(b) *$\mathcal{R}(D_{\tilde{\chi}} G(\tilde{\chi}_0)) = \mathcal{Z}_1$ has codimension one so that there exists a bounded linear functional $\psi^* \in \mathcal{Z}^*$ such that $\mathcal{Z}_1 = \{z \in \mathcal{Z} : \langle z, \psi^* \rangle = 0\}$;*

(c) *$\mathcal{N}(D_{\tilde{\chi}} G(\tilde{\chi}_0) = \tilde{\mathcal{X}}_1$ has dimension two (2), and $\tilde{\mathcal{X}}_2$ is a closed complementing subspace in $\tilde{\mathcal{X}}$.*

Then, if the 2×2 symmetric matrix associated with the quadratic form $D^2_{\tilde{\chi}_1} g(\tilde{\chi}_{01})[v, v]$, where $v \in \mathcal{N}(D_{\tilde{\chi}} G(\tilde{\chi}_0))$, is indefinite, there is an open neighborhood U of $\tilde{\chi}_0$ such that the solution set of $G(\tilde{\chi}) = 0$ consists of two C^{p-1} curves crossing transversally at the $\tilde{\chi}_0$. In a sufficiently small deleted neighborhood of $\tilde{\chi}_0$, these two curves are C^p. If the 2×2 symmetric matrix associated with $D^2_{\tilde{\chi}_1} g(\tilde{\chi}_{01})[v, v]$ is definite, then $\tilde{\chi}_0$ is the only local solution of $G(\tilde{\chi}) = 0$.

We note in passing that this statement of the theorem requires only that $p \geq 2$, whereas that stated in Nirenberg [40] requires $p \geq 3$. The improved version is due to Kuiper [33]. The application of this theorem to the situation in which $\tilde{\mathcal{X}} = \mathcal{X} \times \mathbf{R}$ yields the following theorem.

THEOREM 3.4. *Let $\mathcal{X}$ and $\mathcal{Z}$ be Banach spaces, $\mathcal{B}_1$ and $\mathcal{B}_2$ be open neighborhoods of $\chi_0 \in \mathcal{X}$ and $\alpha_0 \in \mathbf{R}$, respectively, and assume that $G(\chi_0, \alpha_0) = 0$, $G \in \mathcal{C}^p(\mathcal{B}_1 \times \mathcal{B}_2; \mathcal{Z})$ for some fixed $p \geq 2$. Assume that the Fréchet derivative $D_{(\chi,\alpha)} G(\chi_0, \alpha_0)$ is a Fredholm operator of index one with a two-dimensional null space, the dimension of $\mathcal{N}(D_\chi G(\chi_0, \alpha_0))$ is one, and $D_\alpha G(\chi_0, \alpha_0) \in \mathcal{R}(D_\chi G(\chi_0, \alpha_0))$. Let $\phi \in \mathcal{X}$, $\phi^* \in \mathcal{X}^*$, and $\psi^* \in \mathcal{Z}^*$ span $\mathcal{N}(D_\chi G(\chi_0, \alpha_0))$, the one-dimensional space complement of $\mathcal{R}(D_\chi G(\chi_0, \alpha_0)^*)$, and $\mathcal{N}(D_\chi G(\chi_0, \alpha_0)^*)$, respectively, and define*

$$
\begin{aligned}
a &= \langle D^2_\alpha G[1, 1] + 2 D_\alpha D_\chi G[W, 1] + D^2_\chi G[W, W], \psi^* \rangle, \\
b &= \langle D^2_\chi G[\phi, W] + D_\alpha D_\chi G[\phi, 1], \psi^* \rangle, \\
c &= \langle D^2_\chi G[\phi, \phi], \psi^* \rangle, \\
\mathcal{D} &= b^2 - ac,
\end{aligned} \tag{3.5}
$$

where derivatives of G are evaluated at (χ_0, α_0), W is the unique solution of $PW = 0$ and $D_\chi G(\chi_0, \alpha_0) W = -D_\alpha G(\chi_0, \alpha_0)$, and the projection P is defined by $Pu = \frac{\langle u, \phi^* \rangle}{\langle \phi, \phi^* \rangle} \phi$. If $\mathcal{D} > 0$, then there exist open neighborhoods $U \subset \mathcal{B}_1 \times \mathcal{B}_2$ of (χ_0, α_0) and $I \subset \mathbf{R}$ of 0, and two distinct functions $(\chi^\pm, \alpha^\pm) \in \mathcal{C}^{p-1}(I; U)$ such that $(\chi^\pm(0), \alpha^\pm(0)) = (\chi_0, \alpha_0)$ and $G(\chi^\pm(\epsilon), \alpha^\pm(\epsilon)) = 0$ for all $\epsilon \in I$. These two solution manifolds represent the totality of solutions of $G = 0$ in U. Also, if G is an analytic function of (χ, α) in a neighborhood of (χ_0, α_0), then $(\chi^\pm(\epsilon), \alpha^\pm(\epsilon))$ are also analytic on I.

Parameterizations of these manifolds are given as follows:

(i) *If $c \neq 0$ and $\mathcal{D} > 0$, then $\alpha = \alpha^\pm(\epsilon) \equiv \alpha_0 + \epsilon$ so that the two curves can be parameterized by the natural parameter α. These two curves have the structure $\chi = \chi^\pm(\alpha) \equiv \chi_0 + \gamma^\pm(\alpha)\phi + w^\pm(\alpha)$, where $\gamma^\pm(0) = 0$, $\frac{d\gamma^\pm(\alpha_0)}{d\alpha} = \frac{-b \pm \sqrt{\mathcal{D}}}{c}$, $Pw^\pm(\alpha) = 0$, $w^\pm(\alpha_0) = 0$, and $\frac{dw^\pm(\alpha_0)}{d\alpha} \equiv W$ is the unique solution of $PW = 0$ and $D_\chi G(\chi_0, \alpha_0) W = -D_\alpha G(\chi_0, \alpha_0)$.*

(ii) *If $c = 0$ and $\mathcal{D} > 0$, then the two solutions are parameterized as follows:*

(a) *$\alpha = \alpha^-(\epsilon) \equiv \alpha_0 + \epsilon$ so that the natural parameter α can be used and $\chi = \chi^-(\alpha) \equiv \chi_0 + \gamma^-(\alpha)\phi + w^-(\alpha)$, where $\gamma^-(\alpha_0) = 0$, $\frac{d\gamma^-(\alpha_0)}{d\alpha} = \frac{-a}{2b}$, $Pw(\alpha) = 0$, $w(\alpha_0) = 0$, and $\frac{dw^-(\alpha_0)}{d\alpha} \equiv W$ is the unique solution of $PW = 0$ and $D_\chi G(\chi_0, \alpha_0) W = -D_\alpha G(\chi_0, \alpha_0)$;*

(b) *$\alpha = \alpha^+(\epsilon)$ and $\chi = \chi^+(\epsilon) \equiv \chi_0 + \epsilon\phi + w^+(\epsilon)$, where $\alpha^+(0) = \alpha_0$, $\frac{d\alpha^+(0)}{d\epsilon} = 0$, $Pw^+(\epsilon) \equiv 0$, $w^+(0) = 0$, and $\frac{dw^+(0)}{d\epsilon} = 0$.*

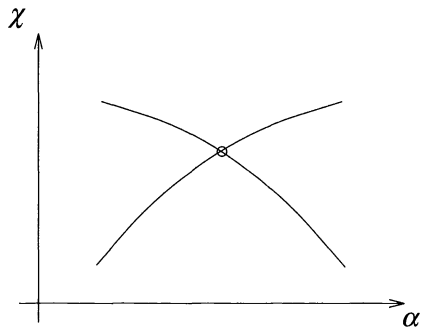


FIG. 3. *Simple bifurcation.*

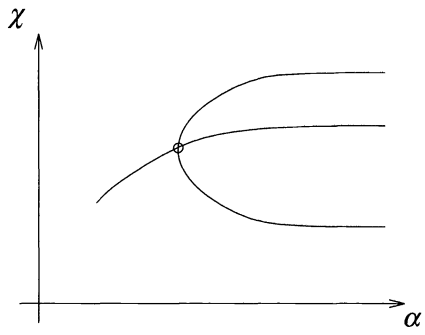


FIG. 4. *Pitchfork bifurcation.*

The simple bifurcation for case (i) in this theorem is illustrated in Figure 3, and case (ii), which is frequently called *pitchfork bifurcation*, in Figure 4. The bifurcation condition $\mathcal{D} > 0$ in this theorem is the same as the condition in Theorem 3.3 that the 2×2 symmetric matrix associated with the quadratic form $D^2_{\tilde{\chi}_1} g(\tilde{\chi}_{01})[v, v]$, where $v \in \mathcal{N}(D_{\tilde{\chi}} G(\tilde{\chi}_0))$, is indefinite.

The quantities a, b, and c in this theorem are significant in that they define a quadratic

$$a\alpha_1^2 + 2b\alpha_1\gamma_1 + c\gamma_1^2 = 0, \tag{3.6}$$

whose solutions indirectly determine the tangents to the two paths at the singularity (χ_0, α_0). For case (i) in which $\mathcal{D} > 0$ and $c \neq 0$, tangents bifurcating curves are

$$T = \begin{pmatrix} W \\ 1 \end{pmatrix} + \frac{-b \pm \sqrt{\mathcal{D}}}{c} \begin{pmatrix} \phi \\ 0 \end{pmatrix}, \tag{3.7}$$

where W is the unique solution of $PW = 0$ and $D_\chi G(\chi_0, \alpha_0)W = -D_\alpha G(\chi_0, \alpha_0)$.

For case (ii) in which $c = 0$ the two tangents for subcases (a) and (b) are, respectively,

$$T = \begin{pmatrix} W \\ 1 \end{pmatrix} - \frac{a}{2b} \begin{pmatrix} \phi \\ 0 \end{pmatrix} \text{ and } T = \begin{pmatrix} \phi \\ 0 \end{pmatrix}, \tag{3.8}$$

where W is as defined in the previous paragraph.

4. Bifurcations for loss of strict complementarity. In this section we relax the strict complementarity condition in Theorem 2.2 but maintain the remaining assumptions. Specifically, we assume that

(a′) strict complementarity is violated by one inequality constraint—say, $\mathcal{A} - \mathcal{A}_s = \{k\}$—so that $f_k(\hat{x}, \hat{\alpha}) = 0$, $\hat{\lambda}_k = 0$, and $l = k - 1$;

(b) the bounded linear transformation $D_x^2 \mathcal{L}_{11} : X_1 \to X_1^*$ is bijective;

(c) The bounded linear transformation $\tilde{L} \equiv (D_x F(\hat{x}, \hat{\alpha}), D_x f_1(\hat{x}, \hat{\alpha}), \ldots, D_x f_l(\hat{x}, \hat{\alpha})) : X \to Y \times \mathbf{R}^l$ is surjective.

If (a′) is replaced by $\mathcal{A} - \mathcal{A}_s = \{l + 1, \ldots, k\}$ for any $l + 1 \leq k$, one can proceed along the lines developed in our earlier work on the finite-dimensional problem [41] to deduce multiple bifurcating branches under conditions similar to those developed there. Although we shall content ourselves with the case (a′), we do note that if (a′) and (b) hold and if in (c) l is replaced by k, Robinson [46] established (in the finite-dimensional case) the existence of a locally unique minimum of (1.1) under a coercivity assumption on $D_x^2 \mathcal{L}_{11}$. Similar results have been obtained by Ito and Kunisch [24] for the Hilbert space setting. The bifurcation analysis below shows that the condition for bifurcation in Theorem 3.4 implies surjectivity of $D_x f_k(\hat{x}, \hat{\alpha})$; however, higher-order bifurcations can occur when there is a loss of this surjectivity.

4.1. Branching for loss of strict complementarity. In the current case the bifurcation Theorem 3.4 applies but Theorem 3.2 does not. As the following theorem shows, the conclusions can be characterized nicely in terms of the problem for which the constraint $f_k \leq 0$ is removed.

THEOREM 4.1. *Let $f_i \in C^2(U \times V; \mathbf{R})$ for $i = 0, 1, \ldots, m$, $F \in C^2(U \times V; Y)$, U and V be open sets in the Banach space X and $\mathbf{R}$, respectively, and Y be a second Banach space containing the range of $F(x, \alpha)$. Let $\mathcal{X} \equiv X \times Y^* \times \mathbf{R}^m \times \mathbf{R}$, and suppose $(\hat{\chi}; \hat{\alpha}) = (\hat{x}, \hat{y}^*, \hat{\lambda}, \hat{\lambda}_0; \hat{\alpha}) \in \mathcal{X} \times \mathbf{R}$ is a solution of $G(\chi; \alpha) = 0$. Suppose that the Fréchet derivative $D_x F(\hat{x}, \hat{\alpha})$ has closed range, its null space is complemented in X, and the aforementioned conditions* (a′), (b), *and* (c) *are valid.*

Then $D_\chi G(\hat{\chi}, \hat{\alpha})$ is a Fredholm operator of index zero with a one-dimensional null space and $D_\alpha G(\hat{\chi}, \hat{\alpha}) \in \mathcal{R}(D_\chi G(\hat{\chi}, \hat{\alpha}))$. If the constraint $f_k \leq 0$ is removed from (1.1), *then conditions* (a)–(c) *in Theorem* 2.2 *are satisfied for the resulting problem. Let $(\bar{x}(\alpha), \alpha)$ be the locally unique solution of* (1.1) *with the constraint $f_k \leq 0$ removed. Then there is a bifurcation provided that $\frac{d}{d\alpha} f_k(\bar{x}(\alpha), \alpha) \neq 0$ at $(\hat{x}, \hat{\alpha})$; i.e., the path $(\bar{x}(\alpha), \alpha)$ is transversal to $f_k(x, \alpha)$ at $(\hat{x}, \hat{\alpha})$.*

Several observations are in order before proceeding to a discussion of the proof. First, if assumption (a) in Theorem 2.2 is relaxed to $\mathcal{A} - \mathcal{A}_s = \{l + 1, \ldots, k\}$ for any $l + 1 \leq k$, assumption (c) implies that the dimension of the null space of the operator $\tilde{L}^*$ is zero so that $\lambda_0 \neq 0$. (Otherwise, all multipliers must be zero.) Also in this case, $D_\alpha G(\hat{\chi}, \hat{\alpha}) \in \mathcal{R}(D_\chi G(\hat{\chi}, \hat{\alpha}))$ so that in the special case of assumption (a′) above, Theorem 3.4 but not Theorem 3.2 is applicable when $\dim \mathcal{N}(D_\chi G(\hat{\chi}, \hat{\alpha})) = 1$. Next, the bifurcation that arises in this "loss of strict complementarity" case is due to the loss surjectivity in the nonlinear system of equations. The relevant part is $\lambda_k f_k(x, \alpha) = 0$. Differentiation of the expression $\lambda_k f_k(x, \alpha)$ shows that it cannot be surjective when both $\lambda_k = 0$ and $f_k(x, \alpha) = 0$; i.e., there is a loss of strict complementarity. Furthermore, since all parametric perturbations occur through the function $f_k(x, \alpha)$, the form of the equation (one of the dependent variables times the constraint) rules out the simple quadratic fold point. Although this is true for the parametric programming problem, the quadratic fold point does arise in numerical methods such as augmented Lagrangian, penalty, and interior point methods because the entire term $\lambda_k f_k(x, \alpha)$ is perturbed (e.g., by the penalty parameter).

We now proceed to the algebraic expressions in Theorem 3.4 for the current problem. Given the assumptions (a′), (b), and (c), $D_\chi G(\hat{\chi}, \hat{\alpha})$ is a Fredholm operator of index zero with a one-dimensional null space. Vectors $\phi \in X \times Y^* \times \mathbf{R}^{k-1} \times \mathbf{R} \times \mathbf{R}^{m-k} \times \mathbf{R}$ and $\psi^* \in X^{**} \times Y^* \times \mathbf{R}^{k-1} \times \mathbf{R} \times \mathbf{R}^{m-k} \times \mathbf{R}$ that span the null spaces of $D_\chi G(\hat{\chi}, \hat{\alpha})$ and $D_\chi G(\hat{\chi}, \hat{\alpha})^*$, respectively, are given by

$$\phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \\ \phi_6 \end{bmatrix} = \begin{bmatrix} \Delta\bar{x}_1 \\ \nu\hat{y}^* + y_p^* \\ \nu\hat{\lambda}_A^* + \lambda_{Ap}^* \\ 1 \\ 0 \\ \nu\lambda_0 \end{bmatrix} \text{ and } \psi^* = \begin{bmatrix} \psi_1^* \\ \psi_2^* \\ \psi_3^* \\ \psi_4^* \\ \psi_5^* \\ \psi_6^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix},$$

where $\Delta\bar{x}_1 \in X_1$ solves $D_x^2 \mathcal{L}_{11} \Delta\bar{x}_1 + P_1^* D_x f_k = 0$, $(\hat{y}^*, \hat{\lambda}_A) \in Y^* \times \mathbf{R}^{k-1}$ solves $L^* y^* + D_x f_A^* \lambda_A + \lambda_0 D_x f_0 = 0$, $(y_p^*, \lambda_{Ap}) \in Y^* \times \mathbf{R}^{k-1}$ solves $L^* y^* + D_x f_A^* \lambda_A + D_x^2 \mathcal{L}_{21} \Delta\bar{x}_1 + P_2^* D_x f_k = 0$, and $\nu = -\langle y_0, \hat{y}^* \rangle \langle y_0, y_p^* \rangle - \langle \hat{\lambda}_A, \lambda_{Ap} \rangle$.

To define the projection used in Theorem 3.4, i.e., $P\chi = \frac{\langle \chi, \phi^* \rangle}{\langle \phi, \phi^* \rangle} \phi$, one needs the element ϕ^*. Now $D_\chi G^* : X^{**} \times Y^* \times \mathbf{R}^m \times \mathbf{R} \to X^* \times Y^{**} \times \mathbf{R}^m \times \mathbf{R} = \mathcal{R}(D_\chi G^*) \oplus [\phi^*]$ and

$D_\chi G : X \times Y^* \times \mathbf{R}^m \times \mathbf{R} \to X^* \times Y \times \mathbf{R}^m \times \mathbf{R} = \mathcal{R}(D_\chi G) \oplus [\psi]$. Thus the required ϕ^* and ψ are given by

$$\phi^* = \begin{bmatrix} \phi_1^* \\ \phi_2^* \\ \phi_3^* \\ \phi_4^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ e_k \\ 0 \end{bmatrix} \text{ and } \psi = \begin{bmatrix} 0 \\ 0 \\ e_k \\ 0 \end{bmatrix},$$

where $e_k \in \mathbf{R}^m$ is the standard basis vector with a one in the kth position and zeros elsewhere.

To evaluate the algebraic expressions for a, b, c, and $\mathcal{D}$ in Theorem 3.4 one needs, in addition to the above expressions, the solution $\frac{dw(\alpha_0)}{d\alpha} \equiv W = (W_1, W_2, W_3, W_4, W_5, W_6) \in X \times Y^* \times \mathbf{R}^{k-1} \times \mathbf{R} \times \mathbf{R}^{m-k} \times \mathbf{R}$ to the problem $D_\chi G W = -D_\alpha G$ and $PW = 0$, where $P\chi \equiv \frac{\langle \chi, \phi^* \rangle}{\langle \phi, \phi^* \rangle}\phi$. The requirement $PW = 0$ forces $W_4 = 0$. To explain the remaining parts of W, note that if the constraint $f_k \leq 0$ and the corresponding multiplier λ_k are deleted from the problem (1.1), then due to the assumptions (a′), (b), and (c) the resulting system is nonsingular at the corresponding critical point $(\hat{x}, \hat{\alpha})$ and $\bar{W} = (W_1, W_2, W_3, W_5, W_6) \in X \times Y^* \times \mathbf{R}^{k-1} \times \mathbf{R}^{m-k} \times \mathbf{R}$ is the solution of the modified system $D_{\bar{\chi}} \bar{G} \frac{d\bar{\chi}}{d\alpha} = -D_\alpha \bar{G}$, where $\bar{\chi}$ is χ with λ_k and the constraint f_k removed. In particular, $W_1 \equiv \frac{d\bar{x}}{d\alpha}$.

The substitution of these into the algebraic expressions for a, b, c, and $\mathcal{D} > 0$ yields $a = 0$, $b = D_\alpha f_k + D_x f_k W_1$, $c = D_x f_k \phi_1$, $\mathcal{D} = b^2$. The bifurcation condition $\mathcal{D} > 0$ in Theorem 3.4 is that $b = \frac{d}{d\alpha} f_k(\bar{x}(\alpha), \alpha) \neq 0$ at $(\hat{x}, \hat{\alpha})$, which says the path $(\bar{x}(\alpha), \alpha)$ is transversal to $f_k(x, \alpha)$ at $(\hat{x}, \hat{\alpha})$.

Returning to Theorem 3.4, we consider the first of two cases presented in that theorem. First, if $c \neq 0$, the two curves can be parameterized by the natural parameter α and have the structure and $x = x^{\pm}(\alpha) \equiv x_0 + \gamma^{\pm}(\alpha)\phi + w^{\pm}(\alpha)$, where $\gamma^{\pm}(\alpha_0) = 0$, and $\frac{d\gamma^{\pm}(\alpha_0)}{d\alpha} = \frac{-b \pm |b|}{c}$, $Pw^{\pm}(\alpha) = 0$, $w^{\pm}(\alpha_0) = 0$, and $\frac{dw^{\pm}(\alpha_0)}{d\alpha} \equiv W$ as defined above. From the definition of the projection P, $\lambda_k = \gamma^{\pm}(\alpha)$. Now one of $\frac{d\gamma^{\pm}(\alpha_0)}{d\alpha} = \frac{-b \pm |b|}{c}$ is zero and one is not. That which is zero corresponds to the aforementioned path $(\bar{x}(\alpha), \alpha)$ about $\alpha = \alpha_0$ in which $\lambda_k \equiv 0$. This path is feasible for α on one side of α_0 and is infeasible on the other side. The second path has $\lambda_k \neq 0$, so the constraint $f_k(x, \alpha) \leq 0$ must be active along this path locally about α_0 since $\lambda_k f_k(x, \alpha) = 0$.

For the second case, if $c = 0$ and $\mathcal{D} > 0$, then the two solutions are parameterized as follows: the first path is parameterized by the natural parameter α as $x = x^-(\alpha) \equiv x_0 + \gamma^-(\alpha)\phi + w^-(\alpha)$, where $\lambda_k = \gamma^-(\alpha)$, $\gamma^-(\alpha_0) = 0$, $\frac{d\gamma^-(\alpha_0)}{d\alpha} = \frac{-a}{2b} = 0$, $Pw^-(\alpha) \equiv 0$, $w^-(\alpha_0) = 0$, and $\frac{dw^-(\alpha_0)}{d\alpha} = W$, where W is as defined above. It is this solution that is transversal to the constraint $f_k = 0$ with $\lambda_k = \gamma^-(\alpha) \equiv 0$. The second path has the parameterization $\alpha = \alpha^+(\epsilon)$ and $x = x^+(\epsilon) \equiv x_0 + \epsilon\phi + w^+(\epsilon)$, where $\alpha^+(0) = \alpha_0$, $\frac{d\alpha^+(0)}{d\epsilon} = 0$, $Pw^+(\epsilon) \equiv 0$, $w^+(0) = 0$, and $\frac{dw^+(0)}{d\alpha} = 0$. In fact, from the definition of the projection P, we conclude that $\epsilon \equiv \lambda_k$ is the parameterization parameter and the constraint $f_k \leq 0$ is active.

In summary, the bifurcation diagram is as in Figure 3. One of the solution branches corresponds to the constraint $f_k \leq 0$ being active, i.e., $f_k = 0$, and the other solution branch crosses this constraint transversally as the parameter α crosses α_0, going from the interior feasible region of the constraint $f_k \leq 0$ to the exterior or vice versa. For a stability result, we have a result similar to the finite-dimensional case.

THEOREM 4.2. *Let $f_i \in C^2(U \times V; \mathbf{R})$ for $i = 0, 1, \ldots, m$, $F \in C^2(U \times V; Y)$, U and V be open sets in the Banach space X and $\mathbf{R}$, respectively, and Y be a second Banach space containing the range of $F(x, \alpha)$. Let $\mathcal{X} \equiv X \times Y^* \times \mathbf{R}^m \times \mathbf{R}$, and suppose that*

$(\hat{\chi};\hat{\alpha}) = (\hat{x}, \hat{y}^*, \hat{\lambda}, \hat{\lambda}_0; \hat{\alpha}) \in \mathcal{X} \times \mathbf{R}^m$ is a solution of $G(\chi;\alpha) = 0$. Suppose the Fréchet derivative $D_x F(\hat{x},\hat{\alpha})$ has closed range, its null space is complemented in X, and the following are valid.

(a′) *Strict complementarity is violated by one inequality constraint—say, $\mathcal{A} - \mathcal{A}_s = \{k\}$—so that $f_k(\hat{x},\hat{\alpha}) = 0$, $\hat{\lambda}_k = 0$, and $l = k - 1$.*

(b′) *$D_x^2\mathcal{L}(\hat{\chi},\hat{\alpha})[h,h] \geq C\|h\|_X^2$ for all $h \in \mathcal{N}(\tilde{L})$.*

(c) *The bounded linear transformation $\tilde{L} \equiv (D_x F(\hat{x},\hat{\alpha}), D_x f_1(\hat{x},\hat{\alpha}), \ldots, D_x f_l(\hat{x},\hat{\alpha})) : X \to Y \times \mathbf{R}^l$ is surjective.*

(d) *The multipliers $\lambda_i > 0$ for $i = 1, \ldots, l$ at $(x,\alpha) = (\hat{x},\hat{\alpha})$.*

On that branch emanating from $(\hat{x},\hat{\alpha})$ that is interior to the constraint $f_k \leq 0$, there exists an interval $[\alpha_0,\hat{\alpha}]$ on which $D_x^2\mathcal{L}(\hat{\chi},\hat{\alpha})[h,h] \geq \bar{C}\|h\|_X^2$ for all $h \in \mathcal{N}(\tilde{L}(\alpha))$. On that branch emanating from $(\hat{x},\hat{\alpha})$, where the constraint $f_k \leq 0$ is active and $\lambda_k > 0$, a similar result holds. Thus and in particular, there is a local persistence of the unique minimizer along these two solution branches.

4.2. An isoperimetric example. A simple example that illustrates the foregoing bifurcation phenomena is

$$\begin{aligned} &\text{minimize} && J(x) = \int_0^1 \left(\frac{dx}{dt}\right)^2 dt \\ &\text{subject to} && x(0) = 0, x(1) = \alpha, \\ &&& \int_0^1 x\,dt \leq l, \qquad x \in C^1([0,1],\mathbf{R}). \end{aligned}$$

For this problem the constraints are affine, so we use the Lagrangian

$$\mathcal{L} = \int_0^1 \left(\frac{dx}{dt}\right)^2 dt + \lambda\left(\int_0^1 x\,dx - l\right) + \mu_1 x(0) + \mu_2\big(x(1) - \alpha\big).$$

If the constraint $\int_0^1 x\,dx \leq l$ is not active, then the solution is given by $\hat{x}(t) = \alpha t$. Furthermore,

$$J(\hat{x}+h) = J(\hat{x}) + J(h) \geq J(x) + \frac{1}{2}\|h\|_{W^{1,2}}^2$$

for all $h \in C^1([0,1],\mathbf{R})$ such that $h(0) = h(1) = 0$. Here, $W^{m,p}[a,b] = \{v \in C^{m-1}[a,b] : v^{(m-1)}$ is absolutely continuous on $[a,b]$, $v^{(m)} \in L^p[a,b]\}$ for $1 \leq p < \infty$ and $\|v\|_{W^{m,p}} = \big(\sum_{k=0}^m \|v^{(k)}\|_{L^p}^p\big)^{1/p}$. (This technique of using two different norms in establishing a minimizer was mentioned at the end of §2.3 and is presented in the work of Ioffe [21] and Maurer [39].) Thus, this solution is a global minimizer as long as the inequality constraint is satisfied and is not active, i.e., $\int_0^1 \hat{x}(t)\,dt = \frac{\alpha}{2} < l$. At $\alpha = 2l$, the inequality constraint becomes active, and as α increases beyond $2l$, the inequality constraint is violated. As discussed in the previous subsection, this path represents one branch in the bifurcation diagram in Figure 3.

The second solution branch emanates from $\alpha = 2l$ and satisfies the active inequality constraint, i.e., $\int_0^1 x\,dt = l$. The solution is $\hat{x}(t) = (\frac{\lambda}{4})t^2 + (\alpha - \frac{\lambda}{4})t$, where α is related to the Lagrange multiplier λ by $\lambda = 12(\alpha - 2l)$, which is zero when $\alpha = 2l$. This solution exists for both α less than and greater than $2l$; however, only for $\alpha > 2l$ is the multiplier positive. In this case, the same inequality $J(\hat{x}+h) = J(\hat{x}) + J(h) \geq J(x) + \frac{1}{2}\|h\|_{W^{1,2}}^2$ holds for all $h \in C^1([0,1],\mathbf{R})$, $h(0) = h(1) = 0$, and $\int_0^1 h(t)\,dt = 0$. Thus for $\alpha > 2l$, which implies that $\lambda > 0$, we have a global minimizer with the inequality constraint being active. The bifurcation diagram is as in Figure 3.

5. Bifurcations for loss of surjectivity. In this section we investigate the problem (2.4) under the same assumptions as in Theorem 2.2, except that the surjectivity assumption is relaxed. In particular, we assume that

(a) strict complementarity holds, i.e., $\mathcal{A}_s = \mathcal{A}$ and $k = l$;

(b) the bounded linear transformation $D_x^2\mathcal{L}_{11} : X_1 \to X_1^*$ is bijective;

(c′) the range of $\tilde{L} = (D_xF(\hat{x}, \hat{\alpha}), D_xf_1(\hat{x}, \hat{\alpha}), \ldots, D_xf_k(\hat{x}, \hat{\alpha})) : X \to Y \times \mathbf{R}^k$ has codimension one in $Y \times \mathbf{R}^k$.

When we use these hypotheses, it is easily verified that the Fréchet derivative $D_\chi G(\hat{x}, \hat{y}^*, \hat{\lambda}, \lambda_0; \hat{\alpha})$ is a Fredholm operator of index zero with a one-dimensional null space. As in the appendix we partition $\lambda \in \mathbf{R}^m$ into $\lambda_A \in \mathbf{R}^k$ and $\lambda_C \in \mathbf{R}^{m-k}$ by $\lambda = \begin{bmatrix} \lambda_A \\ \lambda_C \end{bmatrix}$. ($\lambda_B$ is vacuous since $k = l$.) Likewise, D_xf is to be partitioned into D_xf_A and D_xf_C by

$$D_xf_A = \begin{bmatrix} D_xf_1 \\ \vdots \\ D_xf_k \end{bmatrix} \text{ and } D_xf_C = \begin{bmatrix} D_xf_{k+1} \\ \vdots \\ D_xf_m \end{bmatrix}.$$

The analysis of this situation breaks into two cases: (1) $D_xf_0(\hat{x}, \hat{\alpha}) \notin \mathcal{R}(\tilde{L}^*)$ and (2) $D_xf_0(\hat{x}, \hat{\alpha}) \in \mathcal{R}(\tilde{L}^*)$, as given in §§5.1 and 5.3, respectively. We refer to these two cases as the abnormal and normal cases, respectively. Rather than restating Theorems 3.2 and 3.4 as they relate to the current cases, we adopt an informal style.

5.1. Loss of surjectivity: The abnormal case. In this case, $D_xf_0(\hat{x}, \hat{\alpha}) \notin \mathcal{R}(\tilde{L}^*)$, so $\mathcal{N}(\tilde{L}^*) = \mathcal{R}(\tilde{L})^\perp$ and $\dim \mathcal{N}(\tilde{L}^*) = 1$ imply that the only choice of λ_0 in (2.4) is $\lambda_0 = 0$. Thus, there exists a nonzero multiplier $(y_1^*, \lambda_{A1}) \in \mathcal{N}(\tilde{L}^*)$, where $y_1^* \in Y^*$ and $\lambda_{A1} \in (\mathbf{R}^k)^*$, that spans $\mathcal{N}(\tilde{L}^*)$. We normalize these multipliers by first choosing a $y_0 \in Y$ such that $|\langle y_0, y_1^*\rangle| \geq \frac{1}{2}\|y_1^*\|$ and scaling y_1^* and λ_{A1} so that $|\langle y_0, y_1^*\rangle|^2 + \langle \lambda_{A1}, \lambda_{A1}\rangle = 1$. Then the system of nonlinear equations (2.4) has the following solution $(x, y^*, \lambda_A, \lambda_C, \lambda_0; \alpha) = (\hat{x}, \hat{y}^*, \hat{\lambda}_A, 0, 0; \hat{\alpha})$, where $\hat{y}^* = \tau y_1^*$, $\hat{\lambda}_A = \tau\lambda_{A1}$ and $\tau = \pm 1$.

From the expressions for $D_\chi G(\hat{x}, \hat{y}^*, \hat{\lambda}, \hat{\lambda}_0; \hat{\alpha})$ and its adjoint in the appendix, spanning vectors for the one-dimensional null spaces can be defined as follows. Under the assumed hypotheses, $\Delta x_2 = 0$, $\lambda_0 = 0$, $\hat{\lambda}_C = 0$, and Λ_A is invertible. Let $\Delta\bar{x}_1$ be the unique solution of $D_x^2\mathcal{L}_{11}\Delta\bar{x}_1 + P_1^*(D_xf_0) = 0$. The general solution $(\Delta y^*, \Delta\lambda)$ of $L^*\Delta y^* + D_xf_A^*\Delta\lambda_A = -\Delta\lambda_0 D_x^2\mathcal{L}_{21}\bar{\Delta}x_1 - \Delta\lambda_0 P_2^*(D_xf_0)$ is

$$\begin{bmatrix} \Delta y^* \\ \Delta\lambda_A \end{bmatrix} = \Delta\lambda_0\nu \begin{bmatrix} \hat{y}^* \\ \hat{\lambda}_A \end{bmatrix} + \Delta\lambda_0 \begin{bmatrix} y_p^* \\ \lambda_{Ap} \end{bmatrix},$$

where $\begin{bmatrix} y_p^* \\ \lambda_{Ap} \end{bmatrix}$ is a particular solution with $\Delta\lambda_0 = 1$ and is chosen to be in the space complement of $\mathcal{N}(\tilde{L}^*)$. Also, $\nu = -\langle y_0, \hat{y}^*\rangle\langle y_0, y_p^*\rangle - \langle\hat{\lambda}_A, \hat{\lambda}_{Ap}\rangle$. Thus $\phi \in X \times Y^* \times \mathbf{R}^k \times \mathbf{R}^{m-k} \times \mathbf{R}$ and $\psi^* \in X^{**} \times Y^* \times \mathbf{R}^{k*} \times \mathbf{R}^{(m-k)*} \times \mathbf{R}^*$, given by

$$\phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \end{bmatrix} = \begin{bmatrix} \Delta\bar{x}_1 \\ \nu\hat{y}^* + y_p^* \\ \nu\hat{\lambda}_A + \lambda_{Ap} \\ 0 \\ 1 \end{bmatrix} \text{ and } \psi^* = \begin{bmatrix} \psi_1^* \\ \psi_2^* \\ \psi_3^* \\ \psi_4^* \\ \psi_5^* \end{bmatrix} = \begin{bmatrix} 0 \\ \hat{y}^* \\ \Lambda_A^{*-1}\hat{\lambda}_A \\ 0 \\ 0 \end{bmatrix},$$

span the null spaces of $D_\chi G(\hat{\chi}, \hat{\alpha})$ and $D_\chi G(\hat{\chi}, \hat{\alpha})^*$, respectively.

For evaluation of the expressions in Theorems 3.2 and 3.4, note that $\langle D_\chi^2 G(\hat\chi,\hat\alpha)[u,v],\psi^*\rangle = D_x^2\mathcal{L}(\hat\chi,\hat\alpha)[u_1,v_1]$, $\langle D_\alpha G(\hat x,\hat\alpha)[1],\psi^*\rangle = D_\alpha\mathcal{L}(\hat\chi,\hat\alpha)$, and $D_\alpha D_\chi G(\hat\chi,\hat\alpha)[u],\psi^*\rangle = D_\alpha D_x\mathcal{L}(\hat\chi,\hat\alpha)[u_1]$, where u and $v \in \mathcal{X} = X \times Y^* \times \mathbf{R}^m \times \mathbf{R}$ and u_1 represents the first component of u, i.e., the projection of u onto X.

Theorem 3.2 is applicable when $D_\alpha G \notin \mathcal{R}(D_\chi G)$ or equivalently $D_\alpha\mathcal{L} \neq 0$. Note in particular that $\lambda_0 \equiv \epsilon$ crosses zero with ϵ in Theorem 3.2. If $D_x^2\mathcal{L}(\hat\chi,\hat\alpha)[\phi_1,\phi_1] \neq 0$, then

$$\frac{d^2\alpha(0)}{d\epsilon^2} \equiv -\frac{\langle D_\chi^2 G(\hat\chi,\hat\alpha)[\phi,\phi],\psi^*\rangle}{\langle D_\alpha G(\hat\chi,\hat\alpha)[1],\psi^*\rangle} = -\frac{D_x^2\mathcal{L}(\hat\chi,\hat\alpha)[\phi_1,\phi_1]}{D_\alpha\mathcal{L}(\hat\chi,\hat\alpha)[1]} \neq 0.$$

(Note that the signs of $D_x^2\mathcal{L}(\hat\chi,\hat\alpha)[\phi_1,\phi_1]$ and $D_\alpha\mathcal{L}$ change with τ, but the sign of $\frac{d^2\alpha(0)}{d\epsilon^2}$ does not.)

Theorem 3.4 is applicable when $D_\alpha G \in \mathcal{R}(D_\chi G)$ or equivalently $D_\alpha\mathcal{L}(\hat\chi,\hat\alpha)[1] = 0$. In this case the algebraic expressions for a, b, and c in this theorem reduce to

$$\begin{aligned} a &= D_\alpha^2\mathcal{L}[1,1] + 2D_\alpha D_x\mathcal{L}[W_1,1] + D_x^2\mathcal{L}[W_1,W_1],\\ b &= D_x^2\mathcal{L}[\phi_1,W_1] + D_\alpha D_x\mathcal{L}[\phi_1,1],\\ c &= D_x^2\mathcal{L}[\phi_1,\phi_1] \end{aligned}$$

and depend on τ; however, the sign of $\mathcal{D} = b^2 - ac$, which is crucial for bifurcation, is invariant with respect to τ.

5.2. A fold point example. An example that illustrates the fold point phenomena for this case is Queen Dido's problem of maximizing the area under a curve of fixed length [1]:

$$\begin{aligned} \text{maximize} \quad & \int_{-a}^{a} y(x)\,dx\\ \text{subject to} \quad & \int_{-a}^{a} \sqrt{1+\left(\frac{dy}{dx}\right)^2}\,dx = l,\\ & y(-a) = 0, \qquad y(a) = 0,\\ & y \in C^1([-a,a],\mathbf{R}). \end{aligned}$$

The solution $y \equiv 0$ at $l = 2a$ is a quadratic fold point for the above problem and is an *abnormal case*. To see this, first note that at these values $D_y\big(\int_{-a}^{a}(1+\big(\frac{dy}{dx}\big)^2)^{1/2}\,dx - l\big)[h] = \int_{-a}^{a} 0\cdot h + 0\cdot\frac{dh}{dx}\,dx = 0$ for all $h \in C^1$, so the Fréchet derivative of this constraint is not surjective at $l = 2a$. Since $D_l\big(\int_{-a}^{a}(1+\big(\frac{dy}{dx}\big)^2)^{1/2}\,dx - l\big) = -1$, which is not in the range of the Fréchet derivative of this constraint with respect to y, we are in the abnormal case and the situation is much like that in Theorem 3.2 for the parameter $\alpha = \frac{2a}{l}$. The global solution structure is as follows.

For $\frac{2a}{l} > 1$, there is no feasible solution to the constraints and thus no solution to the problem. For $\frac{2a}{l} = 1$, the unique solution to the constraints (and thus to the problem) is $y \equiv 0$. (This is the fold point.)

For $\frac{2}{\pi} < \frac{2a}{l} < 1$, there are two solutions to the first-order necessary conditions and each is defined implicitly by the circle $x^2 + (y-\kappa)^2 = R^2$, where the center $(0,\kappa)$ and radius R are defined as follows. From the center $(0,\kappa)$ draw two straight lines, one to the origin (0,0) and one to $(a,0)$, and let θ denote the angle between these two lines at $(0,\kappa)$. Then, the relations $2R\theta = l$, $R\cos\theta = |\kappa|$, and $R\sin\theta = a$ yield the equation $\sin\theta = \big(\frac{2a}{l}\big)\theta$ which has a solution in $(0,\pi/2)$, say θ_0, for $\frac{2}{\pi} < \frac{2a}{l} < 1$. The corresponding radius and center are $R = \frac{l}{2\theta_0}$ and

$\kappa = -\frac{l\cos(\theta_0)}{2\theta_0}$, respectively. The solution y is given by $y(x;\kappa,R) = \kappa + \sqrt{R^2 - x^2}$; the second solution is a minimum and is given by the negative of this one. Note also that as $\alpha \equiv \frac{2a}{l} \to 1^-$ the solutions $y \to 0$ and as $\alpha \equiv \frac{2a}{l} \to \frac{2}{\pi}^+$, the solutions $y \to \pm\sqrt{a^2 - x^2}$, either of which has an infinite C^1 norm.

For $\frac{2a}{l} \leq \frac{2}{\pi}$, there are no C^1 solutions in nonparametric form for the above problem. (One can continue the solution in parametric form.) Finally, there is a complete exchange of the stability in the solution as λ_0 crosses the origin. The bifurcation diagram is qualitatively like that in Figure 1.

5.3. Loss of surjectivity: The normal case. In the normal case $D_x f_0(\hat{x},\hat{\alpha}) \in \mathcal{R}(\tilde{L}^*)$, so $\lambda_0 \neq 0$. We modify the normalization in problem (2.4) and replace it with $\lambda_0 = 1$; otherwise, the assumptions are as follows:

(a) strict complementarity holds, i.e., $\mathcal{A}_s = \mathcal{A}$ and $k = l$;

(b) $D_x^2\mathcal{L}_{11} : X_1 \to X_1^*$ is bijective;

(c) the range of $(D_x F(\hat{x},\hat{\alpha}), D_x f_1(\hat{x},\hat{\alpha}), \ldots, D_x f_l(\hat{x},\hat{\alpha})) : X \to Y \times \mathbf{R}^l$ has codimension one in $Y \times \mathbf{R}^l$;

(d′) $D_x f_0(\hat{x},\hat{\alpha}) \in \mathcal{R}(\tilde{L}^*)$.

The situation (d′) occurs for the case (a)–(c) when, for example, the active constraints are affine and consistent (i.e., have a feasible solution).

Since the codimension of the range of the operator $\tilde{L} = \begin{bmatrix} L \\ D_x f_A \end{bmatrix}$ is one, the dimension of the null space of $\tilde{L}^*$ is also one, so there exist nonzero multipliers $(y_1^*, \lambda_1^*) \in \mathcal{N}(\tilde{L}^*)$, where $y_1^* \in Y^*$ and $\lambda_1^* \in (\mathbf{R}^l)^*$ such that $L^* y_1^* + D_x f_A^* \lambda_1^* = 0$ and any solution of this problem is a constant multiple of (y_1^*, λ_1^*). Thus the solution of the (2.4) with the normalization N_3 replaced by N_1 (see equation (2.5)) can be written as

$$\begin{bmatrix} \hat{y}^* \\ \hat{\lambda}_A^* \end{bmatrix} = \gamma \begin{bmatrix} y_1^* \\ \lambda_1^* \end{bmatrix} + \lambda_0 \begin{bmatrix} y_p^* \\ \lambda_{Ap}^* \end{bmatrix}$$

where $\begin{bmatrix} y_p^* \\ \lambda_{Ap}^* \end{bmatrix}$ is a particular solution of (2.4) corresponding to $\lambda_0 = 1$ and is in the space complement of $\mathcal{N}(\tilde{L}^*)$. One can show that the null spaces of $D_\chi G(\hat{\chi},\hat{\alpha})$ and $D_\chi G(\hat{\chi},\hat{\alpha})^*$ are spanned by

$$\phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \end{bmatrix} = \begin{bmatrix} 0 \\ y_1^* \\ \lambda_1^* \\ 0 \\ 0 \end{bmatrix} \text{ and } \psi^* = \begin{bmatrix} \psi_1^* \\ \psi_2^* \\ \psi_3^* \\ \psi_4^* \\ \psi_5^* \end{bmatrix} = \begin{bmatrix} 0 \\ y_1^* \\ \Lambda_1^{*-1}\lambda_1^* \\ 0 \\ 0 \end{bmatrix},$$

respectively.

The bifurcation Theorem 3.2 is applicable when $D_\alpha G(\hat{x},\hat{\alpha}) \notin \mathcal{R}(D_x G(\hat{x},\hat{\alpha}))$, i.e., when $\langle D_\alpha F(\hat{x},\hat{\alpha}), y_1^*\rangle + \langle D_\alpha f_A(\hat{x},\hat{\alpha}), \lambda_1^*\rangle \neq 0$. In this case there is a locally unique solution according to Theorem 3.2 given by

$$\begin{bmatrix} \hat{y}^* \\ \hat{\lambda}_A^* \end{bmatrix} = \gamma \begin{bmatrix} y_1^* \\ \lambda_1^* \end{bmatrix} + \lambda_0 \begin{bmatrix} y_p^* \\ \lambda_{Ap}^* \end{bmatrix}$$

for each γ. Thus, this solution represents the entirety of solutions to the modified (2.4) at $(\hat{x},\hat{\alpha})$.

The bifurcation Theorem 3.4 is applicable when $D_\alpha G(\hat{x}, \hat{\alpha}) \in \mathcal{R}(D_x G(\hat{x}, \hat{\alpha}))$, i.e., when $\langle D_\alpha F(\hat{x}, \hat{\alpha}), y_1^* \rangle + \langle D_\alpha f_A(\hat{x}, \hat{\alpha}), \lambda_1^* \rangle = 0$. However, in this case the algebraic expressions for (a), (b), and (c) now take the values $b = 0$ and $c = 0$, so Theorem 3.4 gives no information about branching.

6. $D_x^2 \mathcal{L}_{11}$ is singular. Finally in this section we relax the condition in Theorem 2.2 that $D_x^2 \mathcal{L}_{11}$ is bijective, but we maintain the remaining assumptions. Specifically, we assume that

(a) strict complementarity holds, i.e., $\mathcal{A} = \mathcal{A}_s$ and $k = l$;

(b′) the bounded linear transformation $D_x^2 \mathcal{L}_{11} : X_1 \to X_1^*$ ($X_1 = \mathcal{N}(\tilde{L})$) is a Fredholm operator of index 0 with a one-dimensional null space;

(c) $\tilde{L} \equiv (D_x F(\hat{x}, \hat{\alpha}), D_x f_1(\hat{x}, \hat{\alpha}), \ldots, D_x f_l(\hat{x}, \hat{\alpha})) : X \to Y \times \mathbf{R}^l$ is surjective.

Before proceeding, it is worth noting that, under the assumptions of $\mathcal{N}(L)$ splitting the space X (Theorem 2.4) and (a), $\mathcal{N}(\tilde{L})$ also splits the space [1, lemma on the closed image]. Thus, there exists a closed linear subspace X_2 such that $X = X_1 \oplus X_2$ with $X_1 = \mathcal{N}(\tilde{L})$. Then by assumption (c), the implicit function Theorem 3.1 implies the existence of a locally unique C^p solution $x_2 = u(x_1, \alpha)$ of $\tilde{F}(x, \alpha) = 0$, where $\tilde{F} = (F, f_1(x, \alpha), \ldots, f_l(x, \alpha)) = 0$. Since $k = l$ by assumption (a), all active constraints in the problem can be removed and one can, for the purposes of local phenomena, consider the equivalent unconstrained optimization problem: minimize $\widetilde{f_0}(x_1, \alpha) \equiv f_0(x_1 + u(x_1, \alpha), \alpha)$. Such unconstrained problems are most efficiently treated by using catastrophe theory as in the book by Poston and Stewart [43], and the examples in this section are for this reduced unconstrained problem. However, we first present the theory within the current framework.

6.1. Branching analysis for a singular $D_x^2 \mathcal{L}_{11}$. Let $\hat{y}^*, \hat{\lambda}_A, \hat{\lambda}_C = 0$ and $\hat{\lambda}_0$ represent the solution of (2.4) at $(x, \alpha) = (\hat{x}, \hat{\alpha})$, and suppose that $\mathcal{N}(D_x^2 \mathcal{L}_{11})$ is spanned by ϕ_1. Finally, let y_p^* and λ_{Ap}^* be the unique solution of $L^* y_p^* + (D_x f_A)^* \lambda_{Ap}^* = -D_x^2 \mathcal{L}_{21} \phi_1$. Then the general solutions of $D_\chi G \phi = 0$ and $D_\chi G^* \psi^* = 0$ are spanned by

$$\phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \end{bmatrix} = \begin{bmatrix} \phi_1 \\ \nu \hat{y}^* + y_p^* \\ \nu \hat{\lambda}_A^* + \lambda_{Ap}^* \\ 0 \\ \nu \hat{\lambda}_0 \end{bmatrix} \quad \text{and} \quad \psi^* = \begin{bmatrix} \psi_1^* \\ \psi_2^* \\ \psi_3^* \\ \psi_4^* \\ \psi_5^* \end{bmatrix} = \begin{bmatrix} \psi_1^* \\ y_a^* \\ \Lambda_A^{*-1} \lambda_{Aa}^* \\ \lambda_{Ca}^* \\ 0 \end{bmatrix}, \tag{6.1}$$

respectively, where the parameter ν is determined from the normalization equation in (2.5c), i.e., $\nu = -\langle y_0, \hat{y}^* \rangle \langle y_0, y_p^* \rangle - \langle \hat{\lambda}_A^*, \lambda_{Ap}^* \rangle$, $\mathcal{N}(D_x^2 \mathcal{L}_{11}^*)$ is spanned by ψ_1^*, y_a^* and λ_{Aa}^* solve $L^* y_a^* + D_x f_A^* \lambda_{Aa}^* = -D_x^2 \mathcal{L}_{12}^* \psi_1^*$, and λ_{Ca}^* solves $\operatorname{diag}(f_{k+1}, \ldots, f_m) \lambda_{Ca}^* = -D_x f_C^{**} P_1^{**} \psi_1^*$. Since the algebraic expressions in Theorems 3.2 and 3.4 do not simplify any further except for the obvious substitutions, we omit a restatement of these theorems incorporating these expressions.

Finally, we remark that the assumptions (a) and (b) imply that $\lambda_0 \neq 0$. Thus, if the normalization $\sum_{i=0}^m \lambda_i^2 + \|y^*\|^2 = 1$ in (2.4) is replaced by $\lambda_0 = 1$, then the eigenvectors are determined as follows: let $y_p^*, \lambda_{Ap}^*, \lambda_{Cp}^* = 0$ solve $L^* y_p^* + D_x f_A^* \lambda_{Ap}^* = -D_x^2 \mathcal{L}_{12} \phi_1$. Then the null spaces of $D_\chi G$ and G_χ^* are spanned by

$$\phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{bmatrix} = \begin{bmatrix} \phi_1 \\ y_p^* \\ \lambda_{Ap}^* \\ 0 \end{bmatrix} \quad \text{and} \quad \psi^* = \begin{bmatrix} \psi_1^* \\ \psi_2^* \\ \psi_3^* \\ \psi_4^* \end{bmatrix} = \begin{bmatrix} \psi_1^* \\ y_a^* \\ \Lambda_A^{*-1} \lambda_{Aa}^* \\ \lambda_{Ca}^* \end{bmatrix}, \tag{6.2}$$

respectively, where $\mathcal{N}(D_x^2\mathcal{L}_{11}^*)$ is spanned by ψ_1^*, y_a^* and λ_{Aa}^* solve $L^*y_a^* + D_x f_A^*\lambda_{Aa}^* = -D_x^2\mathcal{L}_{12}^*\psi_1^*$, and λ_{Ca}^* solves $\operatorname{diag}(f_{k+1}, \ldots, f_m)\lambda_{Ca}^* = -D_x f_C^{**}\psi_1^*$.

6.2. Examples for loss of bijectivity. As remarked above, it suffices to consider the unconstrained problem of minimizing $f(x, \alpha)$, where $f(x, \alpha) : X \times \mathbf{R} \to \mathbf{R}$ and X is a Banach space. The first example illustrates the quadratic fold point phenomena, and the second, pitchfork bifurcation; however, the local bifurcation analysis rests on the extended approach of Zeidler [59, §29.18] and Bobylev and Krasnosel'skii [5].

The problem of finding the minimum surface of revolution connecting two coaxial circular loops each of radius R separated by a distance of $2l$ can be formulated as [16]

$$\begin{aligned} \text{minimize} \quad & 2\pi \int_{-l}^{l} r\sqrt{1 + \left(\frac{dr}{dx}\right)^2}\, dx \\ \text{subject to} \quad & r(-l) = a, \qquad r(l) = a, \\ & r \in C^1([0, 1], \mathbf{R}). \end{aligned}$$

The answer to this problem is as follows. For $0 < \frac{l}{R} < 0.6627$, there are two zeros $\frac{l}{C}$ to the equation $\frac{R}{l}\frac{l}{C} - \cosh\left(\frac{l}{C}\right) = 0$. For either of these two zeros the curve defined by

$$r(x) = C \cosh\left(\frac{x}{C}\right)$$

is an extremal of the above problem: the smaller zero yields a minimum and the larger, a local maximum. In either case the area is given by

$$A_{min} = \begin{cases} 2\pi R^2 \left(\tanh\left(\frac{l}{C}\right) + \frac{l}{C}\operatorname{sech}^2\left(\frac{l}{C}\right)\right) & \text{for } 0 < \frac{l}{C} \le 0.6627, \\ 2\pi R^2 & \text{for } \frac{l}{C} > 0.6627, \end{cases}$$

where the constant C is determined from the above equation.

For $\frac{l}{R} = 0.6627$, there is exactly one zero $\frac{l}{C} = 1.1996786$ to the above equation; the corresponding solution is a quadratic fold point. For $\frac{l}{R} > 0.6627$, there is no C^1 solution. The bifurcation diagram is qualitatively like that in Figure 1.

The next example is that of the equilibrium-deflected shape of an Euler beam as determined from the variational principle [60]:

$$\begin{aligned} \text{minimize} \quad & \int_0^l \frac{EI}{2}u^2 - P(1 - \cos\theta)\, ds \\ \text{subject to} \quad & \frac{d\theta}{ds} = u, \\ & \theta(0) = 0, \qquad \theta(l) = 0, \\ & \theta \in C^1([0, l], \mathbf{R}),\ u \in C([0, l], \mathbf{R}), \end{aligned}$$

where P is the load on the column, E is the modulus of elasticity, I is the second moment of cross-sectional area about the neutral axis, $\tan\theta$ is the slope of the neutral axis, l is the length of the column, s is the arc-length along the neutral axis of the column, and $\frac{d\theta}{ds} = u$ is the curvature. Euler's equation for this problem is

$$\begin{aligned} \frac{d^2\theta}{dx^2} + \omega^2 \sin\theta = 0, \\ \theta(0) = 0, \qquad \theta(1) = 0, \end{aligned}$$

where $\omega^2 = \frac{Pl^2}{EI}$ and $lx = s$. This well-known example in bifurcation theory illustrates the pitchfork bifurcation in Figure 4. Since it is analyzed rather completely in the book by Zeidler [59], we forego any further discussion.

7. Conclusions. The primary objective in this work has been the derivation of conditions under which bifurcation theory is applicable to the parametric optimization problem (1.1) posed in Banach spaces. By using the Fritz John first-order necessary conditions and a nonstandard normalization of the multipliers, this problem has been formulated as a set of equations on Banach spaces. By relaxing the bijectivity assumptions of the implicit function theorem but maintaining the Fredholm property of the Fréchet derivative of the nonlinear system of equations, a rather general framework for the analysis of bifurcation problems has been developed. Singularities generally arise when the strict complementarity fails, the Fréchet derivative of the active constraints fails to be surjective, or a second-order condition fails (see Theorem 2.2). A branching analysis has been provided for each of the generic cases.

With respect to further developments, one can analyze higher-codimension problems as in the books by Golubitsky, Stewart and Schaeffer [17, 18] and by Poston and Stewart [43]. Exploitation of symmetry in the bifurcation analysis has been an active area of research, and the use of this work should prove to be equally productive for the nonlinear program (1.1). Another fundamentally important aspect is the exchange in the stability (i.e., persistence of minima) at the singularities. Similar problems have been investigated for unconstrained problems using monotone operators and spectral theory, as presented in the books of Zeidler [58–60] and references therein. The Hilbert space setting developed by Bobylev and Krasnosel'skii [5] for the unconstrained problem appears to provide a correct framework. These approaches should play a central part in the development of the stability theory for the abstract constrained optimization problem (1.1) and will be investigated in future work.

Finally, we discussed the difficulty of treating problems with general cone constraints in the introduction. To investigate the applicability of bifurcation theory to these problems, one might consider, for example,

$$\begin{aligned} &\text{minimize} && f_0(x, \alpha) \\ &\text{subject to} && F(x, \alpha) = 0, \qquad G(x, \alpha) \in K, \end{aligned}$$

where $f_0 \in C^2(U \times V; \mathbf{R})$, $F \in C^2(U \times V; Y)$, $G \in C^2(U \times V; Z)$, U and V are open subsets of the Banach space X and $\mathbf{R}^r$, respectively, Y and Z are Banach spaces, and $K \subset Z$ is a closed convex cone with a nonempty interior. Under appropriate conditions [34, Thm. 4.5], one has a Fritz John condition in which a constraint qualification is relaxed, and a system of equations somewhat similar to (2.4) can be derived. (The interiority assumption on K again places restrictions on the Banach space Z.) This problem will also be investigated in future work.

Appendix. Proof of Theorems 2.2 and 2.4. Central to the proofs of Theorems 2.2 and 2.4 are expressions for both $D_\chi G$ and $D_\chi G^*$ since we need to investigate solutions of $D_\chi G(\Delta\chi) = b$ and $D_\chi G^* \Delta\chi^* = b^*$, respectively. First, note that $D_\chi G : X \times Y^* \times \mathbf{R}^m \times \mathbf{R} \to X^* \times Y \times \mathbf{R}^m \times \mathbf{R}$ and that the problem $D_\chi G \Delta\chi = b$ can be written as

$$\begin{aligned} D_x^2 \mathcal{L} \Delta x + L^* \Delta y^* + (D_x f)^* \Delta\lambda + (D_x f_0)^* \Delta\lambda_0 &= b_1, \\ L \Delta x &= b_2, \\ \Lambda D_x f \Delta x + \operatorname{diag}(f) \Delta\lambda &= b_3, \\ 2\langle y_0, y^* \rangle \langle y_0, \Delta y^* \rangle + 2\langle \lambda, \Delta\lambda \rangle + 2\lambda_0 \Delta\lambda_0 &= b_4, \end{aligned} \tag{A.1}$$

where $(b_1, b_2, b_3, b_4) \in X^* \times Y \times \mathbf{R}^m \times \mathbf{R}$. The projections P_i and P_j^* defined in §2.1, the definitions $\Lambda_A = \operatorname{diag}(\lambda_1, \ldots, \lambda_l)$, $\Lambda_B = \operatorname{diag}(\lambda_{l+1}, \ldots, \lambda_k)$, $\Lambda_C = \operatorname{diag}(\lambda_{k+1}, \ldots, \lambda_m)$,

$\lambda_A = (\lambda_1, \ldots, \lambda_l)$, $\lambda_B = (\lambda_{l+1}, \ldots, \lambda_k)$, $\lambda_C = (\lambda_{k+1}, \ldots, \lambda_m)$, and similar definitions for f_A, f_B, and f_C allow one to rewrite the system (A.1) as

$$D_x^2\mathcal{L}_{11}\Delta x_1 + D_x^2\mathcal{L}_{12}\Delta x_2 + P_1^*L^*\Delta y^* + P_1^*(D_xf)^*\Delta\lambda + P_1^*(D_xf_0)^*\Delta\lambda_0 = P_1^*b_1, \tag{A.2}$$

$$D_x^2\mathcal{L}_{21}\Delta x_1 + D_x^2\mathcal{L}_{22}\Delta x_2 + P_2^*L^*\Delta y^* + P_2^*(D_xf)^*\Delta\lambda + P_2^*(D_xf_0)^*\Delta\lambda_0 = P_2^*b_1, \tag{A.3}$$

$$L\Delta x_1 + L\Delta x_2 = b_2, \tag{A.4}$$

$$\Lambda_A D_x f_A\Delta x_1 + \Lambda_A D_x f_A\Delta x_2 = b_{31}, \tag{A.5}$$

$$0\Delta\lambda_B = b_{32}, \tag{A.6}$$

$$\operatorname{diag} f_C\Delta\lambda_C = b_{33}, \tag{A.7}$$

$$2\langle y_0, y^*\rangle\langle y_0, \Delta y^*\rangle + 2\langle\lambda_A, \Delta\lambda_A\rangle + 2\lambda_0\Delta\lambda_0 = b_4. \tag{A.8}$$

An important observation about equation (A.3) is that, due to the definition of the projection P_2^* as projecting X^* onto the range of the transformation $\tilde{L}^* = (D_xF^*, D_xf_A^*)$, (A.3) is *always* solvable for $(\Delta y^*, \Delta\lambda_A^*)$.

The problem $D_\chi G^*\Delta\chi^* = b^*$ for the linear transformation $D_\chi G^* : X^{**} \times Y^* \times \mathbf{R}^{m*} \times \mathbf{R}^* \to X^* \times Y^{**} \times \mathbf{R}^{m*} \times \mathbf{R}^*$ can be written as

$$\begin{aligned} D_x^2\mathcal{L}^*\Delta x^{**} + L^*\Delta y^* + (D_xf)^*\Lambda^*\Delta\lambda^* &= b_1^*, \\ L^{**}\Delta x^{**} + 2\langle y_0, y^*\rangle J(y_0)\Delta\lambda_0^* &= b_2^*, \\ D_xf^{**}\Delta x^{**} + \operatorname{diag}(f)\Delta\lambda^* + 2\lambda\Delta\lambda_0^* &= b_3^*, \\ D_xf_0^{**}\Delta x^{**} + 2\lambda_0\Delta\lambda_0^* &= b_4^* \end{aligned} \tag{A.9}$$

and decomposed as

$$D_x^2\mathcal{L}_{11}^*\Delta x_1^{**} + D_x^2\mathcal{L}_{21}^*\Delta x_2^{**} + P_1^*L^*\Delta y^* + P_1^*D_xf_A^*\Lambda_A^*\Delta\lambda_A^* = P_1^*b_1^*, \tag{A.10}$$

$$D_x^2\mathcal{L}_{12}^*\Delta x_1^{**} + D_x^2\mathcal{L}_{22}^*\Delta x_2^{**} + P_2^*L^*\Delta y^* + P_2^*D_xf_A^*\Lambda_A^*\Delta\lambda_A^* = P_2^*b_1^*, \tag{A.11}$$

$$L^{**}P_1^{**}\Delta x_1^{**} + L^{**}P_2^{**}\Delta x_2^{**} + 2\langle y_0, y^*\rangle J(y_0)\Delta\lambda_0^* = b_2^*, \tag{A.12}$$

$$D_xf_A^{**}P_1^{**}\Delta x_1^{**} + D_xf_A^{**}P_2^{**}\Delta x_2^{**} + 2\lambda_A\Delta\lambda_0^* = b_{31}^*, \tag{A.13}$$

$$D_xf_B^{**}P_1^{**}\Delta x_1^{**} + D_xf_B^{**}P_2^{**}\Delta x_2^{**} = b_{32}^*, \tag{A.14}$$

$$D_xf_C^{**}P_1^{**}\Delta x_1^{**} + D_xf_C^{**}P_2^{**}\Delta x_2^{**} + \operatorname{diag}(f_C)\Delta\lambda_C^* = b_{33}^*, \tag{A.15}$$

$$(D_xf_0)^{**}P_1^{**}\Delta x_1^{**} + (D_xf_0)^{**}P_2^{**}\Delta x_2^{**} + 2\lambda_0\Delta\lambda_0^* = b_4^*, \tag{A.16}$$

where $J : Y \to Y^{**}$ is *the* isometric isomorphism of Y onto a closed subspace of Y^{**} defined by $\langle y, y^*\rangle = \langle y^*, J(y)\rangle$ for all $y \in Y$ and $y^* \in Y^*$ [50, p. 95]. As above, note that equation (A.11) is *always* solvable for $(\Delta y^*, \Delta\lambda_A^*)$.

Proof of Theorem 2.2. First assume that (a), (b), and (c) are valid. Assumption (a) implies $k = l$, so equation (A.6) is vacuous. Note that Λ_A and $\operatorname{diag} f_C$ are nonsingular and that $\tilde{L}$ being surjective implies $\tilde{L}^*$ is injective. The latter, along with the first equation in (2.4) and (2.5c), implies that $\lambda_0 \neq 0$ and $D_xf_0(x) \in \mathcal{R}(\tilde{L}^*)$. The proof that $D_\chi G$ is *bijective* is broken into two parts: $D_\chi G$ is injective and then $D_\chi G$ is surjective.

To show that $D_\chi G$ is *injective*, set $b = 0$. Equations (A.4) and (A.5) imply $\Delta x_2 = 0$. Equation (A.7) implies $\Delta\lambda_C = 0$. Equation (A.2) now reduces to $D_x^2\mathcal{L}_{11}\Delta x_1 = 0$, which forces $\Delta x_1 = 0$ by assumption (b). Equation (A.3) with $\Delta x_1 = 0$, $\Delta x_2 = 0$, and $\Delta\lambda_C = 0$

and $(\tilde{L}^*)$ being injective on $Y^* \times \mathbf{R}^l$ implies $\Delta y^* = \gamma y^*$, $\Delta\lambda = \gamma\lambda$, and $\Delta\lambda_0 = \gamma\lambda_0$ for some scalar γ. Equation (A.8) then requires $\gamma = 0$, so $\Delta\chi = 0$.

To show that $D_\chi G$ is *surjective*, let b be an arbitrary member of $X^* \times Y \times \mathbf{R}^m \times \mathbf{R}^1$. Assumption (a) ($k = l$) shows that (A.6) is vacuous. Equation (A.7) is uniquely solvable for $\Delta\lambda_C$. Equations (A.4) and (A.5) are uniquely solvable for Δx_2 since $\tilde{L}$ is surjective and Λ_A is nonsingular. Equation (A.2) is uniquely solvable for Δx_1 since $D_x^2\mathcal{L}_{11} : X_1 \to X_1^*$ is bijective. Since $\tilde{L}^* = (L^*, D_x f_A^*)$ is injective, P_2^* is a projection onto the range of $\tilde{L}^*$, and Δx_1, Δx_2, and Λ_C are now determined, (A.3) is solvable with a solution having the structure $\Delta y^* = \gamma y^* + y_p^*$, $\Delta\lambda_A = \gamma\lambda_A + \hat{\lambda}_{Ap}$, and $\Delta\lambda_0 = \gamma\lambda_0$, where y_p^* and $\hat{\lambda}_{1p}$ are particular solutions. Substitution into the last equation, (A.8), shows that it is now uniquely solvable for γ.

Assume now that $D_\chi G$ is bijective. The objective is to establish (a), (b), and (c). That $D_\chi G$ is surjective implies that (A.6) must be vacuous so that assumption (a) must hold. Condition (c), i.e., surjectivity of $\tilde{L}$, follows from (A.4) and (A.5) and the fact that Λ_A is nonsingular. In particular, $D_x f_0(x) \in \mathcal{R}(\tilde{L}^*)$ and $\lambda_0 \neq 0$ for the same reason as above. Finally, we must show that $D_x^2\mathcal{L}_{11}$ is bijective. Consider the system (A.2)–(A.8) with $b_2 = 0$, $b_{31} = 0$, $b_{32} = 0$, $b_{33} = 0$, and $b_4 = 0$ and with (a) and (c) being valid. Note that $\Delta x_2 = 0$ and $\Delta\lambda_C = 0$. Now (A.2) reduces to $D_x^2\mathcal{L}_{11}\Delta x_1 = P_1^* b_1$, which must be solvable for some Δx_1 since the entire system is uniquely solvable. Thus $D_x^2\mathcal{L}_{11}$ is surjective. We need to show that it is injective. Let Δx_1 be any solution to this equation. Now (A.3) reduces to $L^*\Delta y^* + D_x f_A^*\Delta\lambda_A + (D_x f_0)\Delta\lambda_0 = P_2^* b_1 - D_x^2\mathcal{L}_{21}\Delta x_1$. Thus given b_1 and Δx_1, this equation and (A.8) are uniquely solvable for Δy^*, $\Delta\lambda_A$, and $\Delta\lambda_0$. Hence, if there are two solutions of $D_x^2\mathcal{L}_{11}\Delta x_1 = P_1^* b_1$, there are two distinct solutions of the system (A.2)–(A.8), a contradiction. Thus $D_x^2\mathcal{L}_{11}$ is bijective and (b) must hold. □

Proof of Theorem 2.4. Assuming first that (a) and (b) are valid, we need to demonstrate that the dimension of the null spaces of $D_\chi G(\hat{\chi}, \hat{\alpha})$ and $D_\chi G(\hat{\chi}, \hat{\alpha})^*$ are finite and the range of $D_\chi G(\hat{\chi}, \hat{\alpha}))$ is closed. To determine $\mathcal{N}(D_\chi G(\hat{\chi}, \hat{\alpha}))$, consider equations (A.2)–(A.8) with $b_i = 0$ for $i = 1, \ldots, 4$. Equations (A.4)–(A.6) imply $\Delta x_2 = 0$ and $\Delta\lambda_C = 0$. Equation (A.6) leaves $\Delta\lambda_B$ as a vector of arbitrary constants. Equation (A.2) now reduces to $D_x^2\mathcal{L}_{11}\Delta x_1 = -P_1^*(D_x f_0)\Delta\lambda_0 - P_1^* \sum_{i=l+1}^{k}(D_x f_i)\Delta\lambda_i$. By the Fredholm alternative theorem, this equation has a solution if and only if $\langle P_1^*(D_x f_0)\Delta\lambda_0 + P_1^* \sum_{i=l+1}^{k}(D_x f_i)\Delta\lambda_i, x^{**}\rangle = 0$ for each $x^{**} \in \mathcal{N}(D_x^2\mathcal{L}_{11}^*)$, which is finite dimensional since $D_x^2\mathcal{L}_{11}$ is assumed to be Fredholm. These finite number of conditions are imposed on the quantities $\Delta\lambda_i$ for $i = 0$ and $l+1, \ldots, k$. Thus the solution Δx_1 is in the span of at most a finite number of elements from X. Next, to address (A.11), recall that $\dim\mathcal{N}(L^*) = \dim Y/\mathcal{R}(L)$ [50, p. 112], which is finite since $\mathcal{R}(L)$ has finite codimension. Thus the solution $(\Delta y^*, \Delta\lambda_A)$ to (A.3), which is always solvable, is the linear span of a finite number of elements from $(Y^*, \mathbf{R}^l)$. Finally, (A.8) places one more restriction on these finite number of elements from their respective spaces. This shows that the dimension of the null space of $D_\chi G(\hat{\chi}, \hat{\alpha})$ is finite.

Continuing to assume (a) and (b), we turn to the linear transformation $D_\chi G^* : X^{**} \times Y^* \times \mathbf{R}^m \times \mathbf{R} \to X^* \times Y^{**} \times \mathbf{R}^m \times \mathbf{R}$ and show that it has a finite-dimensional null space. First (a) implies that $D_x^2\mathcal{L}_{11}^*$ is Fredholm; that $\mathcal{R}(L)$ is closed and has finite codimension implies [50, p. 112] that $\mathcal{R}(L^{**})$ is closed and has the same finite codimension in Y^{**}. Furthermore, by our earlier remarks $X^{**} = X_1^{**} \oplus X_2^{**}$, where $X_1^{**} = \mathcal{N}(\tilde{L}^{**})$. Before proceeding, we need to compute $D_x\mathcal{L}$ and its adjoint. Given $\mathcal{L} = \mathcal{L}(x, y^*, \lambda, \lambda_0; \alpha)$ from (2.3), the linear functional $D_x\mathcal{L} : X \to \mathbf{R}$ is bounded so that $D_x\mathcal{L} \in X^*$. Now

$$\begin{aligned} D_x\mathcal{L}\bar{x} &= \langle D_x f_0(x, \alpha)\bar{x}, \lambda_0\rangle + \langle D_x f(x, \alpha)\bar{x}, \lambda\rangle + \langle D_x F(x, \alpha)\bar{x}, y^*\rangle \\ &= \langle \bar{x}, D_x f_0(x, \alpha)^*\lambda_0\rangle + \langle \bar{x}, D_x F(x, \alpha)^* y^*\rangle + \langle \bar{x}, D_x f(x, \alpha)^*\lambda\rangle. \end{aligned}$$

Thus $\langle D_x\mathcal{L}\bar{x}, \gamma\rangle = \langle \bar{x}, D_x\mathcal{L}^*\gamma\rangle$, where $D_x\mathcal{L}^* : \mathbf{R}^* \to X^*$ and

$$D_x\mathcal{L}^* = D_x f_0(x,\alpha)^*\lambda_0 + D_x F(x,\alpha)^* y^* + D_x f(x,\alpha)^*\lambda.$$

Next, we can compute $D_x\mathcal{L}^{**}$ from the requirement $\langle D_x\mathcal{L}^*\gamma, \bar{x}^{**}\rangle = \langle \gamma, D_x\mathcal{L}^{**}\bar{x}^{**}\rangle$ for all $\bar{x}^{**} \in X^{**}$ and $\gamma \in \mathbf{R}^* = \mathbf{R}$. The result is

$$D_x\mathcal{L}^{**}\bar{x}^{**} = \langle \lambda_0, D_x f_0(x,\alpha)^{**}\bar{x}^{**}\rangle + \langle y^*, D_x F(x,\alpha)^{**}\bar{x}^{**}\rangle + \langle \lambda, D_x f(x,\alpha)^{**}\bar{x}^{**}\rangle.$$

Since $D_x\mathcal{L}(\chi,\alpha) = 0$ at $(\chi,\alpha) = (\hat{\chi},\hat{\alpha})$, the same is true of $D_x\mathcal{L}^*(\hat{\chi},\hat{\alpha})$ and $D_x\mathcal{L}^{**}(\hat{\chi},\hat{\alpha})$. Thus, in view of the above, an appropriate combination of (A.12), (A.13), and (A.16) yields $D_x\mathcal{L}^{**}\Delta x^{**} + (|\langle y_0, y^*\rangle|^2 + \langle\lambda_A, \lambda_A\rangle + \lambda_0^2)\Delta\lambda_0^* = 0$; however, $D_x\mathcal{L}^{**}\Delta x^{**} = 0$ and the coefficient of $\Delta\lambda_0^*$ is one, so $\Delta\lambda_0^* = 0$ and thus $\Delta x_2^{**} = 0$. Then (A.10) reduces to $D_x^2\mathcal{L}^*\Delta x_1^{**} = 0$ so that Δx_1^{**} is a linear combination of a finite number of specific elements in X^{**}. Equations (A.14) ($D_x f_B^{**}\Delta x_1^{**} = 0$) and (A.16) ($D_x f_0^{**}\Delta x_1^{**} = 0$) can only further restrict this finite number. Equation (A.15) is solvable for $\Delta\lambda_C^*$ given this Δx_1^{**}. Finally, we come to (A.11), which is always solvable for $(\Delta y^*, \Delta\lambda_A)$. Since the dimension of $\mathcal{N}(L^*)$ is finite, the entire system (A.10)–(A.16) is now solvable for $(\Delta x^{**}, \Delta y^*, \Delta\lambda, \Delta\lambda_0)$ in terms of linear combination of a finite number of elements in $X^{**} \times Y^* \times \mathbf{R}^m \times \mathbf{R}^*$, so the null space of $D_\chi G^*$ is finite dimensional.

It remains to show that $D_\chi G$ has closed range. For this, we make repeated use of the closed-range theorem [50] and the lemma on the closed image [1, p. 80]. Consider first (A.4) and (A.5)–(A.7), i.e., $L\Delta x = b_2$ and $\Lambda D_x f\Delta x + \operatorname{diag}(f)\Delta\lambda = b_3$. Now L as a mapping from $\mathcal{X} = X \times Y^* \times \mathbf{R}^m \times \mathbf{R}$ to Y has closed range since L does as a mapping from X to Y. Then the range of $\Lambda D_x f\Delta x + \operatorname{diag}(f)\Delta\lambda$, where Δx is restricted to $\mathcal{N}(L)$, is a subspace of a finite-dimensional space. Hence the range of the linear transformation defined by (A.5)–(A.7) is a closed subspace of $Y \times \mathbf{R}^m$. Next, as another application of the lemma on the closed image [1, p. 80] consider (A.2) restricted to the kernel of the linear transformation defined by (A.4)–(A.7). This equation reduces to $D_x^2\mathcal{L}_{11}\Delta x_1 + P_1^*(D_x f_B)^*\Delta\lambda_B + P_1^*(D_x f_0)^*\Delta\lambda_0 = 0$. Now by assumption (a), the range of $D_x^2\mathcal{L}_{11}$ is closed and the range of $P_1^*(D_x f_B)^*$ and $P_1^*(D_x f_0)^*$ are finite dimensional. Thus, the range of the algebraic sum is closed [50, p. 32]. Next we turn to (A.3) and apply the same lemma to this equation restricted to the null space of the linear transformation defined by equations (A.2) and (A.4)–(A.7). Now the range of $\tilde{L}$ is all of $P_2^*X^*$, so we have that the range of (A.2)–(A.7) is closed in $X^* \times Y \times \mathbf{R}^m$. Finally, the range of (A.8) restricted to the kernel of (A.2)–(A.7) is either $\mathbf{R}$ or $\{0\}$, both of which are closed. Thus, $D_\chi G$ has closed range. This completes the first half of the theorem.

Next, assume $D_\chi G(\hat{\chi},\hat{\alpha})$ is a Fredholm operator and $D_x^2\mathcal{L}_{11}$ and $D_x F(\hat{x},\hat{\alpha})$ have closed ranges. The objective is to show that (a) and (b) are valid. Thus we start with $\dim\mathcal{N}(D_\chi G(\hat{\chi},\hat{\alpha})) < \infty$ and examine equations (A.2)–(A.8). Equations (A.4) and (A.5) imply $\Delta x_2 = 0$, and equation (A.6) leaves $\Delta\lambda_i$ for $i = l+1, \ldots, k$ as arbitrary constants. Equation (A.6) implies $\Delta\lambda_i = 0$ for $i = k+1, \ldots, m$. Equation (A.2) now takes the form $D_x^2\mathcal{L}_{11}\Delta x_1 = -P_1^*(D_x f_0)\Delta\lambda_0 - P_1^*\sum_{i=l+1}^{k}(D_x f_i)\Delta\lambda_i$, whose solution Δx_1 must belong to a finite-dimensional space. Thus, the dimension of the null space of $D_x^2\mathcal{L}_{11}$ can be at most finite dimensional. Also, (A.3) implies that the dimension of the null space of L^* is at most finite since Δy^* is further constrained only by (A.8). Since $\mathcal{N}(L^*) = \mathcal{R}(L)^\perp$, $Y/\mathcal{R}(L)$ has same finite dimension as $\mathcal{N}(L^*)$, i.e., $\mathcal{R}(L)$ has finite codimension.

Finally, to show the range of $D_x^2\mathcal{L}_{11}$ has finite codimension, we show that $D_x^2\mathcal{L}_{11}^*$ has a finite-dimensional null space by considering the null space of $D_\chi G(\hat{\chi},\hat{\alpha})^*$, which must be finite dimensional since it is Fredholm. Once again considering equations (A.10)–(A.16) with $b^* = 0$, the same argument as above shows that $\Delta\lambda_0^* = 0$ and $\Delta x_2^{**} = 0$. Regardless of Δx_1^{**}, (A.11) is always solvable for $(\Delta y^*, \Delta\lambda_A)$, and (A.15) for $\Delta\lambda_C$. Now Δx_1^{**} by

its definition already satisfies (A.12) and (A.13). Thus we are left with (A.10), (A.14), and (A.16), which reduce to $D_x^2\mathcal{L}_{11}^*\Delta x_1^{**} = 0$, $D_x f_B^{**}\Delta x_1^{**} = 0$, and $D_x f_0^{**}\Delta x_1^{**} = 0$. Since the ranges of $D_x f_B^{**}$ and $D_x f_0^{**}$ are finite dimensional, so is the codimension of the null spaces. Thus $\mathcal{N}(D_x^2\mathcal{L}_{11}^*) \cap \mathcal{N}(D_x f_B^{**}) \cap \mathcal{N}(D_x f_0^{**})$ is finite dimensional if and only if $\mathcal{N}(D_x^2\mathcal{L}_{11}^*)$ is finite dimensional. □

REFERENCES

[1] V. M. ALEKSEEV, V. M. TIKHOMIROV, AND S. V. FOMIN, *Optimal Control*, Consultants Bureau, New York, 1987.

[2] W. ALT, *Stability of solutions for a class of nonlinear cone constrained optimization problems, Part 1: Basic theory*, Numer. Funct. Anal. Optim., 10 (1989), pp. 1053–1064.

[3] ———, *Local stability of solutions to differentiable optimization problems in Banach spaces*, Optim. Theory Appl., 70 (1991), pp. 443–466.

[4] B. BANK, J. GUDDAT, D. KLATTE, B. KUMMER, AND K. TAMMER, *Nonlinear Parametric Optimization*, Birkhäuser-Verlag, Basel, 1983.

[5] N. A. BOBYLEV AND M. A. KRASNOSEL'SKII, *Investigation scheme for extremals of multidimensional variational problems*, Functional Anal. Appl., 28 (1994), pp. 227–237.

[6] J. F. BONNANS, *Directional derivatives of optimal solutions in smooth nonlinear programming*, J. Optim. Theory Appl., 73 (1992), pp. 27–45.

[7] J. F. BONNANS, A. D. IOFFE, AND A. SHAPIRO, *Expansions of exact and approximant solutions in nonlinear programming*, in French-German Conference in Optimization, D. Pallaschke, ed., Lecture Notes in Econom. and Math. Systems, Springer-Verlag, Berlin, 1993.

[8] D. J. BELL AND D. H. JACOBSON, *Singular Optimal Control Problems*, Academic Press, London, 1975.

[9] L. CESARI, *Optimization—Theory and Applications*, Springer-Verlag, New York, 1983.

[10] S.-N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[11] S. DAFERMOS, *Sensitivity analysis in variational inequalities*, Math. Oper. Res., 13 (1988), pp. 421–434.

[12] A. L. DONTCHEV AND W. W. HAGER, *Lipschitzian stability in nonlinear control and optimization*, SIAM J. Control Optim., 31 (1993), pp. 569–603.

[13] A. V. FIACCO, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, New York, Academic Press, 1983.

[14] ———, *Mathematical Programming Study 21: Sensitivity, Stability and Parametric Analysis*, North-Holland, Amsterdam, 1984.

[15] A. V. FIACCO AND Y. ISHIZUKA, *Sensitivity and stability analysis for nonlinear programming*, Ann. Oper. Res., 27 (1991), pp. 215–235.

[16] I. M. GELFAND AND S. V. FOMIN, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

[17] M. GOLUBITSKY AND D. G. SCHAEFFER, *Singularities and Groups in Bifurcation Theory*, Vol. 1, Springer-Verlag, New York, 1985.

[18] M. GOLUBITSKY, I. STEWART, AND D. G. SCHAEFFER, *Singularities and Groups in Bifurcation Theory*, Vol. 2, Springer-Verlag, New York, 1988.

[19] J. GUDDAT, F. GUERRA VAZQUEZ, AND H. TH. JONGEN, *Parametric Optimization: Singularities, Path Following, and Jumps*, John Wiley and Sons, Chichester, England, 1990.

[20] J. GUDDAT, ED., *Parametric Optimization and Related Topics II*, Math. Res. 62, Akademie-Verlag, Berlin, 1991.

[21] A. D. IOFFE, *Necessary and sufficient conditions for a local minimum 3: Second order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266–288.

[22] ———, *On sensitivity analysis of nonlinear programs in Banach spaces: The approach via composite unconstrained optimization*, SIAM J. Optim., 4 (1994), pp. 1–43.

[23] A. D. IOFFE AND V. M. TIKHOMIROV, *Theory of Extremal Problems*, North-Holland, Amsterdam, New York, Oxford, 1979.

[24] K. ITO AND K. KUNISCH, *Sensitivity analysis of solutions to optimization problems in Hilbert spaces with applications to optimal control and estimation*, J. Differential Equations, 99 (1992), pp. 1–40.

[25] H. TH. JONGEN, P. JONKER, AND F. TWILT, *On one-parameter families of sets defined by (in)equality constraints*, Nieuw Arch. Wisk., 3 (1982), pp. 307–322.

[26] ———, *Critical sets in parametric optimization*, Math. Programming, 34 (1986), pp. 333–353.

[27] ———, *One-parameter families of optimization problems: Equality constraints*, J. Optim. Theory Appl., 48 (1986), pp. 141–161.

[28] H. TH. JONGEN AND G. W. WEBER, *On parametric nonlinear programming*, Ann. Oper. Res., 27 (1991), pp. 253–284.

[29] T. KATO, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, 1984.

[30] A. J. KING AND R. T. ROCKAFELLAR, *Sensitivity analysis for nonsmooth generalized equations*, Math. Programming, 55 (1992), pp. 193–212.

[31] J. KOGAN, *Bifurcation of Extremals in Optimal Control*, Lecture Notes in Math. 1216, Springer-Verlag, New York, 1980.
[32] M. KOJIMA AND R. HIRABAYASHI, *Continuous deformation of nonlinear programs*, in Mathematical Programming Study 21: Sensitivity, Stability and Parametric Analysis, A. V. Fiacco, ed., North-Holland, Amsterdam, 1984.
[33] N. H. KUIPER, *C^1 equivalence of functions near isolated critical points*, in Symposium on Infinite-Dimensional Topology, Ann. of Math. Stud. 69, R. D. Anderson, ed., Princeton University Press, Princeton, NJ, 1972.
[34] S. KURCYUSZ, *On the existence and nonexistence of Lagrange multipliers in Banach spaces*, J. Optim. Theory Appl., 20 (1976), pp. 81–110.
[35] J. KYPARISIS, *Sensitivity analysis framework for variational inequalities*, Math. Programming, 38 (1987), pp. 190–203.
[36] E. S. LEVITIN, *Perturbation Theory in Mathematical Programming and Its Applications*, John Wiley & Sons, Chichester, England, 1994.
[37] B. N. LUNDBERG AND A. B. POORE, *Numerical continuation and singularity detection methods for parametric nonlinear programming*, SIAM J. Optim., 3 (1993), pp. 134–154.
[38] K. MALANOWSKI, *Second order conditions and constraint qualifications in stability and sensitivity analysis of solutions to optimization problems in Hilbert spaces*, Appl. Math. Optim., 25 (1992), pp. 51–79.
[39] H. MAURER, *First- and second-order sufficient optimality conditions in mathematical programming and optimal control*, in Mathematical Programming at Overwolfach, H. König, B. Korte, and K. Ritter, eds., Mathematical Programming Study, 14 (1981), pp. 163–177.
[40] L. NIRENBERG, *Topics in Nonlinear Functional Analysis*, Lecture notes, Courant Institute of Mathematical Sciences, New York University, 1974.
[41] A. B. POORE AND C. A. TIAHRT, *Bifurcation problems in nonlinear parametric programming*, Math. Programming, 39 (1987), pp. 189–205.
[42] A. B. POORE, *Bifurcations in parametric nonlinear programming*, Ann. Oper. Res., 27 (1991), pp. 343–370.
[43] T. POSTON AND I. STEWART, *Catastrophe Theory and Its Applications*, Pitman, London, San Francisco, Melbourne, 1978.
[44] Y. QIU AND T. L. MAGNANTI, *Sensitivity analysis for variational inequalities defined on polyhedral sets*, Math. Oper. Res., 14 (1989), pp. 410–432.
[45] P. RABIER, *Lectures on Topics in One-Parameter Bifurcation Problems*, Tata Inst. Fund. Res. Lectures on Math. and Phys., Springer-Verlag, Berlin, 1985.
[46] S. M. ROBINSON, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[47] ———, *Generalized equations and their solutions, Part* II: *Applications to nonlinear programming*, Math. Prog. Stud., 19 (1982), pp. 200–221.
[48] ———, *An implicit-function theorem for a class of nonsmooth functions*, Math. Oper. Res., 16 (1991), pp. 292–309.
[49] W. RUDIN, *Real and Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1987.
[50] ———, *Functional Analysis*, 2nd ed., McGraw-Hill, New York, 1991.
[51] M. SCHECHTER, *Principles of Functional Analysis*, Academic Press, New York, 1971.
[52] A. SHAPIRO, *Sensitivity analysis of parameterized programs via generalized equations*, SIAM J. Control Optim., 32 (1994) pp. 553–571.
[53] ———, *Perturbation analysis of optimization problems in Banach spaces*, Numer. Funct. Anal. Optim., 13 (1992), pp. 97–116.
[54] A. SHAPIRO AND J. F. BONNANS, *Sensitivity analysis of parameterized programs under cone constraints*, SIAM J. Control Optim., 30 (1992) pp. 1409–1422.
[55] C. A. TIAHRT AND A. B. POORE, *A bifurcation analysis of the nonlinear parametric programming problem*, Math. Programming, 47 (1990), pp. 117–141.
[56] V. M. TIKHOMIROV, *Fundamental Principles of the Theory of Extremal Problems*, John Wiley & Sons, New York, 1986.
[57] H. URAKAWA, *Calculus of Variations and Harmonic Maps*, Trans. Math. Monographs 132, American Mathematical Society, Providence, RI, 1993.
[58] E. ZEIDLER, *Nonlinear Functional Analysis and Its Applications* II/A: *Linear Monotone Operators*, Springer-Verlag, Berlin, 1990.
[59] ———, *Nonlinear Functional Analysis and Its Applications* II/B: *Nonlinear Monotone Operators*, Springer-Verlag, Berlin, 1990.
[60] ———, *Nonlinear Functional Analysis and Its Applications* III: *Variational Methods and Optimization*, Springer-Verlag, Berlin, 1985.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 1972–1998, November 1996

007

STABILITY RADII OF SYSTEMS WITH STOCHASTIC UNCERTAINTY AND THEIR OPTIMIZATION BY OUTPUT FEEDBACK*

D. HINRICHSEN† AND A. J. PRITCHARD‡

Abstract. We consider linear plants controlled by dynamic output feedback which are subjected to blockdiagonal stochastic parameter perturbations. The stability radii of these systems are characterized, and it is shown that, for real data, the real and the complex stability radii coincide. A corresponding result does not hold in the deterministic case, even for perturbations of single-output feedback type. In a second part of the paper we study the problem of optimizing the stability radius by dynamic linear output feedback. Necessary and sufficient conditions are derived for the existence of a compensator which achieves a suboptimal stability radius. These conditions consist of a parametrized Riccati equation, a parametrized Liapunov inequality, a coupling inequality, and a number of linear matrix inequalities (one for each disturbance term). The corresponding problem in the deterministic case, the optimal μ-synthesis problem, is still unsolved.

Key words. stability radius, stochastic systems, multiperturbations, state-dependent noise, dynamic output feedback, Riccati inequalities, linear matrix inequalities, scaling

AMS subject classifications. 93C, 93D, 93E

1. Introduction.

One of the main purposes of feedback control is to ensure satisfactory behaviour of a dynamical system in the presence of unforeseen disturbances. This classical problem, which was central to the work of Bode and Nyquist, has seen a vigorous renaissance over the past decade, and recent developments in control theory have been strongly influenced by it. The focus has been on deterministic disturbances: *either* unstructured (additive or multiplicative) perturbations of the plant's transfer function *or* structured perturbations of the parameters of a given nominal state-space model. As examples, we mention two approaches, H^∞ and stability radii. H^∞ theory (see [6], [20]) deals with the problem of minimizing (by feedback compensation) the effect of deterministic disturbances on the variables to be controlled. The results can be applied to maximize robustness of stability with respect to unstructured perturbations of the transfer matrix. On the other hand, the theory of stability radii determines precise robustness measures for stable linear state-space systems subject to different classes of structured parameter perturbations [14]. Surprisingly there is a close relationship between the two theories for the special case where stability radii with respect to *complex* perturbations of single-output feedback type are considered. In fact, in this case the problem of optimizing the stability radius by feedback control is equivalent to a singular H^∞ control problem [13].

In this paper we use the framework of stability radii to study robust stability and robust stabilization problems for systems with *stochastic* uncertainty. Because of the close relationship between the theories of stability radii and H^∞ control, our results can be regarded as an extension of H^∞ control theory to systems with stochastic uncertainty.

We consider the system

$$(1) \qquad dx(t) = Ax(t)\,dt + \sum_{i=1}^{N} D_i \Delta_i(E_i x(t))dw_i(t) + Bu(t)\,dt, \qquad y(t) = Cx(t),$$

where the matrices A, B, C, D_i, and E_i are given and the processes w_i are independent scalar Wiener processes, $i = 1, \ldots, N$. We view the above equations as describing a linear deterministic differentiable system (A, B, C) perturbed by stochastic multiperturbations

*Received by the editors November 11, 1994; accepted for publication (in revised form) August 8, 1995.
†Institut für Dynamische Systeme, Universität Bremen, D-28334 Bremen, Germany.
‡Control Theory Centre, University of Warwick, Coventry CV4 7AL, UK.

$\sum_{i=1}^{N} D_i \Delta_i(E_i x(t))dw_i(t)$. The family $(D_i, E_i)_{i\in\underline{N}}$ of matrix pairs describes the structure of these perturbations, while Δ_i, $i = 1, \ldots, N$, are unknown Lipschitzian nonlinearities. We assume that *all the Wiener processes w_i have zero mean.* In other words, the nominal model (A, B, C) is assumed to be *exact in the mean.* If the system matrix A is also subject to *deterministic* parameter perturbations, the problems of robust stability and robust stabilization are more involved, and so far only estimates are available for the corresponding stability radii; see [15].

Many authors have studied stability and stabilization problems for systems with state-dependent noise; see, for example, [17]. The quadratic optimal control problem was solved in [19], and a collection of papers concerning Liapunov exponents for such systems can be found in [1]. However, there are few papers dealing with robustness issues for this class of systems. An important reference is [18], which, in our terminology, derives necessary and sufficient conditions under which infinite or arbitrarily large stability radii can be achieved by state feedback. Some results on stochastic stability radii defined via Liapunov exponents can be found in [4, §7]. A characterization of the stability radius in the special case where all the E_i are equal was given in [7]. (The mathematical development is essentially the same as in the single-perturbation case $N = 1$.)

Here we study the robust stability and robust stabilization problems under multiperturbations. For deterministic systems the development of such stability radii requires the use of μ-analysis [14], and it is well known that in the presence of more than three perturbation terms ($N > 3$) scaling techniques yield only lower estimates for the complex stability radius. In contrast, we will derive a precise characterization via scaling techniques in this stochastic case. This is based on the analysis of an associated minimax problem for quadratic forms. Moreover we will show that the *real* and the *complex* stability radii coincide for stochastic multiperturbations of the above kind.

The second main contribution of this paper concerns the problem of optimizing the stability radius of systems of the form (1) by dynamic output feedback. We characterize the supremal stability radius by combining a Riccati inequality with a Liapunov inequality, a coupling condition, and a number of additional linear matrix inequalities. Moreover we give explicit formulae for suboptimal controllers. These results are obtained by using an inequality approach to deterministic H^∞ control theory developed by Gahinet and his co-workers; see [10], [11]. Whereas in the deterministic case the suboptimal controllers can be characterized by a pair of Riccati equations and a coupling condition, it is not possible in the stochastic case to replace both the Riccati and the Liapunov inequalities by equalities. This will be illustrated by an example.

We proceed as follows. In the next section we give some results on a minimax problem for quadratic forms. (The proofs are in an appendix.) These results will be instrumental for our characterization of the stability radius relative to stochastic multiperturbations in §3. In §4 the problem of optimizing the stability radius by (linear) feedback is studied, and it is shown that the supremal stability radius can be determined via matrix inequalities. Finally, in §5 we show that in this characterization the Riccati inequality may be replaced by a Riccati equation, whereas the Liapunov inequality *cannot* be transformed into an equality. Moreover the corresponding results for state feedback are derived as corollaries of the previous results on dynamic output feedback.

2. A minimax problem for quadratic forms. Throughout the paper we use the following notation. $\mathbb{K}$ is either the field $\mathbb{R}$ of real numbers or the field $\mathbb{C}$ of complex numbers. For any integer $\ell \geq 1$, $\mathcal{H}_\ell(\mathbb{K})$ is the real vector space of Hermitian matrices in $\mathbb{K}^{\ell\times\ell}$ and $\mathcal{H}_\ell^+(\mathbb{K}) = \{H \in \mathcal{H}_\ell(\mathbb{K});\ H \succeq 0\}$ is the convex cone of positive semidefinite matrices in $\mathcal{H}_\ell(\mathbb{K})$. By $\langle\cdot,\cdot\rangle$ we denote the usual inner product on $\mathbb{K}^\ell$, $\ell \in \mathbb{N}$, and by $\|H\|$, the associated

operator norm or spectral norm of $H \in \mathcal{H}_\ell(\mathbb{K})$:

$$\|H\| = \max_{v \in \mathbb{K}^\ell, \|v\|=1} \langle v, Hv \rangle, \qquad \|v\| = \langle v, v \rangle^{1/2}.$$

We suppose that $N \in \mathbb{N}$ is given and

$$H_{ij} \in \mathcal{H}^+_{\ell_j}(\mathbb{K}), \qquad i, j \in \underline{N} := \{1, \ldots, N\}, \tag{2}$$

is a given family of nonnegative $\ell_j \times \ell_j$ Hermitian matrices. For any set C, we denote by $(0, \infty)^C$ the set of all mappings from C to $(0, \infty)$, the set of all positive real numbers. If C is finite, the elements of $(0, \infty)^C$ are represented by finite families $\alpha = (\alpha_c)_{c \in C}$. In particular the set $(0, \infty)^{\underline{N}}$ consists of all N-tuples $\alpha = (\alpha_1, \ldots, \alpha_N)$, with $\alpha_i > 0$ for all $i \in \underline{N}$. In our later analysis the H_{ij} will be given by

$$H_{ij} = \lambda_j \int_0^\infty D_j^* e^{A^* \tau} E_i^* E_i e^{A\tau} D_j d\tau \succeq 0, \qquad i, j \in \underline{N},$$

and the α_i's are free scaling parameters which are used to improve a lower bound for the stability radius. In fact we will see that the scaling technique applied to multiperturbations leads to the following optimization problem:

$$\text{minimize} \quad \max_{j \in \underline{N}} \left\| \sum_{i=1}^N \left(\frac{\alpha_i}{\alpha_j} \right)^2 H_{ij} \right\| \quad \text{with respect to} \quad \alpha = (\alpha_1, \ldots, \alpha_N) \in (0, \infty)^{\underline{N}}. \tag{3}$$

Let

$$f(\alpha) = \max_{j \in \underline{N}} f_j(\alpha), \quad f_j(\alpha) = \left\| \sum_{i=1}^N \left(\frac{\alpha_i}{\alpha_j} \right)^2 H_{ij} \right\|, \quad \alpha \in (0, \infty)^{\underline{N}}. \tag{4}$$

Then the optimization problem (3) is equivalent to minimizing the function $f(\cdot)$ on the positive orthant $(0, \infty)^{\underline{N}}$. The optimal value of (3) is denoted by

$$\hat{\mu} = \inf_{\alpha \in (0,\infty)^{\underline{N}}} f(\alpha). \tag{5}$$

In this section, in order to maintain the flow of the paper, we give only the pertinent results. The proofs are relegated to the appendix. Nevertheless we would like to stress that these proofs are an important part of the overall proof of our main results in §§3–5.

The solution of (3) depends strongly on the zero block pattern of the compound matrix $H = (H_{ij})_{i,j=1}^N$. To capture this pattern we denote by $\mathcal{G}$ the directed graph [3] with *node set* $\underline{N} = \{1, \ldots, N\}$ and *set of directed arcs*

$$\mathcal{A} = \{(i, j) \in \underline{N}^2; H_{ij} \neq 0\}.$$

$\mathcal{G}$ is said to be *strongly connected* if every node of $\mathcal{G}$ is connected to every distinct node of $\mathcal{G}$ by a directed path in $\mathcal{G}$.

THEOREM 2.1. *Suppose that $\mathcal{G}$ is strongly connected. Then there exists a subset $J \subset \underline{N}$ and a vector $\hat{\alpha} \in (0, \infty)^{\underline{N}}$ satisfying $f(\hat{\alpha}) = \hat{\mu}$, and*

$$\left\| \sum_{i=1}^N \left(\frac{\hat{\alpha}_i}{\hat{\alpha}_j} \right)^2 H_{ij} \right\| = \left\| \sum_{i \in J} \left(\frac{\hat{\alpha}_i}{\hat{\alpha}_j} \right)^2 H_{ij} \right\| = \hat{\mu} \quad \text{if } j \in J,$$

$$\left\| \sum_{i=1}^N \left(\frac{\hat{\alpha}_i}{\hat{\alpha}_j} \right)^2 H_{ij} \right\| < \hat{\mu} \quad \text{if } j \in \underline{N} \backslash J. \tag{6}$$

The theorem shows that in the strongly connected case the optimal value of the minimization problem (3) is equal to the optimal value of the subproblem

$$\text{(7)} \qquad \text{minimize} \quad \max_{j\in J} \left\| \sum_{i\in J} \left(\frac{\alpha_i}{\alpha_j}\right)^2 H_{ij} \right\| \quad \text{subject to } \alpha \in (0,\infty)^J,$$

for which only the data H_{ij}, $i, j \in J$, play a role. According to (6) the reduced scaling vector $\hat{\alpha}_J = (\hat{\alpha}_i)_{i\in J}$ is a minimizer for (7), and each of the norms $\| \sum_{i\in J} (\frac{\hat{\alpha}_i}{\hat{\alpha}_j})^2 H_{ij} \|$, $j \in J$, is equal to $\hat{\mu}$. A natural question is, under which conditions is the subset $J = \underline{N}$? The following proposition gives a sufficient condition.

PROPOSITION 2.2. *Suppose that $\mathcal{G}$ is strongly connected and for every nonempty subset $J \subset \underline{N}$, $J \neq \underline{N}$ there exists $j \in \underline{N}\backslash J$ such that*

$$\text{(8)} \qquad \bigcap_{i\in \underline{N}\backslash J} \ker H_{ij} \cap \left(\bigcap_{i\in J} \ker H_{ij} \right)^{\perp} = \{0\}.$$

Then

$$\text{(9)} \qquad f_1(\hat{\alpha}) = \cdots = f_N(\hat{\alpha}) = \hat{\mu}$$

for all $\hat{\alpha} \in (0,\infty)^{\underline{N}}$ satisfying $f(\hat{\alpha}) = \hat{\mu}$.

Remark 2.3. Suppose that each Hermitian matrix H_{ij}, $i, j \in \underline{N}$, is either invertible or zero. Then assumption (8) is satisfied if $\mathcal{G}$ is strongly connected. In particular, (9) always holds in the scalar case ($\ell_j = 1$, $j \in \underline{N}$) if $\mathcal{G}$ is strongly connected. □

If $\mathcal{G}$ is not strongly connected, there will not, in general, exist a minimum of f. To deal with this case we introduce the following notation. Let C_k, $k = 1, \ldots, K$, be the node sets of the strongly connected components [3] of $\mathcal{G}$ ordered in such a way that for $1 \le h < k \le K$ there is no directed arc $(i, j) \in \mathcal{A}$ such that $i \in C_k$, $j \in C_h$. Then, for all $h, k \in \underline{K}$,

$$\text{(10)} \qquad h < k \implies (i \in C_k \text{ and } j \in C_h \Rightarrow H_{ij} = 0).$$

Since $\underline{N} = \cup_{k\in\underline{K}} C_k$, it follows from (10) that

$$\text{(11)} \qquad f(\alpha) = \max_{h\in\underline{K}} \max_{j\in C_h} \left\| \sum_{i=1}^{N} \left(\frac{\alpha_i}{\alpha_j}\right)^2 H_{ij} \right\| = \max_{h\in\underline{K}} \max_{j\in C_h} \left\| \sum_{k=1}^{K} \sum_{i\in C_k} \left(\frac{\alpha_i}{\alpha_j}\right)^2 H_{ij} \right\|, \qquad \alpha \in (0,\infty)^{\underline{N}}.$$

The next theorem shows that problem (3) can be solved by restricting our considerations to the strongly connected components of $\mathcal{G}$.

THEOREM 2.4. *Let*

$$\text{(12)} \qquad \mu_k = \min_{\alpha\in(0,\infty)^{C_k}} \max_{j\in C_k} \left\| \sum_{i\in C_k} \left(\frac{\alpha_i}{\alpha_j}\right)^2 H_{ij} \right\|, \qquad k \in \underline{K}.$$

Then

$$\hat{\mu} = \max_{k\in\underline{K}} \mu_k.$$

Moreover, if $\hat{k}$ satisfies $\mu_{\hat{k}} = \max_{k \in \underline{K}} \mu_k$ there exist a subset $J \subset C_{\hat{k}}$ and, for every $\delta > 0$, a vector $\alpha \in (0, \infty)^{\underline{N}}$ such that

$$\left\| \sum_{i=1}^{N} \left(\frac{\alpha_i}{\alpha_j} \right)^2 H_{ij} \right\| \leq \hat{\mu} + \delta, \qquad j \in \underline{N},$$

(13)

$$\left\| \sum_{i \in C_{\hat{k}}} \left(\frac{\alpha_i}{\alpha_j} \right)^2 H_{ij} \right\| = \left\| \sum_{i \in J} \left(\frac{\alpha_i}{\alpha_j} \right)^2 H_{ij} \right\| = \hat{\mu}, \qquad j \in J.$$

We see that the solution of problem (3) can always be reduced to the solution of a subproblem corresponding to a strongly connected component of $\mathcal{G}$.

3. Characterization of stability radii. Suppose that $A \in \mathbb{K}^{n \times n}$ is a given matrix with spectrum $\sigma(A)$ in the open left half-plane $\mathbb{C}_- = \{s \in \mathbb{C};\ \mathrm{Re}\, s < 0\}$. Let $N \in \mathbb{N}$, and let $((D_i, E_i))_{i \in \underline{N}}$ be a given family of matrices $D_i \in \mathbb{K}^{n \times \ell_i}$, $E_i \in \mathbb{K}^{q_i \times n}$, $i = 1, \ldots, N$. We will consider uncertain systems described by Ito stochastic differential equations of the form

$$dx(t) = Ax(t)\,dt + \sum_{i=1}^{N} D_i \Delta_i(E_i x(t)) dw_i(t), \tag{14}$$

where $\Delta_1, \ldots, \Delta_N$ are unknown Lipschitzian nonlinearities satisfying

$$\|\Delta_i\|_L < \sigma, \qquad i = 1, \ldots, N. \tag{15}$$

$(w_i(t))_{t \in \mathbb{R}_+}, i = 1, \ldots, N$, are independent zero-mean Wiener processes on a probability space $(\Omega, \mathcal{F}, \mu)$ relative to an increasing family $(\mathcal{F}_t)_{t \in \mathbb{R}_+}$ of σ-algebras $\mathcal{F}_t \subset \mathcal{F}$. Thus, if λ_i denotes the variance of $(w_i(t))_{t \in \mathbb{R}_+}$, we have

$$\mathcal{E}(w_i(t)) = 0, \quad \mathcal{E}((w_i(t) - w_i(s))(w_j(t) - w_j(s))) = \delta_{ij} \lambda_i (t - s), \quad t, s \in \mathbb{R}_+, t > s, i, j \in \underline{N},$$

where δ_{ij} is the Kronecker symbol. For each $i = 1, \ldots, N$, the Euclidean norm is taken on $\mathbb{K}^{q_i}, \mathbb{K}^{\ell_i}$. The disturbances Δ_i vary in

$$\mathrm{Lip}\,(\mathbb{K}^{q_i}, \mathbb{K}^{l_i}) = \{\Delta : \mathbb{K}^{q_i} \longmapsto \mathbb{K}^{\ell_i};\ \Delta(0) = 0 \text{ and } \Delta \text{ is Lipschitzian}\},$$

and the size of each $\Delta_i \in \mathrm{Lip}\,(\mathbb{K}^{q_i}, \mathbb{K}^{l_i})$ is measured by the *Lipschitz norm*

$$\|\Delta_i\|_L = \inf\{\gamma_i > 0;\ \forall y, \hat{y} \in \mathbb{K}^{q_i} : \|\Delta_i(y) - \Delta_i(\hat{y})\|_{\mathbb{K}^{\ell_i}} \leq \gamma_i \|y - \hat{y}\|_{\mathbb{K}^{q_i}}\}.$$

The unknown Δ_i represent uncertainty in the state-dependent gains through which the stationary white noise processes $dw_i(t)$ affect the evolution of the system. The matrix family $((D_i, E_i))_{i \in \underline{N}}$ determines the structure of the perturbations, and $\sigma > 0$ indicates the overall level of the stochastic disturbances. Altogether (14) with constraints (15) describes a *set* of stochastic systems parametrized by $\Delta_i \in \mathrm{Lip}\,(\mathbb{K}^{q_i}, \mathbb{K}^{l_i})$, $\|\Delta_i\|_L < \sigma$ for $i \in \underline{N}$.

Let $L^2(\Omega, \mathbb{K}^m)$ denote the space of square-integrable $\mathbb{K}^m$-valued functions (modulo equivalence) on the probability space $(\Omega, \mathcal{F}, \mu)$. We denote by $L^2_w(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^m))$ the space of nonanticipative stochastic processes $z(\cdot) = (z(t))_{t \in \mathbb{R}_+}$ with respect to $(\mathcal{F}_t)_{t \in \mathbb{R}_+}$ (see, e.g., [9]) satisfying

$$\|z(\cdot)\|^2_{L^2_w} = \mathcal{E}\left(\int_0^\infty \|z(t)\|^2\,dt \right) = \int_0^\infty \mathcal{E}(\|z(t)\|^2)\,dt < \infty, \tag{16}$$

where $\mathcal{E}$ denotes the expectation. For arbitrary $\Delta_i \in \text{Lip}(\mathbb{K}^{q_i}, \mathbb{K}^{l_i})$, $i \in \underline{N}$, and any initial state $x^0 \in \mathbb{K}^n$ there exists a unique solution $x(\cdot) = (x(t))_{t\in\mathbb{R}_+}$ of (14) on $\mathbb{R}_+ = [0, \infty)$ such that $x(0) = x^0$ (see, e.g., [9]). $x(\cdot)$ is a continuous nonanticipative stochastic process with L^2 second moments on every finite interval $[0, T]$:

$$\int_0^T \mathcal{E}(\|x(t)\|^2)\, dt < \infty, \qquad T \geq 0.$$

Many concepts of stability have been studied for stochastic systems. In this paper we consider L^2-stability.

DEFINITION 3.1. *The system* (14) *is said to be* L^2*-stable if, for every* $x^0 \in \mathbb{K}^n$*, the unique solution* $x(\cdot)$ *of* (14) *on* $\mathbb{R}_+ = [0, \infty)$ *with initial value* $x(0) = x^0$ *satisfies*

$$\int_0^\infty \mathcal{E}(\|x(t)\|^2)\, dt < \infty.$$

Our aim is to determine which bounds σ on the perturbations Δ_i ensure that the stability of the deterministic system $\dot{x}(t) = Ax(t)$ is preserved under additive stochastic perturbations of the form $\sum_{i=1}^N D_i \Delta_i(E_i x(t))dw_i(t)$. Let Δ denote the combined perturbation operator

$$\Delta = \bigoplus_1^N \Delta_i \in \text{Lip}(\mathbb{K}^q, \mathbb{K}^\ell), \qquad q = q_1 + \cdots + q_N, \ell = \ell_1 + \cdots + \ell_N. \tag{17}$$

The Lipschitz norm of Δ is given by

$$\|\Delta\|_L = \max_{i\in\underline{N}} \|\Delta_i\|_L.$$

Note that because $\Delta_i(0) = 0$, we have

$$\|\Delta(y)\|^2_{\mathbb{K}^\ell} = \sum_1^N \|\Delta_i(y_i)\|^2_{\mathbb{K}^{\ell_i}} \leq \sum_1^N \|\Delta_i\|_L^2 \|y_i\|^2_{\mathbb{K}^{q_i}} \leq \|\Delta\|_L^2 \|y\|^2_{\mathbb{K}^q}, \qquad y = (y_i)_{i\in\underline{N}} \in \mathbb{K}^q. \tag{18}$$

The maximum $\sigma > 0$ for which all the systems in (14) are L^2-stable is called the stability radius of (14).

DEFINITION 3.2. *The* stochastic stability radius *of* $A \in \mathbb{K}^{n\times n}$ *with respect to the perturbation structure* $((D_i, E_i))_{i\in\underline{N}}$ *and the Wiener processes* $(w_i)_{i\in\underline{N}}$ *is*

$$r^w_{\mathbb{K}}(A; (D_i, E_i)_{i\in\underline{N}}) = \inf\left\{ \left\| \bigoplus_1^N \Delta_i \right\|_L ; \Delta_i \in \text{Lip}(\mathbb{K}^{q_i}, \mathbb{K}^{\ell_i}) \text{ such that (14) is not } L^2\text{-stable} \right\}. \tag{19}$$

Remark 3.3. (i) We have chosen $x^0 \in \mathbb{K}^n$ since we regard (14) as a stochastic perturbation of a deterministic system. However, it is straightforward to extend the theory to any $\mathcal{F}_0$-measurable initial state $x^0 \in L^2(\Omega, \mathbb{K}^n)$.

(ii) A stability radius with respect to *linear* perturbations can be defined analogously by restricting the perturbations Δ_i in (19) to be linear, i.e., $\Delta_i \in \mathcal{L}(\mathbb{K}^{q_i}, \mathbb{K}^{\ell_i})$. It is an open question whether this restriction leads to a different stability radius.

(iii) If the data A, D_i, E_i are real, two stability radii are obtained according to whether one chooses $\mathbb{K} = \mathbb{C}$ (complex perturbations) or $\mathbb{K} = \mathbb{R}$ (only real perturbations) in (19). In a deterministic framework the real and the complex stability radii are, in general, distinct; see [14]. We will show later that they are equal in the present stochastic framework. □

In order to characterize the stochastic stability radius we need the following lemmas.

LEMMA 3.4. *Suppose that* $E \in \mathbb{K}^{q\times n}$ *and*

$$y(t) = Ee^{At}x^0 + \sum_{i=1}^{N} \int_0^t Ee^{A(t-s)} D_i v_i(s) dw_i(s),$$

where $v_i \in L_w^2(\mathbb{R}_+; L_2(\Omega, \mathbb{K}^{\ell_i}))$, $i \in \underline{N}$, *and* $x^0 \in \mathbb{K}^n$. *Then*

$$\mathcal{E}(\|y(t)\|^2) = \|Ee^{At}x^0\|^2 + \sum_{i=1}^{N} \lambda_i \int_0^t \mathcal{E}(\|Ee^{A(t-s)} D_i v_i(s)\|^2)\, ds, \qquad t \in \mathbb{R}_+. \tag{20}$$

Moreover, $y(\cdot) \in L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^q))$ *and*

$$\begin{aligned} \|y(\cdot)\|_{L_w^2}^2 &= \int_0^\infty \mathcal{E}(\|y(t)\|^2)\, dt \\ &= \int_0^\infty \|Ee^{At}x^0\|^2\, dt + \sum_{i=1}^{N} \lambda_i \int_0^\infty \mathcal{E}(\langle D_i v_i(s), P D_i v_i(s)\rangle)\, ds, \end{aligned} \tag{21}$$

where

$$PA + A^*P + E^*E = 0. \tag{22}$$

Proof. The first part is a standard result for stochastic integrals [9]. Now since $\sigma(A) \subset \mathbb{C}_-$, the following integrals are well defined, and we have

$$\int_0^\infty \mathcal{E}(\|y(t)\|^2)\, dt = \int_0^\infty \|Ee^{At}x^0\|^2\, dt + \sum_{i=1}^{N} \lambda_i \int_0^\infty \int_0^t \mathcal{E}(\|Ee^{A(t-s)} D_i v_i(s)\|^2)\, ds dt.$$

By Fubini's theorem

$$\begin{aligned} &\int_0^\infty \int_0^t \mathcal{E}(\|Ee^{A(t-s)} D_i v_i(s)\|^2)\, ds dt \\ &\quad = \int_0^\infty \int_s^\infty \mathcal{E}(\|Ee^{A(t-s)} D_i v_i(s)\|^2)\, dt ds \\ &\quad = \int_0^\infty \mathcal{E}\left(\left\langle D_i v_i(s), \int_s^\infty e^{A^*(t-s)} E^* E e^{A(t-s)}\, dt\, D_i v_i(s)\right\rangle\right) ds. \end{aligned}$$

The result follows since the unique solution $P = P^*$ of (22) is given by

$$P = \int_0^\infty e^{A^*\tau} E^* E e^{A\tau}\, d\tau. \qquad \square \tag{23}$$

Consider the map $\mathbb{L} : L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell)) \to L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^q))$ defined by

$$(\mathbb{L}v(\cdot))(t) = \left(\mathbb{L}\begin{bmatrix} v_1(\cdot) \\ \vdots \\ v_N(\cdot) \end{bmatrix}\right)(t) = \sum_{i=1}^{N} \int_0^t Ee^{A(t-s)} D_i v_i(s) dw_i(s). \tag{24}$$

$\mathbb{L}v(\cdot) \in L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^q))$ for all $v(\cdot) \in L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell))$ by the previous lemma.

LEMMA 3.5. *The linear map* $\mathbb{L} : L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell)) \to L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^q))$ *defined by* (24) *has the operator norm*

$$\|\mathbb{L}\| = \max_{i \in \underline{N}} \left(\lambda_i \|D_i^* \left[\int_0^\infty e^{A^*\tau} E^* E e^{A\tau} \, d\tau \right] D_i \| \right)^{1/2} = \max_{i \in \underline{N}} (\lambda_i \|D_i^* P D_i\|)^{1/2} \tag{25}$$

where P *satisfies* (22).

Proof. If $v(\cdot) \in L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell))$ and $y = \mathbb{L}v$, we have by Lemma 3.4

$$\|\mathbb{L}v\|_{L_w^2}^2 = \int_0^\infty \mathcal{E}(\|y(t)\|^2) \, dt = \sum_{i=1}^N \lambda_i \int_0^\infty \mathcal{E}(\langle D_i v_i(s), P D_i v_i(s) \rangle) \, ds,$$

where P satisfies (22). Hence

$$\|\mathbb{L}v\|_{L_w^2}^2 \le \max_{i \in \underline{N}} (\lambda_i \|D_i^* P D_i\|) \sum_{i=1}^N \int_0^\infty \mathcal{E}(\|v_i(s)\|^2) \, ds = \max_{i \in \underline{N}} (\lambda_i \|D_i^* P D_i\|) \|v\|_{L_w^2}^2.$$

So $\|\mathbb{L}\| \le \max_{i \in \underline{N}} (\lambda_i \|D_i^* P D_i\|)^{1/2}$. Now suppose that $\max_{i \in \underline{N}} (\lambda_i \|D_i^* P D_i\|)^{1/2}$ is achieved for $i = j$ and $v_j \in \mathbb{K}^{\ell_j}$ satisfies $\|v_j\|_{\mathbb{K}^{\ell_j}} = 1$ and $\langle v_j, D_j^* P D_j v_j \rangle = \|D_j^* P D_j\|$. Let $v_i(t) = 0$, $t \in \mathbb{R}_+$, $i \ne j$, and $v_j(\cdot) = \beta(\cdot) v_j$, where $\beta(\cdot) \in L^2(\mathbb{R}_+; \mathbb{R})$, $\|\beta(\cdot)\|_{L^2} = 1$ is chosen arbitrarily. Then $v(\cdot) = (v_i(\cdot))_{i \in \underline{N}} \in L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell))$, $\|v(\cdot)\|_{L_w^2} = 1$, and

$$\begin{aligned} \|\mathbb{L}v\|_{L_w^2}^2 &= \lambda_j \int_0^\infty \mathcal{E}(\langle D_j v_j \beta(s), P D_j v_j \beta(s) \rangle) \, ds \\ &= \lambda_j \|D_j^* P D_j\| \int_0^\infty |\beta(s)|^2 \, ds = \max_{i \in \underline{N}} (\lambda_i \|D_i^* P D_i\|). \end{aligned}$$

This completes the proof. □

THEOREM 3.6. *Suppose that* A *is stable and there exist* $\alpha = (\alpha_i)_{i \in \underline{N}} \in (0, \infty)^{\underline{N}}$, $P \in \mathcal{H}_n^+(\mathbb{K})$ *satisfying*

$$A^* P + P A + \sum_{i=1}^N \alpha_i^2 E_i^* E_i = 0, \tag{26}$$

$$I_{\ell_j} - (\sigma/\alpha_j)^2 \lambda_j D_j^* P D_j \succeq 0, \qquad j \in \underline{N}. \tag{27}$$

Then $r_{\mathbb{K}}^w(A; (D_i, E_i)_{i \in \underline{N}}) \ge \sigma$.

Proof. Let $x^0 \in \mathbb{K}^n$, $\Delta_i \in \text{Lip}(\mathbb{K}^{q_i}, \mathbb{K}^{l_i})$, $i \in \underline{N}$, $\Delta = \oplus_1^N \Delta_i$, and $\|\Delta\|_L < \sigma$, and suppose that $\alpha \in (0, \infty)^{\underline{N}}$, $P \in \mathcal{H}_n^+(\mathbb{K})$ are such that (26), (27) hold. The unique solution $x(\cdot)$ of (14) with initial condition $x(0) = x^0$ satisfies the scaled integral equation

$$x(t) = e^{At} x^0 + \sum_{i=1}^N \int_0^t e^{A(t-s)} D_i^{\alpha_i} \Delta_i^{\alpha_i} (E_i^{\alpha_i} x(s)) dw_i(s), \qquad t \in \mathbb{R}_+, \tag{28}$$

where

$$D_i^{\alpha_i} = \alpha_i^{-1} D_i, \quad E_i^{\alpha_i} = \alpha_i E_i, \quad \Delta_i^{\alpha_i}(\cdot) = \alpha_i \Delta_i(\alpha_i^{-1} \cdot), \quad i \in \underline{N}. \tag{29}$$

The input-output operator $\mathbb{L}^\alpha : L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell)) \to L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^q))$ of the scaled system $(A, (D_i^{\alpha_i}, E_i^{\alpha_i})_{i \in \underline{N}})$ is given by

$$(\mathbb{L}^\alpha v(\cdot))(t) = \left(\mathbb{L}^\alpha \begin{bmatrix} v_1(\cdot) \\ \vdots \\ v_N(\cdot) \end{bmatrix} \right)(t) = \sum_{i=1}^N \int_0^t \begin{bmatrix} E_1^{\alpha_1} \\ \vdots \\ E_N^{\alpha_N} \end{bmatrix} e^{A(t-s)} D_i^{\alpha_i} v_i(s) dw_i(s), \qquad t \in \mathbb{R}_+.$$

Let $u_i^{\alpha_i}(t) = \Delta_i^{\alpha_i}(E_i^{\alpha_i}x(t))$, $y_i^{\alpha_i}(t) = E_i^{\alpha_i}x(t)$, $t \in \mathbb{R}_+$, and

$$
\begin{aligned}
E(\alpha) &= \begin{bmatrix} E_1^{\alpha_1} \\ \vdots \\ E_N^{\alpha_N} \end{bmatrix}, \quad \Delta^\alpha = \bigoplus_1^N \Delta_i^{\alpha_i}; \\
y^\alpha(t) &= \begin{bmatrix} y_1^{\alpha_1}(t) \\ \vdots \\ y_N^{\alpha_N}(t) \end{bmatrix}, \quad u^\alpha(t) = \begin{bmatrix} u_1^{\alpha_1}(t) \\ \vdots \\ u_N^{\alpha_N}(t) \end{bmatrix}, \quad t \in \mathbb{R}_+.
\end{aligned} \tag{30}
$$

Then (28) implies

$$
y^\alpha(t) = E(\alpha)e^{At}x^0 + \sum_{i=1}^N \int_0^t E(\alpha)e^{A(t-s)}D_i^{\alpha_i}u_i^{\alpha_i}(s)dw_i(s), \qquad t \in \mathbb{R}_+. \tag{31}
$$

For every $T > 0$, define the truncations $u_{i,T}^{\alpha_i} \in L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^{\ell_i}))$, $i \in \underline{N}$, and $u_T^\alpha \in L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell))$ by

$$
u_{i,T}^{\alpha_i}(t) = \begin{cases} u_i^{\alpha_i}(t) = \Delta_i^{\alpha_i}(y_i^{\alpha_i}(t)) & \text{if } t \in [0, T], \\ 0 \quad \text{if } t > T, \end{cases} \qquad u_T^\alpha(t) = \begin{bmatrix} u_{1,T}^{\alpha_1}(t) \\ \vdots \\ u_{N,T}^{\alpha_N}(t) \end{bmatrix}.
$$

Then

$$
\begin{aligned}
\|u_T^\alpha\|_{L_w^2}^2 &= \int_0^T \left(\sum_{i=1}^N \int_\Omega \|u_i^{\alpha_i}(t,\omega)\|_{\mathbb{K}^{\ell_i}}^2 \mu(d\omega) \right) dt \\
&\le \int_0^T \left(\sum_{i=1}^N \int_\Omega \|\Delta_i^{\alpha_i}\|_L^2 \|y_i^{\alpha_i}(t,\omega)\|_{\mathbb{K}^{q_i}}^2 \mu(d\omega) \right) dt \\
&\le \|\Delta^\alpha\|_L^2 \int_0^T \mathcal{E}(\|y^\alpha(t)\|^2)\, dt.
\end{aligned} \tag{32}
$$

Let y_T^α denote the output of the scaled system $(A, (D_i^\alpha, E_i^\alpha)_{i \in \underline{N}})$ generated by the input u_T^α with initial condition $x(0) = x^0$:

$$
\begin{aligned}
y_T^\alpha(t) &= E(\alpha)e^{At}x^0 + \sum_{i=1}^N E(\alpha) \int_0^t e^{A(t-s)}D_i^{\alpha_i}u_{i,T}^{\alpha_i}(s)dw_i(s) \\
&= E(\alpha)e^{At}x^0 + (\mathbb{L}^\alpha u_T^\alpha(\cdot))(t), \qquad t \in \mathbb{R}_+.
\end{aligned} \tag{33}
$$

It follows from (31)–(33) that

$$
\begin{aligned}
\left(\int_0^T \mathcal{E}(\|y^\alpha(t)\|^2)\, dt \right)^{1/2} &\le \|y_T^\alpha(\cdot)\|_{L_w^2} \le \|E(\alpha)e^{A(\cdot)}x^0\|_{L_w^2} + \|\mathbb{L}^\alpha\| \|u_T^\alpha(\cdot)\|_{L_w^2} \\
&\le \|E(\alpha)e^{A(\cdot)}x^0\|_{L_w^2} + \|\mathbb{L}^\alpha\|\, \|\Delta^\alpha\|_L \left(\int_0^T \mathcal{E}(\|y^\alpha(t)\|^2)\, dt \right)^{1/2}.
\end{aligned} \tag{34}
$$

By Lemma 3.5

$$\|\mathbb{L}^\alpha\| = \max_{i\in\underline{N}} (\lambda_i \|(D_i^{\alpha_i})^* P D_i^{\alpha_i}\|)^{1/2},$$

where P satisfies (26). Thus from (27) we have $\|\mathbb{L}^\alpha\| \le \sigma^{-1}$, and since $\|\Delta^\alpha\|_L = \|\Delta\|_L < \sigma$, the operator $\mathbb{L}^\alpha\Delta^\alpha$ is a contraction on $L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^q))$ with $\gamma := \|\mathbb{L}^\alpha\|\,\|\Delta^\alpha\|_L < 1$. Hence from (34) for all $T > 0$,

$$\left(\int_0^T \mathcal{E}(\|y^\alpha(t)\|^2)\,dt\right)^{1/2} \le (1-\gamma)^{-1}\|E(\alpha)e^{A(\cdot)}x^0\|_{L_w^2}.$$

Therefore $y^\alpha(\cdot) \in L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^q))$ and $u^\alpha(\cdot) = \Delta^\alpha(y^\alpha(\cdot)) \in L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^\ell))$. Applying Lemma 3.4 (with $E = I_n$, D_i^α instead of D_i and $u^\alpha(\cdot)$ instead of $v(\cdot)$) it follows from (28) that $x(\cdot) \in L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^n))$, and this completes the proof. □

Remark 3.7. Suppose the $\Delta_i(t, y)$ are *time-varying* Lipschitzian nonlinearities, measurable in $(t, y) \in \mathbb{R}_+ \times \mathbb{K}^{q_i}$, satisfying $\Delta_i(t, 0) = 0$, $t \ge 0$, and

$$\|\Delta_i\|_L = \inf\{\gamma_i > 0;\ \forall y, \hat{y} \in \mathbb{K}^{q_i} \forall t \in \mathbb{R}_+ : \|\Delta_i(t, y) - \Delta_i(t, \hat{y})\|_{\mathbb{K}^{\ell_i}} \le \gamma_i \|y - \hat{y}\|_{\mathbb{K}^{q_i}}\} < \sigma.$$

Then the previous proof carries through, showing that no time-varying Lipschitzian perturbations $\Delta_i(t, y)$ of Lipschitz norm smaller than σ can destabilize the system. □

If in the previous theorem condition (27) is satisfied with $\succ$ instead of $\succeq$, then clearly $r_{\mathbb{K}}^w(A; (D_i, E_i)_{i\in\underline{N}}) > \sigma$ follows. Similarly, the equality in (26) may be replaced by an inequality $\preceq$.

COROLLARY 3.8. *Suppose that A is stable and there exist $\alpha \in (0, \infty)^{\underline{N}}$, $P \in \mathcal{H}_n^+(\mathbb{K})$ satisfying*

$$A^*P + PA + \sum_{i=1}^N \alpha_i^2 E_i^* E_i \preceq 0, \tag{35}$$

$$I_{\ell_j} - (\sigma/\alpha_j)^2 \lambda_j D_j^* P D_j \succeq 0 \quad (\textit{resp.}, I_{\ell_j} - (\sigma/\alpha_j)^2\lambda_j D_j^* P D_j \succ 0), \qquad j \in \underline{N}. \tag{36}$$

Then $r_{\mathbb{K}}^w(A; (D_i, E_i)_{i\in\underline{N}}) \ge \sigma$ (resp., $r_{\mathbb{K}}^w(A; (D_i, E_i)_{i\in\underline{N}}) > \sigma$).

Proof. Let $P(\alpha)$ denote the solution of the Liapunov equation (26). Then $0 \preceq P(\alpha) \preceq P$. Hence $P(\alpha)$ satisfies (26) and (27) and $r_{\mathbb{K}}^w(A; (D_i, E_i)_{i\in\underline{N}}) \ge \sigma$ follows from Theorem 3.6. □

For any $J \subset \underline{N}$ and any scaling vector $\alpha^J \in (0, \infty)^J$, define

$$P(\alpha^J) = \int_0^\infty e^{A^*\tau}\left(\sum_{i\in J} \alpha_i^2 E_i^* E_i\right) e^{A\tau}\,d\tau; \tag{37}$$

i.e., $P(\alpha^J) \in \mathcal{H}_n^+(\mathbb{K})$ is the unique solution of

$$A^*P + PA + \sum_{i\in J} \alpha_i^2 E_i^* E_i = 0. \tag{38}$$

Then

$$\begin{aligned}(\lambda_j/\alpha_j^2) D_j^* P(\alpha^J) D_j &= \sum_{i\in J} (\alpha_i/\alpha_j)^2 \lambda_j \int_0^\infty D_j^* e^{A^*\tau} E_i^* E_i e^{A\tau} D_j\,d\tau \\ &= \sum_{i\in J} (\alpha_i/\alpha_j)^2 H_{ij}, \qquad j \in J,\end{aligned} \tag{39}$$

where

$$H_{ij} = \lambda_j \int_0^\infty D_j^* e^{A^*\tau} E_i^* E_i e^{A\tau} D_j d\tau \geq 0, \qquad i, j \in \underline{N}. \tag{40}$$

We are now in a position to prove our main theorem by applying the results of §2 to the family of positive semidefinite matrices (40).

THEOREM 3.9. *Given* $(A, (D_i, E_i)_{i\in\underline{N}})$, $\sigma(A) \subset \mathbb{C}_-$, *and* $(w_i)_{i\in\underline{N}}$ *as in* (14), *the associated stability radius is determined by*

$$r_{\mathbb{K}}^w(A; (D_i, E_i)_{i\in\underline{N}}) = \sup_{\alpha\in(0,\infty)^{\underline{N}}} \left(\max_{j\in\underline{N}} \|(\lambda_j/\alpha_j^2) D_j^* P(\alpha) D_j\|\right)^{-1/2}, \tag{41}$$

where $P(\alpha)$ *is the unique solution of* (26). *If* $r_{\mathbb{K}}^w(A; (D_i, E_i)_{i\in\underline{N}}) < \infty$, *there exists a minimum norm destabilizing perturbation* $\Delta = \oplus_1^N \Delta_i \in \text{Lip}\,(\mathbb{K}^q, \mathbb{K}^l)$, $\|\Delta\|_L = r_{\mathbb{K}}^w$. *Moreover, there exist a subset* $J \subset \underline{N}$ *and a scaling vector* $\alpha^J \in (0, \infty)^J$ *such that*

$$r_{\mathbb{K}}^w(A; (D_i, E_i)_{i\in\underline{N}}) = \left(\max_{j\in J} \|(\lambda_j/\alpha_j^2) D_j^* P(\alpha^J) D_j\|\right)^{-1/2} = r_{\mathbb{K}}^w(A; (D_i, E_i)_{i\in J}), \tag{42}$$

where $P(\alpha^J) \in \mathcal{H}_n^+(\mathbb{K})$ *is the unique solution of* (38).

Proof. By (39)

$$\begin{aligned} \hat{\mu} &:= \inf_{\alpha\in(0,\infty)^{\underline{N}}} \max_{j\in\underline{N}} \|(\lambda_j/\alpha_j^2) D_j^* P(\alpha) D_j\| \\ &= \inf_{\alpha\in(0,\infty)^{\underline{N}}} \max_{j\in\underline{N}} \left\| \sum_{i=1}^N (\alpha_i/\alpha_j)^2 H_{ij} \right\| \geq 0. \end{aligned} \tag{43}$$

If $\hat{\mu} = 0$, then $r_{\mathbb{K}}^w(A; (D_i, E_i)_{i\in\underline{N}}) = \infty$ by Theorem 3.6; hence (41) is satisfied. Moreover, in this trivial case,

$$E_j e^{A\tau} D_j = 0, \qquad \tau \in R_+, j \in \underline{N},$$

so that (42) is satisfied for every singleton $J = \{j\} \subset \underline{N}$ and all $\alpha^J \in (0, \infty)$.

Now assume that $\hat{\mu} > 0$. For every $\alpha \in (0, \infty)^{\underline{N}}$ let $\sigma(\alpha)$ be the largest σ for which (26) and (27) have a joint solution $P \in \mathcal{H}_n(\mathbb{K})$, i.e.,

$$\sigma(\alpha)^2 = \left(\max_{j\in\underline{N}} \|(\lambda_j/\alpha_j^2) D_j^* P(\alpha) D_j\|\right)^{-1} = \left(\max_{j\in\underline{N}} \left\| \sum_{i=1}^N (\alpha_i/\alpha_j)^2 H_{ij} \right\|\right)^{-1} \tag{44}$$

(see (39)), whence $\sup_{\alpha\in(0,\infty)^{\underline{N}}} \sigma(\alpha)^2 = \hat{\mu}^{-1}$. By Theorem 2.4 there exist $J \subset \underline{N}$ and a vector $\alpha^J = (\alpha_j)_{j\in J} \in (0, \infty)^J$ satisfying

$$\hat{\mu} = \left\| \sum_{i\in J} (\alpha_i/\alpha_j)^2 H_{ij} \right\| = \|(\lambda_j/\alpha_j^2) D_j^* P(\alpha^J) D_j\|, \qquad j \in J, \tag{45}$$

where $P(\alpha^J) \in \mathcal{H}_n^+(\mathbb{K})$ is the unique solution of (38). Hence there are $v_j \in \mathbb{K}^{\ell_j}$, $j \in J$, $\|v_j\|_{\mathbb{K}^{\ell_j}} = 1$ such that

$$\left\langle v^j, \left(\sum_{i\in J} (\alpha_i/\alpha_j)^2 H_{ij} \right) v^j \right\rangle = \langle v^j, (\lambda_j/\alpha_j^2) D_j^* P(\alpha^J) D_j v^j \rangle = \hat{\mu}, \qquad j \in J.$$

Setting $\hat{\sigma}^2 = \hat{\mu}^{-1}$ we obtain

$$\langle v_j, (\lambda_j/\alpha_j^2)\hat{\sigma}^2 D_j^* P(\alpha^J) D_j v_j\rangle = 1, \qquad j \in J. \tag{46}$$

Define $\Delta_j(\cdot) \in \text{Lip}\,(\mathbb{K}^{q_j}, \mathbb{K}^{l_j})$ for $j \in \underline{N}$ by

$$\begin{aligned} \Delta_j(y_j) &= \hat{\sigma}\|y_j\| v_j, \qquad y_j \in \mathbb{K}^{q_j}, j \in J, \\ \Delta_i &= 0, \qquad i \in \underline{N}\backslash J. \end{aligned} \tag{47}$$

Then $\|\Delta_j\|_L = \hat{\sigma}$, $j \in J$, and hence $\|\Delta\|_L = \hat{\sigma}$ for $\Delta = \oplus_1^N \Delta_i$. We will show that for this Δ (14) cannot be stable. Assume the contrary; then, for all $x^0 \in \mathbb{K}^n$, the solution $x(\cdot)$ of (14) with $x(0) = x^0$ must satisfy $\int_0^\infty \mathcal{E}(\|x(t)\|^2)\,dt < \infty$. $x(\cdot)$ satisfies the reduced scaled integral equation

$$x(t) = e^{At}x^0 + \sum_{j\in J} \int_0^t e^{A(t-s)} D_j^{\alpha_j} \Delta_j^{\alpha_j}(E_j^{\alpha_j} x(s)) dw_j(s), \tag{48}$$

where $D_j^{\alpha_j}$, $E_j^{\alpha_j}$, and $\Delta_j^{\alpha_j}(\cdot)$ are defined by (29). By assumption $y_j^{\alpha_j}(\cdot) = E_j^{\alpha_j} x(\cdot) \in L_w^2(\mathbb{R}_+; L^2(\Omega, \mathbb{K}^{q_j}))$, $j \in J$. Now

$$\Delta_j^{\alpha_j}(y_j) = \alpha_j \Delta_j(\alpha_j^{-1} y_j) = \hat{\sigma}\|y_j\| v_j, \qquad y_j \in \mathbb{K}^{q_j}, j \in J.$$

Defining $y^{\alpha^J}(\cdot)$ and E^{α^J} by

$$\begin{aligned} y^{\alpha^J}(\cdot) &= (y_j^{\alpha_j}(\cdot))_{j\in J}, \\ E^{\alpha^J} x &= (E_j^{\alpha_j} x)_{j\in J} \in \bigoplus_{j\in J} \mathbb{K}^{q_j}, \qquad x \in \mathbb{K}^n, \end{aligned}$$

we get

$$y^{\alpha^J}(t) = E^{\alpha^J} e^{At} x^0 + \hat{\sigma} \sum_{j\in J} E^{\alpha^J} \int_0^t e^{A(t-s)} D_j^{\alpha_j} v_j \|y_j^{\alpha_j}(s)\| dw_j(s). \tag{49}$$

Application of Lemma 3.4 to (49) yields

$$\begin{aligned} \int_0^\infty \mathcal{E}(\|y^{\alpha^J}(t)\|^2)\,dt &= \int_0^\infty \|E^{\alpha^J} e^{At} x^0\|^2\,dt \\ &\quad + \hat{\sigma}^2 \sum_{j\in J} \lambda_j \langle D_j^{\alpha_j} v_j, P(\alpha^J) D_j^{\alpha_j} v_j\rangle \int_0^\infty \mathcal{E}(\|y_j^{\alpha_j}(s)\|^2)\,ds, \end{aligned}$$

where $P(\alpha^J) \in \mathcal{H}_n^+(\mathbb{K})$ is the unique solution of (38). But then by (46), we have

$$\begin{aligned} \int_0^\infty \mathcal{E}(\|y^{\alpha^J}(t)\|^2)\,dt &= \int_0^\infty \|E^{\alpha^J} e^{At} x^0\|^2\,dt + \sum_{j\in J} \int_0^\infty \mathcal{E}(\|y_j^{\alpha_j}(s)\|^2)\,ds \\ &= \int_0^\infty \|E^{\alpha^J} e^{At} x^0\|^2\,dt + \int_0^\infty \mathcal{E}(\|y^{\alpha^J}(s)\|^2)\,ds \end{aligned}$$

for all $x^0 \in \mathbb{K}^n$. This would imply that $E_j = 0$ for every $j \in J$; hence $P(\alpha^J) = 0$ and $\hat{\mu} = 0$, contrary to our assumption. Therefore, there exists $x^0 \in \mathbb{K}^n$ such that $\int_0^\infty \mathcal{E}(\|x(t)\|^2)\,dt = \infty$ and neither of the two stochastic systems (14) and (48) is L^2-stable. It follows that

$$r_{\mathbb{K}}^w(A; (D_i, E_i)_{i\in\underline{N}}) \le r_{\mathbb{K}}^w(A; (D_i, E_i)_{i\in J}) \le \left(\max_{j\in J} \|\lambda_j D_j^* P(\alpha^J) D_j/\alpha_j^2\|\right)^{-1/2} = \hat{\mu}^{-1/2}.$$

On the other hand, for every $\sigma < \hat{\sigma} = \hat{\mu}^{-1/2}$, i.e., $\sigma^{-2} > \hat{\mu}$, there exists $\alpha \in (0, \infty)^{\underline{N}}$ such that

$$\max_{j \in \underline{N}} \|(\lambda_j/\alpha_j^2) D_j^* P(\alpha) D_j\| = \max_{j \in \underline{N}} \left\| \sum_{i=1}^{N} (\alpha_i/\alpha_j)^2 H_{ij} \right\| < \sigma^{-2}.$$

But this implies $\sigma \leq r_{\mathbb{K}}^w(A; (D_i, E_i)_{i \in \underline{N}})$ by Theorem 3.6 and concludes the proof. □

Note that the destabilizing disturbance Δ defined by (47) is real when the data $(A, (D_i, E_i)_{i \in \underline{N}})$ are real. In this case we can choose $\mathbb{K} = \mathbb{R}$ in (41) and obtain a formula for the real stability radius. However, we may also choose $\mathbb{K} = \mathbb{C}$ so that we obtain the same formula for the complex stability radius since the right-hand side of (41) does not depend on the choice of the field $\mathbb{K}$. Thus we obtain the following corollary.

COROLLARY 3.10. *Under the conditions of Theorem* 3.9, *if the data* $(A, (D_i, E_i)_{i \in \underline{N}})$ *are real, then the complex and the real stability radii coincide:*

$$r_{\mathbb{R}}^w(A; (D_i, E_i)_{i \in \underline{N}}) = r_{\mathbb{C}}^w(A; (D_i, E_i)_{i \in \underline{N}}).$$

Remark 3.11. In the deterministic case, stability radii for *complex* and *real* multiperturbations are, in general, not the same, not even in the single-perturbation case ($N = 1$); see [14]. Moreover, the scaling technique does not provide a characterization of the complex stability radius but yields only a lower bound; see [16]. In the stochastic case, however, the scaling technique works and we have, as a consequence of (41),

$$r_{\mathbb{K}}^w(A; (D_i, E_i)_{i \in \underline{N}}) = \sup_{\alpha \in (0,\infty)^{\underline{N}}} \|\mathbb{L}^\alpha\|^{-1} = \sup_{\alpha \in (0,\infty)^{\underline{N}}} r_{\mathbb{K}}^w(A; D^\alpha, E^\alpha)$$

for $\mathbb{K} = \mathbb{R}$ and $\mathbb{K} = \mathbb{C}$. The application of the theorems of §2 to obtain this result is based on the simple formula (25) for the norm of $\mathbb{L}$. The corresponding characterization in the deterministic case is much more complicated and involves a parametrized Riccati equation (see [14]) instead of the single Liapunov equation (22) without parameters.

One reason for the basic difference between the deterministic and the stochastic case lies in the fact that there is no deterministic counterpart to the fundamental equation (20) (on which all our results are built). □

For later use we note the following characterization of the stability radius in terms of strict inequalities.

COROLLARY 3.12. *Given* $(A, (D_i, E_i)_{i \in \underline{N}})$ *and* $(w_i)_{i \in \underline{N}}$ *as in* (14), *the following statements are equivalent for* $\sigma \geq 0$:

(i) $\sigma(A) \subset \mathbb{C}_-$ *and* $r_{\mathbb{K}}^w(A; (D_i, E_i)_{i \in \underline{N}}) > \sigma$;

(ii) *there exists* $\alpha_i > 0$, $i \in \underline{N}$, *and* $X \in \mathcal{H}_n^+(\mathbb{K})$ *satisfying*

$$XA + A^*X + E(\alpha)^*E(\alpha) \prec 0, \tag{50}$$

$$I_{\ell_i} - \lambda_i(\sigma/\alpha_i)^2 D_i^* X D_i \succ 0, \qquad i = 1, \ldots, N. \tag{51}$$

Proof. Suppose (i) and choose $\sigma' \in (\sigma, r_{\mathbb{K}}^w(A; (D_i, E_i)_{i \in \underline{N}}))$. By Theorem 3.9 there exists $\alpha \in (0, \infty)^{\underline{N}}$ such that

$$(\sigma'/\alpha_i)^2 \lambda_i D_i^* \left[\int_0^\infty e^{A^*\tau} E(\alpha)^* E(\alpha) e^{A\tau} d\tau \right] D_i \preceq I_{\ell_i}, \qquad i \in \underline{N}.$$

But then

$$(\sigma/\alpha_i)^2 \lambda_i D_i^* \left[\int_0^\infty e^{A^*\tau} E(\alpha)^* E(\alpha) e^{A\tau} d\tau \right] D_i \prec I_{\ell_i}, \qquad i \in \underline{N},$$

and hence there exists $\varepsilon > 0$ such that

$$(\sigma/\alpha_i)^2 \lambda_i D_i^* \left[\int_0^\infty e^{A^*\tau}(E(\alpha)^* E(\alpha) + \varepsilon I_n) e^{A\tau} d\tau \right] D_i \prec I_{\ell_i}, \qquad i \in \underline{N}.$$

Setting $X = \int_0^\infty e^{A^*\tau}(E(\alpha)^* E(\alpha) + \varepsilon I_n) e^{A\tau} d\tau$ it follows that

$$XA + A^* X + E(\alpha)^* E(\alpha) + \varepsilon I_n = 0.$$

Hence $X \succ 0$ satisfies (ii).

Conversely, (ii) $\Rightarrow$ (i) follows first from the fact that (50) implies $\sigma(A) \subset \mathbb{C}_-$ and then from application of Corollary 3.8. □

Remark 3.13. Let $\mathcal{G}$ denote the directed graph with node set $\underline{N}$ and set of directed arcs $\mathcal{A} = \{(i, j) \in \underline{N}^2;\ H_{ij} \neq 0\}$, where the Hermitian matrices $H_{ij} \in \mathcal{H}_n^+(\mathbb{K})$ are defined by (40). By Theorem 2.4

$$\hat{\mu} = \max_{k \in \underline{K}} \mu_k \quad \mu_k = \min_{\alpha \in (0,\infty)^{C_k}} \max_{j \in C_k} \left\| \sum_{i \in C_k} \left(\frac{\alpha_i}{\alpha_j} \right)^2 H_{ij} \right\|, \quad k \in \underline{K},$$

where $\hat{\mu}$ is defined by (43) and C_k, $k \in \underline{K}$, are the strongly connected components of $\mathcal{G}$. Since $\mu_k^{-1/2} = r_{\mathbb{K}}^w(A; (D_i, E_i)_{i \in C_k})$ we have

$$r_{\mathbb{K}}^w(A; (D_i, E_i)_{i \in \underline{N}}) = \max_{k \in \underline{K}} r_{\mathbb{K}}^w(A; (D_i, E_i)_{i \in C_k}).$$

Thus the L^2-stability of the uncertain stochastic system (14) is equivalent to the L^2-stability of each uncertain stochastic system corresponding to the connected components of $\mathcal{G}$:

$$dx(t) = Ax(t)\,dt + \sum_{i \in C_k} D_i \Delta_i(E_i x(t)) dw_i(t), \qquad \|\Delta_i\|_L < \sigma, i \in C_k, k = 1, \ldots, K.$$

This reduces our original problem to the separate investigation of K uncertain stochastic systems with strongly connected perturbation structures. In particular, the subset J in Theorem 3.9 can be chosen in a strongly connected component C_k with $\mu_k = \hat{\mu}$. The question of determining conditions under which no further reduction beyond the connected components is possible, i.e., $J = C_k$, has been dealt with in Proposition 2.2. □

4. Maximizing the stability radius by dynamic output feedback. In this section we investigate how the stability radius of a stochastically perturbed system can be improved by dynamic output feedback. For this we introduce a control term into the system equation (14) and add a measurement equation. We consider controlled stochastic systems described by Ito equations of the form

$$(52)\quad dx(t) = Ax(t)dt + \sum_{i=1}^{N} D_i \Delta_i(E_i x(t)) dw_i(t) + Bu(t)dt, \quad y(t) = Cx(t), \quad t \in \mathbb{R}_+,$$

where $B \in \mathbb{K}^{n \times m}$ and $C \in \mathbb{K}^{p \times n}$ are the input and output matrices, respectively, and the other variables and matrices are of the form specified in the previous section.

Remark 4.1. Extensions of this problem, including control-dependent noise and noise corrupting the output, should be considered in the framework of a general stochastic H^∞ control. While the development of a comprehensive H^∞ control theory for stochastic systems requires substantial new work, it can be built on the results presented here. This will be the subject of future work. □

The compensator takes the form

$$d\hat{x}(t) = H\hat{x}(t)dt + Gy(t)dt, \qquad u(t) = F\hat{x}(t) + Ky(t), \tag{53}$$

where $(H, G, F, K) \in \mathbb{K}^{\hat{n}\times\hat{n}} \times \mathbb{K}^{\hat{n}\times p} \times \mathbb{K}^{m\times\hat{n}} \times \mathbb{K}^{m\times p}$ and the dimension $\hat{n} \geq 0$ is arbitrary. The resulting overall system is

$$\begin{bmatrix} dx(t) \\ d\hat{x}(t) \end{bmatrix} = \begin{bmatrix} A + BKC & BF \\ GC & H \end{bmatrix} \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix} dt + \sum_{i=1}^{N} \begin{bmatrix} D_i \\ 0 \end{bmatrix} \Delta_i \left([E_i \;\; 0] \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix} \right) dw_i(t).$$

We will use the notation

$$\mathcal{A} = \begin{bmatrix} A + BKC & BF \\ GC & H \end{bmatrix} \in \mathbb{K}^{(n+\hat{n})\times(n+\hat{n})}, \mathcal{D}_i = \begin{bmatrix} D_i \\ 0 \end{bmatrix} \in \mathbb{K}^{(n+\hat{n})\times\ell_i},$$

$$\mathcal{E}_i = [E_i \;\; 0] \in \mathbb{K}^{q_i\times(n+\hat{n})},$$

and $\overline{x} = \left[\begin{smallmatrix} x \\ \hat{x} \end{smallmatrix}\right] \in \mathbb{K}^{n+\hat{n}}$. Then the above system can be written

$$d\overline{x}(t) = \mathcal{A}\overline{x}(t)dt + \sum_{i=1}^{N} \mathcal{D}_i \Delta_i(\mathcal{E}_i\overline{x}(t))dw_i(t). \tag{54}$$

For all compensators (53), arbitrary $\Delta_i \in \text{Lip}\,(\mathbb{K}^{q_i}, \mathbb{K}^{l_i})$, $i \in \underline{N}$, and any $\overline{x}^0 \in \mathbb{K}^{n+\hat{n}}$ there exists a unique solution $\overline{x}(\cdot) = (\overline{x}(t))_{t\in\mathbb{R}_+}$ of (54) on $\mathbb{R}_+$ such that $\overline{x}(0) = \overline{x}_0$ [9].

Our aim is to determine conditions for the existence of dynamic compensators of the form (53) that stabilize the system and achieve a stability radius $r^w_{\mathbb{K}}(\mathcal{A}; (\mathcal{D}_i, \mathcal{E}_i)_{i\in\underline{N}}) > \sigma$ for a given $\sigma > 0$. We follow an approach based on inequalities similar to that which Gahinet developed in his approach to the H^∞ control problem; see [10], [11]. We proceed in two steps. First we derive some necessary conditions, and then we show that these conditions are also sufficient for building a stabilizing compensator of dimension n which achieves the required stability radius. We will make use of the following criterion for the positive definiteness of Hermitian block matrices.

LEMMA 4.2. *Let* $\mathcal{X} = \left[\begin{smallmatrix} S & N \\ N^* & Q \end{smallmatrix}\right]$, *where* $S \in \mathcal{H}_k(\mathbb{K})$, $Q \in \mathcal{H}_\ell(\mathbb{K})$, $N \in \mathbb{K}^{k\times\ell}$. *Then*

$$\mathcal{X} \succ 0 \quad \Leftrightarrow \quad Q \succ 0 \;\; \text{and} \;\; S - NQ^{-1}N^* \succ 0.$$

THEOREM 4.3. *Given* $\sigma > 0$ *and a compensator* (53) *such that* $\sigma(\mathcal{A}) \subset \mathbb{C}_-$ *and* $r^w_{\mathbb{K}}(\mathcal{A}; (\mathcal{D}_i, \mathcal{E}_i)_{i\in\underline{N}}) > \sigma$. *Then there exist* $\alpha \in (0, \infty)^{\underline{N}}$, $\gamma, \delta > 0$, $R, S \in \mathcal{H}^+_n(\mathbb{K})$ *such that*

$$AR + RA^* + RE(\alpha)^*E(\alpha)R - BB^*/\gamma^2 \prec 0, \tag{55}$$

$$SA + A^*S + E(\alpha)^*E(\alpha) - C^*C/\delta^2 \prec 0, \tag{56}$$

$$R \succ 0 \;\; \text{and} \;\; S \succ R^{-1}, \tag{57}$$

$$I_{\ell_i} - \lambda_i(\sigma/\alpha_i)^2 D_i^* S D_i \succ 0, \qquad i = 1, \ldots, N, \tag{58}$$

where (*see* (30)) $E(\alpha)^*E(\alpha) = \sum_{i=1}^N \alpha_i^2 E_i^* E_i$.

Proof. By Corollary 3.12 there exists $\alpha \in (0, \infty)^{\underline{N}}$, $\mathcal{X} \in \mathcal{H}^+_{n+\hat{n}}(\mathbb{K})$, $\mathcal{X} \succ 0$, such that

$$\mathcal{X}\mathcal{A} + \mathcal{A}^*\mathcal{X} + \mathcal{E}(\alpha)^*\mathcal{E}(\alpha) \prec 0, \tag{59}$$

$$I_{\ell_i} - \lambda_i(\sigma/\alpha_i)^2 \mathcal{D}_i^* \mathcal{X} \mathcal{D}_i \succ 0, \qquad i = 1, \ldots, N. \tag{60}$$

Writing

$$\mathcal{X} = \begin{bmatrix} S & N \\ N^* & Q \end{bmatrix}, \qquad \mathcal{X}^{-1} = \begin{bmatrix} R & M \\ M^* & P \end{bmatrix}$$

with $R, S \in \mathcal{H}^+_n(\mathbb{K})$, we obtain

$$SR + NM^* = I_n, \qquad N^*R + QM^* = 0.$$

Since $\mathcal{X} \succ 0$, we have

$$S \succ 0, \quad Q \succ 0, \quad S - NQ^{-1}N^* \succ 0,$$
$$R \succ 0, \quad P \succ 0, \quad R - MP^{-1}M^* \succ 0.$$

Now $SR - NQ^{-1}N^*R = I_n$ and $\mathcal{D}_i^* = [D_i^* \ \ 0]$. Hence

$$\begin{aligned} &0 \prec R^{-1} = S - NQ^{-1}N^* \preceq S, \\ &I_{\ell_i} - \lambda_i(\sigma/\alpha_i)^2 D_i^* S D_i \succ 0, \qquad i = 1, \ldots, N, \end{aligned} \tag{61}$$

because of (60). Moreover,

$$\mathcal{X} \begin{bmatrix} I_n & R \\ 0 & M^* \end{bmatrix} = \begin{bmatrix} S & I_n \\ N^* & 0 \end{bmatrix}. \tag{62}$$

Multiplying (59) on the left by $\left[\begin{smallmatrix} I_n & 0 \\ R & M \end{smallmatrix}\right]$ and on the right by $\left[\begin{smallmatrix} I_n & R \\ 0 & M^* \end{smallmatrix}\right]$ yields

$$\begin{aligned} &\begin{bmatrix} S & N \\ I_n & 0 \end{bmatrix} \begin{bmatrix} A + BKC & BF \\ GC & H \end{bmatrix} \begin{bmatrix} I_n & R \\ 0 & M^* \end{bmatrix} \\ &\quad + \begin{bmatrix} I_n & 0 \\ R & M \end{bmatrix} \begin{bmatrix} (A + BKC)^* & C^*G^* \\ F^*B^* & H^* \end{bmatrix} \begin{bmatrix} S & I_n \\ N^* & 0 \end{bmatrix} \\ &\quad + \begin{bmatrix} I_n & 0 \\ R & M \end{bmatrix} \begin{bmatrix} E(\alpha)^*E(\alpha) & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I_n & R \\ 0 & M^* \end{bmatrix} \\ &\quad + \begin{bmatrix} I_n & 0 \\ R & M \end{bmatrix} \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix} \begin{bmatrix} I_n & R \\ 0 & M^* \end{bmatrix} = 0 \end{aligned} \tag{63}$$

for some

$$\Pi = \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix} \succ 0.$$

Writing out these equations, we obtain

$$(A + BKC)^*S + C^*G^*N^* + S(A + BKC) + NGC + E(\alpha)^*E(\alpha) + \Pi_{11} = 0, \tag{64}$$

(65)
$$R(A+BKC)^*+MF^*B^*+(A+BKC)R+BFM^*+RE(\alpha)^*E(\alpha)R+[R\ M]\Pi\begin{bmatrix} R \\ M^* \end{bmatrix}=0,$$

(66)
$$(A+BKC)^*+SAR+NGCR+SBFM^*+NHM^*+E(\alpha)^*E(\alpha)R+\Pi_{11}R+\Pi_{12}M^*=0.$$

Equation (64) is equivalent to

$$A^*S+SA+E(\alpha)^*E(\alpha)-C^*C/\delta^2+(C^*/\delta+\delta(NG+SBK))(C^*/\delta+\delta(NG+SBK))^*$$
$$-\delta^2(NG+SBK)(NG+SBK)^*=-\Pi_{11},$$

and (65) is equivalent to

$$AR+RA^*+RE(\alpha)^*E(\alpha)R$$
$$-BB^*/\gamma^2+(B^*/\gamma+\gamma(FM^*+KCR))^*(B^*/\gamma+\gamma(FM^*+KCR))$$
$$-\gamma^2(FM^*+KCR)^*(FM^*+KCR)=-[R\ M]\Pi\begin{bmatrix} R \\ M^* \end{bmatrix}.$$

Note that $[R\ \ M]$ has full row rank so that $[R\ \ M]\Pi\left[\begin{smallmatrix} R \\ M^* \end{smallmatrix}\right]\succ 0$. Therefore, choosing γ and δ sufficiently small we obtain (55) and (56). These inequalities and (61) are still satisfied if for $\varepsilon>0$ sufficiently small we replace S with $S+\varepsilon I_n$. Denoting the modified S by the same symbol we get $0\prec R^{-1}\prec S$, and thus (55)–(58) are satisfied. □

Remark 4.4. Equations (55) and (56) hold for some $\gamma,\delta>0$ if and only if

$$AR+RA^*+RE(\alpha)^*E(\alpha)R\prec 0 \quad \text{on } \ker B^*,$$
$$SA+A^*S+E(\alpha)^*E(\alpha)\prec 0 \quad \text{on } \ker C,$$

so it is possible to state an equivalent theorem which does not involve γ and δ. □

THEOREM 4.5. *Suppose that* (55)–(58) *hold for some* $\alpha\in(0,\infty)^{\underline{N}}$, $\sigma>0$, $\gamma>0$, $\delta>0$, $R,S\in\mathcal{H}_n^+(\mathbb{K})$. *Then there exists an n-dimensional compensator* $(H,G,F,K)\in\mathbb{K}^{n\times n}\times\mathbb{K}^{n\times p}\times\mathbb{K}^{m\times n}\times\mathbb{K}^{m\times p}$ *such that* $\sigma(\mathcal{A})\subset\mathbb{C}_-$ *and* $r_{\mathbb{K}}^w(\mathcal{A};(\mathcal{D}_i,\mathcal{E}_i)_{i\in\underline{N}})>\sigma$.

Proof. Choose

(67) $$K=0,\quad F=-B^*R^{-1}/\gamma^2,\quad G=-N^{-1}C^*/\delta^2,\quad N=R^{-1}-S,$$

(68) $$M=R,\quad \Pi_{11}=-(A^*S+SA+E(\alpha)^*E(\alpha)-2C^*C/\delta^2),\quad \Pi_{12}=-\Pi_{11},$$

(69) $$\Pi_{22}=\Pi_{11}-R^{-1}[AR+RA^*+RE(\alpha)^*E(\alpha)R-2BB^*/\gamma^2]R^{-1}.$$

An easy calculation shows that (64) and (65) hold. Moreover, it follows from (55) and (56) that $0\prec\Pi_{11}\prec\Pi_{22}$ and $\Pi_{11}-\Pi_{12}\Pi_{22}^{-1}\Pi_{12}=\Pi_{11}-\Pi_{11}\Pi_{22}^{-1}\Pi_{11}\succ 0$; hence

$$\Pi=\begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{12} & \Pi_{22} \end{bmatrix}\succ 0.$$

Finally, we have by assumption $0\prec R^{-1}\prec S$ so that $N=(I-SR)R^{-1}$ and $M=R$ are invertible. Therefore,

(70) $$H=-N^{-1}[A^*R^{-1}+SA+NGC+SBF+E(\alpha)^*E(\alpha)]$$

is the unique $H \in \mathbb{K}^{n\times n}$ satisfying (66) (with specifications (67)–(69)). Altogether we see that, with the above choices, (63) holds with $\Pi \succ 0$. Now $\left[\begin{smallmatrix} I_n & R \\ 0 & M^* \end{smallmatrix}\right] = \left[\begin{smallmatrix} I_n & R \\ 0 & R \end{smallmatrix}\right]$ is invertible. Multiplication of (63) on the left by $\left[\begin{smallmatrix} I_n & 0 \\ R & M \end{smallmatrix}\right]^{-1}$ and on the right by $\left[\begin{smallmatrix} I_n & R \\ 0 & M^* \end{smallmatrix}\right]^{-1}$ yields equation (59) with

$$\mathcal{X} = \begin{bmatrix} S & I_n \\ N^* & 0 \end{bmatrix} \begin{bmatrix} I_n & R \\ 0 & M^* \end{bmatrix}^{-1} = \begin{bmatrix} S & N \\ N & -N \end{bmatrix}. \tag{71}$$

□

Remark 4.6. (i) The above theorems show that if a compensator of any order $\hat{n}$ stabilizes the system with a stability radius greater than σ, then this can always be achieved by a compensator of order n. Moreover, for this compensator the feedthrough matrix K may be taken to be zero. We do not address the problem of reduced-order observers but expect that a development similar to [11] (in the deterministic case) is possible.

(ii) For $\sigma \geq 0$, let $\mathcal{A}_\sigma$ denote the set of all pairs $(R, S) \in \mathcal{H}_n(\mathbb{K}) \times \mathcal{H}_n(\mathbb{K})$ such that (55)–(58) hold for some $\gamma > 0$, $\delta > 0$, $\alpha \in (0, \infty)^{\underline{N}}$. By the construction in the proof of Theorem 4.3, for any stabilizing compensator (H, G, F, K) which achieves a stability radius $r^w_{\mathbb{K}}(\mathcal{A}; (\mathcal{D}_i, \mathcal{E}_i)_{i\in\underline{N}}) > \sigma$, there exists an $(R, S) \in \mathcal{A}_\sigma$ obtained from the solutions $\mathcal{X}$ of the matrix inequalities (59) and (60). Conversely, by the construction in the proof of Theorem 4.5, for every given $(R, S) \in \mathcal{A}_\sigma$ there is an associated stabilizing compensator (H, G, F, K) which achieves a stability radius greater than σ. □

For later use we add another remark in which an alternative formula for the system matrix H of the observer is derived; cf. (70).

Remark 4.7. The gap between the Riccati inequality (55) and the corresponding Riccati equation is measured by the operator

$$\Pi_R = -[AR + RA^* + RE(\alpha)^*E(\alpha)R - BB^*/\gamma^2] \succ 0. \tag{72}$$

Using specifications (67)–(69) we get

$$SA + SBF = S(A - BB^*R^{-1}/\gamma^2) = -S[RA^*R^{-1} + RE(\alpha)^*E(\alpha) + \Pi_R R^{-1}],$$

and so

$$\begin{aligned} &A^*R^{-1} + SA + SBF + E(\alpha)^*E(\alpha) \\ &\quad = (I_n - SR)A^*R^{-1} + (I_n - SR)E(\alpha)^*E(\alpha) - S\Pi_R R^{-1}. \end{aligned}$$

Using this equation and $N^{-1} = R(I - SR)^{-1}$ to transform (70) we get

$$-H = RA^*R^{-1} + RE(\alpha)^*E(\alpha) - R(I_n - SR)^{-1}S\Pi_R R^{-1} + GC.$$

It follows from (72) and $F = -B^*R^{-1}/\gamma^2$ that

$$\begin{aligned} H &= [AR - BB^*/\gamma^2 + \Pi_R]R^{-1} - GC + R(I_n - SR)^{-1}S\Pi_R R^{-1} \\ &= [A - BB^*R^{-1}/\gamma^2] - GC + [I_n + R(I_n - SR)^{-1}S]\Pi_R R^{-1} \\ &= A + BF - GC + (R^{-1} - S)^{-1}R^{-1}\Pi_R R^{-1}. \end{aligned} \tag{73}$$

Thus H is the sum of the usual matrix of the observer-based compensator system $A+BF-GC$ and a correction term $(R^{-1} - S)^{-1}R^{-1}\Pi_R R^{-1}$ depending on the gap between the Riccati inequality (55) and the corresponding equality. □

Theorems 4.3 and 4.5 together yield a complete characterization of the stability radii which can be achieved by dynamic output feedback applied to the system (52). We conclude this section by discussing this point.

For any $\alpha \in (0, \infty)^{\underline{N}}$, $\sigma \geq 0$, $\gamma > 0$, $\delta > 0$, define

$$\mathcal{A}_{\sigma,\gamma,\delta,\alpha} = \{(R, S) \in \mathcal{H}_n(\mathbb{K}) \times \mathcal{H}_n(\mathbb{K});\ (55)\text{–}(58) \text{ hold}\},$$

$$\mathcal{A}_{\sigma,\gamma,\delta} = \bigcup_{\alpha \in (0,\infty)^{\underline{N}}} \mathcal{A}_{\sigma,\gamma,\delta,\alpha},$$

$$\sigma(\gamma, \delta) = \sup\{\sigma : \mathcal{A}_{\sigma,\gamma,\delta} \neq \emptyset\}.$$

In the following remark we collect some properties of $\mathcal{A}_{\sigma,\gamma,\delta,\alpha}$ and $\sigma(\gamma, \delta)$.

Remark 4.8. $\mathcal{A}_{\sigma,\gamma,\delta,\alpha}$ is an open subset of $\mathcal{H}_n(\mathbb{K}) \times \mathcal{H}_n(\mathbb{K})$ for all $\sigma \geq 0$, $\gamma > 0$, $\delta > 0$, $\alpha \in (0, \infty)^{\underline{N}}$. If $\alpha \in (0, \infty)^{\underline{N}}$ is fixed, the set $\mathcal{A}_{\sigma,\gamma,\delta,\alpha}$ increases as the parameters $\sigma, \gamma, \delta > 0$ decrease:

$$\sigma_2 \geq \sigma_1 \geq 0, \quad \gamma_2 \geq \gamma_1 > 0, \quad \text{and } \delta_2 \geq \delta_1 > 0 \quad \Rightarrow \quad \mathcal{A}_{\sigma_2,\gamma_2,\delta_2,\alpha} \subseteq \mathcal{A}_{\sigma_1,\gamma_1,\delta_1,\alpha}.$$

As a consequence we have

$$0 < \gamma_1 \leq \gamma_2 \quad \text{and} \quad 0 < \delta_1 \leq \delta_2 \quad \Rightarrow \quad \sigma(\gamma_1, \delta_1) \geq \sigma(\gamma_2, \delta_2). \qquad \square$$

DEFINITION 4.9. $\bar{r}(A; (D_i, E_i)_{i \in \underline{N}}; B, C) = \sup\{r^w_{\mathbb{K}}(\mathcal{A}; (\mathcal{D}_i, \mathcal{E}_i)_{i \in \underline{N}});$ the compensator (H, G, F, K) is stabilizing$\}$ *is said to be the* supreme stability radius *for the uncertain stochastic system* (52).

As a consequence of the previous two theorems we obtain the following characterization of the supreme stability radius.

COROLLARY 4.10.

$$\bar{r}(A; (D_i, E_i)_{i \in \underline{N}}; B, C) = \lim_{(\gamma,\delta) \downarrow (0,0)} \sigma(\gamma, \delta) = \sup\{\sigma(\gamma, \delta);\ \gamma > 0, \delta > 0\}.$$

For the computation of $\sigma(\gamma, \delta)$ the following description of the sets $\mathcal{A}_{\sigma,\gamma,\delta,\alpha}$ in terms of linear matrix inequalities is useful.

LEMMA 4.11. *Given* $\alpha \in (0, \infty)^{\underline{N}}$, $\sigma \geq 0$, $\gamma > 0$, $\delta > 0$, *the set* $\mathcal{A}_{\sigma,\gamma,\delta,\alpha}$ *consists of all Hermitian matrix pairs* $(R, S) \in \mathcal{H}_n(\mathbb{K}) \times \mathcal{H}_n(\mathbb{K})$ *which satisfy the following linear matrix inequalities:*

$$\begin{bmatrix} AR + RA^* - BB^*/\gamma^2 & RE(\alpha)^* \\ E(\alpha)R & -I_q \end{bmatrix} \prec 0, \tag{74}$$

$$SA + A^*S + E(\alpha)^*E(\alpha) - C^*C/\delta^2 \prec 0, \tag{75}$$

$$R \succ 0, \quad S \succ 0, \quad \begin{bmatrix} R & I_n \\ I_n & S \end{bmatrix} \succ 0, \tag{76}$$

$$I_{\ell_i} - \lambda_i(\sigma/\alpha_i)^2 D_i^* S D_i \succ 0 \qquad i = 1, \ldots, N. \tag{77}$$

In particular, $\mathcal{A}_{\sigma,\gamma,\delta,\alpha}$ *is convex.*

Proof. The equivalence of (55)–(58) and (74)–(77) is obtained by application of Lemma 4.2. The convexity of $\mathcal{A}_{\sigma,\gamma,\delta,\alpha}$ then follows. $\square$

For a discussion of linear matrix inequalities and their numerical solution, see [2].

Remark 4.12. In the case of *single complex* perturbations ($N = 1$) the deterministic counterpart of the problem considered in this section leads to a singular H^∞-optimal control problem which can be solved via Riccati equations (see [13]) or linear matrix inequalities (see [11]). The maximization of the *real* stability radius by dynamic output feedback is still an unsolved problem, even for $N = 1$. In the (complex) multiperturbation case, the deterministic version of our problem leads to an optimal μ-synthesis problem since the stability radius with respect to multiperturbations can be characterized via the μ-function [14]. To our knowledge this problem is still unsolved. Our solution of the stochastic problem is based on the fact that the scaling technique works and yields a characterization of the stochastic stability radius in terms of matrix inequalities; see Theorem 3.9. A similar result is not available for deterministic multiperturbations; see Remark 3.11. □

5. Replacing Riccati inequalities with Riccati equations. In this section we explore the possibility of replacing the Riccati inequality (55) with a Riccati equation. *Throughout the section we will assume that* (A, B) *is stabilizable and* (A, C) *is detectable.*

As a starting point we take the Riccati inequality

$$PA + A^*P + E(\alpha)^*E(\alpha) - PBB^*P/\gamma^2 \prec 0. \tag{78}$$

Since (A, B) is assumed to be stabilizable, this inequality always has positive definite solutions. Moreover, every such solution $P \in \mathcal{H}_n(\mathbb{K})$ is invertible and solves (78) if and only if $R = P^{-1}$ solves (55). We will also consider the usual linear-quadratic control Riccati equation,

$$XA + A^*X + E(\alpha)^*E(\alpha) - XBB^*X/\gamma^2 = 0, \tag{$\mathrm{ARE}_{\alpha,\gamma}$}$$

and its ε-approximations,

$$XA + A^*X + E(\alpha)^*E(\alpha) - XBB^*X/\gamma^2 + \varepsilon^2 I_n = 0. \tag{79}$$

The following lemma summarizes some useful and well-known properties of these equations; see, e.g., [12], [5].

LEMMA 5.1. *Suppose that* (A, B) *is stabilizable and* $\gamma > 0$, $\alpha \in (0, \infty)^{\underline{N}}$. *Then*

(i) *for each* $\varepsilon > 0$ (79) *has a unique solution* $X_{\alpha,\gamma}(\varepsilon)$ *in* $\mathcal{H}_n^+(\mathbb{K})$, *and* $X_{\alpha,\gamma}(\varepsilon) \succ 0$. *For every solution* $P \succ 0$ *of* (78) *there exists* ε *such that* $X_{\alpha,\gamma}(\varepsilon) \prec P$.

(ii) $(\mathrm{ARE}_{\alpha,\gamma})$ *has a unique maximal solution* $X_{\alpha,\gamma} \in \mathcal{H}_n^+(\mathbb{K})$, *and this solution is characterized among all other Hermitian solutions of* $(\mathrm{ARE}_{\alpha,\gamma})$ *by the property*

$$\sigma(A - BB^*X_{\alpha\gamma}/\gamma^2) \subset \overline{\mathbb{C}_-}.$$

Moreover, $X_{\alpha,\gamma} \prec X_{\alpha,\gamma}(\varepsilon)$, *for all* $\varepsilon > 0$.

(iii) *If* $\varepsilon \downarrow 0$, *then* $X_{\alpha,\gamma}(\varepsilon) \downarrow X_{\alpha,\gamma}$.

Remark 5.2. It can be shown that the maximal solution $X_{\alpha,\gamma}$ of $(\mathrm{ARE}_{\alpha,\gamma})$ is stabilizing if and only if

$$\ker(\iota\omega I_n - A) \cap \ker E_1 \cap \cdots \cap \ker E_N = \{0\}, \qquad \omega \in \mathbb{R}. \tag{80}$$

□

Consider the following set of conditions for $S \in \mathcal{H}_n(\mathbb{K})$:

$$SA + A^*S + E(\alpha)^*E(\alpha) - C^*C/\delta^2 \prec 0, \tag{81}$$

$$S \succ X_{\alpha,\gamma}, \tag{82}$$

$$I_{\ell_i} - \lambda_i(\sigma/\alpha_i)^2 D_i^* S D_i \succ 0, \qquad i = 1, \ldots, N, \tag{83}$$

where $X_{\alpha,\gamma}$ denotes the maximal solution of $(\mathrm{ARE}_{\alpha,\gamma})$. For $\sigma \geq 0$, $\gamma > 0$, $\delta > 0$, let

$$\mathcal{A}^R_{\sigma,\gamma,\delta,\alpha} = \{S \in \mathcal{H}_n(\mathbb{K}); \quad (81)\text{–}(83) \text{ hold}\},$$

$$\mathcal{A}^R_{\sigma,\gamma,\delta} = \bigcup_{\alpha \in (0,\infty)^{\underline{N}}} \mathcal{A}^R_{\sigma,\gamma,\delta,\alpha},$$

$$\sigma^R(\gamma,\delta) = \sup\{\sigma : \mathcal{A}^R_{\sigma,\gamma,\delta} \neq \emptyset\}.$$

The following proposition is derived by means of Lemma 5.1.

PROPOSITION 5.3. *Suppose that* (A, B) *is stabilizable and* $\gamma > 0$, $\delta > 0$. *Then*

$$\sigma^R(\gamma,\delta) = \sigma(\gamma,\delta).$$

Proof. Let $\sigma < \sigma(\gamma,\delta)$. Then there exist $\alpha \in (0,\infty)^{\underline{N}}$, $(R, S) \in \mathcal{H}_n(\mathbb{K}) \times \mathcal{H}_n(\mathbb{K})$ such that (55)–(58) hold. By Lemma 5.1 $R^{-1} \succ X_{\alpha,\gamma}$ since R^{-1} solves (78). But $S \succ R^{-1}$ and hence $S \succ X_{\alpha,\gamma}$. So $S \in \mathcal{A}^R_{\sigma,\gamma,\delta}$ and thus $\sigma \leq \sigma^R(\gamma,\delta)$. This shows that $\sigma^R(\gamma,\delta) \geq \sigma(\gamma,\delta)$.

Conversely, suppose that $\sigma < \sigma^R(\gamma,\delta)$. Then there exist $\alpha \in (0,\infty)^{\underline{N}}$, $S \in \mathcal{H}_n(\mathbb{K})$ such that (81)–(83) hold. By Lemma 5.1 and (82) there exists $\varepsilon > 0$ such that $X = X_{\alpha,\gamma}(\varepsilon) \prec S$. Hence $(X^{-1}, S) \in \mathcal{A}_{\sigma,\gamma,\delta}$ and thus $\sigma \leq \sigma(\gamma,\delta)$. This shows that $\sigma(\gamma,\delta) \geq \sigma^R(\gamma,\delta)$ and concludes the proof. □

For any $\sigma < \sigma^R(\gamma,\delta)$ we now construct, *via the Riccati equation* $(\mathrm{ARE}_{\alpha,\gamma})$, a compensator of order n so that the stability radius of the overall system is greater than σ. For this, the maximal solution of $(\mathrm{ARE}_{\alpha,\gamma})$ must be stabilizing; i.e., condition (80) must be satisfied.

PROPOSITION 5.4. *Suppose that* (80), (81)–(83) *hold for some* $S \in \mathcal{H}_n(\mathbb{K})$ *and given* $\sigma \geq 0$, $\gamma > 0$, $\delta > 0$, $\alpha \in (0,\infty)^{\underline{N}}$. *Then the compensator* (53) *defined by*

$$(84)\quad H = A + BF - GC, \quad K = 0, \quad F = -B^*X_{\alpha,\gamma}/\gamma^2, \quad G = -(X_{\alpha,\gamma} - S)^{-1}C^*/\delta^2$$

achieves a stability radius $r^w_{\mathbb{K}}(\mathcal{A}; (\mathcal{D}_i, \mathcal{E}_i)_{i\in\underline{N}}) > \sigma$.

Proof. Replacing R^{-1} with $X_{\alpha,\gamma}$ in the formulas (67) and (73), then using equation (72) for Π_R, we see that the compensator (H, G, F, K) in (84) coincides with the compensator defined in the proof of Theorem 4.5 by (67) and (70). Let

$$\mathcal{X} = \begin{bmatrix} S & N \\ N & -N \end{bmatrix}.$$

Then the same calculation as in the proof of Theorem 4.5 yields

$$(85)\qquad \mathcal{X}\mathcal{A} + \mathcal{A}^*\mathcal{X} + \mathcal{E}(\alpha)^*\mathcal{E}(\alpha) = -\Pi,$$

where now

$$\Pi = \begin{bmatrix} \Pi_{11} & -\Pi_{11} \\ -\Pi_{11} & \Pi_{11} + X_{\alpha,\gamma}BB^*X_{\alpha,\gamma}/\gamma^2 \end{bmatrix}, \quad \Pi_{11} = -(SA + A^*S + E(\alpha)^*E(\alpha) - 2C^*C/\delta^2).$$

$\Pi_{11} \succ 0$, but note that in contrast to the development in the proof of Theorem 4.5 we only have $\mathcal{X} \succeq 0$ and $\Pi \succeq 0$. Now $\overline{x} \in \ker \Pi$ if and only if $\overline{x} = \begin{bmatrix} x_1 \\ x_1 \end{bmatrix}$ with $x_1 \in \ker B^*X_{\alpha,\gamma}$. We will now show that $\sigma(\mathcal{A}) \subset \mathbb{C}_-$. Suppose there exist $\overline{x} \in \mathbb{C}^{2n}$, $\overline{x} \neq 0$, and $\lambda \in \mathbb{C}$ with $\operatorname{Re}\lambda \geq 0$ such that $\mathcal{A}\overline{x} = \lambda\overline{x}$. Multiplying (85) on the right by $\overline{x}$ and on the left by $\overline{x}^*$, we obtain

$$2(\operatorname{Re}\lambda)\langle\overline{x}, \mathcal{X}\overline{x}\rangle + \langle\overline{x}, \Pi\overline{x}\rangle \leq 0.$$

This can only be the case if $\overline{x} \in \ker \Pi$, i.e., $\overline{x} = \left[\begin{smallmatrix} x_1 \\ x_1 \end{smallmatrix}\right]$ with $x_1 \in \ker B^* X_{\alpha,\gamma}$. But then the first component of the equation $\mathcal{A}\overline{x} = \lambda \overline{x}$ reads $(A - BB^* X_{\alpha,\gamma}/\gamma^2)x_1 = \lambda x_1$, and this contradicts assumption (80), which implies that $(A - BB^* X_{\alpha,\gamma}/\gamma^2) \subset \mathbb{C}_-$. So $\sigma(\mathcal{A}) \subset \mathbb{C}_-$, and we conclude from Corollary 3.8 and

$$\mathcal{X}\mathcal{A} + \mathcal{A}^*\mathcal{X} + \mathcal{E}(\alpha)^*\mathcal{E}(\alpha) \preceq 0,$$

$$I_{\ell_i} - \lambda_i(\sigma/\alpha_i)^2 \mathcal{D}_i^* \mathcal{X} \mathcal{D}_i = I_{\ell_i} - \lambda_i(\sigma/\alpha_i)^2 D_i^* S D_i \succ 0, \qquad i = 1, \ldots, N,$$

that $r_{\mathbb{K}}^w(\mathcal{A}; (\mathcal{D}_i, \mathcal{E}_i)_{i \in \underline{N}}) > \sigma$. □

In the deterministic case the Liapunov inequality (81) takes the form of a Riccati inequality, and it is possible to replace this by a Riccati equation. Here this is not possible, in general, since the requirements $S \succ R^{-1}$ and $I_{\ell_i} - \lambda_i(\sigma/\alpha_i)^2 D_i^* S D_i \succ 0$, $i = 1, \ldots, N$, work against each other as S increases or decreases. This is an essential difference between the deterministic and stochastic cases. We illustrate it in the following example.

Example 5.5. Consider the perturbed stochastic system

$$\begin{aligned} dx_1(t) &= (-x_1(t) + x_2(t))dt + \Delta_1(x_1(t))dw_1(t), \\ dx_2(t) &= (x_2(t) + u(t))dt + \Delta_2(x_1(t))dw_2(t), \\ y(t) &= x_2(t), \end{aligned}$$

i.e.,

$$A = \begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = [0 \;\; 1], \quad E_1 = E_2 = [1 \;\; 0],$$

$$D_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Note that (A, B) is controllable, (A, C) is detectable, and $\ker(\iota\omega I_2 - A) \cap \ker E_1 = \{0\}$ for all $\omega \in \mathbb{R}$. We write

$$R = \begin{bmatrix} r_{11} & r_{12} \\ r_{12}^* & r_{22} \end{bmatrix}, \qquad S = \begin{bmatrix} s_{11} & s_{12} \\ s_{12}^* & s_{22} \end{bmatrix}.$$

Equality in (56) or (81) takes the form

$$-2s_{11} + (\alpha_1^2 + \alpha_2^2) = 0, \tag{86}$$

$$s_{11} = 0, \tag{87}$$

$$s_{12} + s_{12}^* + 2s_{22} - 1/\delta^2 = 0. \tag{88}$$

We see that, for arbitrary parameters $\alpha_1 > 0$, $\alpha_2 > 0$, $\delta > 0$, there does not exist any solution $S \in \mathcal{H}_2(\mathbb{K})$ of (86)–(88). Therefore, we cannot replace the inequalities in (56) or (81) with equalities.

We will now determine $\overline{r} = \overline{r}(A; (D_i, E_i)_{i=1,2}; B, C)$ for the present example and do so via inequalities rather than the Riccati equation ($\mathrm{ARE}_{\alpha,\gamma}$) and (81)–(83). Suppose $\sigma < \overline{r}$. Since $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ is a basis for both $\ker B^*$ and $\ker C$, by Theorem 4.3 and Remark 4.4, there exist $R, S \in \mathcal{H}_2(\mathbb{K})$ and $\alpha_1 > 0$, $\alpha_2 > 0$ such that

$$-2s_{11} + (\alpha_1^2 + \alpha_2^2) < 0, \qquad -2r_{11} + r_{12}^* + r_{12} + (\alpha_1^2 + \alpha_2^2)(r_{11}^2 + |r_{12}|^2) < 0, \tag{89}$$

(90) $$1 - \lambda_1\sigma^2 s_{11}/\alpha_1^2 > 0, \qquad 1 - \lambda_2\sigma^2 s_{22}/\alpha_2^2 > 0.$$

Therefore, $\sigma^2 < \alpha_1^2/(\lambda_1 s_{11}) < 2\alpha_1^2/[\lambda_1(\alpha_1^2 + \alpha_2^2)] < 2/\lambda_1$ and hence $\bar{r} \leq [2/\lambda_1]^{1/2}$.

We will now demonstrate that equality holds. For this let $\lambda_1\sigma^2 = 2(1-\varepsilon) < 2, 1 > \varepsilon > 0$. We must show that there exist $\alpha_1 > 0, \alpha_2 > 0$ such that (89) and (90) hold and $S \succ R^{-1} \succ 0$. Normalizing $\alpha_1^2 + \alpha_2^2 = 1$, (89) becomes

(91) $$s_{11} > 1/2, \qquad (r_{11} - 1)^2 + |r_{12} + 1|^2 - 2 < 0.$$

Choose

$$\alpha_2 = \varepsilon, \quad s_{11} = (1 + \varepsilon/2)/2, \quad r_{12} = s_{12} = 0,$$
$$r_{11} = 2(1 + \varepsilon/3)^{-1}, \quad s_{22} = \varepsilon^2/(2\lambda_2\sigma^2), \quad r_{22} = 3/(2s_{22}).$$

Then it is easy to see that (91) is satisfied. Since

$$1 - \lambda_1\sigma^2 s_{11}/\alpha_1^2 = 1 - (1 + \varepsilon/2)(1 + \varepsilon)^{-1}, \qquad 1 - \lambda_2\sigma^2 s_{22}/\alpha_2^2 = 1 - 1/2,$$

(90) also holds. So all we need to show is $S \succ R^{-1} \succ 0$. But

$$S - R^{-1} = \begin{bmatrix} (1+\varepsilon/2)/2 & 0 \\ 0 & \varepsilon^2/(2\lambda_2\sigma^2) \end{bmatrix} - \begin{bmatrix} (1+\varepsilon/3)/2 & 0 \\ 0 & \varepsilon^2/(3\lambda_2\sigma^2) \end{bmatrix} \succ 0. \qquad \square$$

We conclude the paper with a proposition concerning the *state feedback case*.

PROPOSITION 5.6. *Suppose $p = n$ and C is invertible. Then the following conditions are equivalent for $\sigma \geq 0$:*

(i) *There exists a static feedback matrix $K \in \mathbb{K}^{m\times n}$ such that $\sigma(A + BKC) \subset \mathbb{C}_-$ and $r_{\mathbb{K}}^w(A + BKC; (D_i, E_i)_{i\in\underline{N}}) > \sigma$.*

(ii) *There exists a dynamic output feedback of the form* (53) *such that $\sigma(\mathcal{A}) \subset \mathbb{C}_-$ and $r_{\mathbb{K}}^w(\mathcal{A}; (\mathcal{D}_i, \mathcal{E}_i)_{i\in\underline{N}}) > \sigma$.*

(iii) *There exist $\alpha \in (0,\infty)^{\underline{N}}$, $\gamma > 0$, $R \in \mathcal{H}_n^+(\mathbb{K})$, $R \succ 0$, such that*

$$AR + RA^* + RE(\alpha)^*E(\alpha)R - BB^*/\gamma^2 \prec 0,$$
$$I_{\ell_i} - \lambda_i(\sigma/\alpha_i)^2 D_i^* R^{-1} D_i \succ 0, \qquad i = 1, \ldots, N.$$

(iv) *There exist $\alpha \in (0,\infty)^{\underline{N}}$, $\gamma > 0$ such that the maximal solution $X_{\alpha,\gamma}$ of* ($\mathrm{ARE}_{\alpha,\gamma}$) *satisfies*

$$I_{\ell_i} - \lambda_i(\sigma/\alpha_i)^2 D_i^* X_{\alpha,\gamma} D_i \succ 0, \qquad i = 1, \ldots, N.$$

Proof. Clearly (i) $\Rightarrow$ (ii). (ii) $\Rightarrow$ (iii) follows from Theorem 4.3, and (iii) $\Rightarrow$ (iv) follows from Lemma 5.1. Now suppose (iv). Select an $S \in \mathcal{H}_n(\mathbb{K})$ such that $S \succ X_{\alpha,\gamma}$ and (83) are satisfied. Then choose $\delta > 0$ sufficiently small so that (81) holds. Then (81)–(83) hold, and (ii) follows from Proposition 5.3 and Theorem 4.3. Hence (ii), (iii), and (iv) are equivalent.

It remains to prove, e.g., (iii) $\Rightarrow$ (i). Suppose (iii). Then there exist $\alpha \in (0,\infty)^{\underline{N}}$, $\gamma > 0$, $P \in \mathcal{H}_n^+(\mathbb{K})$, $P \succ 0$, such that

$$PA + A^*P + E(\alpha)^*E(\alpha) - PBB^*P/\gamma^2 \prec 0,$$
$$I_{\ell_i} - \lambda_i(\sigma/\alpha_i)^2 D_i^* P D_i \succ 0, \qquad i = 1, \ldots, N.$$

From the first inequality it follows that

$$P(A - BB^*P/\gamma^2) + (A - BB^*P/\gamma^2)^*P + E(\alpha)^*E(\alpha) \prec 0.$$

Applying Corollary 3.12 we get $r_{\mathbb{K}}^w(A - BB^*P/\gamma^2; (D_i, E_i)_{i\in\underline{N}}) > \sigma$. Thus is suffices to choose $K = -B^*PC^{-1}/\gamma^2$ in order to obtain (i). $\square$

In particular, the proposition shows that it is not possible to obtain a larger stability radius by *dynamic state feedback* than that which can be achieved by *static state feedback*.

In [8] the equivalence (i) $\Leftrightarrow$ (iv) was proven for the special case $E_i = E, i \in \underline{N}$.

6. Appendix. We use the notation introduced in §2. In order to prove Theorem 2.1 we need the following lemmas. First note that the functions f_j, $j \in \underline{N}$, and f in (4) are constant on rays in $(0, \infty)^{\underline{N}}$:

$$f(r\alpha) = f(\alpha), \qquad \alpha \in (0, \infty)^{\underline{N}}, r > 0.$$

LEMMA 6.1. *Suppose that $(\alpha^k)_{k\in N}$ is a sequence in $S_+^N = \{\alpha \in (0, \infty)^{\underline{N}}; \|\alpha\| = 1\}$ and $(f(\alpha^k))_{k\in\mathbb{N}}$ is bounded. If $j \in \underline{N}$ can be reached from $i \in \underline{N}$ via a directed path in $\mathcal{G}$, then*

$$\lim_{k\to\infty} \alpha_j^k = 0 \quad \Rightarrow \quad \lim_{k\to\infty} \alpha_i^k = 0. \tag{92}$$

Proof. By induction it suffices to prove (92) for the case where $(i, j) \in \mathcal{A}$. But in this case $H_{ij} \neq 0$, and thus (92) follows from the boundedness of $(f(\alpha^k))_{k\in\mathbb{N}}$, because

$$f(\alpha^k) \geq f_j(\alpha^k) \geq \left(\frac{\alpha_i^k}{\alpha_j^k}\right)^2 \|H_{ij}\|. \qquad \square$$

Applying this lemma we obtain the following existence result.

PROPOSITION 6.2. *Suppose that $\mathcal{G}$ is strongly connected. Then there exists $\hat{\alpha} \in S_+^N$ such that*

$$f(\hat{\alpha}) = \hat{\mu}. \tag{93}$$

Proof. Since f is constant on rays, it suffices to consider f on S_+^N. Let (α^k) be a minimizing sequence for f on S_+^N which converges toward some limit $\hat{\alpha}$ in the closure of S_+^N. We have only to prove that $\hat{\alpha} \in S_+^N$, i.e., $\hat{\alpha}_i > 0$ for all $i \in \underline{N}$. But if $\hat{\alpha}_j = 0$ for some $j \in \underline{N}$, then since $(f(\alpha^k))$ is bounded and j can be reached from every $i \in \underline{N}$ via a directed path in $\mathcal{G}$, we must have $\hat{\alpha} = 0$ by the previous lemma. On the other hand $\hat{\alpha} \neq 0$ because (α^k) is a sequence in S_+^N. The contradiction shows that $\hat{\alpha}_i > 0$ for all $i \in \underline{N}$. $\square$

LEMMA 6.3. *Suppose H_0, $H \in \mathcal{H}_\ell^+(\mathbb{K})$, and $r_1 < r_2$ are such that*

$$\|H_0 + r_1 H\| = \|H_0 + r_2 H\|.$$

Then

$$\|H_0 + rH\| = \|H_0\|, \qquad 0 \leq r \leq r_2.$$

Proof. Let $v \in \mathbb{K}^\ell$, $\|v\| = 1$, be such that $\langle v, (H_0 + r_1 H)v\rangle = \|H_0 + r_1 H\|$. Then

$$\langle v, (H_0 + r_1 H)v\rangle \leq \langle v, (H_0 + r_2 H)v\rangle \leq \|H_0 + r_2 H\| = \langle v, (H_0 + r_1 H)v\rangle.$$

Hence $Hv = 0$. So $\|H_0 + r_2 H\| = \langle v, H_0 v\rangle \leq \|H_0\|$. But $r \mapsto \|H_0 + rH\|$ is increasing, and hence the result follows. $\square$

Proof of Theorem 2.1. By Proposition 6.2 there is a vector $z \in (0, \infty)^{\underline{N}}$ such that

$$\max_{j\in\underline{N}} \left\| \sum_{i\in\underline{N}} \left(\frac{z_i}{z_j}\right)^2 H_{ij} \right\| = \hat{\mu}.$$

Among all these minimizing vectors we choose one, denoted by $\hat{z}$, for which the number of $j \in \underline{N}$ satisfying

$$\left\| \sum_{i\in\underline{N}} \left(\frac{\hat{z}_i}{\hat{z}_j}\right)^2 H_{ij} \right\| = \hat{\mu} \tag{94}$$

is minimal. Let J be the set of these $j \in \underline{N}$. Then

$$\left\| \sum_{i \in \underline{N}} \left(\frac{\hat{z}_i}{\hat{z}_j} \right)^2 H_{ij} \right\| < \hat{\mu}, \qquad j \in \underline{N} \backslash J. \tag{95}$$

For $r \in [0, 1]$, define $\hat{z}(r) \in (0, \infty)^{\underline{N}}$ by setting $\hat{z}_j(r) = \hat{z}_j$ if $j \in J$ and $\hat{z}_j(r) = r\hat{z}_j$ if $j \in \underline{N} \backslash J$. For r sufficiently close to 1, say $r \in [\hat{r}, 1]$ with $\hat{r} < r$, the inequalities (95) still hold when $\hat{z}$ is replaced by $\hat{z}(r)$. For these r,

$$\left\| \sum_{i \in \underline{N}} \left(\frac{\hat{z}_i(r)}{\hat{z}_j(r)} \right)^2 H_{ij} \right\| = \left\| \sum_{i \in J} \left(\frac{\hat{z}_i}{\hat{z}_j} \right)^2 H_{ij} + r^2 \sum_{i \in \underline{N} \backslash J} \left(\frac{\hat{z}_i}{\hat{z}_j} \right)^2 H_{ij} \right\| \leq \hat{\mu}, \qquad j \in J.$$

But by the minimality assumption on J, none of the above inequalities can be strict for $r \in [\hat{r}, 1]$. Applying Lemma 6.3 with $H_0 = \sum_{i \in J} \left(\frac{\hat{z}_i}{\hat{z}_j} \right)^2 H_{ij}$ and $H = \sum_{i \in \underline{N} \backslash J} \left(\frac{\hat{z}_i}{\hat{z}_j} \right)^2 H_{ij}$ we conclude that

$$\left\| \sum_{i \in \underline{N}} \left(\frac{\hat{z}_i(r)}{\hat{z}_j(r)} \right)^2 H_{ij} \right\| = \left\| \sum_{i \in J} \left(\frac{\hat{z}_i}{\hat{z}_j} \right)^2 H_{ij} \right\| = \hat{\mu}, \qquad r \in [0, 1], \qquad j \in J. \tag{96}$$

Setting $\hat{\alpha}_i = \hat{z}_i(\hat{r})$, $i \in \underline{N}$ it follows from (96), (95) that (6) is satisfied. □

Proof of Proposition 2.2. Let $\hat{\alpha}$ be a minimum such that $J = \{j \in \underline{N};\ f_j(\hat{\alpha}) = \hat{\mu}\}$ has a minimum number of elements, and suppose $J \neq \underline{N}$. Arguing as in the proof of Theorem 2.1 we obtain (see (96))

$$\left\| \sum_{i \in \underline{N}} \left(\frac{\hat{\alpha}_i}{\hat{\alpha}_j} \right)^2 H_{ij} \right\| = \left\| \sum_{i \in J} \left(\frac{\hat{\alpha}_i}{\hat{\alpha}_j} \right)^2 H_{ij} \right\| = \hat{\mu}, \qquad j \in J. \tag{97}$$

If $\hat{\mu} = 0$, (9) is trivially satisfied. Therefore we may assume $\hat{\mu} \neq 0$. Choose $v^j \in \mathbb{K}^{\ell_j}$, $\|v^j\| = 1$, such that

$$\left\| \sum_{i \in J} \left(\frac{\hat{\alpha}_i}{\hat{\alpha}_j} \right)^2 H_{ij} \right\| = \left\langle v^j, \left(\sum_{i \in J} \left(\frac{\hat{\alpha}_i}{\hat{\alpha}_j} \right)^2 H_{ij} \right) v^j \right\rangle.$$

Then

$$v^j \in \left(\ker \left(\sum_{i \in J} \left(\frac{\hat{\alpha}_i}{\hat{\alpha}_j} \right)^2 H_{ij} \right) \right)^{\perp} = \left(\bigcap_{i \in J} \ker H_{ij} \right)^{\perp}, \qquad j \in J,$$

$$\left\langle v^j, \left(\sum_{i \in \underline{N} \backslash J} \left(\frac{\hat{\alpha}_i}{\hat{\alpha}_j} \right)^2 H_{ij} \right) v^j \right\rangle = 0.$$

It follows that $v^j \in \bigcap_{i \in \underline{N} \backslash J} \ker H_{ij} \cap (\bigcap_{i \in J} \ker H_{ij})^{\perp}$ for all $j \in J$, and this contradicts assumption (8). Thus $J = \underline{N}$, and the proposition is proven. □

In the case where $\mathcal{G}$ is not strongly connected, for every $\varepsilon > 0$ we define the set

$$X(\varepsilon) = \left\{ \alpha \in (0, \infty)^{\underline{N}};\ \forall k \in \underline{K-1} : i \in C_k \wedge j \in C_{k+1} \Rightarrow \frac{\alpha_i}{\alpha_j} < \varepsilon \right\}. \tag{98}$$

The proof of the following statement is straightforward.

LEMMA 6.4. *Suppose that, for each* $k \in \underline{K}$, *a positive vector* $z^k \in (0, \infty)^{C_k}$ *is given. Choose* $r_k > 0$ *for all* $k \in \underline{K}$ *such that*

$$\max_{i \in C_k} r_k (z^k)_i < \min_{j \in C_{k+1}} r_{k+1} (z^{k+1})_j, \qquad k = 1, \ldots, K-1,$$

and define $\alpha \in (0, \infty)^{\underline{N}}$ *by*

$$\alpha_i = r_k \varepsilon^{K-k} (z^k)_i, \qquad k \in \underline{K}, i \in C_k.$$

Then

$$\alpha \in X(\varepsilon) \quad \textit{and} \quad \forall i, j \in C_k : \frac{\alpha_i}{\alpha_j} = \frac{(z^k)_i}{(z^k)_j}. \tag{99}$$

LEMMA 6.5. *Given a family of vectors* $z^k \in (0, \infty)^{C_k}$, $k \in \underline{K}$, *satisfying*

$$\max_{j \in C_k} \left\| \sum_{i \in C_k} \left(\frac{(z^k)_i}{(z^k)_j} \right)^2 H_{ij} \right\| = \mu_k, \qquad k \in \underline{K}, \tag{100}$$

(*where* μ_k *is defined by* (12)), *there exists, for every* $\delta > 0$, *a vector* $\alpha \in (0, \infty)^{\underline{N}}$ *such that*

$$f(\alpha) \leq \max_{k \in \underline{K}} \mu_k + \delta \quad \textit{and} \quad \forall k \in \underline{K} \ \forall i, j \in C_k : \frac{\alpha_i}{\alpha_j} = \frac{(z^k)_i}{(z^k)_j}. \tag{101}$$

Proof. Suppose that $z^k \in (0, \infty)^{C_k}$, $k \in \underline{K}$, satisfy (100), and let $\delta > 0$. Then

$$\max_{k \in \underline{K}} \max_{j \in C_k} \left\| \sum_{i \in C_k} \left(\frac{(z^k)_i}{(z^k)_j} \right)^2 H_{ij} \right\| = \max_{k \in \underline{K}} \mu_k.$$

For every $\varepsilon > 0$, there exists, by Lemma 6.4, a vector $\alpha = \alpha(\varepsilon) \in (0, \infty)^{\underline{N}}$ such that (99) is satisfied. It follows that

$$\begin{aligned} f(\alpha(\varepsilon)) &= \max_{k \in \underline{K}} \max_{j \in C_k} \left\| \sum_{h=1}^{k} \sum_{i \in C_h} \left(\frac{\alpha(\varepsilon)_i}{\alpha(\varepsilon)_j} \right)^2 H_{ij} \right\| \\ &\leq \max_{k \in \underline{K}} \max_{j \in C_k} \left\| \sum_{h=1}^{k-1} \sum_{i \in C_h} \left(\frac{\alpha(\varepsilon)_i}{\alpha(\varepsilon)_j} \right)^2 H_{ij} \right\| + \max_{k \in \underline{K}} \mu_k. \end{aligned}$$

But for all $\varepsilon \in (0, 1)$ we have

$$i \in C_h, \quad j \in C_k, \quad \text{and} \quad h < k \quad \Rightarrow \quad \frac{\alpha(\varepsilon)_i}{\alpha(\varepsilon)_j} \leq \varepsilon.$$

Choosing $\varepsilon \in (0, 1)$ such that

$$\max_{k \in \underline{K}} \max_{j \in C_k} \left\| \sum_{h=1}^{k-1} \sum_{i \in C_h} \varepsilon^2 H_{ij} \right\| \leq \delta$$

we obtain

$$f(\alpha(\varepsilon)) \leq \max_{k \in \underline{K}} \mu_k + \delta.$$

This concludes the proof. □

Proof of Theorem 2.4. By Proposition 6.2 there always exists a family of vectors z^k, $k \in \underline{K}$, satisfying (100). Hence by Lemma 6.5 for $\delta > 0$ there exists $\alpha \in (0, \infty)^{\underline{N}}$ such that

$$f(\alpha) \leq \max_{k \in \underline{K}} \mu_k + \delta.$$

But for every $\alpha \in (0, \infty)^{\underline{N}}$,

$$f(\alpha) = \max_{k \in \underline{K}} \max_{j \in C_k} \left\| \sum_{h=1}^{k} \sum_{i \in C_h} \left(\frac{\alpha_i}{\alpha_j} \right)^2 H_{ij} \right\| \geq \max_{k \in \underline{K}} \max_{j \in C_k} \left\| \sum_{i \in C_k} \left(\frac{\alpha_i}{\alpha_j} \right)^2 H_{ij} \right\| \geq \max_{k \in \underline{K}} \mu_k.$$

Since $\hat{\mu} = \inf_{\alpha \in (0,\infty)^{\underline{N}}} f(\alpha)$, we conclude that $\hat{\mu} = \max_{k \in \underline{K}} \mu_k$. The second part of Theorem 2.4 follows from this and Theorem 2.1. □

REFERENCES

[1] L. ARNOLD AND V. WIHSTUTZ, EDS., *Lyapunov Exponents, Proceedings of a Workshop*, Lecture Notes in Math. 1186, Springer-Verlag, New York, 1986.

[2] S. BOYD, L. EL GHAOUI, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, Society for Industrial and Applied Mathematics, Philadelphia, 1994.

[3] W.-K. CHEN, *Applied Graph Theory*, 2nd ed., North-Holland, Amsterdam, New York, Oxford, 1970.

[4] F. COLONIUS AND W. KLIEMANN, *Stability radii and Lyapunov exponents*, in Control of Uncertain Systems, Progr. Systems Control Theory 6, D. Hinrichsen and B. Mårtensson, eds., Birkhäuser, Basel, 1990, pp. 19–55.

[5] D. F. DELCHAMPS, *A note on the analyticity of the Riccati metric*, in Algebraic and Geometric Methods in Linear Systems Theory, Lecture Notes in Appl. Math. 18, C. I. Byrnes and C. F. Martin, eds., American Mathematical Society, Providence, RI, 1980, pp. 37–41.

[6] J. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. FRANCIS, *State space solutions to standard H_2 and H^∞ control problems*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 831–847.

[7] A. EL BOUHTOURI AND A. J. PRITCHARD, *Stability radius of linear systems with respect to stochastic perturbations*, Systems Control Lett., 19 (1992), pp. 29–33.

[8] ———, *A Riccati equation approach to maximizing the stability radius of a linear system by state feedback under structured stochastic Lipschitzian perturbations*, Systems Control Lett., 21 (1993), pp. 475–484.

[9] A. FRIEDMAN, *Stochastic Differential Equations and Applications*, Probab. Math. Stat. 28, Academic Press, Boston, 1975.

[10] P. GAHINET, *A convex parametrization of suboptimal H_∞ controllers*, in Proc. CDC, 1992, pp. 937–942.

[11] P. GAHINET AND P. APKARIAN, *A linear matrix inequality approach to H_∞ control*, International Journal of Robust and Nonlinear Control, 4 (1994), pp. 421–448.

[12] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *On the Hermitian solutions of the symmetric algebraic Riccati equation*, SIAM J. Control Optim., 24 (1986), pp. 1323–1334.

[13] D. HINRICHSEN, A. J. PRITCHARD, AND S. TOWNLEY, *Riccati equation approach to maximizing the complex stability radius by state feedback*, Internat. J. Control, 52 (1990), pp. 769–794.

[14] D. HINRICHSEN AND A. J. PRITCHARD, *Real and complex stability radii: A survey*, in Control of Uncertain Systems, Progr. Systems Control Theory, D. Hinrichsen and B. Mårtensson, eds., Birkhäuser, Basel, 1990, pp. 119–162.

[15] ———, *Stability margins for systems with deterministic and stochastic uncertainty*, in Proc. 33rd Conf. Decision and Control, Florida, 1994, pp. 3825–3836.

[16] A. PACKARD AND J. DOYLE, *The complex structured singular value*, Automatica, 29 (1993), pp. 71–109.

[17] J. J. WILLEMS AND J. C. WILLEMS, *Feedback stabilizability of stochastic systems with state and control dependent noise*, Automatica, 12 (1976), pp. 277–283.

[18] ———, *Robust stabilization of uncertain systems*, SIAM J. Control Optim., 21 (1983), pp. 352–374.

[19] W. M. WONHAM, *Optimal stationary control of a linear system with state dependent noise*, SIAM J. Control Optim., 5 (1967), pp. 486–500.

[20] G. ZAMES, *Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control AC-26 (1981), pp. 301–320.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 1999–2023, December 1996

008

LINEARIZATION OF DISCRETE-TIME SYSTEMS*

E. ARANDA-BRICAIRE†, Ü. KOTTA‡, AND C. H. MOOG§

Abstract. The algebraic formalism developed in this paper unifies the study of the accessibility problem and various notions of feedback linearizability for discrete-time nonlinear systems. The accessibility problem for nonlinear discrete-time systems is shown to be easy to tackle by means of standard linear algebraic tools, whereas this is not the case for nonlinear continuous-time systems, in which case the most suitable approach is provided by differential geometry. The feedback linearization problem for discrete-time systems is recasted through the language of differential forms. In the event that a system is not feedback linearizable, the largest feedback linearizable subsystem is characterized within the same formalism using the notion of derived flag of a Pfaffian system. A discrete-time system may be linearizable by dynamic state feedback, though it is not linearizable by static state feedback. Necessary and sufficient conditions are given for the existence of a so-called linearizing output, which in turn is a sufficient condition for dynamic state feedback linearizability.

Key words. nonlinear discrete-time systems, algebraic methods, accessibility, feedback linearization, differential forms, Pfaffian systems

AMS subject classifications. 93C10, 93C55, 93B25, 93B05, 93B18, 58A10, 58A17

1. Introduction.

Suppose one is given a discrete-time nonlinear (analytic) system Σ. The goal of this paper is to develop a formalism that provides answers to the following four questions.

Question 1. Is there a neighborhood from which the system Σ is (forward) accessible?

Question 2. Does there exist a regular static state feedback and a state diffeomorphism such that in new coordinates the system Σ reads as a linear controllable system in Brunovsky canonical form?

Question 3. If the answer to Question 2 is negative, then which is the largest feedback linearizable subsystem contained in Σ?

Question 4. If the answer to Question 2 is negative, then does there exist a dynamic state feedback such that the extended system becomes linearizable by static state feedback and state diffeomorphism?

Question 1 has originated a great amount of work; cf. [2, 29, 30] and the references therein. Current literature characterizes pointwise accessibility, whereas we will consider accessibility in a generic sense. Another difference from our work is that we replaced the assumption of invertibility of the discrete-time dynamics by the weaker assumption of submersivity, which is, by the way, invariant under regular static state feedback. On the other hand, our algebraic formalism seems to be a natural tool for the analysis of discrete-time systems.

Question 2 has also received a lot of attention; cf. [21, 27, 31, 34, 35, 45, 46]. The problem consists of finding candidates for output functions whose feedback linearization (as done in [36, 37, 40, 41, 42, 43]) fully linearizes the state equation. The interest of recasting the problem of static state feedback linearization is that it fits nicely in our formalism. For example, it can be viewed as a special case of the answer to Questions 3 and 4.

*Received by the editors May 6, 1994; accepted for publication (in revised form) August 16, 1995. This research was performed while E. Aranda-Bricaire and Ü. Kotta were at Laboratoire d'Automatique de Nantes.

†Department of Electrical Engineering, CINVESTAV-IPN, Apartado Postal 14-740, 07000 México, D.F., México (aranda@ctrl.cinvestav.mx).

‡Institute of Cybernetics, Estonian Academy of Sciences, Akadeemia tee 21, EE0026 Tallinn, Estonia (kotta@ioc.ee).

§Laboratoire d'Automatique de Nantes (Unité de Recherche Associée, Centre National de la Recherche Scientifique 823), Ecole Centrale de Nantes–Université de Nantes, 1 rue de la Noë, 44072 Nantes cedex 03, France (Claude.Moog@lan.ec-nantes.fr).

For nonlinear continuous-time systems, Question 3 was solved in [38]. However, this problem has never been studied for discrete-time systems. The solution that we provide to this problem can be seen as a dual version of the differential geometric approach used in [38].

The dynamic feedback linearization problem is a challenging research problem. Although in the continuous-time case it has been tackled by several authors (cf. [5, 8, 9, 17, 18, 19, 28, 47, 48, 49]), a complete answer is still missing. To our best knowledge, this problem has never been addressed for discrete-time systems. It should be acknowledged, however, that the idea of using a dynamic compensator in order to linearize fully a discrete-time system with outputs was outlined in [44]. We give necessary and sufficient conditions for a nonlinear discrete-time system to admit a so-called linearizing output [17]. This is a sufficient condition for dynamic feedback linearizability.

Preliminary results of this work have been presented in [3, 4].

The contributions of this paper are organized around the classification of one-forms (not necessarily exact) with respect to their relative degree. Our mathematical formalism employs several results both from difference algebra and from exterior differential systems. For the reader's convenience, these results are briefly summarized in §2.

The paper is organized in the following manner. In §3 we develop the algebraic formalism that will allow us to tackle the problems stated above. In §4 we give three equivalent conditions for a discrete-time system to be (forward) accessible. The feedback linearization problem is addressed in §5. Necessary and sufficient conditions are derived under which a discrete-time system is fully linearizable by regular static state feedback and coordinates transformation. When these conditions are not met, it is interesting to characterize the largest feedback linearizable subsystem. This is done in §6. In §7 we define the notion of linearizing output and show its relation with the dynamic feedback linearization problem. Finally, concluding remarks are offered in §8.

2. Mathematical preliminaries.

2.1. Difference algebra. J. F. Ritt founded the branch of mathematics known as difference algebra in the late thirties. His aim was to provide difference equations with a formalism as powerful as commutative algebra is for algebraic equations. Fifty years later, M. Fliess used this formalism for the analysis of discrete-time nonlinear systems [14]. Some basic definitions of difference algebra are recalled in this section. For an introductory exposition of difference algebra we refer the reader to [14], and for a complete panorama of the subject to [11].

A difference ring $\mathcal{B}$ is a pair consisting of a commutative ring $\mathcal{U}$—called the underlying ring—and a monomorphism τ of $\mathcal{U}$ onto a subring $\mathcal{U}'$—called the transforming operator. If $a \in \mathcal{U}$ and $\tau^{-1}a$ is defined, it is unique and is called the inverse transform of a. If $\tau^{-1}a$ is defined for all $a \in \mathcal{U}$, then $\mathcal{B}$ is said to be inversive. $\mathcal{B}$ is inversive if and only if $\mathcal{U}' = \mathcal{U}$.

A difference ring $\mathcal{D} = (\mathcal{S}, \sigma)$ is called a difference overring of the difference ring $\mathcal{B} = (\mathcal{U}, \tau)$ if $\mathcal{S}$ is an overring of $\mathcal{U}$ in the sense of ring theory and if $\tau : \mathcal{U} \to \mathcal{U}$ is a contraction of $\sigma : \mathcal{S} \to \mathcal{S}$. A pair such as $(\mathcal{D}, \mathcal{B})$ is called a difference ring extension and is denoted by the symbol $\mathcal{D}/\mathcal{B}$. An isomorphism of the difference extension $\mathcal{D}/\mathcal{B}$ into the difference extension $\mathcal{D}'/\mathcal{B}$ is an isomorphism of difference rings that leaves fixed every element of the difference ring $\mathcal{B}$.

If rings are replaced by fields in the preceding discussion, one obtains the definitions of difference field, inversive difference field, difference overfield, and difference field extension.

It is often helpful to work with difference rings that are inversive. The inversive closure of a difference ring $\mathcal{B}$ is defined to be a difference overring $\mathcal{D}$ that is inversive and is such that, for every $a \in \mathcal{D}$, there exists an integer $r \geq 0$ such that $\tau^r a \in \mathcal{B}$.

It is proved [11] that every difference ring admits an inversive closure. The inversive closure is not unique, but any two inversive closures of a given difference ring are isomorphic. Finally notice that the inversive closure of a difference field is a difference field as well.

In §3 we define a difference field associated to a discrete-time nonlinear system. A practical procedure for the construction of the inversive closure of this difference field is given in Appendix A.

2.2. Exterior differential systems. The material of this section has been borrowed from [6, 10]. Henceforth, it is assumed that the reader is comfortable with the notions of analytic manifold, vector fields, differential forms, Lie derivative, and Lie bracket [1, 50].

Given an analytic manifold M, which in general will be $\mathbb{R}^N$ for some positive integer N, we adopt the following notation. TM denotes the tangent bundle, T^*M the cotangent bundle, $C^\omega(M)$ the ring of analytic functions defined over M, $V(M)$ the set of analytic vector fields defined over M, and $\Omega^p(M)$ the set of differential p-forms. In particular, $\Omega^0(M) = C^\omega(M)$ and $\Omega^1(M) = T^*M$. $V(M)$ and $\Omega^p(M)$ are $C^\omega(M)$-modules.

Let $\{x_i\}$ be a system of local coordinates of M. Thus $\{\mathrm{d}x_i\}$ is a system of local coordinates of the cotangent bundle. In this frame, every p-form $\alpha \in \Omega^p(M)$ has a unique representation of the form

$$\alpha = \sum_{i_1<\cdots<i_p} a_{i_1,\ldots,i_p}\mathrm{d}x_{i_1} \wedge \cdots \wedge \mathrm{d}x_{i_p}, \qquad a_{i_1,\ldots,i_p} \in C^\omega(M).$$

The exterior product of a p-form α and a q-form β is a mapping

$$\wedge : \Omega^p(M) \times \Omega^q(M) \to \Omega^{(p+q)}(M),$$

which is bilinear and associative. In general, the exterior product is not commutative. Instead, it satisfies the relation $\alpha \wedge \beta = (-1)^{pq}\beta \wedge \alpha$. This relation implies that

1. if α is an odd form, then $\alpha \wedge \alpha \equiv 0$.
2. the exterior product of a p-form α and a q-form β commutes if pq is even.

The exterior differential d is an $\mathbb{R}$-linear operator

$$\mathrm{d} : \Omega^p(M) \to \Omega^{p+1}(M),$$

which satisfies the following properties.

1. $\mathrm{d}(\alpha \wedge \beta) = \mathrm{d}\alpha \wedge \beta + (-1)^p\alpha \wedge \mathrm{d}\beta$, where p is the degree of α.
2. If $f \in C^\omega(M)$ f, then $\mathrm{d}f$ coincides with ordinary differential.
3. $\mathrm{d}^2 = 0$.

These properties uniquely define the operator d.

A differential p-form $\alpha \in \Omega^p(M)$ is said to be closed if $\mathrm{d}\alpha = 0$. It is said to be exact if there exists a differential $(p-1)$-form $\beta \in \Omega^{(p-1)}(M)$ such that $\alpha = \mathrm{d}\beta$. An exact differential form is closed. Poincaré's lemma states that the converse holds locally.

A Pfaffian system I is a $C^\omega(M)$-submodule of the cotangent bundle. The rank of a Pfaffian system I at point x is the dimension of the submodule $I_x \subset T_x^*M$. The rank depends in general on the point x, but if it is maximal at point x, then it is constant in a neighborhood of x.

An algebraic ideal $\mathcal{I}$ is a $C^\omega(M)$-module of differential forms that is also an ideal with respect to the exterior product. An exterior differential system is an algebraic ideal $\mathcal{I}$ that is stable with respect to exterior differentiation.

Frobenius's theorem states that in the event that the algebraic ideal and the exterior differential system generated by a Pfaffian system I coincide, then there exists a collection of exact forms that generates I.

THEOREM 2.1 (Frobenius). *Let I be a Pfaffian system generated by the one-forms $\{\omega_1 \dots, \omega_s\}$. Suppose that the condition*

$$d\omega_k \wedge \omega_1 \wedge \cdots \wedge \omega_s = 0, \qquad k = 1, \dots, s,$$

is satisfied. Then, there exists locally a system of coordinates $\{x_i\}$ such that I is generated by $\{dx_1, \dots, dx_s\}$. In this case, the Pfaffian system I is said to be completely integrable.

The interior product by a vector field $X \in V(M)$ is an $\mathbb{R}$-linear operator

$$X\lrcorner : \Omega^p(M) \to \Omega^{p-1}(M),$$

which satisfies the following properties.

1. $X \lrcorner (\alpha \wedge \beta) = (X \lrcorner \alpha) \wedge \beta + (-1)^p \alpha \wedge (X \lrcorner \beta)$, where p is the degree of α.
2. $\forall f \in \Omega^0(M),\ X \lrcorner f = 0$.
3. $X \lrcorner dx_i = X_i$, where X_i is the i-th component of X.

The characteristic vector fields associated with an exterior differential system $\mathcal{I}$ are the elements of the set

$$A(\mathcal{I}) = \{X \in V(M) \mid X \lrcorner \mathcal{I} \subset \mathcal{I}\}.$$

The annihilator $C(\mathcal{I})$ of $A(\mathcal{I})$ is the characteristic system of $\mathcal{I}$. The characteristic system is completely integrable.

The first derived system of a Pfaffian system $\mathcal{I}$ is defined by

$$I^{(1)} = \{\omega \in I \mid d\omega \equiv 0 \bmod I\},$$

where mod I means modulo the algebraic ideal generated by I. The first derived system is again a Pfaffian system. Higher order derived systems are defined by $I^{(k+1)} = (I^{(k)})^{(1)}$ and yield the filtration $I \supset I^{(1)} \supset \cdots \supset I^{(k)}$, which is called the derived flag. Let K be the smallest integer such that $I^{(K+1)} = I^{(K)}$. It is proved that such an integer exists. $I^{(K)}$ is called the bottom derived system, and K the derived length. The bottom derived system is the largest completely integrable subsystem contained in I.

3. Algebraic formalism. The material that we shall develop in this section is related to the linear algebraic approach introduced by Grizzle [22] and with the difference algebraic approach introduced by Fliess [13, 14] for the analysis of discrete-time systems. The ground field $\mathcal{K}^*$ that we consider below can be viewed as a field extension of the field $\mathcal{K}$ of meromorphic functions that was already defined in [22]. As a matter of fact, $\mathcal{K}^*$ is the inversive closure of $\mathcal{K}$ that is unique up to an isomorphism.

Working with $\mathcal{K}$ instead of $\mathcal{K}^*$ leads to sets of one-forms whose dimension or whose integrability properties are not the appropriate tool for characterizing accessibility or for solving the various feedback linearization problems.

Consider the discrete-time nonlinear system

$$\Sigma : x(t+1) = f(x(t), u(t)), \qquad t = 0, 1, \dots, \tag{1}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and the map

$$f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$$

is supposed to be analytic and generically to define a submersion, i.e.,

$$\operatorname{rank} \frac{\partial f}{\partial (x, u)} = n.$$

The assumption that the discrete-time system Σ admits the (global) analytic representation (1) may be a disadvantage in some cases. This may be circumvented by stating our results locally. As in [2], equations (1) are assumed to hold only for positive time, which is not the case in [15]. Throughout the paper it is also assumed that

$$\operatorname{rank} \frac{\partial f}{\partial u} = m.$$

Except in §6, where dynamic state feedbacks will be considered, through the paper we shall consider regular static state feedbacks, which are defined as follows.

DEFINITION 3.1. *A regular static state feedback is a mapping*

$$\begin{aligned} u : \mathbb{R}^n \times \mathbb{R}^m &\to \mathbb{R}^m, \\ (x(t), v(t)) &\mapsto u(t) = \varphi(x(t), v(t)) \end{aligned} \tag{2}$$

such that

$$\operatorname{rank} \frac{\partial \varphi}{\partial v} = m.$$

Let $\mathcal{R}$ denote the ring of analytic functions in a finite number of the variables $\{x(0), u(t), t \geq 0\}$; and let $\mathcal{K}$ be its associated quotient field, i.e., the field of meromorphic functions in a finite number of the variables $\{x(0), u(t), t \geq 0\}$. The forward-shift operator $\delta : \mathcal{R} \to \mathcal{R}$ is defined by

$$\delta \varphi(x(0), u(0), \ldots, u(N)) = \varphi(f(x(0), u(0)), u(1), \ldots, u(N+1)).$$

The mapping $\delta : \mathcal{R} \to \mathcal{R}$ is injective, thanks to the following technical lemma [22].

LEMMA 3.2. *Assume that the system Σ is submersive. Hence the kernel of the endomorphism $\delta : \mathcal{R} \to \mathcal{R}$ is trivial.*

Thus, the pair $(\mathcal{R}, \delta)$ is a difference ring [11, 14]. Moreover, Lemma 3.2 also implies that the mapping $\delta : \mathcal{K} \to \mathcal{K}$ is well defined. Hence the pair $(\mathcal{K}, \delta)$ is a difference field [11, 14]. $(\mathcal{K}, \delta)$ is not inversive in general. Nevertheless, it is always possible to embed $\mathcal{K}$ into an inversive difference overfield $\mathcal{K}^*$, called the *inversive closure* [11, 14] of $\mathcal{K}$. With a slight abuse of notation, we denote by $\delta : \mathcal{K}^* \to \mathcal{K}^*$ the forward-shift operator that extends $\delta : \mathcal{K} \to \mathcal{K}$. Sometimes the abridged notation $\varphi^+(\cdot) = \delta\varphi(\cdot)$ and $\varphi^-(\cdot) = \delta^{-1}\varphi(\cdot)$ are used.

The inversive closure of $\mathcal{K}$ is unique up to an isomorphism [11]. From now, we assume that the inversive closure $\mathcal{K}^*$ is given. A practical procedure for the construction of $\mathcal{K}^*$ is given in Appendix A.

Let $\mathcal{F} = \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}\varphi \mid \varphi \in \mathcal{K}^*\}$. The operators δ and δ^{-1} induce, respectively, the operators $\Delta : \mathcal{F} \to \mathcal{F}$ and $\Delta^{-1} : \mathcal{F} \to \mathcal{F}$ by

$$\begin{aligned} \Delta(\textstyle\sum_i a_i \mathrm{d}\varphi_i) &\mapsto \textstyle\sum_i a_i^+ \mathrm{d}\varphi_i^+, \\ \Delta^{-1}(\textstyle\sum_i a_i \mathrm{d}\varphi_i) &\mapsto \textstyle\sum_i a_i^- \mathrm{d}\varphi_i^-. \end{aligned}$$

With some abuse of notation, sometimes we write $\omega^+ = \Delta\,\omega$ and $\omega^- = \Delta^{-1}\omega$. The elements of $\mathcal{F}$ are called one-forms. Hereafter, for any set $\mathcal{W}$ of one-forms, the notation $\mathcal{W}^+$ should be understood elementwise.

Let us investigate the following introductory example.

Example 3.3. Consider the nonlinear control system

$$\begin{aligned} x_1(t+1) &= u_1(t), \\ x_2(t+1) &= x_3(t)u_1(t), \\ x_3(t+1) &= u_2(t). \end{aligned} \tag{3}$$

For system (3), one can readily check that

$$\begin{aligned}\operatorname{span}_{\mathcal{K}}\{\mathrm{d}x(1)\} \cap \operatorname{span}_{\mathcal{K}}\{\mathrm{d}x(0)\} &= \operatorname{span}_{\mathcal{K}}\{\mathrm{d}x_2(1) - x_3(0)\mathrm{d}x_1(1)\} \\ &= \operatorname{span}_{\mathcal{K}}\{\mathrm{d}x_3(0)\}.\end{aligned}$$

More precisely, $x_3(0) = x_2(1)/x_1(1)$. Thus, a pre-image in $\mathcal{K}$ (through δ) of $x_3(0)$ is $x_2(0)/x_1(0)$. Whereas $x_3(0)$ has a pre-image in $\mathcal{K}$, $x_1(0)$ and $x_2(0)$ have none. According to the procedure given in Appendix A (see also the Proof of Theorem II in pp. 66–67 of [11]), the construction of the inversive closure of $\mathcal{K}$ amounts to embed $\mathcal{K}$ into an overfield $\mathcal{K}^*$ and extend δ in such a way that $\delta : \mathcal{K}^* \to \mathcal{K}^*$ becomes an automorphism. For our example, $\mathcal{K}^*$ is nothing but the field of meromorphic functions in a finite number of the variables $\{x(0), u(t), u(-k), z(-k)\ t \geq 0,\ k \geq 1\}$, where $z(-k) = [z_1(-k), z_2(-k)]$, subject to $\delta(z_1(-1)) = x_1(0)$ and $\delta(z_2(-1)) = x_2(0)$.

We now continue with the analysis of the general case. The relative degree r of a one-form $\omega(0) \in \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0), \mathrm{d}u(0)\}$ is defined to be

$$r = \min\{k \geq 0 \mid w(k) = \Delta^k \omega(0) \notin \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0)\}\}.$$

If such an integer does not exist, set $r = \infty$. The relative degree of a meromorphic function $\varphi(x(0), u(0))$ is defined to be the relative degree of the one-form $\mathrm{d}\varphi(x(0), u(0))$.

Introduce the sequence of subspaces $\mathcal{H}_0 \supset \mathcal{H}_1 \supset \cdots \supset \mathcal{H}_k$ of $\mathcal{F}$ defined by

$$\begin{aligned}\mathcal{H}_0 &= \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0), \mathrm{d}u(0)\}, \\ \mathcal{H}_k &= \Delta^{-k}\left(\mathcal{H}_0 \cap \Delta\mathcal{H}_0 \cap \cdots \cap \Delta^k \mathcal{H}_0\right),\end{aligned} \tag{4}$$

where $\Delta\mathcal{H}_0 = \operatorname{span}_{\mathcal{K}^*}\{\omega^+ \mid \omega \in \mathcal{H}_0\}$ and $\Delta^k\mathcal{H}_0 = \operatorname{span}_{\mathcal{K}^*}\{\omega^+ \mid \omega \in \Delta^{k-1}\mathcal{H}_0\}$, $k \geq 1$.

PROPOSITION 3.4. 1. *For $k \geq 0$, $\mathcal{H}_k$ is the space of one-forms whose relative degree is greater than or equal to k.*

2. *There exists an integer $k^* \leq n$ such that, for $0 \leq k \leq k^*$, $\mathcal{H}_{k+1} \subset \mathcal{H}_k$ but $\mathcal{H}_{k+1} \neq \mathcal{H}_k$ and $\mathcal{H}_{k^*+1} = \mathcal{H}_{k^*+2} = \cdots = \mathcal{H}_\infty$.*

3. *The subspaces $\mathcal{H}_k$ are invariant under regular static state feedback and under state diffeomorphism.*

Proof. Point 1 is clear because the $\mathcal{H}_k$'s can alternatively be defined by

$$\mathcal{H}_k = \operatorname{span}_{\mathcal{K}^*}\{\omega \in \mathcal{H}_{k-1} \mid \omega^+ \in \mathcal{H}_{k-1}\}, \qquad k \geq 1.$$

In particular, $\mathcal{H}_1 = \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0)\}$. Existence of the integer k^* comes from the fact that each $\mathcal{H}_k$ is a finite-dimensional $\mathcal{K}^*$-vector space so that, at each step, either its dimension decreases or $\mathcal{H}_{k+1} = \mathcal{H}_k$. Moreover, $k^* \leq n = \dim_{\mathcal{K}^*} \mathcal{H}_1$. Feedback invariance comes from the fact that the relative degree is obviously invariant under regular static state feedback. □

THEOREM 3.5. *Suppose $\mathcal{H}_\infty = 0$. Then there exists a list of integers $r_1, \ldots, r_m$, invariant under regular static state feedback, and m one-forms $\omega_1(0), \ldots, \omega_m(0) \in \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0)\}$ whose relative degrees are, respectively, $r_1, \ldots, r_m$ such that*

1. $\operatorname{span}_{\mathcal{K}^*}\{\omega_i(k), 1 \leq i \leq m, 0 \leq k \leq r_i - 1\} = \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0)\}$.
2. $\operatorname{span}_{\mathcal{K}^*}\{\omega_i(k), 1 \leq i \leq m, 0 \leq k \leq r_i\} = \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0), \mathrm{d}u(0)\}$.
3. *The one-forms $\{\omega_i(k), 1 \leq i \leq m, k \geq 0\}$ are linearly independent; in particular, $\sum_i r_i = n$.*

Proof. Let $\mathcal{W}_{k^*}$ be a basis for $\mathcal{H}_{k^*}$. By definition, $\mathcal{W}_{k^*}$ and $\mathcal{W}_{k^*}^+$ are in $\mathcal{H}_{k^*-1}$. We next prove that $\mathcal{W}_{k^*}$ and $\mathcal{W}_{k^*}^+$ are linearly independent. Let $\mathcal{W}_{k^*} = \{\eta_1, \ldots, \eta_{\rho_{k^*}}\}$; then $\mathcal{W}_{k^*}^+ = \{\eta_1^+, \ldots, \eta_{\rho_{k^*}}^+\}$. Suppose, contrary to our claim, that $\mathcal{W}_{k^*}$ and $\mathcal{W}_{k^*}^+$ are linearly dependent. This means that there exist some coefficients λ_i, μ_i, $1 \leq i \leq \rho_{k^*}$, which are not all zero, such that

$$\sum_{i=1}^{\rho_{k^*}} (\lambda_i \eta_i + \mu_i \eta_i^+) = 0. \tag{5}$$

The linear independence of the η_i's implies that not all the μ_i's vanish. Now, consider the one-form $\omega = \sum_i \mu_i^- \eta_i \in \mathcal{H}_{k^*}$ whose forward-time shift is, by (5),

$$\omega^+ = \sum_i \mu_i \eta_i^+ = -\sum_i \lambda_i \eta_i \in \mathcal{H}_{k^*}.$$

So, equality (5) implies that $\omega \in \mathcal{H}_{k^*+1} = \mathcal{H}_\infty$, which contradicts our assumption $\mathcal{H}_\infty = 0$. Hence, it is always possible to choose a set (possibly empty) $\mathcal{W}_{k^*-1}$ such that $\mathcal{W}_{k^*} \cup \mathcal{W}_{k^*}^+ \cup \mathcal{W}_{k^*-1}$ is a basis for $\mathcal{H}_{k^*-1}$. Repeating this procedure $k^* - 1$ times we obtain

$$\mathcal{H}_k = \operatorname{span}_{\mathcal{K}^*}\{\mathcal{W}_i(j),\ k \le i \le k^*,\ 0 \le j \le i-k\}, \qquad 0 \le k \le k^*.$$

The assumption $\operatorname{rank} \frac{\partial f}{\partial u} = m$ implies $\mathcal{W}_0 = \emptyset$. Finally, set

$$\{\omega_1(0), \ldots, \omega_m(0)\} = \mathcal{W}_{k^*} \cup \cdots \cup \mathcal{W}_1.$$

Since the subspaces $\mathcal{H}_k$ are invariant under regular static state feedback, the invariance of the integers $\{r_i\}$ is obvious from the construction. □

COROLLARY 3.6. *Suppose $\mathcal{H}_\infty = 0$. Then there exists a basis*

$$\{\omega_{i,j}(0),\ 1 \le i \le m,\ 1 \le j \le r_i\}$$

of $\operatorname{span}_{\mathcal{K}^*}\{dx(0)\}$ *such that the first-order approximation of* Σ *yields the infinitesimal Brunovsky form*

$$\begin{aligned}
\omega_{i,1}(t+1) &= w_{i,2}(t),\\
\omega_{i,2}(t+1) &= w_{i,3}(t),\\
&\vdots\\
\omega_{i,r_i-1}(t+1) &= w_{i,r_i}(t),\\
\omega_{i,r_i}(t+1) &= \textstyle\sum_{s=1}^m \sum_{j=1}^{r_s} a_{s,j}^i w_{s,j}(t) + \sum_{j=1}^m b_j^i du_j(t), \qquad 1 \le i \le m,
\end{aligned} \tag{6}$$

where $a_{s,j}^i$, $b_j^i \in \mathcal{K}^*$ and $[b_j^i]$ has an inverse in the ring of $m \times m$ matrices with entries in $\mathcal{K}^*$.

Proof. For $1 \le i \le m$ and $1 \le j \le r_i$, take $\omega_{i,j}(0) = \omega_i(j-1)$. □

Example 3.7. Consider the nonlinear discrete-time system

$$\begin{aligned}
x_1(t+1) &= u_1(t), & x_4(t+1) &= x_5(t)u_1(t),\\
x_2(t+1) &= x_3(t)u_1(t), & x_5(t+1) &= x_6(t)u_1(t),\\
x_3(t+1) &= x_4(t)u_1(t), & x_6(t+1) &= u_2(t).
\end{aligned} \tag{7}$$

For system (7), one has

$$\begin{aligned}
\operatorname{span}_{\mathcal{K}^*}\{dx(0)\} \cap \operatorname{span}_{\mathcal{K}^*}\{dx(1)\} &= \operatorname{span}_{\mathcal{K}^*}\{dx_i(1) - x_{i+1}(0)dx_1(1),\ i = 2, \ldots, 5\}\\
&= \operatorname{span}_{\mathcal{K}^*}\{u_1(0)dx_i(0),\ i = 3, \ldots, 6\}.
\end{aligned}$$

In fact, it is straightforward to check that

$$x_i(0) = \frac{x_{i-1}(1)}{x_1(1)} = \delta\left(\frac{x_{i-1}(0)}{x_1(0)}\right), \qquad i = 3, \ldots, 6. \tag{8}$$

This shows that $x_i(-1)$, for $i = 3, \ldots, 6$, must not be considered as independent variables in the sense that they can be expressed as functions of $x(0)$. This is not the case for $x_1(-1)$ and $x_2(-1)$, which cannot be expressed as functions of $x(0)$. Thus we can choose, according to the procedure given in Appendix A, $z(0) = (x_1(0), x_2(0))$.

We carry out the computations necessary to bring forward the infinitesimal Brunovsky form associated with the system described by (7). We know that $\mathcal{H}_1 = \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0)\}$ and we already computed

$$\operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0)\} \cap \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(1)\} = \operatorname{span}_{\mathcal{K}}\{\mathrm{d}x_i(1) - x_{i+1}(0)\mathrm{d}x_1(1),\ i = 2, \ldots, 5\}.$$

Thus, using (8), one obtains

$$\mathcal{H}_2 = \Delta^{-1}(\mathcal{H}_1 \cap \Delta\,\mathcal{H}_1) = \operatorname{span}_{\mathcal{K}^*}\{x_1(0)\mathrm{d}x_i(0) - x_i(0)\mathrm{d}x_1(0),\ i = 2, \ldots, 5\}.$$

In a similar vein, one may check that

$$\begin{aligned}
\mathcal{H}_3 &= \operatorname{span}_{\mathcal{K}^*}\{x_2(0)\mathrm{d}x_3(0) - x_3(0)\mathrm{d}x_2(0),\ x_2(0)\mathrm{d}x_4(0) - x_4(0)\mathrm{d}x_2(0)\},\\
\mathcal{H}_4 &= \operatorname{span}_{\mathcal{K}^*}\{x_2(0)\mathrm{d}x_3(0) - x_3(0)\mathrm{d}x_2(0)\},\\
\mathcal{H}_5 &= 0.
\end{aligned}$$

Up to multiplication by a nonzero function, the choice of ω_1 is unique. Namely

$$\omega_1 = x_2(0)\mathrm{d}x_3(0) - x_3(0)\mathrm{d}x_2(0),$$

whose relative degree is $r_1 = 4$. For ω_2 it suffices to pick any one-form $\omega_2 \in \mathcal{H}_2$ independent of $\omega_1(0)$, $\omega_1(1)$, $\omega_1(2)$. Let

$$\omega_2 = x_1(0)\mathrm{d}x_2(0) - x_2(0)\mathrm{d}x_1(0),$$

whose relative degree is $r_2 = 2$. Finally set $\omega_{i,j} = \omega_i(j-1)$, for $i = 1, 2$, $j = 1, \ldots, r_i$. In coordinates $\omega_{i,j}$, the first-order approximation of (7) takes the infinitesimal Brunovsky form (6).

4. Accessibility. Following the notation in [30] we shall denote by $A_k(x)$ the set of points reachable from x in k forward steps using arbitrary sequences of controls

$$\mathbf{u} = (u(0), \ldots, u(k-1)) \in (\mathbb{R}^m)^k.$$

Denote by $A(x)$ the set of points reachable from x in any number of forward steps using arbitrary sequences of controls. That is, $A(x) = \bigcup_{k\geq 0} A_k(x)$. The system Σ is forward accessible from x if its reachable set $A(x)$ has nonempty interior. A generic notion of accessibility can be derived from this pointwise definition as in [2, §5].

DEFINITION 4.1 (see [2]). *System Σ is said to be (forward) accessible if its reachable set $A(x)$ has a nonempty interior in $\mathbb{R}^n$ for almost all $x \in \mathbb{R}^n$.*

Associated with Σ there is a family of maps

$$f_u = f(\cdot, u) : \mathbb{R}^n \to \mathbb{R}^n, \qquad u \in \mathbb{R}^m.$$

If we apply a sequence of controls $\mathbf{u}$, then we obtain the composition of such maps, which is denoted by

$$f_{\mathbf{u}} = f_{u(k-1)} \circ \cdots \circ f_{u(0)} : \mathbb{R}^n \to \mathbb{R}^n.$$

Conversely, for each fixed state x define the transition map

$$\begin{aligned}
\Gamma(k, x)(\mathbf{u}) : (\mathbb{R}^m)^k &\to \mathbb{R}^n,\\
\mathbf{u} &\mapsto f_{\mathbf{u}}(x).
\end{aligned}$$

By definition, $A_k(x) = \operatorname{Im}\Gamma(k, x)(\mathbf{u})$.

PROPOSITION 4.2 (see [30]). *The interior of the reachable set $A_k(x)$ is nonempty if and only if*

$$\sup\left\{\operatorname{rank}\frac{\partial}{\partial \mathbf{u}}\Gamma(k,x)(\mathbf{u}),\ \mathbf{u}\in(\mathbb{R}^m)^k\right\}=n.$$

Thus, system Σ is (forward) accessible if and only if

$$\sup_{x\in\mathbb{R}^n}\left\{\operatorname{rank}\frac{\partial}{\partial \mathbf{u}}\Gamma(k,x)(\mathbf{u}),\ \mathbf{u}\in(\mathbb{R}^m)^k,\ k\geq 1\right\}=n.$$

It was pointed out in [30] that Proposition 4.2 does not provide a practical test of accessibility mainly because from a computational point of view the composition of functions is a difficult task. In the rest of this section we develop alternative accessibility criteria that require only a reduced number of algebraic operations.

From this point, we make extensive use of the results on exterior differential systems presented in §1. However, we need to state some facts in order to employ such results. Notice that the elements of $\mathcal{K}^*$ are defined over a space that is isomorphic to $\mathbb{R}^\infty$ and hence is not a Banach space. It turns out that the results of §1 do not apply to general subspaces of $\mathcal{F}$, viewed as Pfaffian systems. Nevertheless, the filtration

$$\operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0)\}=\mathcal{H}_1\supset\mathcal{H}_2\supset\cdots\supset\mathcal{H}_{k^*}\supset\mathcal{H}_{k^*+1}=\mathcal{H}_\infty \tag{9}$$

has some nice properties that we examine next.

LEMMA 4.3. *For $1\leq k\leq k^*+1$, there exist ρ_k one-forms $\omega_1,\ldots,\omega_{\rho_k}$ that depend only on the variables $\{x(0),u(-j),z(-j),\ j\leq k-1\}$ and constitute a basis for $\mathcal{H}_k$.*

Proof. We proceed by induction. Lemma 4.3 is evidently true for $k=1$. Suppose it is also true for some integer $k\geq 1$. Let $\{\eta_1,\ldots,\eta_{\rho_k},\mu_1,\ldots,\mu_{\rho_{k-1}-\rho_k}\}$ and $\{\eta_1,\ldots,\eta_{\rho_k}\}$ be, respectively, bases of $\mathcal{H}_{k-1}$ and $\mathcal{H}_k$. An arbitrary element $\omega=\sum_i a_i\eta_i\in\mathcal{H}_k$ belongs to $\mathcal{H}_{k+1}$ if and only if $\omega^+=\sum_i a_i^+\eta_i^+\in\mathcal{H}_k$. Notice, since $\eta_i\in\mathcal{H}_k$, it follows that $\eta_i,\ \eta_i^+\in\mathcal{H}_{k-1}$. Hence,

$$\omega^+=\sum_i a_i^+\left(\sum_j b_{ij}\eta_j+\sum_\ell c_{i\ell}\mu_\ell\right).$$

Thus, $\omega\in\mathcal{H}_{k+1}$ if and only if the a_i^+'s satisfy the system of linear equations

$$\sum_i a_i^+c_{i\ell}=0,\qquad 1\leq\ell\leq\rho_{k-1}-\rho_k. \tag{10}$$

To each nontrivial solution a_i^+ of (10) corresponds a form $\omega=\sum_i a_i\eta_i$ that belongs to $\mathcal{H}_{k+1}$. More precisely, $\dim_{\mathcal{K}^*}\mathcal{H}_{k+1}=\rho_k-\operatorname{rank}_{\mathcal{K}^*}[c_{ij}]$. Now, notice that from the induction assumption the a_i^+'s may be chosen to depend only on the variables $\{x(0),u(-j),z(-j),\ j\leq k-1\}$. Finally notice that, since $a_i=\delta^{-1}a_i^+$, the a_i's depend only on the variables $\{x(0),u(-j),z(-j),\ j\leq k\}$. □

Lemma 4.3 is crucial because it allows consideration of the filtration (9) as a nested sequence of Pfaffian systems defined over $\mathbb{R}^N$, for some integer N large enough. Moreover, the proof of Lemma 4.3 provides a systematic procedure to compute bases of the subspaces $\mathcal{H}_k$.

PROPOSITION 4.4. *Let $\{\alpha_1,\ldots,\alpha_{\rho_\infty}\}$ be a basis for $\mathcal{H}_\infty$. Then the Frobenius condition*

$$\mathrm{d}\alpha_i\wedge\alpha_1\wedge\cdots\wedge\alpha_{\rho_\infty}=0,\qquad i=1,\ldots,\rho_\infty,$$

is satisfied. In other words, there exists (locally) a basis for $\mathcal{H}_\infty$ composed of exact one-forms.

Proof. The proof is largely technical and is given in Appendix B. □

Theorem 4.5 is our main result concerning the accessibility property. It should be strengthened that the characterization of this property through condition 2 is very natural. In plain words, it means that every nonconstant function of the state is eventually influenced by the input of the system and hence cannot satisfy any autonomous difference equation. On the other hand, from a computational point of view, condition 3 can be checked rather easily.

THEOREM 4.5 (Accessibility criteria). *The following statements are equivalent.*

1. *System* Σ *is (forward) accessible.*
2. *Any nonconstant function* $\varphi(x(0))$ *has finite relative degree.*
3. $\mathcal{H}_\infty = 0$.
4. *Define* $\mathcal{X}_k = \text{span}_{\mathcal{K}^*}\{\text{d}x(k)\}$. *Then, there exists an integer* $k \geq 0$ *such that*

$$\dim_{\mathcal{K}^*} \frac{\mathcal{X}_0 + \mathcal{X}_k}{\mathcal{X}_0} = n.$$

Proof. We show (1) $\Leftrightarrow$ (4) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (1).

(1) $\Leftrightarrow$ (4) By definition, $x(k) = \Gamma(k, x)(\mathbf{u})$. Hence

$$\text{d}x(k) = \frac{\partial \Gamma(k, x)}{\partial x(0)} \text{d}x(0) + \frac{\partial \Gamma(k, x)}{\partial \mathbf{u}} \text{d}\mathbf{u}$$

and thus

$$\frac{\mathcal{X}_0 + \mathcal{X}_k}{\mathcal{X}_0} \sim \text{span}_{\mathcal{K}^*} \left\{ \frac{\partial \Gamma(k, x)}{\partial \mathbf{u}} \text{d}\mathbf{u} \right\}.$$

The equivalence follows from Proposition 4.2.

(4) $\Rightarrow$ (2) Suppose that there exists a function $\varphi(x(0))$ whose relative degree is infinite. $\varphi(x(0))$ can always be completed to define a diffeomorphism $z(k) = \phi(x(k))$. Define $\mathcal{Z}_k = \text{span}_{\mathcal{K}^*}\{\text{d}z(k)\}$. The fact that $\varphi(x(0))$ has infinite relative degree implies that, for $k \geq 0$, $\dim_{\mathcal{K}^*} (\mathcal{Z}_0 + \mathcal{Z}_k)/\mathcal{Z}_0 = \dim_{\mathcal{K}^*} \mathcal{Z}_k - \dim_{\mathcal{K}^*} (\mathcal{Z}_0 \cap \mathcal{Z}_k) < n$. The implication follows because, for $k \geq 0$, $\mathcal{X}_k \sim \mathcal{Z}_k$.

(2) $\Rightarrow$ (3) Suppose $\mathcal{H}_\infty \neq 0$. By Proposition 4.4, this implies the existence of ρ_∞ functions whose relative degrees are infinite.

(3) $\Rightarrow$ (1) Suppose that system Σ is not accessible. Since (1) $\Leftrightarrow$ (4), this implies that $\dim_{\mathcal{K}^*} (\mathcal{X}_0 + \mathcal{X}_k)/\mathcal{X}_0 = \dim_{\mathcal{K}^*} \mathcal{X}_0 - \dim_{\mathcal{K}^*} (\mathcal{X}_0 \cap \mathcal{X}_k) < n$, for $k > 0$. Thus $\mathcal{X}_0 \cap \mathcal{X}_k \neq 0$ and hence there exists a nonzero one-form $\omega \in \mathcal{X}_0$ whose relative degree is infinite, i.e., $\mathcal{H}_\infty \neq 0$. □

Remark 4.6. For linear time-invariant systems, condition 4 of Theorem 4.5 coincides with the celebrated *Kalman controllability criterion,* i.e.,

$$\dim_{\mathcal{K}^*} \frac{\mathcal{X}_0 + \mathcal{X}_k}{\mathcal{X}_0} = \text{rank}\left[B \mid AB \mid \cdots \mid A^{k-1}B\right].$$

Example 4.7 shows the application of the results of this section. It has been borrowed from [30].

Example 4.7 (see [30]). Consider the discrete-time polynomial system

$$\begin{aligned} x_1(t+1) &= x_1(t)(x_3^2(t)+1)^2, \\ x_2(t+1) &= x_2(t)(x_3^2(t)+1)^3, \\ x_3(t+1) &= x_3(t) + u(t). \end{aligned} \tag{11}$$

Clearly, for system (11) one has $\mathcal{H}_2 = \text{span}_{\mathcal{K}^*}\{dx_1(0), dx_2(0)\}$. Straightforward computations also show that $\mathcal{H}_3 = \text{span}_{\mathcal{K}^*}\{\eta\}$, where

$$\eta = 3x_2(0)dx_1(0) - 2x_1(0)dx_2(0).$$

Moreover, $\mathcal{H}_3$ is closed under forward-shifting. In fact, $\eta^+ = (x_3^2(0) + 1)^5\eta$. Thus system (11) is not accessible.

Note that although η is not an exact one-form, Proposition 4.4 guarantees the existence of an integrating factor, i.e., a nonzero function $a \in \mathcal{K}^*$ such that $a\,\eta$ becomes integrable. Taking $a = x_1^2(0)/x_2^3(0)$ one obtains

$$a\,\eta = 3\frac{x_1^2(0)}{x_2^2(0)}dx_1(0) - 2\frac{x_1^3(0)}{x_2^3(0)}dx_1(0) = d\left(\frac{x_1^3(0)}{x_2^2(0)}\right).$$

The relative degree of the function $\varphi(x(t)) = x_1^3(t)/x_2^2(t)$ is infinite.

5. Static state feedback linearization. The static state feedback linearization problem has already been considered by several authors; cf. [21, 27, 31, 34, 35, 45, 46]. The problem can be thought of as consisting of finding suitable output functions (without zero dynamics and with vector relative degree) and of applying standard input-output feedback linearization techniques [41, 40, 42, 43]. The interest of recasting this problem is that our solution naturally fits in the more general frame of dynamic feedback linearization. In the rest of this paper, the results on the various feedback linearization problems are local results.

DEFINITION 5.1. *System Σ is said to be linearizable by static state feedback if there exist a state diffeomorphism*

$$\tilde{x}(t) = \phi(x(t)) \tag{12}$$

and a regular static state feedback (2) *such that, in new coordinates, the compensated system reads*

$$\tilde{x}(t+1) = A\tilde{x}(t) + Bv(t), \tag{13}$$

where the pair (A, B) is in Brunovsky canonical form.

THEOREM 5.2. *System Σ is linearizable by regular static state feedback if and only if*

1. $\mathcal{H}_\infty = 0$.
2. *For $1 \le k \le k^*$, $\mathcal{H}_k$ is completely integrable.*

Proof. Sufficiency. If $\mathcal{H}_\infty = 0$, then by Corollary 3.6 there exists a basis

$$\{\omega_{i,j}(0), 1 \le i \le m, 1 \le j \le r_i\} \tag{14}$$

of $\text{span}_{\mathcal{K}^*}\{dx(0)\}$ such that in this basis the first-order approximation of system Σ, i.e.,

$$dx(t+1) = \frac{\partial f}{\partial x}(x(t), u(t))dx(t) + \frac{\partial f}{\partial u}(x(t), u(t))du(t),$$

takes the infinitesimal Brunovsky form (6). By Frobenius's Theorem, there is no loss of generality if we assume that the basis (14) is composed of exact one-forms. Thus, every $\omega_{i,j}(0)$, $1 \le i \le m$, $1 \le j \le r_i$, can be integrated; i.e., there exist $\phi_{i,j}(x(0))$ such that $\omega_{i,j}(0) = d\phi_{i,j}(x(0))$. In coordinates $\tilde{x}_{i,j} = \phi_{i,j}(x)$, $1 \le i \le m$, $1 \le j \le r_i$, system Σ reads

$$\begin{aligned}
\tilde{x}_{i,1}(t+1) &= \tilde{x}_{i,2}(t),\\
\tilde{x}_{i,2}(t+1) &= \tilde{x}_{i,3}(t),\\
&\vdots\\
\tilde{x}_{i,r_i-1}(t+1) &= \tilde{x}_{i,r_i}(t),\\
\tilde{x}_{i,r_i}(t+1) &= \tilde{f}_i(\tilde{x}(t), u(t)), \quad 1 \le i \le m,
\end{aligned}$$

with $\frac{\partial \tilde{f}_i}{\partial \tilde{x}_{s,j}} = a^i_{s,j}$, $\frac{\partial \tilde{f}_i}{\partial u_j} = b^i_j$. The proof is completed by showing that under the feedback $u(t) = \varphi(\tilde{x}(t), v(t))$ where

$$\tilde{f}_i(\tilde{x}(t), \varphi(\tilde{x}(t), v(t))) = v_i(t), \ 1 \leq i \leq m, \tag{15}$$

the system Σ takes the Brunovsky canonical form (13) with controllability indices $r_1, \ldots, r_m$. Note that by the implicit function theorem (15) has a local solution with respect to $u(t) = \varphi(\tilde{x}(t), v(t))$ since the matrix with elements b^i_j is invertible by Corollary 3.6.

Necessity. This is clear because for a linear system the $\mathcal{H}_k$'s are completely integrable and this property is invariant under regular static state feedback and state diffeomorphism. □

Remark 5.3. Theorem 5.2 gives an alternative solution to the static linearization problem, considered earlier in [21, 27, 31, 34, 35, 45, 46]. One can see a particular close connection with Lemma 5 of [35]. The proof of our theorem indicates how actually to find the functions $h_1, \ldots, h_m$ in terms of which Lemma 5 is formulated.

Example 5.4 (Example 3.7 continued). We have computed, for the system (7), the one-forms

$$\omega_1 = x_2(0)\mathrm{d}x_3(0) - x_3(0)\mathrm{d}x_2(0),$$
$$\omega_2 = x_1(0)\mathrm{d}x_2(0) - x_2(0)\mathrm{d}x_1(0)$$

that generate the infinitesimal Brunovsky form (6). The one-forms ω_1, ω_2 are not exact. However, one can readily verify that

$$\mathrm{d}\omega_1 \wedge \omega_1 = \mathrm{d}\omega_2 \wedge \omega_2 = 0,$$

so that there exist integrating factors a_i such that $\tilde{\omega}_i = a_i\omega_i$ are exact one-forms for $i = 1, 2$. This implies that the conditions of Theorem 5.2 are satisfied and hence system (7) is linearizable by regular static state feedback. Taking $a_1 = 1/x_2^2(0)$ and $a_2 = 1/x_1^2(0)$ one obtains

$$\tilde{\omega}_1 = \mathrm{d}\left(\frac{x_3(0)}{x_2(0)}\right), \qquad \tilde{\omega}_2 = \mathrm{d}\left(\frac{x_2(0)}{x_1(0)}\right).$$

Define now the diffeomorphism

$$\tilde{x}(t) = \left(\frac{x_3(t)}{x_2(t)}, \frac{x_4(t)}{x_3(t)}, \frac{x_5(t)}{x_4(t)}, \frac{x_6(t)}{x_5(t)}, \frac{x_2(t)}{x_1(t)}, x_3(t)\right).$$

In coordinates $\tilde{x}(t)$, the system (7) reads

$$\begin{aligned}
\tilde{x}_1(t+1) &= \tilde{x}_2(t), & \tilde{x}_4(t+1) &= u_2(t)/(\tilde{x}_2(t)\tilde{x}_3(t)\tilde{x}_4(t)\tilde{x}_6(t)u_1(t)),\\
\tilde{x}_2(t+1) &= \tilde{x}_3(t), & \tilde{x}_5(t+1) &= \tilde{x}_6(t),\\
\tilde{x}_3(t+1) &= \tilde{x}_4(t), & \tilde{x}_6(t+1) &= \tilde{x}_2(t)\tilde{x}_6(t)u_1(t).
\end{aligned}$$

Finally, the state feedback

$$u_1(t) = \frac{v_1(t)}{\tilde{x}_2(t)\tilde{x}_6(t)}, \qquad u_2(t) = \tilde{x}_3(t)\tilde{x}_4(t)v_1(t)v_2(t)$$

yields the Brunovsky form (13) with controllability indices $\{4, 2\}$.

6. The largest feedback linearizable subsystem. When system Σ cannot be fully linearized using static state feedback, it is interesting to characterize the largest feedback linearizable subsystem. This is done in this section. Throughout this section, we assume that system Σ is accessible, i.e., $\mathcal{H}_\infty = 0$.

DEFINITION 6.1. *System Σ is said to be partially linearizable with controllability indices $n_1 \geq \cdots \geq n_m$ if there exists a state diffeomorphism* (12) *and a regular static state feedback* (2) *such that, in new coordinates, the compensated system reads*

$$\begin{aligned} \tilde{x}^1(t+1) &= A^1\tilde{x}^1(t) + B^1 v(t), \\ \tilde{x}^2(t+1) &= f^2(\tilde{x}^1(t), \tilde{x}^2(t), v(t)), \end{aligned} \tag{16}$$

where the pair (A^1, B^1) is in Brunovsky canonical form with controllability indices $n_1 \geq \cdots \geq n_m$.

The sequence of subspaces $\mathcal{H}_0 \supset \mathcal{H}_1 \supset \cdots \supset \mathcal{H}_{k^*}$ can be viewed as a nested sequence of Pfaffian systems. For $k = 1, \ldots, k^*$, the bottom derived system of $\mathcal{H}_k$ is denoted $\bar{\mathcal{H}}_k$. Also, introduce the list of integers $\{p'_k\}$, defined by

$$p'_k = \dim_{\mathcal{K}^*} \frac{\bar{\mathcal{H}}_k + \mathcal{H}_{k+1}}{\mathcal{H}_{k+1}},$$

and its dual list $\{n^*_j\}$, defined by $n^*_j = \operatorname{card}\{p'_k \mid p'_k \geq j\}$.

Our claim is that $\{n^*_j\}$ is the list of controllability indices of the largest feedback linearizable subsystem.

THEOREM 6.2. *System Σ is partially feedback linearizable with controllability indices $n^*_1 \geq \cdots \geq n^*_m$.*

Proof. Let $\bar{\mathcal{W}}_{k^*}$ be a basis for $\bar{\mathcal{H}}_{k^*}$, and choose a set $\hat{\mathcal{W}}_{k^*}$ such that $\bar{\mathcal{W}}_{k^*} \cup \hat{\mathcal{W}}_{k^*}$ is a basis for $\mathcal{H}_{k^*}$. Let $\bar{\mathcal{W}}_{k^*-1}$ be a set such that

$$\bar{\mathcal{W}}_{k^*} \cup \bar{\mathcal{W}}^+_{k^*} \cup \hat{\mathcal{W}}_{k^*} \cup \bar{\mathcal{W}}_{k^*-1}$$

is a basis for $\bar{\mathcal{H}}_{k^*-1} + \mathcal{H}_{k^*}$, and choose a set $\hat{\mathcal{W}}_{k^*-1}$ such that

$$\bar{\mathcal{W}}_{k^*} \cup \bar{\mathcal{W}}^+_{k^*} \cup \bar{\mathcal{W}}_{k^*-1} \cup \hat{\mathcal{W}}_{k^*-1}$$

is a basis for $\mathcal{H}_{k^*-1}$. This procedure can be repeated $k^* - 1$ times in such a way that

$$\left\{ \bigcup_{k+1 \leq i \leq k^*} \bigcup_{0 \leq j \leq i-k} \bar{\mathcal{W}}_i(j) \right\} \cup \hat{\mathcal{W}}_{k+1} \cup \bar{\mathcal{W}}_k$$

is a basis for $\bar{\mathcal{H}}_k + \mathcal{H}_{k+1}$ and

$$\left\{ \bigcup_{k \leq i \leq k^*} \bigcup_{0 \leq j \leq i-k} \bar{\mathcal{W}}_i(j) \right\} \cup \hat{\mathcal{W}}_k$$

is a basis for $\mathcal{H}_k$. For $k = 1, \ldots, k^*$, denote $p_k = \operatorname{card} \bar{\mathcal{W}}_k$. Thus, the k^* sets we have constructed (some of them possibly empty) satisfy

$$p'_k = \dim_{\mathcal{K}^*} \frac{\bar{\mathcal{H}}_k + \mathcal{H}_{k+1}}{\mathcal{H}_{k+1}} = \sum_{i \geq k} p_i. \tag{17}$$

Clearly, the $\bar{\mathcal{W}}_k$'s can be chosen to be composed by exact one-forms. Say, for $k = 1, \ldots, k^*$, $\bar{\mathcal{W}}_k = \{\mathrm{d}\theta_{k,1}, \ldots, \mathrm{d}\theta_{k,p_k}\}$. Moreover, the fact that each $\bar{\mathcal{W}}_k$ can be completed into a basis for $\mathcal{H}_k$ implies that the $\mathrm{d}\theta_{k,j}$'s and their corresponding forward shifts are linearly independent.

Introduce the set of indices $\mathcal{I}$ such that $\forall k \in \mathcal{I}$, $\bar{\mathcal{W}}_k \neq \emptyset$. For $k \in \mathcal{I}$ and $1 \leq j \leq k$, define $\phi_{k,j} = (\theta_{k,1}(j-1), \ldots, \theta_{k,p_k}(j-1))$. Let $\bar{n} = \sum_{k\in\mathcal{I}} k\, p_k$. Complete the $\phi_{k,j}$'s by a set of arbitrary functions $\psi = (\psi_{\bar{n}+1}, \ldots, \psi_n)$ such that $(\tilde{x}^1_{k,j}, \tilde{x}^2) = (\phi_{k,j}, \psi)$ defines a diffeomorphism. In new coordinates, system Σ reads

$$
\begin{aligned}
\tilde{x}^1_{k,1}(t+1) &= \tilde{x}^1_{k,2}(t),\\
\tilde{x}^1_{k,2}(t+1) &= \tilde{x}^1_{k,3}(t),\\
&\vdots\\
\tilde{x}^1_{k,k-1}(t+1) &= \tilde{x}^1_{k,k}(t),\\
\tilde{x}^1_{k,k}(t+1) &= \tilde{f}^1_k(\tilde{x}^1(t), \tilde{x}^2(t), u(t)), \qquad k \in \mathcal{I};\\
\tilde{x}^2(t+1) &= \tilde{f}^2(\tilde{x}^1(t), \tilde{x}^2(t), u(t)).
\end{aligned} \tag{18}
$$

Under the state feedback $u = \varphi(\tilde{x}^1(t), \tilde{x}^2(t), v(t))$ where

$$
\tilde{f}^1_k(\tilde{x}^1(t), \tilde{x}^2(t), u(t)) = v_k(t), \qquad \dim v_k(t) = p_k, \qquad k \in \mathcal{I}, \tag{19}
$$

system Σ takes the form (16). Next we prove that (19) has a local solution. The fact that the $\mathrm{d}\theta_{k,j}$'s are linearly independent implies

$$
\operatorname{rank} \frac{\partial \tilde{f}^1}{\partial u} = \sum_{k\in\mathcal{I}} \dim\, \tilde{x}^1_{k,k} = \sum_{k\in\mathcal{I}} p_k,
$$

where $\tilde{f}^1 = [(\tilde{f}^1_k)^T,\ k \in \mathcal{I}]^T$. On the other hand, by (17), one has

$$
\sum_{k\in\mathcal{I}} p_k = p'_1 = \dim_{\mathcal{K}^*} \frac{\bar{\mathcal{H}}_1 + \mathcal{H}_2}{\mathcal{H}_2} = m.
$$

Then, our assertion holds by the Implicit Function Theorem.

The proof is completed by showing that the linear part of (18) under the state feedback (19) has $n^*_1 \geq \cdots \geq n^*_m$ as controllability indices.

The controllability indices of (18) are $\{k_i \mid k_i \in \mathcal{I}\}$ with multiplicities p_{k_i}, i.e., $\{n^j_k = k \mid k \in \mathcal{I},\ 1 \leq j \leq p_k\}$. Up to reordering one has, by (17), $p'_k = \sum_{i\geq k} p_i = \operatorname{card}\{n^j_i \geq k\}$. The result follows because both lists $\{n^*_i\}$, $\{n^j_i\}$ are dual to the list $\{p'_k\}$ and then they are the same. □

We state the following theorem without proof. It can be easily demonstrated by contradiction.

THEOREM 6.3. *If system Σ is feedback linearizable with controllability indices $n_1 \geq \cdots \geq n_m$, then $n_i \leq n^*_i$, $i = 1, \ldots, m$.*

Example 6.4. Consider the nonlinear discrete-time system

$$
\begin{aligned}
x_1(t+1) &= x_2(t) + u_1(t),\\
x_2(t+1) &= x_3(t)u_1(t),\\
x_3(t+1) &= x_3(t)u_2(t),\\
x_4(t+1) &= x_4(t) + u_1(t).
\end{aligned} \tag{20}
$$

For system (20) one can readily check that

$$\frac{\operatorname{span}_{\mathcal{K}}\{\mathrm{d}x(0)\}}{\operatorname{span}_{\mathcal{K}}\{\mathrm{d}x(0)\} \cap \operatorname{span}_{\mathcal{K}}\{\mathrm{d}x(1)\}} \sim \operatorname{span}_{\mathcal{K}}\{\mathrm{d}x_1(0), \mathrm{d}x_3(0)\}.$$

Thus one can set $z(0) = (x_1(0), x_3(0))$. This allows us to compute

$$\mathcal{H}_2 = \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x_4(0) - \mathrm{d}x_1(0),\ \mathrm{d}x_2(0) - z_2(-1)\mathrm{d}x_1(0)\},$$
$$\mathcal{H}_3 = 0.$$

Clearly, $\bar{\mathcal{H}}_2 = \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x_4(0) - \mathrm{d}x_1(0)\}$; thus set $\bar{\mathcal{W}}_2 = \{\mathrm{d}(x_4(0) - x_1(0))\}$ and $\hat{\mathcal{W}}_2 = \{\mathrm{d}x_2(0) - z_2(-1)\mathrm{d}x_1(0)\}$. According to the proof of Theorem 6.2, $\bar{\mathcal{W}}_1$ has to be chosen in such a way that $\bar{\mathcal{W}}_2 \cup \bar{\mathcal{W}}_2^+ \cup \hat{\mathcal{W}}_2 \cup \bar{\mathcal{W}}_1$ is a basis for $\bar{\mathcal{H}}_1 + \mathcal{H}_2$; set $\bar{\mathcal{W}}_1 = \{\mathrm{d}x_3\}$. One concludes that system (20) is partially feedback linearizable with controllability indices $\{2, 1\}$. Consider the diffeomorphism

$$\tilde{x}(t) = (x_4(t) - x_1(t), x_4(t) - x_2(t), x_3(t), x_4(t))\,.$$

In new coordinates, system (20) reads

$$\tilde{x}_1(t+1) = \tilde{x}_2(t), \qquad \tilde{x}_3(t+1) = \tilde{x}_3(t)u_2(t),$$
$$\tilde{x}_2(t+1) = \tilde{x}_4(t) - \tilde{x}_3(t)u_1(t), \qquad \tilde{x}_4(t+1) = \tilde{x}_4(t) + u_1(t).$$

Under the state feedback

$$u_1(t) = \frac{\tilde{x}_4(t) - v_1(t)}{\tilde{x}_3(t)}, \qquad u_2(t) = \frac{v_2(t)}{\tilde{x}_3(t)},$$

system (20) takes form (16) with controllability indices $\{2, 1\}$.

7. Dynamic state feedback linearization. For continuous-time systems it is well known [7, 24, 25] that a sufficient condition for a nonlinear system with outputs to be linearizable by dynamic state feedback is that

1. the system be right-invertible; and
2. the system has no zero dynamics, in the sense of the dynamics of the reduced inverse system [26].

In [44] it was pointed out that a similar property can be deduced for discrete-time systems. However, if a system without outputs is given (or if the outputs of the system do not satisfy the properties above), then it is interesting to decide whether these outputs exist or not. This question is addressed in this section. To be more precise, the existence of outputs satisfying properties 1 and 2 is shown to be equivalent to the existence of a transformal operator that maps a certain Pfaffian system into another one that is completely integrable.

Consider system Σ and suppose that the output function $y(t) = h(x(t))$, $y(t) \in \mathbb{R}^m$, has been specified. Then one has the following definitions and results.

Define a chain of subspaces $\mathcal{E}_0 \subset \mathcal{E}_1 \subset \cdots \subset \mathcal{E}_n$ of $\mathcal{F}$ by

$$\mathcal{E}_k = \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0), \mathrm{d}y(0), \ldots, \mathrm{d}y(k)\} \tag{21}$$

and the associated list of dimensions $\rho_k = \dim_{\mathcal{K}^*} \mathcal{E}_k$ [22].

DEFINITION 7.1. 1. *For $k = 0, \ldots, n$, $\sigma_k = \rho_k - \rho_{k-1}$ is the number of zeros at infinity of order less than or equal to k, with the convention that $\rho_{-1} = n$.*

2. *The rank ρ^* of the system is the total number of zeros at infinity, i.e., $\rho^* = \sigma_n = \rho_n - \rho_{n-1}$.*

3. *The system is said to be invertible if $\rho^* = m$.*

Definition 7.1 gives an abstract characterization of the rank [12, 14, 22] and the structure at infinity [22] for discrete-time nonlinear systems. These notions have, however, a meaningful interpretation in terms of the inversion algorithm [22, 32]. The integer σ_k represents the number of independent scalar inputs that can be recovered at the k-th step of the inversion algorithm. In a similar vein, invertibility of the system means that it is possible to recover the complete input $u(t)$ as a function of the output $y(t)$ and its forward-shifts.

The structure at infinity can be expressed in different manners, which are of course equivalent and suitable for different tasks. For instance, the list $\{n'_1, \ldots, n'_{\rho^*}\}$ of the orders of the zeros at infinity is the list of integers k such that $\sigma_k - \sigma_{k-1} \neq 0$, each one repeated $\sigma_k - \sigma_{k-1}$ times.

In the rest of this section we are concerned with dynamic compensators of the type

$$\mathcal{C} : \begin{cases} \xi(t+1) = a(x(t), \xi(t), v(t)), \\ u(t) = b(x(t), \xi(t), v(t)), \end{cases} \tag{22}$$

where $\xi(t) \in \mathbb{R}^q$ and $v(t) \in \mathbb{R}^m$. Like in the continuous-time case [20, 23], a dynamic compensator $\mathcal{C}$ is said to be regular if it is invertible when $v(t)$ is viewed as input and $u(t)$ is viewed as output. This leads to the characterization of the regularity of the compensator $\mathcal{C}$ as

$$\dim_{\mathcal{K}^*} \frac{\operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0), \ldots, \mathrm{d}x(q), \mathrm{d}\xi(0), \mathrm{d}u(0), \ldots, \mathrm{d}u(q)\}}{\operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0), \ldots, \mathrm{d}x(q), \mathrm{d}\xi(0), \mathrm{d}u(0), \ldots, \mathrm{d}u(q-1)\}} = m.$$

For invertible systems it is always possible to construct a dynamic compensator $\mathcal{C}$ in such a way that

1. noninteracting control is achieved; i.e., for the compensated system $\Sigma \circ \mathcal{C}$ one has $y_i(t + \delta_i) = v_i(t)$, for some integers δ_i.
2. the dimension of the compensated system $\Sigma \circ \mathcal{C}$ is $n + \sum_i (\delta_i - n'_i)$ [46].

Set $n_1 = \sum_i \delta_i$, $n_2 = n - \sum_i n'_i$. The n_1 functions

$$\{y_i(t+j),\ 1 \leq i \leq m,\ 0 \leq j \leq \delta_i - 1\}$$

can be completed by n_2 functions φ_k such that $(\tilde{x}_1, \tilde{x}_2) = (y_i(t+j), \varphi_k)$ defines a diffeomorphism on $\mathbb{R}^{n_1+n_2}$. In new coordinates $\tilde{x}$ the compensated system $\Sigma \circ \mathcal{C}$ reads

$$\Sigma \circ \mathcal{C} : \begin{cases} \tilde{x}^1(t+1) = A^1 \tilde{x}^1(t) + B^1 v(t), \\ \tilde{x}^2(t+1) = f^2(\tilde{x}^1(t), \tilde{x}^2(t), v(t)), \\ y(t) = C^1 \tilde{x}^1(t), \end{cases}$$

where $\dim \tilde{x}^1 = \sum_i \delta_i$, $\dim \tilde{x}^2 = n - \sum_i n'_i$, and the pair (A^1, B^1) is in Brunovsky canonical form with controllability indices $\{\delta_i\}$. The dynamics

$$\tilde{x}^2(t+1) = f^2(\tilde{x}^1(t), \tilde{x}^2(t), v(t))$$

is the zero dynamics in the sense of the reduced inverse system.

DEFINITION 7.2. *System Σ is said to be linearizable by dynamic state feedback if there exist a regular dynamic compensator* (22) *and an extended coordinates transformation $\tilde{x}(t) = \phi(x(t), \xi(t))$ such that, in new coordinates, the compensated system reads*

$$\tilde{x}(t+1) = A\tilde{x}(t) + Bv(t), \qquad t = 0, 1, \ldots, \tag{23}$$

where $\tilde{x} \in \mathbb{R}^{n+q}$ and the pair (A, B) is in Brunovsky canonical form.

From our previous analysis, one concludes that the existence of an output $y(t) = h(x(t))$, $y(t) \in \mathbb{R}^m$, which defines an invertible system without zero dynamics, is a sufficient condition for dynamic feedback linearizability. This property was already recognized in [44, §4]. One can also see that the sum of the orders of the zeros at infinity will play an important role. Lemma 7.3 gives a characterization of this quantity. Define the subspaces

$$\mathcal{X} = \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0)\},\ \mathcal{Y} = \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}y(k),\ k \geq 0\}.$$

LEMMA 7.3. *Suppose that the output $y(t) = h(x(t))$, $y(t) \in \mathbb{R}^m$, defines an invertible system. Then*

$$\dim_{\mathcal{K}^*}(\mathcal{X} \cap \mathcal{Y}) = \sum_i p_i' = \sum_i n_i',$$

where $p_i' = m - \sigma_i$ is the number of zeros at infinity of order greater than or equal to $i + 1$.

Proof. The proof is a direct consequence of the inversion algorithm [22]; see also [32]. An application of the algorithm for invertible systems gives, for each $0 \leq k \leq n$,

$$\dim_{\mathcal{K}^*}\left(\mathcal{X} \cap \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}y(\ell),\ 0 \leq \ell \leq k\}\right) = \sum_{i=0}^{k}(m - \sigma_i) = \sum_{i=0}^{k}(p_1' - \sigma_i) = \sum_{i=0}^{k} p_i'.$$

On the other hand, one has $\sum_i n_i' = \sum_i i(p_i' - p_{i+1}') = \sum_i p_i'$. □

For continuous-time systems, Fliess et al. [17] have shown that a more general formulation consists of allowing the output function to depend explicitly on the input and a finite number of its time-derivatives. Accordingly, for discrete-time systems, a more general problem statement consists of allowing the output $y(t)$ to depend explicitly on $u(t)$ and a finite number $\nu - 1$ of its forward shifts. Within our framework, this implies the statement of a more general definition of the structure at infinity, which includes possibly nonproper discrete-time systems. This can be done by applying the usual definition to an extended system for which the forward-shifts of the input that explicitly appear in the output equation are considered as states. See for example [39] for the continuous-time case. We do not pursue this line further because, as we explain next, it is possible to state the problem without making explicit reference to the extended system. More precisely, we seek an output $y(t) = h(x(t), u(t), \ldots, u(t + \nu - 1))$, $y(t) \in \mathbb{R}^m$, such that

$$\dim_{\mathcal{K}^*}(\mathcal{X}_\nu \cap \mathcal{Y}) = n + m\nu,$$

where $\mathcal{X}_\nu = \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0), \mathrm{d}u(0), \ldots, \mathrm{d}u(\nu - 1)\}$. For square invertible systems one has

$$\mathcal{X}_\nu + \mathcal{Y} = \mathcal{X} + \mathcal{Y}$$

and hence

$$\dim_{\mathcal{K}^*}(\mathcal{X}_\nu \cap \mathcal{Y}) = \dim_{\mathcal{K}^*}(\mathcal{X} \cap \mathcal{Y}) + m\nu.$$

The term *linearizing output* in Definition 7.4 below is borrowed from [17].

DEFINITION 7.4. *A linearizing output is an output function*

$$y(t) = h(x(t), u(t), \ldots, u(t + \nu - 1)), \qquad y(t) \in \mathbb{R}^m,$$

that satisfies the following properties.

1. *$y(t) = h(x(t), u(t), \ldots, u(t + \nu - 1))$ defines an invertible system.*
2. *$\dim_{\mathcal{K}^*}(\mathcal{X} \cap \mathcal{Y}) = n$.*

Remark 7.5. For continuous-time systems [17, 19, 39], a linearizing output has the important property that any variable of the system can be expressed as a function of the linearizing output and a finite number of its time-derivatives, i.e., without integrating any differential equation. The corresponding property for the discrete-time case should be that any variable of the system can be expressed as a function of the linearizing output and a finite number of its forward-shifts, i.e., without solving any difference equation. As a matter of fact, this property can be deduced from Definition 7.4. First, recall that invertibility of the system means that one is able to recover the complete input of the system as a function of the output, a finite number of its forward-shifts, and the state. Second, notice that the second property of Definition 7.4 implies that each state variable can also be expressed as a function of the linearizing output and its forward-shifts.

It is also possible to show that the conditions of Definition 7.4 are independent of the system of coordinates.

Let $\mathcal{K}^*[\Delta]$ denote the set of polynomials in the operator Δ with coefficients in $\mathcal{K}^*$. One can give $\mathcal{K}^*[\Delta]$ the structure of a noncommutative ring with the addition defined in the usual manner and the multiplication defined by the noncommutative operation

$$\Delta\, p = p^{+}\Delta, \qquad \text{for all } p \in \mathcal{K}^*,$$

which corresponds to operators composition. Let $\mathcal{K}^{*m\times m}[\Delta]$ denote the set of $m \times m$ matrices whose entries belong to $\mathcal{K}^*[\Delta]$. The set $\mathcal{K}^{*m\times m}[\Delta]$ is also a noncommutative ring. The elements of $\mathcal{K}^{*m\times m}[\Delta]$ are called transformal operators.

Let $\mathcal{F}^m$ denote the $\mathcal{K}^*$-vector space spanned by m-tuples of one-forms. Every transformal operator $P \in \mathcal{K}^{*m\times m}[\Delta]$ defines a mapping from $\mathcal{F}^m$ to $\mathcal{F}^m$ in the following way. Let $P = \sum_i P_i\,(\Delta)^i \in \mathcal{K}^{*m\times m}[\Delta]$, and let $\Omega(0) = (\omega_1(0), \ldots, \omega_m(0)) \in \mathcal{F}^m$. Then define

$$\begin{aligned} P \;:\; & \mathcal{F}^m \to \mathcal{F}^m, \\ & \Omega \mapsto P\,\Omega = \textstyle\sum_i P_i \Omega(i), \end{aligned}$$

where $\Omega(i) = (\omega_1(i), \ldots, \omega_m(i))$. A transformal operator $P \in \mathcal{K}^{*m\times m}[\Delta]$ is said to be invertible if there exists another transformal operator $Q \in \mathcal{K}^{*m\times m}[\Delta]$ such that

$$\text{for all } \Omega \in \mathcal{F}^m, \qquad P \circ Q\,(\Omega) = Q \circ P\,(\Omega) = \Omega,$$

or, equivalently, $P \circ Q = Q \circ P = I_m$, the identity matrix in $\mathbb{R}^{m\times m}$. The only invertible elements of $\mathcal{K}^*[\Delta]$ are the nonzero polynomials of degree zero.

The concept of infinite zero structure can be generalized in a natural way to arbitrary m-tuples of one-forms $\Omega(t) = (\omega_1(t), \ldots, \omega_m(t)) \in \mathcal{F}^m$, which are not necessarily exact. The m-tuple $\Omega(t)$ is said to have

$$\sigma_k = \dim_{\mathcal{K}^*}\left(\frac{\operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0), \Omega(0), \ldots, \Omega(k)\}}{\operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0), \Omega(0), \ldots, \Omega(k-1)\}}\right) \tag{24}$$

zeros at infinity of order less than or equal to k. If $\Omega(t)$ is a set of exact one-forms ($\omega_i(t) = \mathrm{d}h_i(t)$), then (24) coincides with the definition of structure at infinity given above. In particular, Lemma 7.3 is also valid for m-tuples of one-forms, provided that one defines invertibility of a set of one-forms $\Omega(t)$ by condition $\sigma_n = m$.

Proposition 7.6 below states that the sum of the orders of the zeros at infinity of a system of one-forms cannot increase under the action of a transformal operator $P \in \mathcal{K}^{*m\times m}[\Delta]$.

PROPOSITION 7.6. *Consider the m-tuple of one-forms $\Omega(t) = (\omega_1(t), \ldots \omega_m(t))$ and the polynomial matrix operator $P \in \mathcal{K}^{*m\times m}[\Delta]$. Define $\tilde{\Omega}(t) = P\Omega(t)$. Then*

$$\dim_{\mathcal{K}^*}\left(\mathcal{X} \cap \operatorname{span}_{\mathcal{K}^*}\{\tilde{\Omega}(k), k \geq 0\}\right) \leq \dim_{\mathcal{K}^*}\left(\mathcal{X} \cap \operatorname{span}_{\mathcal{K}^*}\{\Omega(k), k \geq 0\}\right).$$

The proof is straightforward and is left to the reader. Note that Proposition 7.6 establishes that the sum of the orders of the zeros at infinity of an m-tuple of one-forms remains constant under the action of invertible transformal operators $P \in \mathcal{K}^{*m\times m}[\Delta]$.

Let $\Omega(t) = (\omega_1(t), \ldots \omega_m(t))$ be a system of one-forms satisfying the conditions of Theorem 3.5. Then $\Omega(t)$ satisfies the properties

1. $\sigma_n = m$.
2. $\dim_{\mathcal{K}^*}\left(\mathcal{X} \cap \operatorname{span}_{\mathcal{K}}\{\Omega(k),\ k \geq 0\}\right) = n$.

In the case when $\Omega(t)$ is composed of exact one-forms ($\omega_i(t) = \mathrm{d}h_i(t)$), these properties coincide with the conditions in Definition 7.4. Therefore, an m-tuple of one-forms that satisfies the conditions of Theorem 3.5 is called a *system of linearizing one-forms.*

THEOREM 7.7. *Suppose $\mathcal{H}_\infty = 0$, and let $\Omega(t)$ be a system of linearizing one-forms. Then, there locally exists a system of linearizing outputs if and only if there exists an invertible transformal operator $P \in \mathcal{K}^{*m\times m}[\Delta]$ such that*

$$\mathrm{d}(P\Omega) = 0. \tag{25}$$

Hence (25) *is a sufficient condition for system Σ to be dynamic feedback linearizable.*

Proof. Necessity. Suppose $y(t) = h(x(t), u(t), \ldots, u(t+\nu-1))$ is a linearizing output. Definition 7.4 implies that $\mathcal{F} = \mathcal{Y}$. Theorem 3.5 implies that $\mathcal{F} = \operatorname{span}_{\mathcal{K}^*}\{\Omega(k), k \geq 0\}$. Thus there exist transformal operators P, Q such that $\mathrm{d}y(t) = P\,\Omega(t)$ and $\Omega(t) = Q\,\mathrm{d}y(t)$. Clearly $PQ = QP = I_m$, and hence P is invertible. Moreover $\mathrm{d}(P\Omega(t)) = \mathrm{d}(\mathrm{d}y(t)) = 0$.

Sufficiency. Let

$$N = \dim_{\mathcal{K}^*}\left(\mathcal{X} \cap \operatorname{span}_{\mathcal{K}^*}\{\Omega(k),\ k \geq 0\}\right),$$
$$\tilde{N} = \dim_{\mathcal{K}^*}\left(\mathcal{X} \cap \operatorname{span}_{\mathcal{K}^*}\{\tilde{\Omega}(k),\ k \geq 0\}\right),$$

where $\tilde{\Omega}(t) = P\Omega(t)$. Theorem 3.5 implies that $N = n$. Existence of the operator P implies $\tilde{N} \leq N$. Invertibility of P implies the existence of an operator Q such that $\Omega(t) = Q\,\tilde{\Omega}(t)$; i.e., $N \leq \tilde{N}$ and hence $\tilde{N} = N$. The result follows because one can assume, without loss of generality, that $\tilde{\Omega}(t) = \mathrm{d}\psi(x(t), u(t), \ldots, u(t+\nu-1))$. $\psi(x(t), u(t), \ldots, u(t+\nu-1))$ is a linearizing output. □

COROLLARY 7.8. *Suppose $\mathcal{H}_\infty = 0$, and let $\Omega(t) = (\omega_1(t), \ldots, \omega_m(t))$ be a system of linearizing one-forms. Further, suppose that the Frobenius condition*

$$\mathrm{d}\omega_k(t) \wedge \omega_1(t) \wedge \cdots \wedge \omega_m(t) = 0, \qquad k = 1, \ldots, m,$$

is satisfied. Then there exists a linearizing output. The relative degrees of the $\omega_k(t)$'s coincide with the orders of the zeros at infinity of the linearizing output.

Proof. The Frobenius condition implies that the Pfaffian system generated by $\Omega(t) = (\omega_1(t), \ldots, \omega_m(t))$ admits a basis composed of exact one-forms. Therefore there exists an invertible matrix (with entries in $\mathcal{K}^*$) relating this basis to $\{\omega_1(t), \ldots, \omega_m(t)\}$. □

Example 7.9. We already showed that system (20) cannot be fully linearized using regular static state feedback. However, we shall show that system (20) is linearizable by dynamic state feedback. For system (20) we have already computed

$$\mathcal{H}_2 = \operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x_4(0) - \mathrm{d}x_1(0),\ \mathrm{d}x_2(0) - z_2(-1)\mathrm{d}x_1(0)\},$$
$$\mathcal{H}_3 = 0;$$

hence

$$\Omega(t) = \begin{pmatrix} \omega_1(t) \\ \omega_2(t) \end{pmatrix} = \begin{pmatrix} \mathrm{d}x_2(t) - z_2(-1)\mathrm{d}x_1(t) \\ \mathrm{d}x_4(t) - \mathrm{d}x_1(t) \end{pmatrix}$$

is a system of linearizing one-forms. Consider now the transformal operators

$$P(\Delta) = \begin{pmatrix} \frac{1}{1-z_2(-1)} & \frac{z_2(-1)\Delta - z_2(-1)}{1-z_2(-1)} \\ 0 & 1 \end{pmatrix},$$

$$Q(\Delta) = \begin{pmatrix} 1 - z_2(-1) & z_2(-1) - z_2(-1)\Delta \\ 0 & 1 \end{pmatrix}.$$

Straightforward computations show that $P \circ Q = Q \circ P = I_2$ and that

$$\begin{pmatrix} \mathrm{d}x_2(t) \\ \mathrm{d}(x_4(t) - x_1(t)) \end{pmatrix} = P(\Delta) \begin{pmatrix} \omega_1(t) \\ \omega_2(t) \end{pmatrix}.$$

Therefore,

$$y(t) = \begin{pmatrix} x_2(t) \\ x_4(t) - x_1(t) \end{pmatrix} \tag{26}$$

is a linearizing output.

Consider the dynamic state feedback

$$\begin{aligned} \xi(t+1) &= \tilde{u}_1(t), \\ u_1(t) &= \xi(t), \\ u_2(t) &= \tilde{u}_2(t) \end{aligned} \tag{27}$$

and the extended diffeomorphism

$$\tilde{x}(t) = (x_2(t), x_3(t)\xi(t), x_4(t) - x_1(t), x_4(t) - x_2(t), x_4(t) - \xi(t) - x_3(t)\xi(t)).$$

In new coordinates, the extended system (20)–(27) reads

$$\begin{aligned} \tilde{x}_1(t+1) &= \tilde{x}_2(t), \\ \tilde{x}_2(t+1) &= \frac{\tilde{x}_2(t)\tilde{u}_1(t)\tilde{u}_2(t)}{\tilde{x}_2(t) + \tilde{x}_5(t) - \tilde{x}_1(t) - \tilde{x}_4(t)}, \\ \tilde{x}_3(t+1) &= \tilde{x}_4(t), \\ \tilde{x}_4(t+1) &= \tilde{x}_5(t), \\ \tilde{x}_5(t+1) &= \tilde{x}_2(t) + \tilde{x}_5(t) + \tilde{u}_1(t) - \frac{\tilde{x}_2(t)\tilde{u}_1(t)\tilde{u}_2(t)}{\tilde{x}_2(t) + \tilde{x}_5(t) - \tilde{x}_1(t) - \tilde{x}_4(t)}, \end{aligned} \tag{28}$$

and under the state feedback

$$\begin{aligned} \tilde{u}_1(t) &= v_1(t) + v_2(t) - \tilde{x}_2(t) - \tilde{x}_5(t), \\ \tilde{u}_2(t) &= \frac{v_2(t)(\tilde{x}_2(t) + \tilde{x}_5(t) - \tilde{x}_1(t) - \tilde{x}_4(t))}{\tilde{x}_2(t)(v_1(t) + v_2(t) - \tilde{x}_2(t) - \tilde{x}_5(t))} \end{aligned}$$

the extended system (28) reads as a linear system in Brunovsky canonical form with controllability indices $\{3, 2\}$.

The linearizing output (26) is not unique. For instance, the output function

$$\tilde{y}(t) = \begin{pmatrix} x_2(t) \\ x_3(t)u_1(t) + x_4(t) - x_1(t) \end{pmatrix} \tag{29}$$

(which depends on the input) is a linearizing output as well in the sense of Definition 7.4. Both linearizing outputs are related by the invertible transformal operator

$$\tilde{P} = \begin{pmatrix} 1 & 0 \\ \Delta & 1 \end{pmatrix}. \tag{30}$$

Remark 7.10. For continuous-time systems, it has been shown in [8] that any single-input system is dynamic feedback linearizable if and only if it is linearizable by static state feedback. We can state an analogous result for discrete-time systems. Let Σ be a single-input system, and suppose $\mathcal{H}_\infty = 0$. Then a system of linearizing one-forms reduces to any single one-form $\omega_1(t)$ such that $\mathcal{H}_n = \operatorname{span}_{\mathcal{K}^*}\{\omega_1(0)\}$.

THEOREM 7.11. *Let Σ be a single-input system, and suppose $\mathcal{H}_\infty = 0$. Then, the following statements are equivalent.*

1. *Σ admits a linearizing output.*
2. *Σ is linearizable by static state feedback.*
3. *$\mathrm{d}\omega_1(t) \wedge \omega_1(t) = 0$, where $\omega_1(t)$ is such that $\mathcal{H}_n = \operatorname{span}_{\mathcal{K}^*}\{\omega_1(0)\}$.*

The proof of Theorem 7.11 is an immediate consequence of Theorems 5.2 and 7.7 and is left to the reader.

8. Conclusion. The accessibility problem and three different notions of feedback linearizability for discrete-time analytic systems have been addressed in this paper. Our main contribution has been to show that these issues can be organized around the classification of differential forms with respect to their relative degree. The solutions that we have stated along the paper fit within a single linear algebraic framework.

The notion of dynamic feedback linearizability defined in §7 relies on the use of a discrete-time version of a so-called *Singh compensator* [23]. An open problem for further research is to determine to what extent the existence of a linearizing output is necessary for dynamic feedback linearizability using a more general class of dynamic compensators.

A. Construction of $\mathcal{K}^*$. As pointed out in §2, every difference field $\mathcal{K}$ can be embedded into an inversive difference overfield $\mathcal{K}^*$, which is called the inversive closure of $\mathcal{K}$. The inversive closure is unique up to an isomorphism. A precise statement of these properties can be found on pp. 66–67 of Cohn's monograph [11].

We next give an explicit construction of $\mathcal{K}^*$. This construction allows us to carry out the practical computations.

Assume that system Σ is given and is submersive. Hence $(\mathcal{K}, \delta)$ is a difference field. Denote by $\mathcal{E}$ the $\mathcal{K}$-vector space spanned by $\{\mathrm{d}\varphi \mid \varphi \in \mathcal{K}\}$. The operator δ induces a forward-shift operator $\Delta : \mathcal{E} \to \mathcal{E}$ by

$$\Delta\left(\sum_i a_i \mathrm{d}\varphi_i\right) \mapsto \sum_i a_i^+ \mathrm{d}\varphi_i^+,$$

where $a_i, \varphi_i \in \mathcal{K}$. With some abuse of notation, sometimes we write $\omega^+ = \Delta\,\omega$. The pair $(\mathcal{E}, \Delta)$ is a difference vector space [14]. Introduce a (nonunique) vector-valued function $z(0) = \varphi(x(0))$ such that $\mathrm{d}\varphi \in \operatorname{span}_{\mathcal{K}}\{\mathrm{d}x(0)\}$ and

$$\operatorname{span}_{\mathcal{K}}\{\mathrm{d}\bar{z}(0)\} \sim \frac{\operatorname{span}_{\mathcal{K}}\{\mathrm{d}x(0)\}}{\operatorname{span}_{\mathcal{K}}\{\mathrm{d}x(0)\} \cap \operatorname{span}_{\mathcal{K}}\{\mathrm{d}x(1)\}},$$

where $\mathrm{d}\bar{z}(0)$ denotes the coset associated to the element $\mathrm{d}z(0)$. Consequently,

$$\operatorname{span}_{\mathcal{K}}\{\mathrm{d}x(0)\} \subset \operatorname{span}_{\mathcal{K}}\{\mathrm{d}x(1)\} + \operatorname{span}_{\mathcal{K}}\{\mathrm{d}z(0)\}.$$

This implies that (locally) there exist a vector-valued function $\psi(x(1), z(0))$ such that $x(0) = \psi(x(1), z(0))$. Finally, let $\mathcal{K}^*/\mathcal{K}$ be the field extension of meromorphic functions in a finite number of the independent variables

$$\{x(0), u(t), u(-k), z(-k),\ t \geq 0\ k \geq 1\}.$$

Notice that although the choice of a variable $z(0) = \varphi(x(0))$ is not unique, each possible choice brings up a field extension that is isomorphic to $\mathcal{K}^*/\mathcal{K}$.

The pair $(\mathcal{K}^*, \delta)$ can be given the structure of an inversive difference field using the usual rules and noting that

$$\delta^{-1}x(0) = \delta^{-1}\psi(x(1), z(0)) = \psi(x(0), z(-1)).$$

B. Integrability of $\mathcal{H}_\infty$. Let $\mathcal{F}^*$ be the algebraic dual of $\mathcal{F}$. More precisely, $\mathcal{F}^*$ is the space of linear mappings from $\mathcal{F}$ to $\mathcal{K}^*$. As pointed out in §4 (see Lemma 4.3), the filtration

$$\operatorname{span}_{\mathcal{K}^*}\{\mathrm{d}x(0)\} = \mathcal{H}_1 \supset \cdots \supset \mathcal{H}_{k^*} \supset \mathcal{H}_\infty$$

can be viewed as a nested sequence of Pfaffian systems defined over $\mathbb{R}^N$, for some integer N large enough. Therefore, the elements of $\mathcal{F}^*$ can be viewed as sections or vector fields of the tangent bundle $T\mathbb{R}^N$. These vector fields can be written in the manner

$$X = \sum_{i,j} a_i \frac{\partial}{\partial x_i(0)} + \sum_{i,j} b_{i,j} \frac{\partial}{\partial u_i(j)} + \sum_{i,j} c_{i,j} \frac{\partial}{\partial u_i(-j)} + \sum_{i,j} d_{i,j} \frac{\partial}{\partial z_i(-j)},$$

where $a_i, b_{i,j}, c_{i,j}, d_{i,j} \in \mathcal{K}^*$, all sums are finite, and the set

$$\left\{ \frac{\partial}{\partial x_i(0)}, \frac{\partial}{\partial u_i(j)}, \frac{\partial}{\partial u_i(-j)}, \frac{\partial}{\partial z_i(-j)} \right\}$$

is defined to be a dual basis of the canonical basis $\{\mathrm{d}x_i(0), \mathrm{d}u_i(j), \mathrm{d}u_i(-j), \mathrm{d}z_i(-j)\}$ of $\mathcal{F}$. Given a vector field $X \in \mathcal{F}^*$, its forward-shift X^+ is defined by

$$\langle X, \omega\rangle^+ = \langle X^+, \omega^+\rangle \quad \text{for all } \omega \in \mathcal{F}. \tag{31}$$

Formula (31) has to be interpreted in the following manner. First notice that $\langle X, \omega\rangle \in \mathcal{K}^*$, so that $\langle X, \omega\rangle^+ = \delta\langle X, \omega\rangle$ is well defined. Therefore, evaluating (31) with different choices of ω, we obtain a system of equations that uniquely defines X^+. Let us investigate this by an example.

Example B.1 (Example 4.7 continued). Consider the nonlinear system described by (11), and let $X = \frac{\partial}{\partial x_3(0)}$ be an element of $\mathcal{F}^*$. Assume that X^+ has the form

$$X^+ = \sum_{i=1}^{3} a_i \frac{\partial}{\partial x_i(0)} + \sum_{j=0}^{N} b_j \frac{\partial}{\partial u(j)},$$

where $a_i, b_j \in \mathcal{K}^*$. Evaluating (31) with different choices of ω yields the system of equations

$$\begin{aligned}
&\langle X, \mathrm{d}x_1(0)\rangle^+ = 0 = \langle X^+, \mathrm{d}x_1(1)\rangle = (x_3^2(0)+1)^2 a_1 + 4x_1(0)x_3(0)(x_3^2(0)+1)a_3,\\
&\langle X, \mathrm{d}x_2(0)\rangle^+ = 0 = \langle X^+, \mathrm{d}x_2(1)\rangle = (x_3^2(0)+1)^3 a_2 + 6x_2(0)x_3(0)(x_3^2(0)+1)^2 a_3,\\
&\langle X, \mathrm{d}x_3(0)\rangle^+ = 1 = \langle X^+, \mathrm{d}x_3(1)\rangle = a_3 + b_0,\\
&\langle X, \mathrm{d}u(j)\rangle^+ = 0 = \langle X^+, \mathrm{d}u(j+1)\rangle = b_{j+1}, \qquad j \geq -1.
\end{aligned} \tag{32}$$

Equations (32) have the unique solution $b_j = 0$, $a_3 = 1$, $a_2 = -6x_2(0)x_3(0)/(x_3^2(0)+1)$, $a_1 = -4x_1(0)x_3(0)/(x_3^2(0)+1)$. Therefore, one concludes that

$$X^+ = -4\left[\frac{x_1(0)x_3(0)}{(x_3^2(0)+1)}\right]\frac{\partial}{\partial x_1(0)} - 6\left[\frac{x_2(0)x_3(0)}{(x_3^2(0)+1)}\right]\frac{\partial}{\partial x_2(0)} + \frac{\partial}{\partial x_3(0)}.$$

We now prove that $\mathcal{H}_\infty$, when viewed as a Pfaffian system defined on $T^*\mathbb{R}^N$, is completely integrable.

Let $\{\alpha_1, \ldots, \alpha_{\rho_\infty}\}$ be a basis for $\mathcal{H}_\infty$. The characteristic vector fields $\mathcal{H}_\infty$ are the elements of the set

$$\mathcal{G}_\infty = \{X \in \mathcal{F}^* \mid \langle X, \omega\rangle = 0,\ \alpha_1 \wedge \cdots \wedge \alpha_{\rho_\infty} \wedge X \lrcorner d\omega = 0\ \forall\, \omega \in \mathcal{H}_\infty\}.$$

The characteristic system of $\mathcal{H}_\infty$ is defined to be $C(\mathcal{H}_\infty) = \mathcal{G}_\infty^\perp$. It is clear that $\mathcal{H}_\infty \subset C(\mathcal{H}_\infty)$. The rest of the proof consists of showing the converse inclusion.

First we show that $\mathcal{G}_\infty$ is invariant under forward-shifting. Let $X \in \mathcal{G}_\infty$, and $\omega \in \mathcal{H}_\infty$. Therefore one has

$$\begin{aligned} \langle X, \omega\rangle^+ &= \langle X^+, \omega^+\rangle = 0, \\ (\alpha_1 \wedge \cdots \wedge \alpha_{\rho_\infty} \wedge X \lrcorner d\omega)^+ &= \alpha_1^+ \wedge \cdots \wedge \alpha_{\rho_\infty}^+ \wedge X^+ \lrcorner d\omega^+ = 0. \end{aligned} \tag{33}$$

The fact that $\mathcal{H}_\infty$ is closed under forward-shifting implies that $\alpha_i^+, \omega^+ \in \mathcal{H}_\infty$. Hence, $\mathcal{G}_\infty$ is closed under forward-shifting because (33) hold for any $\omega \in \mathcal{H}_\infty$.

Next we show that $C(\mathcal{H}_\infty)$ is also closed under forward-shifting. Let $\eta \in C(\mathcal{H}_\infty)$ and $X \in \mathcal{G}_\infty$. Then one has

$$\langle X, \eta\rangle^+ = \langle X^+, \eta^+\rangle = 0. \tag{34}$$

The fact that $\mathcal{G}_\infty$ is closed under forward-shifting implies that $X^+ \in \mathcal{G}_\infty$. Hence $C(\mathcal{H}_\infty)$ is closed under forward-shifting because (34) holds for any $X \in \mathcal{G}_\infty$.

Finally note that $\mathcal{H}_\infty$ is the largest subspace of $\mathcal{H}_1$ that is closed under forward-shifting so that $C(\mathcal{H}_\infty) \subset \mathcal{H}_\infty$. We have shown that $\mathcal{H}_\infty = C(\mathcal{H}_\infty)$, and the result follows.

Acknowledgments. The authors are greatly indebted to the anonymous referees for their insightful comments. The first and the third authors would also like to thank Professor J. W. Grizzle for his valuable comments on the subject of this paper.

This work was performed in part with the auspices of the GR Automatique of CNRS, France. The work of E. Aranda-Bricaire was done while he was with Laboratoire d'Automatique de Nantes with the support of CONACYT and CINVESTAV-IPN, Mexico. The work of Ü. Kotta was done during her visit at Laboratoire d'Automatique de Nantes with the support of the European Union.

REFERENCES

[1] R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis and Applications*, 2nd ed., Appl. Math. Sci., 75, Springer, New York, 1988.

[2] F. Albertini and E. D. Sontag, *Discrete-time transitivity and accessibility: Analytic systems*, SIAM J. Control Optim., 31 (1993), pp. 1599–1622.

[3] E. Aranda-Bricaire, Ü. Kotta, and C. H. Moog, *A linear algebraic framework for feedback linearization of discrete-time nonlinear systems*, in Proc. 1st IFAC Workshop on New Trends in Design of Control Systems, Smolenice, Slovakia, June 1994, pp. 35–40.

[4] ———, *Accessibility and feedback linearization of discrete-time systems*, in Proc. 33rd IEEE Conference on Decision Control, Lake Buena Vista, 1994, pp. 1627–1632.

[5] E. Aranda-Bricaire, C. H. Moog, and J.-B. Pomet, *A linear algebraic framework for dynamic feedback linearization*, IEEE Trans. Automat. Control, 40 (1995), pp. 127–132.

[6] R. L. Bryant, S. S. Chern, R. B. Gardner, H. L. Goldschmidt, and P. A. Griffiths, *Exterior Differential Systems*, Math. Sci. Res. Inst. Publ., 18, Springer, New York, 1991.

[7] C. I. Byrnes and A. Isidori, *Exact linearization and zero dynamics*, in Proc. 29th IEEE Conference on Decision Control, Honolulu, 1990, pp. 2080–2084.

[8] B. Charlet, J. Lévine, and R. Marino, *On dynamic feedback linearization*, Systems Control Lett., 13 (1989), pp. 143–151.

[9] D. CHENG, *Linearization with dynamic compensation*, J. Systems Sci. Math. Sci., 7 (1987), pp. 63–83.
[10] Y. CHOQUET-BRUHAT, C. DEWITT-MORETTE, AND M. DILLARD-BLEICK, *Analysis, Manifolds and Physics, Part I: Basics*, North-Holland, Amsterdam, 1989.
[11] R. M. COHN, *Difference Algebra*, Wiley-Interscience, New York, 1965.
[12] S. EL ASMI AND M. FLIESS, *Invertibility of discrete-time systems*, in Proc. 2nd IFAC Symposium on Nonlinear Control Systems Design, Bordeaux, 1992, pp. 192–196.
[13] M. FLIESS, *Should the theories for continuous-time and discrete-time linear and nonlinear systems really look alike?*, in Proc. IFAC Symposium on Nonlinear Control Systems Design, Capri, 1989, pp. 186–191.
[14] ———, *Automatique en temps discret et algèbre aux différences*, Forum Math., 2 (1990), pp. 213–232.
[15] ———, *Reversible linear and nonlinear discrete-time dynamics*, IEEE Trans. Automat. Control, 37 (1992), pp. 1144–1153.
[16] ———, *Invertibility of causal discrete time dynamical systems*, J. Pure Appl. Algebra, 86 (1993), pp. 173–179.
[17] M. FLIESS, J. LÉVINE, P. MARTIN, AND P. ROUCHON, *Sur les systèmes non linéaires différentiellement plats*, C. R. Acad. Sci. Paris, Sér. I Math., 315 (1992), pp. 619–624.
[18] ———, *Linéarisation par bouclage dynamique et transformations de Lie-Bäcklund*, C. R. Acad. Sci., Paris, Sér. I Math., 317 (1993), pp. 981–986.
[19] ———, *Flatness and defect of nonlinear systems: Introductory and examples*, Internat. J. Control, to appear.
[20] A. GLUMINEAU, *Solutions Algébriques pour l'Analyse et le Contrôle des Systèmes Non Linéaires*, Thèse de Docteur ès Sciences, École Centrale de Nantes, 1992.
[21] J. W. GRIZZLE, *Feedback Linearization of Discrete-Time Systems*, Lecture Notes in Control and Inform. Sci., Springer, New York, 83 (1986), pp. 273–281.
[22] ———, *A linear algebraic framework for the analysis of discrete-time nonlinear systems*, SIAM J. Control Optim., 31 (1993), pp. 1026–1044.
[23] H. J. C. HUIJBERTS, H. NIJMEIJER, AND L. L. M. VAN DER WEGEN, *Minimality of dynamic input-output decoupling for nonlinear systems*, Systems Control Lett., 18 (1992), pp. 435–443.
[24] A. ISIDORI, *Nonlinear Control Systems*, 2nd ed., Springer, Berlin, 1989.
[25] A. ISIDORI, C. H. MOOG, AND A. DE LUCA, *A sufficient condition for full linearization via dynamic state feedback*, in Proc. 25th IEEE Conference on Decision Control, Athens, 1986, pp. 203–208.
[26] A. ISIDORI AND C. H. MOOG, *On the nonlinear equivalence of the notions of transmission zeros*, in Modelling and Adaptive Control, C. I. Byrnes and A. Kurzhanski, eds., Lecture Notes in Control and Inform. Sci. 105, Springer, New York, 1988, pp. 146–158.
[27] B. JAKUBCZYK, *Feedback linearization of discrete-time systems*, Systems Control Lett., 9 (1987), pp. 411–416.
[28] ———, *Invariants of dynamic feedbacks and free systems*, in Proc. 2nd European Control Conf., Groningen, 1993, pp. 1510–1513.
[29] B. JAKUBCZYK AND D. NORMAND-CYROT, *Orbites de pseudo-groupes de difféomorphismes et commandabilité des systèmes non linéaires en temps discret*, C. R. Acad. Sci. Paris Sér. I Math., 298 (1984), pp. 257–260.
[30] B. JAKUBCZYK AND E. D. SONTAG, *Controllability of nonlinear discrete-time systems: A Lie-algebraic approach*, SIAM J. Control Optim., 28 (1990), pp. 1–33.
[31] G. JAYARAMAN AND H. J. CHIZECK, *Feedback linearization of discrete-time systems*, in Proc. 32nd IEEE Conference on Decision Control, San Antonio, 1993, pp. 2972–2977.
[32] Ü. KOTTA, *Right inverse of a discrete-time non-linear system*, Internat. J. Control, 51 (1990), pp. 1–9.
[33] Ü. KOTTA AND H. NIJMEIJER, *On dynamic input-output linearization of discrete-time nonlinear systems*, Internat. J. Control, 60 (1994), pp. 1319–1337.
[34] H. G. LEE, A. ARAPOSTATIS, AND S. I. MARCUS, *Linearization of discrete-time systems*, Internat. J. Control, 45 (1987), pp. 1803–1882.
[35] H. G. LEE AND S. I. MARCUS, *Approximate and local linearization of nonlinear discrete-time systems*, Internat. J. Control, 44 (1986), pp. 1103–1124.
[36] ———, *On input-output linearization of discrete-time nonlinear systems*, Systems Control Lett., 8 (1987), pp. 249–259.
[37] ———, *Immersion and immersion by nonsingular feedback of a discrete-time nonlinear system into a linear system*, IEEE Trans. Automat. Control, 33 (1988), pp. 479–483.
[38] R. MARINO, *On the largest feedback linearizable subsystem*, Systems Control Lett., 6 (1986), pp. 345–351.
[39] P. MARTIN, *Contribution à l'étude des systèmes differentiellement plats*, Thèse de Doctorat, Ecole des Mines de Paris, 1992.
[40] S. MONACO AND D. NORMAND-CYROT, *Sur la subordination d'un système non linéaire discret à un système linéaire*, in Outils et Modèles Mathématiques pour L'Automatique, l'Analyse des systèmes et le Traitement du Signal, I. Landau, ed., CNRS, Paris, 1982, pp. 609–621.
[41] ———, *The immersion under feedback of a multidimensional discrete time nonlinear system into a linear system*, Internat. J. Control, 38 (1983), pp. 245–261.
[42] ———, *Formal power series and input-output linearization of nonlinear discrete time systems*, in Proc. 22nd IEEE Conference on Decision Control, San Antonio, 1983, pp. 665–670.

[43] S. MONACO AND D. NORMAND-CYROT, *Sur la commande non interactive des systèmes non linéaires en temps discret*, in Proc. 6th International Conference on Analysis and Optimization of Systems, Nice, June 1984, Analysis and Optimization of Systems 63, A. V. Balakrishnan and M. Thoma, eds., Springer-Verlag, Berlin, 1984, pp. 364–377.

[44] ———, *Minimum-phase nonlinear discrete-time systems and feedback stabilization*, in Proc. 26th IEEE Conference on Decision Control, Los Angeles, 1987, pp. 979–986.

[45] K. NAM, *Linearization of discrete-time nonlinear systems and a canonical structure*, IEEE Trans. Automat. Control, 34 (1989), pp. 119–121.

[46] H. NIJMEIJER AND A. VAN DER SCHAFT, *Nonlinear Dynamical Control Systems*, Springer, New York, 1990.

[47] P. ROUCHON, *Necessary condition and genericity of dynamic feedback linearization*, J. Math. Systems, Estimation and Control, 4 (1994), pp. 257–260.

[48] W. F. SHADWICK, *Absolute equivalence and dynamic feedback linearization*, Systems Control Lett., 15 (1990), pp. 35–39.

[49] W. M. SLUIS, *A necessary condition for dynamic feedback linearization*, Systems Control Lett., 21 (1993), pp. 277–283.

[50] M. SPIVAK, *A Comprehensive Introduction to Differential Geometry*, Publish or Perish, Houston, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 2024–2050, November 1996

009

NUMERICAL STABILIZATION OF BILINEAR CONTROL SYSTEMS*

LARS GRÜNE†

Abstract. Extremal Lyapunov exponents for bilinear control systems with constrained control values are computed numerically by solving discounted optimal control problems. Based on this computation a numerical algorithm to calculate stabilizing control functions is developed.

Key words. stabilization, bilinear control systems, Lyapunov exponents, discounted optimal control problems, Hamilton–Jacobi–Bellman equation

AMS subject classifications. 93D22, 49L25

1. Introduction.

In this paper we present numerical algorithms for the calculation of extremal Lyapunov exponents and stabilization of bilinear control systems in $\mathbb{R}^d$, i.e., systems of the form

$$\dot{x}(t) = \Big(A_0 + \sum_{i=1}^{m} u_i(t) A_i\Big) x(t), \quad x(0) = x_0 \in \mathbb{R}^d \setminus \{0\} \tag{1.1}$$

with $A_j \in \mathbb{R}^{d\times d}$, $j = 0, \dots, m$, $u(\cdot) \in \mathcal{U} := \{u : \mathbb{R} \to U,\ u \text{ measurable}\}$ with a compact and convex set of control values $U \subset \mathbb{R}^m$ with nonvoid interior. The Lyapunov exponent of (1.1) with respect to an initial value $x_0 \in \mathbb{R}^d$ and a control function $u(\cdot) \in \mathcal{U}$ is given by

$$\lambda(x_0, u(\cdot)) := \limsup_{t\to\infty} \frac{1}{t} \ln \|x(t, x_0, u(\cdot))\|,$$

where $x(t, x_0, u(\cdot))$ denotes the trajectory of (1.1).

Bilinear control systems arise, e.g., by linearization of a nonlinear control system with a common fixed point x^* for all control values $u \in U$ with respect to x. They were first studied systematically by Mohler [18] in 1973. Lyapunov exponents were introduced by A.V. Lyapunov in 1892 (under the name of order numbers) as a tool to study nonlinear differential equations via their linearizations along trajectories. Recent results about the Lyapunov spectrum of families of time-varying matrices (cf. Colonius and Kliemann [11]) made it possible to characterize the domain of null controllability of bilinear systems using Lyapunov exponents (cf. Colonius and Kliemann [10]). A basic property of the Lyapunov exponents is that $\lambda(x, u(\cdot)) < 0$ iff $x(t, x_0, u(t))$ converges to zero faster than any exponential e^{at} with $\lambda(x_0, u(\cdot)) < a < 0$. As an easy consequence $\inf_{u(\cdot)\in\mathcal{U}} \lambda(x_0, u(\cdot)) < 0$ implies that there exists a control function such that the corresponding trajectory converges to zero. The domain of null controllability—the set of all points x_0 with negative minimal Lyapunov exponent—may be only a part of $\mathbb{R}^d$ and as a consequence stabilization may only be possible for subsets of $\mathbb{R}^d$. Null controllability in this context always means asymptotical null controllability since the origin is not reachable in finite time from any other point of the state space. This implies that an approach via the minimum time function (cf. e.g., Bardi and Falcone [1]) does not apply here.

In contrast to the direct approach to this stabilization problem via Lyapunov functions (cf., e.g., Chabour, Sallet, and Vivalda [5]) the method developed here is in some sense an indirect approach:

*Received by the editors August 3, 1994; accepted for publication (in revised form) September 8, 1995.

†Institut für Mathematik, Universität Augsburg, Universitätsstr. 8, 86135 Augsburg, Germany (Lars.Gruene@Math.Uni-Augsburg.de).

First a numerical approximation of the extremal Lyapunov exponents of (1.1) is calculated. This enables us to characterize the stability properties of (1.1). Once this approximation is known we stabilize the system (i.e., we find control functions such that the corresponding trajectories converge to zero) by searching for control functions such that the corresponding Lyapunov exponent is close to the minimal exponent or at least negative. In §2 these problems are discussed in terms of optimal control theory. We show that the problem of calculating extremal Lyapunov exponents—which can be expressed as an average yield optimal control problem—can be approximated by discounted optimal control problems.

If we look at the uncontrolled system with $U = \{0\}$ it turns out that the Lyapunov exponents are just the real parts of the eigenvalues of A_0. Together with the corresponding eigenspaces they determine the stability properties of the system. For the controlled system we need suitable generalizations of eigenspaces associated with the Lyapunov exponents. The basic ideas of this concept are presented in §3, followed by an interpretation of the results of §2 in terms of calculating extremal Lyapunov exponents.

Section 4 presents algorithms to solve discounted optimal control problems numerically based on a discretization scheme by Capuzzo Dolcetta [2], Capuzzo Dolcetta and Ishii [4], and Falcone [12], [13] connected to the framework of dynamic programming (cf. [3]). Section 5 contains several numerical examples calculated with these algorithms.

2. Discounted and average cost optimal control problem. In this section we will show that average yield optimal control problems can be approximated by discounted optimal control problems.

Consider a control system on a *connected n-dimensional C^∞-manifold M* given by

$$\dot{x}(t) = X(x(t), u(t)) \quad \text{for all } t \in \mathbb{R}, \tag{2.1}$$

$$x(0) = x_0 \in M, \tag{2.2}$$

$$u(\cdot) \in \mathcal{U} := \{u : \mathbb{R} \to U \mid u \text{ measurable}\}, \tag{2.3}$$

with

$$U \subseteq \mathbb{R}^m \text{ compact}, \tag{2.4}$$

$$X(\cdot, u) \text{ is a } C^\infty\text{−vector field on } M, \text{ continuous on } M \times U, \tag{2.5}$$

$$\text{for all } x \in M,\ u(\cdot) \in \mathcal{U} \text{ the trajectory } \varphi(t, x, u(\cdot)) \text{ exists for all } t \in \mathbb{R}. \tag{2.6}$$

We now consider the following two optimal control problems given by the control system (2.1)–(2.6) and a cost function g satisfying

$$g : M \times U \to \mathbb{R} \text{ continuous on } M \times U, \tag{2.7}$$

$$|g(x, u)| \le M_g \quad \text{for all } (x, u) \in M \times U. \tag{2.8}$$

The δ-discounted cost for $\delta > 0$ and the average cost are defined by

$$J_\delta(x, u(\cdot)) := \int_0^\infty e^{-\delta t} g(\varphi(t, x, u(\cdot)), u(t))dt, \tag{2.9}$$

$$J_0(x, u(\cdot)) := \limsup_{T\to\infty} \frac{1}{T} \int_0^T g(\varphi(t, x, u(\cdot)), u(t))dt. \tag{2.10}$$

The associated optimal value functions are

$$v_\delta(x) := \inf_{u(\cdot)\in\mathcal{U}} J_\delta(x, u(\cdot)), \tag{2.11}$$

$$v_0(x) := \inf_{u(\cdot)\in\mathcal{U}} J_0(x, u(\cdot)). \tag{2.12}$$

A basic property of the discounted optimal value function is Bellman's optimality principle: for any $t > 0$ we have

$$v_\delta(x) = \inf_{u(\cdot)\in\mathcal{U}} \left\{ \int_0^t e^{-\delta s} g(\varphi(s, x, u(\cdot)), u(s))ds \;+\; e^{-\delta t} v_\delta(\varphi(t, x, u(\cdot))) \right\}. \tag{2.13}$$

For the average cost a similar estimate is valid: for any $t > 0$ we have

$$v_0(x) = \inf_{u(\cdot)\in\mathcal{U}} \{v_0(\varphi(t, x, u(\cdot)))\}. \tag{2.14}$$

Results about the relation between discounted and average cost optimal control problems as the discount rate tends to zero have been developed by Colonius [6] and Wirth [20]. Here we will first show the relation between the values of δJ_δ and J_0 along certain trajectories. Then we will use similar techniques as in [6] and [20] to obtain convergence results for the optimal value functions. The first theorem shows that J_0 is bounded if δJ_δ is bounded. Since J_0 has an infinite time horizon it is not sufficient that δJ_δ is bounded for the initial value. It has to be bounded for all $\varphi(t, x, u(\cdot))$, $t > 0$, and the corresponding shifted control function.

THEOREM 2.1 (approximation theorem I). *Consider optimal control systems on M given by* (2.1)–(2.6) *and* (2.7)–(2.10)*, a discount rate $\delta > 0$, $x \in M$, $u(\cdot) \in \mathcal{U}$, $C \in \mathbb{R}$, and $\alpha > 0$ such that $\delta J_\delta(\varphi(t, x, u(\cdot)), u(t + \cdot)) \le C - \alpha$ for all $t \ge 0$. Then*

$$J_0(x, u(\cdot)) < C.$$

Proof. We may assume $C = 0$ by using $g - C$ instead of g. In the first step we show that for every $t > 0$ there exists a $\tilde\tau(t)$ such that

$$\int_t^{\tilde\tau(t)} g(\varphi(s, x, u(\cdot)), u(s))ds \le -\frac{\alpha}{2\delta}. \tag{2.15}$$

Abbreviate $f(s) := e^{-\delta(s-t)} g(\varphi(s, x, u(\cdot)), u(s))$. Obviously there exists a $\tilde\tau(t)$ such that (2.15) is true for the shifted discounted functional $\int_t^{\tilde\tau(t)} f(s)ds \le -\frac{\alpha}{2\delta}$. Choose $\tilde\tau(t)$ minimal with this property. Since g is bounded there exist constants $a, b > 0$ such that $\tilde\tau(t) - t \in [a, b]$ for all $t > 0$, $a = \frac{\alpha}{2\delta M_g}$. In the case of $\int_t^{\tilde\tau(t)} f^+(s)ds = 0$, (2.15) is immediately implied. In the case of $\int_t^{\tilde\tau(t)} f^+(s)ds > 0$ it follows that $\int_t^{\tilde\tau(t)} f^-(s)ds < -\frac{\alpha}{2\delta}$, and we can choose $\gamma > 0$ maximal such that $\int_t^{t+\gamma} f^-(s)ds = -\frac{\alpha}{2\delta}$. Hence we have

$$\int_t^\tau f^+(s)ds - \int_{t+\gamma}^\tau f^-(s)ds > 0 \quad \text{for all } \tau \in [t + \gamma, \tau(\tilde t))$$

and

$$\int_t^{\tilde\tau(t)} f^+(s)ds - \int_{t+\gamma}^{\tilde\tau(t)} f^-(s)ds = 0.$$

Fixing $\varepsilon > 0$ we can define a monotone increasing sequence (τ_i), $i \in \mathbb{N}$ by $\tau_1 := t$, $\tau_2 := t + \gamma$,

$$\tau_{i+1} := \max \left\{ \tau \in [\tau_i, \tilde\tau(t) \mid \int_{\tau_{i-1}}^{\tau_i} f^+(s)ds = \int_{\tau_i}^{\tau_{i+1}} f^-(s)ds \right\}.$$

From the construction of this sequence it follows that τ_i converges to $\tilde{\tau}(t)$, and we may truncate the sequence by choosing $k \in \mathbb{N}$ such that $|\tau_{k-1} - \tilde{\tau}(t)| < \varepsilon$ and set $\tau_k := \tilde{\tau}(t)$. Now we can estimate

$$\begin{aligned}
\int_t^{\tilde{\tau}(t)} g(\varphi(s,x,u(\cdot)),u(s))ds &= \int_t^{\tilde{\tau}(t)} e^{\delta(s-t)} f(s)ds \\
&\leq \sum_{i=2}^{n-1} \left(\int_{\tau_{i-1}}^{\tau_i} e^{\delta(s-t)} f^+(s)ds - \int_{\tau_i}^{\tau_{i+1}} e^{\delta(s-t)} f^-(s)ds \right) + M_g\varepsilon - \frac{\alpha}{2\delta} \\
&\leq \sum_{i=2}^{n-1} \left(\underbrace{\int_{\tau_{i-1}}^{\tau_i} e^{\delta(\tau_i-t)} f^+(s)ds - \int_{\tau_i}^{\tau_{i+1}} e^{\delta(\tau_i-t)} f^-(s)ds}_{=0} \right) + M_g\varepsilon - \frac{\alpha}{2\delta} \\
&= M_g\varepsilon - \frac{\alpha}{2\delta},
\end{aligned}$$

which proves (2.15) since $\varepsilon > 0$ was arbitrary.

To prove the theorem we first fix $T > 0$ and define a sequence $(\tilde{\tau}_i)$, $1 \leq i \leq k$ by $\tilde{\tau}_0 := 0$, $\tilde{\tau}_{i+1} := \tilde{\tau}(\tilde{\tau}_i)$, as long as $\tilde{\tau}(\tilde{\tau}_i) \leq T$, $\tilde{\tau}_k := T$. Then we have $a \leq \tilde{\tau}_{i+1} - \tilde{\tau}_i \leq b$ for all $i = 0, \ldots, k-1$ and hence $\frac{T}{b} \leq k \leq \frac{T}{a}$. By definition of $\tilde{\tau}(t)$ it follows that $\int_{\tilde{\tau}_i}^{\tilde{\tau}_{i+1}} g(\varphi(t,x,u(\cdot)),u(t))dt < -\frac{\alpha}{2\delta}$ for all $i = 0, \ldots, k-2$. This yields

$$\begin{aligned}
&\int_0^T g(\varphi(t,x,u(\cdot)),u(t))dt \\
&= \sum_{i=0}^{k-2} \int_{\tilde{\tau}_i}^{\tilde{\tau}_{i+1}} g(\varphi(t,x,u(\cdot)),u(t))dt + \int_{\tilde{\tau}_{k-1}}^{\tilde{\tau}_k} g(\varphi(t,x,u(\cdot)),u(t))dt \\
&\leq -\frac{k\alpha}{2\delta} + (\tilde{\tau}_k - \tilde{\tau}_{k-1})M_g \leq -\frac{T\alpha}{2b\delta} + bM_g \qquad (2.16)
\end{aligned}$$

and as a conclusion

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T g(\varphi(t,x,u(\cdot)),u(t))dt \leq \limsup_{T\to\infty} -\frac{\alpha}{2b\delta} + \frac{bM_g}{T} = -\frac{\alpha}{2b\delta} < 0,$$

which finishes the proof. □

Note that it is possible just to replace $\leq$ by $\geq$ and $-\alpha$ by $+\alpha$ to obtain the analogous result for a lower bound of J_0.

THEOREM 2.2 (approximation theorem II). *Consider optimal control systems on M given by* (2.1)–(2.6) *and* (2.7)–(2.10).

Assume there exists a control function $u(\cdot) \in \mathcal{U}$ such that $J_0(x,u(\cdot)) \leq C - \alpha$ for constants $C \in \mathbb{R}$, $\alpha > 0$. Then there exists a constant $R = R(x,u(\cdot),\alpha) > 0$ such that

$$\delta J_\delta(x,u(\cdot)) < C \quad \text{for all } \delta < R.$$

Proof. We may again assume $C = 0$. Hence it follows that there exists $T_0 \geq 0$ such that

$$\int_0^T g(\varphi(t,x,u(\cdot)),u(t))dt \leq T\left(-\frac{\alpha}{2}\right) \quad \text{for all } T \geq T_0. \qquad (2.17)$$

Now assume that $\delta J_\delta(x, u(\cdot)) \geq 0$ for arbitrarily small $\delta > 0$. The first step of the proof of Theorem 2.1 for the opposite inequality with $t = 0$ applied to $g + \frac{\alpha}{2}$ then yields that there exist arbitrarily large times $\tilde{T} > 0$ such that

$$\int_0^{\tilde{T}} g(\varphi(t, x, u(\cdot)), u(t))dt + \tilde{T}\frac{\alpha}{2} > 0,$$

which contradicts (2.17). Hence the assertion follows. □

In contrast to the first approximation theorem here it is not possible simply to replace $\leq$ by $\geq$ and $-\alpha$ by $+\alpha$ to obtain an analogous result for the lower bound. Estimate (2.17) does only hold for the reverse inequality if in (2.10) the lim sup is replaced by the lim inf.

We will now combine these two theorems with controllability properties to obtain results about the relation between δv_δ and v_0 as δ tends to zero. To do this we first introduce some definitions.

DEFINITION 2.3. *The* positive orbit *of $x \in M$ up to the time T is defined by*

$$O_T^+(x) := \{y \in M \mid \text{there is } 0 \leq t \leq T \text{ and } u(\cdot) \in \mathcal{U}, \text{ such that } \varphi(t, x, u(\cdot)) = y\}.$$

The positive orbit *of $x \in M$ is defined by*

$$O^+(x) := \bigcup_{T \geq 0} O_T^+(x).$$

The negative orbits $O_T^-(x)$ and $O^-(x)$ are defined similarly by using the time-reversed system.

For a subset $D \subset M$ we define $O_T^+(D) := \bigcup_{x \in D} O_T^+(x)$ and $O^+(D)$, $O_T^-(D)$, $O^-(D)$ analogously.

DEFINITION 2.4. *A subset $D \subseteq M$ is called a* control set *if*

(i) $D \subseteq \overline{O^+(x)}$ *for all $x \in D$,*

(ii) *for every $x \in D$ there is $u(\cdot) \in \mathcal{U}$ such that the corresponding trajectory $\varphi(t, x, u(\cdot))$ stays in D for all $t \geq 0$,*

(iii) *D is maximal with the properties* (i) *and* (ii).

A control set C is called invariant *if*

$$\overline{C} = \overline{O^+(x)} \text{ for all } x \in C.$$

A noninvariant control set is called variant.

In order to avoid degenerate situations we need the following setup: Let $L = \mathcal{LA}\{X(\cdot, u), u \in U\}$ denote the Lie algebra generated by the vector fields $X(\cdot, u)$. Let Δ_L denote the distribution generated by L in TM, the tangent space of M. Assume that

$$\dim \Delta_L(x) = \dim M \text{ for all } x \in M. \tag{2.18}$$

This assumption guarantees that the positive and negative orbits of any point $x \in M$ up to any time $T \neq 0$ have nonvoid interior. Note that the definition of control sets demands only approximate reachability (i.e., existence of controls steering into any neighborhood of a given point); as a consequence of assumption (2.18) we have exact controllability in the interior of control sets, more precisely $\text{int}D \subset O^+(x)$ for all $x \in D$.

The following proposition shows—as an extension of [7, Prop. 2.3]—that we have exact controllability in *finite time* on certain compact subsets.

PROPOSITION 2.5. *Consider a control system on M given by* (2.1)–(2.6) *and satisfying* (2.18). *Let $D \subset M$ be a control set and consider compact sets $K_1 \subset O^-(D)$, $K_2 \subset \text{int}D$.*

Then there exists a constant $r > 0$ such that for every $x \in K_1$, $y \in K_2$ there exists a control function $u(\cdot) \in \mathcal{U}$ with $\varphi(t_0, x, u(\cdot)) = y$ for some $t_0 \leq r$.

Proof. (i) We first show that for every $x \in K_1$, $z \in K_2$ there is an open neighborhood $U(x)$ such that all $y \in U(x)$ can be steered to z in bounded time t_0. By (2.18) there is $T < \infty$ and $z_1 \in \text{int}D \cap O^-_{\leq T}(z)$ and an open neighborhood $U(z_1) \subset \text{int}D \cap O^-_{\leq T}(z)$. For $x \in K$ there exists a control $u(\cdot) \in \mathcal{U}$ and a time $t_1 < \infty$ such that $\varphi(t_1, x, u(\cdot)) = z_1$ (as a consequence of exact controllability in the interior of control sets). Since the solutions of the system depend continuously on the initial value, there is an open neighborhood $U(x)$ with $\varphi(t_1, x_1, u(\cdot)) \in U(z_1)$ for all $x_1 \in U(x)$. Putting this together yields $U(x) \subset O^-_{\leq t_1+T}(y)$, which proves the assertion with $t_0 \leq t_1 + T$.

(ii) For $x \in K_1$, $y \in K_2$ we now show that there exists a time $t_y < \infty$ such that all z in some open neighborhood of y can be reached from x in time t_y. Let $x_1 \in \text{int}D$ and $u_1(\cdot) \in \mathcal{U}$, $t_1 < \infty$ such that $\varphi(t_1, x, u(\cdot)) = x_1$ (the existence of x_1, $u_1(\cdot)$, t_1 follows from (2.18)). Again by (2.18) there exists $T < \infty$ and $y_1 \in \text{int}D \cap \mathcal{O}^-_{\leq T}(x_1)$; let $U(y_1)$ be an open neighborhood of y_1 contained in $\text{int}D \cap \mathcal{O}^-_{\leq T}(x_1)$. Now because of the exact controllability there exists $u_2(\cdot) \in \mathcal{U}$, $t_2 < \infty$ with $\varphi(t_2, y_1, u_2) = y$. Since the solution of the control system using the control $u_2(\cdot)$ defines a semigroup of homeomorphisms on M, the open neighborhood $U(y_1)$ is mapped onto some open neighborhood $U(y)$ and $U(y) \subset \mathcal{O}^+_{\leq t_1+T+t_2}(x)$. This means that all $z \in U(y)$ can be reached from x in time $t_y = t_1 + T + t_2$.

(iii) Because of the compactness of K_1 and K_2 now the proof of the proposition follows. □

The following proposition summarizes the consequences of these controllability properties for the optimal value functions.

PROPOSITION 2.6. *Consider optimal control systems on M given by* (2.1)–(2.6) *and* (2.7)–(2.10) *and satisfying* (2.18). *Let $D \subset M$ be a control set and consider compact sets $K_1 \subset O^-(D)$, $K_2 \subset \text{int}D$. Then the following estimates hold:*

(i) $v_0(x) = v_0(y)$ *for all* $x, y \in \text{int}D$.

(ii) $v_0(x) \leq v_0(y)$ *for all* $x \in O^-(D)$, $y \in \text{int}D$.

(iii) $|\delta v_\delta(x) - \delta v_\delta(y)| \leq \varepsilon(\delta)$ *for all* $x, y \in K_2$.

(iv) $\delta v_\delta(x) \leq \delta v_\delta(y) + \varepsilon(\delta)$ *for all* $x \in K_1$, $y \in K_2$

and $\varepsilon(\delta) \to 0$ as δ tends to zero.

Proof. Just combine (2.13) and (2.14) with the controllability properties stated above. □

Now we can formulate the results about the relation between the optimal value functions.

PROPOSITION 2.7. *Consider optimal control systems on M given by* (2.1)–(2.6) *and* (2.7)–(2.10) *and satisfying* (2.18). *Then*

$$\limsup_{\delta \to 0} \delta v_\delta(x) \leq v_0(x) \quad \textit{for all } x \in M.$$

Proof. Fix $\varepsilon > 0$. Choose a control function $u(\cdot)$ such that $|v_0(x) - J_0(x, u(\cdot))| \leq \frac{\varepsilon}{2}$. Using Theorem 2.2 with $\alpha = \frac{\varepsilon}{2}$ yields a $R_1 > 0$ such that for all $\delta \in (0, R_1]$: $\delta v_\delta(x) \leq \delta J_\delta(x, u(\cdot)) \leq J_0(x, u(\cdot)) + \frac{\varepsilon}{2} \leq v_0(x) + \varepsilon$. It follows that $\limsup_{\delta \to 0} \delta v_\delta(x) \leq v_0(x)$ since $\varepsilon > 0$ was arbitrary. □

PROPOSITION 2.8. *Consider optimal control systems on M given by* (2.1)–(2.6) *and* (2.7)–(2.10) *and satisfying* (2.18). *Let $D \subseteq M$ be a control set. Then for every compact $Q \subset \text{int}D$ and every $\varepsilon > 0$ there exists a $R_0 > 0$ such that*

$$\delta v_\delta(x) \leq v_0(x) + \varepsilon \quad \textit{for all } \delta \in (0, R_0], \quad x \in Q.$$

Proof. Fix $x_0 \in Q$. Using Proposition 2.7 we know that there exists a constant $R_1 > 0$ such that $\delta v_\delta(x_0) \leq v_0(x_0) + \frac{\varepsilon}{2}$ for all $\delta \in (0, R_1]$. Now choose $R_2 > 0$ such that

Proposition 2.6 (iii) holds with $\varepsilon(\delta) < \frac{\varepsilon}{2}$ for $\delta < R_2$. Since v_0 is constant on Q now the assertion holds for all $x \in Q$ with $R_0 := \min\{R_1, R_2\}$. □

LEMMA 2.9 (pointwise convergence). *Consider optimal control systems on M given by* (2.1)–(2.6) *and* (2.7)–(2.10). *Assume there exists $x \in M$, $R \in \mathbb{R}$, and a set $B \subset M$ such that $\delta v_\delta(y) \le \delta v_\delta(x) + \alpha(\delta)$ for all $y \in B$, $\delta \in (0, R]$, and constants $\alpha(\delta) \ge 0$. Assume there exist optimal controls $u_\delta(\cdot) \in \mathcal{U}$ for all $\delta \in (0, R]$ such that $\varphi(t, x, u_\delta(\cdot)) \in B$ for all $t \ge 0$. Then for every $\varepsilon > 0$ there exists $R_0 > 0$ such that*

$$|\delta v_\delta(x) - v_0(x)| \le \max\{\varepsilon, \alpha(\delta)\} \text{ for all } \delta \in (0, R_0].$$

In particular if $\alpha(\delta) \to 0$ as $\delta \to 0$ the convergence $\delta v_\delta(x) \to v_0(x)$ is implied.

Proof. From Theorem 2.1 it is clear that $v_0(x) \le \delta v_\delta(x) + \alpha(\delta)$ for all $\delta < R$. Now choose a control function $u(\cdot)$ such that $|v_0(x) - J_0(x, u(\cdot))| \le \frac{\varepsilon}{2}$. Using Theorem 2.2 with $\alpha = \frac{\varepsilon}{2}$ yields $R_0 > 0$ such that for all $\delta < R_0$: $\delta v_\delta(x) \le \delta J_\delta(x, u(\cdot)) \le J_0(x, u(\cdot)) + \frac{\varepsilon}{2} \le v_0(x) + \varepsilon$. Combining these inequalities finishes the proof. □

By using the estimate of Proposition 2.6, two results on uniform convergence can be obtained.

THEOREM 2.10 (uniform convergence). *Consider optimal control systems on M given by* (2.1)–(2.6) *and* (2.7)–(2.10) *and satisfying* (2.18). *Let $D \subseteq M$ be a control set and assume there exist $x_0 \in \mathrm{int}D$, a compact subset $K \subseteq D$, and optimal controls $u_\delta(\cdot)$ such that*

$$\varphi(t, x_0, u_\delta(\cdot)) \in K \text{ for all } t \ge 0 \text{ for all } \delta \in (0, R]$$

for some constant $R > 0$. Then

$$\delta v_\delta \to v_0 \text{ uniformly on compact subsets of } \mathrm{int}D.$$

Proof. By Proposition 2.6 (iii), on any compact subset Q of $\mathrm{int}D$ we have $|\delta v_\delta(x) - \delta v_\delta(y)| \le \varepsilon(\delta) \to 0$ uniformly for all $x, y \in Q$ as δ tends to zero. By Proposition 2.6 (iv), x_0 and K fulfill the conditions of Lemma 2.9 with $\alpha(\delta) = \varepsilon(\delta)$ since $K \subseteq D \subseteq O^-(D)$. Hence pointwise convergence follows. Since v_0 is constant on $\mathrm{int}D$, uniform convergence on Q follows. □

THEOREM 2.11 (uniform convergence in compact invariant control sets). *Consider optimal control systems on M given by* (2.1)–(2.6) *and* (2.7)–(2.10) *and satisfying* (2.18). *Let $C \subseteq M$ be a compact invariant control set. Then for $\delta \to 0$*

(i) $\delta v_\delta(x) \to v_0(x)$ *for all* $x \in \mathrm{int}C$,

(ii) $\delta v_\delta \to v_0$ *uniformly on compact subsets of* $\mathrm{int}C$,

(iii) *if M is compact and C is the unique invariant control set we have* $\sup_{x \in M} \delta v_\delta(x) \to \sup_{x \in M} v_0(x)$.

Proof. Since C is a compact subset of C and no trajectory can leave C, the conditions of Theorem 2 (with $K = C$) are fulfilled. Hence the assertions (i) and (ii) follow.

If M is compact and C is the unique invariant control set it follows that $O^-(C) = M$ [16, proof of Lem. 2.2 (i)].

From Proposition 2.6 (ii) and (iv) and the compactness of $M = O^-(C)$, it follows for any compact subset $Q \subset \mathrm{int}C$ that $v_0(x) \le v_0(y)$ and $\delta v_\delta(x) \le \delta v_\delta(y) + \varepsilon(\delta)$ for all $x \in M$, $y \in Q$. Since we have uniform convergence on Q assertion (iii) of Theorem 2.11 is proved. □

Remark 2.12. Note that these results are not valid in general for the corresponding maximization problems, since the second approximation theorem is not valid for the reverse inequality. However some of the results remain valid and others are valid under additional conditions.

(i) The application of the results to the maximization problems is possible if the lim sup in (2.10) can be replaced by a lim inf without changing the value of v_0. This is possible if there exist approximately optimal trajectories and controls—with respect to the maximization problem—such that the lim sup is a limit. From [20, proof of Prop. 1.4 (a)] it is clear that this is the fact if there exist approximately optimal trajectories and controls which are periodic. A sufficient condition for this is that there exists an optimal trajectory that stays inside some compact subset $K \subset \operatorname{int} D$ (cf. [20, Prop. 2.7]).

(ii) Adding this condition to the assumptions of Theorem 2.10 we obtain Theorem 2.10 from Wirth [20] under the weaker condition that the optimal trajectories with respect to the discounted problems stay inside a compact subset of a control set instead of a compact subset of the *interior* of a control set.

(iii) For invariant control sets C we can use [7, Cor. 4.3] to conclude that for any initial value $x_0 \in \operatorname{int} C$ there exist approximately optimal periodic control functions and trajectories. Hence Theorem 3 remains valid for the maximization problem without any additional assumptions.

3. Lyapunov exponents of bilinear control systems. We will now return to the *bilinear control systems* in $\mathbb{R}^d$, i.e., systems of the form

$$\dot{x}(t) = \Big(A_0 + \sum_{i=1}^{m} u_i(t) A_i\Big) x(t), \quad x(0) = x_0 \in \mathbb{R}^d \setminus \{0\} \tag{3.1}$$

with $A_j \in \mathbb{R}^{d\times d}$, $j = 0, \dots, m$, $u(\cdot) \in \mathcal{U} := \{u : \mathbb{R} \to U,\ u \text{ measurable}\}$ with a compact and convex set of control values $U \subset \mathbb{R}^m$ with nonvoid interior.

We denote the unique trajectory for any initial value $x_0 \in \mathbb{R}^d$ and any control function $u(\cdot) \in \mathcal{U}$ by $x(t, x_0, u(\cdot))$.

In order to characterize the exponential growth rate of the solutions of (3.1) we define the Lyapunov exponent of a solution by

$$\lambda(x_0, u(\cdot)) := \limsup_{t\to\infty} \frac{1}{t} \ln \|x(t, x_0, u(\cdot))\|. \tag{3.2}$$

The minimal Lyapunov exponent with respect to $x_0 \in \mathbb{R}^n \setminus \{0\}$ is defined by

$$\lambda^*(x_0) := \inf_{u(\cdot)\in\mathcal{U}} \lambda(x_0, u(\cdot)), \tag{3.3}$$

and the extremal Lyapunov exponents of the control system are defined by

$$\kappa^* := \inf_{x_0\neq 0} \inf_{u(\cdot)\in\mathcal{U}} \lambda(x_0, u(\cdot)), \tag{3.4}$$

$$\kappa := \sup_{x_0\neq 0} \sup_{u(\cdot)\in\mathcal{U}} \lambda(x_0, u(\cdot)), \tag{3.5}$$

$$\tilde{\kappa} := \sup_{x_0\neq 0} \inf_{u(\cdot)\in\mathcal{U}} \lambda(x_0, u(\cdot)). \tag{3.6}$$

The Lyapunov exponent can be interpreted as a measure for the exponential growth of trajectories. Our aim is to calculate numerical approximations of the minimal and maximal Lyapunov exponents with respect to the initial values. If $\lambda^*(x_0) < 0$ the system can be steered asymptotically to the origin from x_0. Using the approximation of the Lyapunov exponents we then are able to calculate controls that stabilize the system.

For a bilinear control system (3.1) the following identity is obvious:

$$\lambda(x_0, u(\cdot)) = \lambda(\alpha x_0, u(\cdot)) \text{ for all } x_0 \in \mathbb{R}^d \setminus \{0\},\ \alpha \in \mathbb{R} \setminus \{0\},\ u \in \mathcal{U}.$$

Due to this observation we can identify all $x \neq 0$ lying on a straight line through the origin. Hence it is sufficient to consider initial values s_0 in $\mathbb{P}^{d-1}$, the real projective space. To calculate the Lyapunov exponents we can project the system onto the unit sphere $\mathbb{S}^{d-1}$ via $s_0 := x_0/\|x_0\|$. This yields the projection onto $\mathbb{P}^{d-1}$ by identifying opposite points. A simple application of the chain rule shows that the projected system can be written as

$$\dot{s}(t) = h_0(s(t)) + \sum_{i=1}^{m} u_i(t) h_i(s(t)) \tag{3.7}$$

where

$$h_i(s) = [A_i - s^t A_i s \cdot \mathrm{Id}]s \quad \text{for all } i = 0, \ldots, m.$$

The Lyapunov exponent (3.2) with respect to $s_0 = x_0/\|x_0\|$ can be written as

$$\lambda(x_0, u(\cdot)) = \lambda(s_0, u(\cdot)) = \limsup_{t\to\infty} \frac{1}{t} \int_0^t q(s(\tau, s_0, u(\cdot)), u(\tau)) d\tau \tag{3.8}$$

where

$$q(s, u) = s^t \Big(A_0 + \sum_{i=0}^{m} u_i A_i \Big) s. \tag{3.9}$$

We recall some facts about projected bilinear control systems and their Lyapunov exponents.

For the projected bilinear system assumption (2.18) reads

$$\dim \Delta_L(p) = d - 1 \text{ for all } p \in \mathbb{P}^{d-1}, \quad L = \mathcal{LA}\{h(\cdot, u), u \in U\}$$

where $h(\cdot, u) := h_0(\cdot) + \sum_{i=1}^{m} u_i h_i(\cdot)$. Under this assumption the following facts hold (cf. [9, Cor. 4.4], [8, Thm. 3.10]):

If κ_1 denotes the maximal Lyapunov exponent of the original system and κ_2^* the minimal exponent of the time-reversed system the identity $\kappa_1 = -\kappa_2^*$ holds.

For the projected system there exist k control sets with nonvoid interior where $1 \leq k \leq d$. These are called the main control sets. They are linearly ordered by $D_i < D_j \Leftrightarrow$ there exists $p_i \in D_i$, $p_j \in D_j$, $t > 0$, and $u(\cdot) \in \mathcal{U}$ such that $\varphi(t, p_i, u) = p_j$.

The control set D_1 is open; the control set $C := D_k$ is closed and invariant. All other control sets are neither open nor closed. Furthermore we have $O^-(p) = \mathbb{P}^{d-1}$ for all $p \in \mathrm{int}C$.

The linear order of the control sets implies a linear order on the minimal Lyapunov exponents (which can easily be proved using Proposition 2.6): $\lambda^*(p_i) \leq \lambda^*(p_j)$ for $p_i \in D_i$, $p_j \in D_j$, and $i < j$. Furthermore, $\lambda^*(p)$ is constant on the interior of control sets.

Under the following condition there is a stronger relation between the control sets of the projected and the Lyapunov exponents of the bilinear system. Considering the set of control values $\rho U := \{\rho u \mid u \in U\}$ for $\rho \geq 0$ and the corresponding set of control functions $\mathcal{U}^\rho$ we assume the following ρ–ρ' inner pair condition:

> For all $0 \leq \rho \leq \rho'$ and all $(u(\cdot), p) \in \mathcal{U}^\rho \times \mathbb{P}^{d-1}$ there exist $T > 0$ and $S > 0$ such that $\varphi(T, p, u(\cdot)) \in \mathrm{int} O_{S+T}^{\rho'+}(p)$ (the positive orbit corresponding to $\mathcal{U}^{\rho'}$).

Let D^ρ be a main control set corresponding to $\mathcal{U}^\rho$. We define the *Lyapunov spectrum of* (3.1) *over* $\overline{D^\rho}$ by

$$\Sigma_{Ly}^\rho(\overline{D^\rho}) := \{\lambda(p, u(\cdot)) \mid \varphi(t, p, u) \in \overline{D^\rho} \text{ for all } t \geq T \text{ for some } T \geq 0\}$$

and the *Lyapunov spectrum of* (3.1) by

$$\Sigma_{Ly}^{\rho} := \{\lambda(p, u(\cdot)) \mid u(\cdot) \in \mathcal{U},\ p \in \mathbb{P}^{d-1}\}.$$

Under the ρ–ρ' inner pair condition we know that

$$\Sigma_{Ly} = \bigcup_{i=1}^{k(\rho)} \Sigma_{Ly}(\overline{D_i^{\rho}}) \tag{3.10}$$

for all except at most countably many $\rho < \rho'$, where $k(\rho)$ is the number of main control sets D_i^{ρ} corresponding to $\mathcal{U}^{\rho}$ [11, Cor. 5.6].

Furthermore $\Sigma_{Ly}^{\rho}(\overline{D_i^{\rho}})$ are closed intervals and thus it is sufficient to calculate the minima and the maxima of $\Sigma_{Ly}(\overline{D_i^{\rho}})$ to obtain the whole Lyapunov spectrum of the system. These maxima and minima can be approximated by periodic trajectories with initial values in $\operatorname{int} D_i^{\rho}$.

In the case $d = 2$ these results hold for all $\rho > 0$ without assuming the ρ–ρ' inner pair condition [11, Cor. 4.9].

We will now give an interpretation of the results of §2 in terms of calculating Lyapunov exponents and stabilization. Since we are going to solve the discounted optimal control problem numerically we cannot expect to calculate optimal control functions but only ε-optimal control functions. We call a control function $u_x(\cdot) \in \mathcal{U}$ *uniformly ε-optimal* with respect to $x \in M$ iff $|\delta J_\delta(\varphi(t, x, u_x(\cdot)), u_x(t + \cdot)) - \delta v_\delta(\varphi(t, x, u_x(\cdot)))| < \varepsilon$ for all $t \geq 0$.

THEOREM 3.1. *Consider a bilinear control system* (3.1) *and the related optimal control system on* $\mathbb{P}^{d-1}$ *given by* (3.7) *and* (3.8) *with cost function q from* (3.9). *Assume* (2.18) *is satisfied. Let*

$$v_\delta(x) := \inf_{u(\cdot)\in\mathcal{U}} J_\delta(x, u(\cdot)) \quad \text{and} \quad \bar{v}_\delta(x) := \sup_{u(\cdot)\in\mathcal{U}} J_\delta(x, u(\cdot)).$$

Then the following estimates hold with $\varepsilon \to 0$ as δ tends to zero.

(i) $\delta v_\delta(x) \leq \lambda^*(x) + \varepsilon$ *for all* $x \in M$.

(ii) $\delta v_\delta(x) \leq \lambda^*(x) + \varepsilon$ *uniformly on compact subsets Q of the interior of control sets.*

(iii) $|\delta v_\delta(x) - \lambda^*(x)| \leq \varepsilon$ *uniformly on compact subsets Q of the interior of control sets under the conditions of Theorem* 2.10.

(iv) $|\delta v_\delta(x) - \lambda^*(x)| \leq \varepsilon$ *uniformly on compact subsets Q of the interior of the invariant control set.*

(v) $\sup_{x\in M} \delta v_\delta(x) \to \tilde{\kappa}$ *as δ tends to zero.*

(vi) $\inf_{x\in M} \delta \bar{v}_\delta(x) \to \kappa$ *as δ tends to zero.*

(vii) *If $\tilde{\kappa} < 0$ and $u_s(\cdot)$ is uniformly ε-optimal with respect to s then $\varphi(t, x, u_s(\cdot))$ is asymptotically stable for all $x \in \mathbb{R}^d$ with $s = x/\|x\|$ provided δ and ε are sufficiently small.*

(viii) *If $\lambda^* < 0$ in the interior of some control set D and $u_s(\cdot)$ is uniformly ε-optimal with respect to s and $\varphi(t, s, u_s(\cdot))$ stays inside a compact subset of $O^-(D)$ for all times, then $\varphi(t, x, u_s(\cdot))$ is asymptotically stable for all $x \in \mathbb{R}^d$ with $s = x/\|x\|$ provided δ and ε are sufficiently small.*

Proof. All assertions follow directly from the results in §2. Assertion (iv) is true since the invariant control set of the projected system is compact. Assertions (v) and (vi) are proved using the fact that the projective space is compact and that there exists a unique invariant control set for the projected system. □

Remark 3.2. Knowing the facts cited in this section we can see that even more can be calculated.

(i) Property (vi) can be used to calculate κ^* by calculating κ of the time reversed system. Hence it is possible to approximate κ, κ^*, and $\tilde{\kappa}$ for any bilinear control system satisfying (2.18) by solving discounted optimal control problems.

(ii) For all main control sets D_i we can approximate the minimal Lyapunov exponent over $\operatorname{int} D_i$ as follows: Proposition 2.8 yields that $\delta v_\delta < \lambda^* + \varepsilon$ uniformly on compact subsets of $\operatorname{int} D_i$. If we find control functions as described in Theorem 3.1 (viii) for $\varepsilon > 0$ we know that there exists a Lyapunov exponent $\lambda^* < \delta v_\delta + \varepsilon$; hence $\lambda^* \in [\delta v_\delta - \varepsilon, \delta v_\delta + \varepsilon]$. However, the existence of such control functions is not guaranteed; nevertheless for all examples discussed in §5 it was possible to find them.

(iii) For systems with $d = 2$ or systems with $d > 2$ satisfying the ρ–ρ' inner pair condition we are also able to compute $\Sigma_{Ly}^{\rho}(\overline{D})$ for $D = C$ and $D = D_1$ at least for all but countably many $\rho > 0$, since in this case the upper and lower bounds of this interval coincide with κ and $\tilde{\kappa}$ of the original or of the time-reversed system, respectively. For all other main control sets we can apply the technique from (ii) to both the original and the time-reversed system to calculate $\Sigma_{Ly}^{\rho}(\overline{D_i})$.

(iv) In the case that $d > 2$ and $\rho > 0$ is one of the (at most countably many) exceptional points of the spectrum (3.10) we can use the monotonicity of v_δ and Σ_{Ly}^{ρ} in ρ. This implies that there exist values $\rho_1 < \rho < \rho_2$ arbitrarily close to ρ such that the approximated spectrum contains $\Sigma_{Ly}^{\rho_1}$ and is contained in $\Sigma_{Ly}^{\rho_2}$.

4. Numerical solution of the discounted optimal control problem. A discretization scheme to solve discounted optimal control problems in $\mathbb{R}^n$ has been developed by Capuzzo Dolcetta [2], Capuzzo Dolcetta and Falcone [3], Capuzzo Dolcetta and Ishii [4], and Falcone [12], [13]. The algorithm used here to solve these problems is based on this discretization. We will first describe this discretization scheme and then present the modifications for our case, where the system is given on $\mathbb{P}^{d-1}$ instead of $\mathbb{R}^n$.

Hence we first assume that we have a discounted optimal control problem defined by (2.1)–(2.6) and (2.8) with $M = \mathbb{R}^n$. In addition we need the following conditions on X and g:

$$\|X(x,u) - X(y,u)\| \le L_X \|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n \text{ for all } u \in U \text{ for an } L_X \in \mathbb{R}, \tag{4.1}$$

$$\|X(x,u)\| \le M_X \quad \text{for all } (x,u) \in \mathbb{R}^n \times U \text{ for an } M_X \in \mathbb{R}, \tag{4.2}$$

$$|g(x,u) - g(y,u)| \le L_g \|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n \text{ for all } u \in U \text{ for an } L_g \in \mathbb{R}. \tag{4.3}$$

The δ discounted cost functional J_δ and the optimal value function v_δ are defined as in (2.9) and (2.11).

Under the assumptions made above the value function v_δ satisfies

$$|v_\delta(x)| \le \frac{M_g}{\delta} \quad \text{and} \quad |v_\delta(x) - v_\delta(y)| \le C|x - y|^\gamma \tag{4.4}$$

for all $x, y \in \mathbb{R}^n$ (cf. [4]; the second estimate can be proved by using [4, Lem. 4.1]). For small $\delta > 0$ we have $C = \frac{M}{\delta}$ for a constant M independent on δ, and γ is a constant satisfying $\gamma = 1$ for $\delta > L_X$, $\gamma = \frac{\delta}{L_X}$ for $\delta < L_X$, and $\gamma \in (0,1)$ arbitrary for $\delta = L_X$.

Furthermore (cf. [17]) v_δ is the unique bounded and uniformly continuous viscosity solution of the Hamilton–Jacobi–Bellman equation

$$\sup_{u \in U} \{\delta v_\delta(x_0) - g(x_0,u) - Dv_\delta(x_0) X(x_0,u)\} = 0. \tag{4.5}$$

The first discretization step is a discretization in time. By replacing Dv_δ by the difference quotient with time step h one obtains

$$\sup_{u \in U} \{v_h(x) - \beta v_h(x + hX(x,u)) - hg(x,u)\} = 0 \tag{4.6}$$

with $\beta := 1 - \delta h$.

It turns out that the unique bounded solution of this equation is the optimal value function v_h of the *discretized optimal control system* with respect to the space $\mathcal{U}_h$ of all controls constant on each interval $[jh, (j+1)h)$, $j \in \mathbb{N}$:

$$x_0 := x, \quad x_{j+1} := x_j + hX(x_j, u_j), \quad j = 0, 1, 2, \ldots, \tag{4.7}$$

with running cost

$$J_h(x, u(\cdot)) := h \sum_{j=0}^{\infty} \beta^j g(x_j, u_j).$$

Furthermore for all $p \in \mathbb{N}$, v_h satisfies

$$v_h(x) = \inf_{u(\cdot) \in \mathcal{U}_h} \left\{ h \sum_{j=0}^{p-1} \beta^j g(x_j, u_j) + \beta^p v_h(x_p) \right\} \tag{4.8}$$

and the estimates (4.4) also apply to v_h.

The discretization error can be estimated as follows [4, Thm. 3.1]:

$$\sup_{x \in \mathbb{R}^n} |(v_\delta - v_h)(x)| \le C h^{\frac{\gamma}{2}} \tag{4.9}$$

for all $h \in (0, \frac{1}{\delta})$. Here we have $C = \frac{M}{\delta^2}$ for small $\delta > 0$ and γ is the constant from (4.4).

The discretization error of the functionals for any $u(\cdot) \in \mathcal{U}_h$ can be estimated as

$$\sup_{x \in \mathbb{R}^n,\, u(\cdot) \in \mathcal{U}_h} |J_h(x, u(\cdot)) - J_\delta(x, u(\cdot))| \le C h^\gamma \tag{4.10}$$

where $C = \frac{M}{\delta}$ for small $\delta > 0$ and γ as above [4, Lem. 4.1].

In order to reduce (4.6) to a finite-dimensional problem we apply a finite difference technique. To do this we assume the existence of an open, bounded, and convex subset Ω of the state space $\mathbb{R}^n$ which is invariant for (2.1). Thus a triangulation of Ω into a finite number P of simplices S_j with N nodes x_i can be constructed (cf. [13, Prop. 2.5]) such that $\Omega^k := \cup_{j=1,\ldots,P} S_j$ is invariant with respect to the discretized trajectories (4.7). Here $k := \sup\{\|x - y\| \mid x \text{ and } y \text{ are nodes of } S_j,\ j = 1, \ldots, P\}$. We are now looking for a solution of (4.6) in the space of piecewise affine functions $\mathcal{W} := \{w \in C(\Omega^k) \mid Dw(x) = c_j \text{ in } S_j\}$.

Every point $x_i + hf(x_i, u)$ can be written as a convex combination of the nodes of the simplex containing it with coefficients $\lambda_{ij}(u)$. Let $\Lambda(u) := [\lambda_{ij}(u)]_{i,j=1,\ldots,N}$ be the matrix containing the coefficients and $G(u) := [g(x_i, u)]_{i=1,\ldots,N}$ an N-dimensional vector containing the values of g with control value u at the nodes of the triangulation. Now we can rewrite (4.6) as a fixed point equation

$$V = T_h^k(V), \quad T_h^k(V) := \inf_{u \in U} (\beta \Lambda(u) V + hG(u)). \tag{4.11}$$

It follows that T_h^k is a contraction in $\mathbb{R}^N$ with contraction factor $\beta := 1 - \delta h$ and therefore has a unique fixed point V^*. If v_h^k denotes the function obtained by $v_h^k(x_i) := [V^*]_i$ and linear interpolation between the nodes, the discretization error can be estimated by

$$\sup_{x \in \Omega^k} |(v_h^k - v_h)(x)| \le C \frac{k^\gamma}{h} \tag{4.12}$$

with γ as in (4.4) and $C = \frac{M}{\delta^2}$ for small $\delta > 0$ (cf. [13, corrigenda]).

For the whole discretization error we obtain the following estimate:

$$\sup_{x\in\Omega^k} |(v_h^k - v_\delta)(x)| \le C\left(h^{\frac{\gamma}{2}} + \frac{k^\gamma}{h}\right) \tag{4.13}$$

with the constants from (4.9) and (4.12).

Remark 4.1. These results have been improved by Gonzales and Tidball [14]. From [14, Lem. 3.4] in connection with [4, Lem. 4.1] it follows that

$$\sup_{x\in\Omega^k} |(v_h^k - v_h)(x)| \le C\left(\frac{k}{\sqrt{h}}\right)^\gamma; \tag{4.14}$$

[14, Thm. 3.1] yields

$$\sup_{x\in\Omega^k} |(v_h^k - v_\delta)(x)| \le C\left(\sqrt{h} + \frac{k}{\sqrt{h}}\right)^\gamma \tag{4.15}$$

with similar constants C and γ.

Remark 4.2. Note that the convergence becomes slow if the discount rate δ becomes small. For the approximation of the average cost functional as described in §2 it is nevertheless necessary to calculate v_δ for small $\delta > 0$. This means that for this purpose we need a fine discretization in time and space to get reliable results.

If one uses estimate (4.13) we obtain as an additional condition that k should be smaller than h, using (4.15) convergence for the case $k = h$ is guaranteed.

To handle the optimal control problem on $\mathbb{P}^{d-1}$ we use the following modifications on this scheme.

We first consider the optimal control problem on $\mathbb{S}^{d-1}$ defined by the projected system (3.7). The optimal value function v_δ then again satisfies (4.4) and is the unique bounded and uniformly continuous viscosity solution of (4.5). This can be proved exactly the same way as in the $\mathbb{R}^n$ case by using the metric on $\mathbb{S}^{d-1}$ induced by the norm on $\mathbb{R}^d$.

We have seen that the discretization in time of (4.5) corresponds to the Euler discretization of the control system. Hence here we use the following Euler method on $\mathbb{S}^{d-1}$. For $h > 0$ and any $s \in \mathbb{S}^{d-1}$ we define

$$\Phi_h(s,u) := \frac{s + hX(s,u)}{\|s + hX(s,u)\|}; \tag{4.16}$$

i.e., we perform an Euler step in $\mathbb{R}^d$ and project the solution back to $\mathbb{S}^{d-1}$. With this (4.6) reads

$$\sup_{u\in U}\{v_h(s) - \beta v_h(\Phi_h(s,u)) - hg(s,u)\} = 0 \tag{4.17}$$

and (4.7) translates to

$$s_0 := s, \quad s_{j+1} := \Phi_h(s_j, u), \quad j = 0, 1, 2, \ldots. \tag{4.18}$$

The estimates (4.8)–(4.10) remain valid; again all proofs from the $\mathbb{R}^n$ case apply by using the metric on $\mathbb{S}^{d-1}$ induced by the norm on $\mathbb{R}^d$.

We will now use the fact that this discrete time control system on $\mathbb{S}^{d-1}$ defines a (well-defined) control system on $\mathbb{P}^{d-1}$ by identifying s and $-s$ on $\mathbb{S}^{d-1}$. Let $W \subset \mathbb{S}^{d-1}$ be an open set in $\mathbb{S}^{d-1}$ such that it contains the upper half of the sphere. Any discrete time trajectory

$(s_i)_{i\in\mathbb{N}_0} \subset \mathbb{S}^{d-1}$ as defined in (4.18) can be mapped on a trajectory $(\tilde{s}_i)_{i\in\mathbb{N}_0} \subset W$ by $\tilde{s}_i := s_i$ if $s_i \in W$, $\tilde{s}_i := -s_i$ if $s_i \notin W$. Since $X(s,u) = -X(-s,u)$ this mapping is well defined and $g(s,u) = g(-s,u)$ implies that v_h does not change if we only consider trajectories in W. Hence we can define a discrete time optimal control problem on W via

$$\tilde{\Phi}_h(s) = \begin{cases} \Phi_h(s), & \Phi_h(s) \in W, \\ -\Phi_h(s), & \Phi_h(s) \notin W \end{cases}$$

without changing v_h.

To obtain a region $\Omega \subset \mathbb{R}^{d-1}$ suitable for the space discretization we use a parametrization Ψ of $\mathbb{S}^{d-1}$ which is invertible on W such that Ψ^{-1} maps W to an open and bounded set $\Omega \subset \mathbb{R}^{d-1}$. (The parametrizations used in our examples are given in §5.) Now we can project the system on W to a system on Ω and compute v_h on Ω. The system on Ω is then given by

$$\Phi_{h,\Omega}(x,u) := \Psi^{-1}(\tilde{\Phi}_h(\Psi(x),u)), \quad g_\Omega(x,u) := g(\Psi(x),u),$$

and by definition of $\tilde{\Phi}_h$ the set Ω is invariant for this discrete time system. We can rewrite (4.17) by using $\Phi_{h,\Omega}$ and g_Ω and denoting the solution by $v_{h,\Omega}$. This solution satisfies $v_h(\Psi(x)) = v_{h,\Omega}(x)$ and, since Ψ is Lipschitz continuous, estimate (4.4) remains valid for $v_{h,\Omega}$.

Thus we can proceed as in the $\mathbb{R}^n$ case described above. Keeping in mind that there exists a one-to-one relation between the system on W and the system on Ω we can simplify the notation by writing Φ_h, g, and v_h instead of $\Phi_{h,\Omega}$, g_Ω, and $v_{h,\Omega}$.

We will now turn to the problem of how the fixed point equation (4.11) can be solved numerically. In order to do this it is possible to use the contraction T_h^k to construct an iteration scheme, but since the contraction factor $\beta = 1-\delta h$ is close to one this iteration converges rather slowly. An acceleration method for this iteration scheme has been proposed by Falcone [12]. Falcone uses the set $\mathcal{V}$ of monotone convergence of T_h^k given by $\mathcal{V} := \{V \in \mathbb{R}^N \mid T_h^k(V) \geq V\}$ where "$\geq$" denotes the componentwise order. A simple computation shows that $\mathcal{V}$ is a convex closed subset of $\mathbb{R}^N$. Given a $V_0 \in \mathcal{V}$ the operator T_h^k is used to determine an initial direction. The algorithm follows this direction until it crosses the boundary of $\mathcal{V}$; then it determines a new direction using T_h^k and continues the same way.

A different algorithm to calculate V^* can be developed by observing that V^* is the componentwise maximum of $\mathcal{V}$ and that $\mathcal{V}$ can be written as

$$\mathcal{V} = \left\{ V \in \mathbb{R}^N \;\middle|\; [V]_i \leq \min_{u\in U} \left\{ \frac{\beta \sum_{\substack{j=1,\dots,N \\ j\neq i}} \lambda_{ij}(u)[V]_j + hG_i(u)}{1-\beta\lambda_{ii}(u)} \right\} \text{ for all } i \in \{1,\dots,N\} \right\}. \tag{4.19}$$

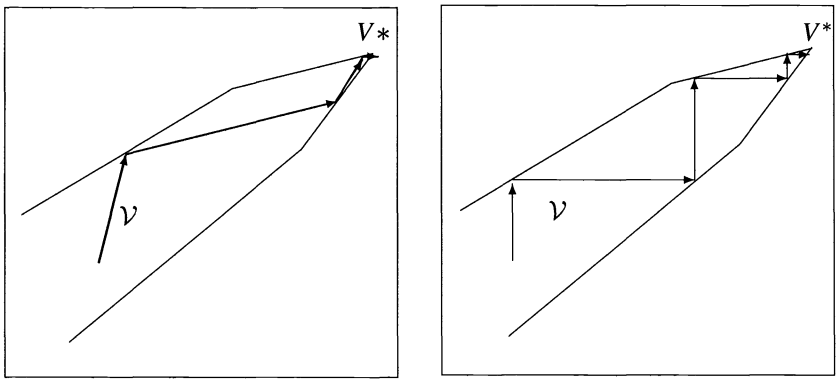
Note that the fraction on the right side does not depend on $[V]_i$. Thus we can construct the *increasing coordinate algorithm*:

Step 1: take $V \in \mathcal{V}$ (e.g., $V = (-\frac{M_g}{\delta}, \dots, -\frac{M_g}{\delta})^T$)
Step 2: compute sequentially

$$[V]_i = \min_{u\in U} \left\{ \frac{\beta \sum_{\substack{j=1,\dots,N \\ j\neq i}} \lambda_{ij}(u)[V]_j + hG_i(u)}{1-\beta\lambda_{ii}(u)} \right\} \quad \text{for all } i \in \{1,\dots,N\}.$$

Step 3: continue with Step 2 and the new vector V.

Figure 4.1 shows an illustration of the algorithms for $N = 2$.



Falcone's accelerated method Increasing coordinate algorithm

FIG. 4.1. *Algorithms.*

Note that for every arrow in the left picture the intersection between the initial direction and the boundary of $\mathcal{V}$ has to be determined. To do this, e.g., by bisection as in the implementation used here, the operator T_h^k has to be evaluated several times to decide if a point is inside or outside $\mathcal{V}$. In the increasing coordinate algorithm N arrows (i.e., two arrows in Figure 4.1) are calculated by N evaluations of the fraction in Step 2. These N evaluations are about as expensive as one evaluation of T_h^k. This means that one iteration in the increasing coordinate algorithm corresponds to one evaluation of T_h^k in the acceleration method.

The convergence of this algorithm is guaranteed by the following lemma.

LEMMA 4.3. *Let V_1 be the vector obtained by applying Step 2 for $i = 1, \dots, N$ to a vector $V_0 \in \mathcal{V}$. Then*

$$[V_1]_i - [V_0]_i \;\geq\; [T_h^k(V_0)]_i - [V_0]_i.$$

Proof. Because of $V_0 \in \mathcal{V}$ and (4.19) it follows that $[V_1]_i \geq [V_0]_i$ for all $i = 1, \dots, N$. Hence

$$\begin{aligned}
[V_1]_i - [V_0]_i &= \min_{u\in U} \left\{ \frac{\beta \sum_{\substack{j=1,\dots,N \\ j\neq i}} \lambda_{ij}(u)[V_1]_j + hG_i(u) - (1-\beta\lambda_{ii}(u))[V_0]_i}{1-\beta\lambda_{ii}(u)} \right\} \\
&\geq \min_{u\in U} \left\{ \beta \sum_{\substack{j=1,\dots,N \\ j\neq i}} \lambda_{ij}(u)[V_0]_j + hG_i(u) - (1-\beta\lambda_{ii}(u))[V_0]_i \right\} \\
&= \min_{u\in U} \left\{ \beta \sum_{j=1}^{N} \lambda_{ij}(u)[V_0]_j + hG_i(u) - [V_0]_i \right\} \;=\; [T_h^k(V_0)]_i - [V_0]_i. \qquad \square
\end{aligned}$$

The convergence of the increasing coordinate algorithm therefore is a consequence of the monotone convergence of the iteration scheme using the contraction T_h^k.

All iteration methods described here have in common that during the iteration a minimum over all $u \in U$ has to be calculated. The following lemma shows that this can be done by minimizing over a finite set $U_\varepsilon \subset U$.

LEMMA 4.4. *Assume that X and g are uniformly Lipschitz continuous in the control $u \in U$ with Lipschitz constant L_u. Let $U_\varepsilon \subset U$ such that for all $u \in U$ there exists $\bar{u} \in U_\varepsilon$ with $\|u - \bar{u}\| < \varepsilon$. Let $\mathcal{U}_\varepsilon$ denote the corresponding set of control functions. Then for all $s \in \mathbb{S}^{d-1}$*

it holds that

$$\| \inf_{u(\cdot)\in\mathcal{U}} J_\delta(s,u(\cdot)) - \inf_{\bar u(\cdot)\in\mathcal{U}_\varepsilon} J_\delta(s,\bar u(\cdot))\| < C\varepsilon^\eta$$

where for $\delta < L_X + 1$ *we have* $\eta = \frac{L_X+1}{\delta}$.

Proof. For all $u(\cdot) \in \mathcal{U}$ there exists $\bar u(\cdot) \in \mathcal{U}_\varepsilon$ such that $\|u(t) - \bar u(t)\| < \varepsilon$ for almost all $t \in \mathbb{R}$. Hence we have

$$\|\varphi(t,s,u(\cdot)) - \varphi(t,s,\bar u(\cdot))\| < L_u\varepsilon t + \int_0^t L_X\|\varphi(\tau,s,u(\cdot)) - \varphi(\tau,s,\bar u(\cdot))\|d\tau$$

where $\|\cdot\|$ denotes the norm on $\mathbb{R}^d$. Now the Gronwall lemma and [4, Lem. 4.1] can be used to estimate this integral equation and the assertion follows. □

For the projected bilinear control system with cost function $g = q$ the assumptions of Lemma 4.4 are fulfilled and hence we may use a finite set of control values to calculate v_h^k.

Once v_h^k is calculated it can be used to construct ε-optimal control functions:

Step 1: Let $x_0 = x$, $n = 0$.

Step 2: Choose a control value $\tilde u^k_{x_n,h} \in U$ such that $\beta v_h^k(\Phi_h(x_n, \tilde u^k_{x_n,h})) + hg(x_n, \tilde u^k_{x_n,h})$ becomes minimal.

Step 3: Let $u^k_{x,h}(t) = \tilde u^k_{x_n,h}$ for all $t \in [nh, (n+1)h]$.

Step 4: Let $x_{n+1} = \Phi_h(x_n, \tilde u^k_{x_n,h})$, $n = n + 1$ and continue with Step 2.

In Step 2 a unique $\tilde u^k_{x_n,h} \in U$ may be found, e.g., by using a lexicographic order on U.

THEOREM 4.5. *Let* $u^k_{x,h}$ *denote the control function defined above. Then for every* $\varepsilon > 0$ *there exist* $H > 0$, $K(h) > 0$, *such that for all* $h < H$, $k \le K(h)$:

$$|J_\delta(x, u^k_{x,h}(\cdot)) - v_\delta(x)| \le \varepsilon \quad \textit{for all } x \in \Omega.$$

Proof. Using (4.12) or (4.14) and the definition of $u^{k,i}_{x,h} := u^k_{x,h}|_{[ih,(i+1)h)}$ we have the following for sufficiently small k and x_i from (4.7):

$$\begin{aligned} hg(x_i, u^{k,i}_{x,h}) + \beta v_h^k(\Phi_h(x_i, u^{k,i}_{x,h})) &\ge hg(x_i, u^{k,i}_{x,h}) + \beta v_h(\Phi_h(x_i, u^{k,i}_{x,h})) - \frac{\varepsilon}{2} \\ &\ge v_h(x_i) - \frac{\varepsilon}{2} \ \ge\ v_h^k(x_i) - \varepsilon \end{aligned}$$

and with $u^{0,i}_{x,h} \in U$ denoting the value, where $hg(x_i,u) + \beta v_h(\Phi_h(x_i,u))$, $u \in U$ attains its minimum:

$$\begin{aligned} hg(x_i, u^{k,i}_{x,h}) + \beta v_h^k(\Phi_h(x_i, u^{k,i}_{x,h})) &\le hg(x_i, u^{0,i}_{x,h}) + \beta v_h^k(\Phi_h(x_i, u^{0,i}_{x,h})) \\ &\le hg(x_i, u^{0,i}_{x,h}) + \beta v_h(\Phi_h(x_i, u^{0,i}_{x,h})) + \frac{\varepsilon}{2} \\ &= v_h(x_i) + \frac{\varepsilon}{2} \ \le\ v_h^k(x_i) + \varepsilon. \end{aligned}$$

Putting this together yields

$$|hg(x_i, u^{k,i}_{x,h}) + \beta v_h^k(\Phi_h(x_i, u^{k,i}_{x,h})) - v_h^k(x_i)| \le \varepsilon \quad \text{for all } x \in \overline{\Omega^k}. \tag{4.20}$$

By induction we can conclude that for every $\varepsilon > 0$, $p \in \mathbb{N}$, $h > 0$ there exists $k > 0$ such that

$$\left| h\sum_{j=0}^{p} \beta^j g(x_j, u^{k,j}_{x,h}) + \beta^{p+1} v_k^h(x_{p+1}) - v_h^k(x) \right| \le \frac{\varepsilon}{2} \quad \text{for all } x \in \overline{\Omega^k}. \tag{4.21}$$

Since $\beta < 1$ for all $h > 0$ and g and v_h^k are bounded on $\overline{\Omega^k}$, for every $\varepsilon > 0$ we may find a $p_h \in \mathbb{N}$ such that

$$\left| h \sum_{j=0}^{\infty} \beta^j g(x, u_{x,h}^{k,j}) - h \sum_{j=0}^{p_h} \beta^j g(x, u_{x,h}^{k,j}) - \beta^{p_h+1} v_h^k(x) \right| < \frac{\varepsilon}{2} \quad \text{for all } x \in \overline{\Omega^k},\ u \in \mathcal{U}_h. \tag{4.22}$$

Combining (4.12) or (4.14), (4.21), and (4.22) yields

$$|J_h(x, u_{x,h}^k(\cdot)) - v_h(x)| \le \varepsilon \quad \text{for all } x \in \Omega.$$

Using estimates (4.10) and (4.9) the assertion follows. □

Remark 4.6. The proof also shows how k and h have to be chosen: first choose h such that (4.10) and (4.9) hold for the desired accuracy; then choose k dependent on p_h from (4.22) such that (4.21) is fulfilled.

To construct a control function that is uniformly ε-optimal we can put together the ε-optimal control functions according to the following definition and lemma.

DEFINITION 4.7. *Let $u_x(\cdot) \in \mathcal{U}$ be control functions for every $x \in \overline{\Omega^k}$. Let $(\tau_i)_{i\in\mathbb{N}}$ be a real sequence of* switching times *satisfying $\tau_1 = 0$, $\tau_{i+1} > \tau_i$ and $a \le \tau_{i+1} - \tau_i \le b$ for all $i \in \mathbb{N}$ for positive constants a, $b \in \mathbb{R}$, $a \le b$. Then we define control functions $\bar{u}_x(\cdot) \in \mathcal{U}$ by*

$$\bar{u}_x|_{[\tau_i,\tau_{i+1})} \equiv u_{\varphi(x,\tau_i,\bar{u}_x(\cdot))}|_{[0,\tau_{i+1}-\tau_i)} \quad \text{for all } i \in \mathbb{N}.$$

LEMMA 4.8. *Assume for every $x \in \overline{\Omega^k}$ there exists a control function $u_x(\cdot) \in \mathcal{U}$ such that $|J_\delta(x, u_x(\cdot)) - v_\delta(x)| < \varepsilon$. Then for $\bar{u}_x(\cdot) \in \mathcal{U}$ from Definition* 4 *the following estimate holds:*

$$J_\delta(\varphi(\sigma, x, \bar{u}_x(\cdot)), \bar{u}_x(\sigma + \cdot)) \le v_\delta(\varphi(\sigma, x, \bar{u}_x(\cdot))) + \frac{e^{\delta b}}{\delta a}\varepsilon \quad \text{for all } \sigma \ge 0.$$

Proof. For all $t > 0$ it holds that

$$\begin{aligned} v_\delta(x) &\ge J_\delta(x, u_x(\cdot)) - \varepsilon \\ &\ge \int_0^t e^{-\delta\tau} g(\varphi(x, \tau, u_x(\cdot)), u_x(\tau)) d\tau + e^{-\delta t} v_\delta(\varphi(x, t, u_x(\cdot)) - \varepsilon. \end{aligned} \tag{4.23}$$

By induction with $t = \tau_i$ it follows that

$$J_\delta(x, \bar{u}_x(\cdot)) \le v_\delta(x) + \sum_{i=0}^{\infty} e^{-\delta\tau_i}\varepsilon$$

and for $0 < 1 - \delta a < 1$ this sum can be estimated by

$$\sum_{i=0}^{\infty} e^{-\delta\tau_i} \le \sum_{i=0}^{\infty} e^{-\delta a i} \le \sum_{i=0}^{\infty} (1 - \delta a)^i \le \frac{1}{\delta a}.$$

Together with the definition of the $\bar{u}_x(\cdot)$ this implies

$$J_\delta(\varphi(\tau_i, x, \bar{u}_x(\cdot)), \bar{u}_x(\tau_i + \cdot)) \le v_\delta(\varphi(\tau_i, x, \bar{u}_x(\cdot))) + \frac{\varepsilon}{\delta a}$$

for all $i \in \mathbb{N}$.

For the times between τ_i let $\sigma > 0, \tilde{\varepsilon} > 0$ and consider $u_{x_0}(\cdot) \in \mathcal{U}$ such that $|J_\delta(x_0, u_{x_0}(\cdot)) - v_\delta(x_0)| \leq \tilde{\varepsilon}$:

$$\begin{aligned} v_\delta(x_0) + \tilde{\varepsilon} &\geq \int_0^\infty e^{-\delta t} g(\varphi(t, x_0, u_{x_0}(\cdot)), u_{x_0}(t)) dt \\ &= \int_0^\sigma e^{-\delta t} g(\varphi(t, x_0, u_{x_0}(\cdot)), u_{x_0}(t)) dt \\ &\quad + e^{-\delta\sigma} \int_0^\infty e^{-\delta t} g(\varphi(t, \varphi(\sigma, x_0, u_{x_0}(\cdot)), u_{x_0}(\sigma + \cdot)), u_{x_0}(\sigma + t)) dt \\ &= \int_0^\sigma e^{-\delta t} g(\varphi(t, x_0, u_{x_0}(\cdot)), u_{x_0}(t)) dt \\ &\quad + e^{-\delta\sigma} J_\delta(\varphi(\sigma, x_0, u_{x_0}(\cdot)), u_{x_0}(\sigma + \cdot)) \\ &\geq \int_0^\sigma e^{-\delta t} g(\varphi(t, x_0, u_{x_0}), u_{x_0}(t)) dt \quad + \quad e^{-\delta\sigma} v_\delta(\varphi(\sigma, x_0, u_{x_0})) \\ &\geq v_\delta(x_0). \end{aligned}$$

From this inequality it follows that

$$|v_\delta(\varphi(\sigma, x_0, u_{x_0}(\cdot))) - J_\delta(\varphi(\sigma, x_0, u_{x_0}(\cdot)), u_{x_0}(\sigma + \cdot))| \quad \leq \quad e^{\delta\sigma} \tilde{\varepsilon}.$$

Choosing $i \in \mathbb{N}$ maximal with $\tau_i \leq \sigma$ and $x_0 := \varphi(\tau_i, x, \bar{u}_x(\cdot))$ it follows that

$$|v_\delta(\varphi(\sigma, x, \bar{u}_x(\cdot))) - J_\delta(\varphi(\sigma, x, \bar{u}_x(\cdot)), \bar{u}_x(\sigma + \cdot))| \leq e^{\delta(\sigma - \tau_i)} \frac{1}{\delta a} \varepsilon \leq e^{\delta b} \frac{1}{\delta a} \varepsilon = \frac{e^{\delta b}}{\delta a} \varepsilon$$

which finishes the proof. □

Remark 4.9. This lemma does not answer the question of which switching times τ_i are optimal. In estimate (4.23) we have to assume the worst case, i.e., that the error up to the time t

$$\varepsilon(t) := \left| v_\delta(x) - \int_0^t e^{-\delta\tau} g(\varphi(x, \tau, u_x(\cdot)), u_x(\tau)) d\tau - e^{-\delta t} v_\delta(\varphi(x, t, u_x(\cdot)) \right|$$

may be equal to ε for all $t > 0$ and hence the error becomes large if $a = \min(\tau_{i+1} - \tau_i)$ becomes small. The numerical examples discussed in the next section show that good results can be obtained for small a.

Using the results from Theorem 3.1 we can use the control functions constructed here to develop an *algorithm to stabilize bilinear control systems*:

Step 1: Calculate v_h^k, the approximation of the optimal value function for small discount rate $\delta > 0$ to approximate the minimal Lyapunov exponents of the systems (under the assumptions of Theorem 3.1).

Step 2: Given an initial value $x \in \mathbb{R}^d$ with $\lambda^*(x) < 0$ compute the control function that is ε-optimal along its trajectory according to Definition 4 (using the projected system). The trajectory of the bilinear system using this control is asymptotically stable under the assumptions of Theorem 3.1 provided h and k are small enough.

Note that the main numerical expense lies in the calculation of the approximated optimal value function v_h^k. Once this function is known the algorithm to calculate the control functions is numerically simple and quite fast.

For this algorithm only the information $x(t, x_0, u(\cdot))/\|x(t, x_0, u(\cdot))\|$ of the bilinear system is needed. In particular the calculated control functions are exactly the same for all x_1, $x_2 \in \mathbb{R}^d$ with $x_1/\|x_1\| = x_2/\|x_2\|$ and hence the algorithm works for arbitrarily large or small $\|x\|$. It is not necessary to discretize the trajectory of the bilinear system or to lift the discretized solution from $\mathbb{S}^{d-1}$ to $\mathbb{R}^d$ which then would imply that small discretization errors on $\mathbb{S}^{d-1}$ could become large in $\mathbb{R}^d$.

The value function and the corresponding optimal control values for each point can also be used to "verify" the assumptions of Theorem 3.1 (viii) numerically: if there exists a set such that $v_h^k < 0$ and this set is invariant with respect to the numerically computed optimal controls, the corresponding trajectory will tend to zero for any initial value from this set, provided the discretization is fine enough (see also Remark 3.2).

Remark 4.10. The way the stabilizing control functions are constructed leads to the question of whether v_h^k can be used to construct a stabilizing feedback for the bilinear control system. This question is closely related to the optimal switching times τ_i. If it is possible to choose $(\tau_{i+1} - \tau_i)$ arbitrarily small it could also be possible to obtain an ε-optimal feedback, e.g., by linear interpolation or averaging of the feedback for the discrete time system.

The main problem in proving this property of the switching times lies in the fact that the Euler method yields only linear convergence in h, hence quadratic convergence for one time step. Thus the difference between $v_h(\varphi(h, x, u))$ (the value that can be reached after the first time step) and $v_h(\Phi_h(x, u))$ (the value that is supposed to be reached) is of the order $h^{2\gamma}$. For $\gamma < \frac{1}{2}$ this error will accumulate and convergence is no longer guaranteed. However, there is hope to overcome this difficulty by using a higher-order method to calculate $\Phi_h(x, u)$ which then will require a different proof of the convergence of v_h.

5. Numerical examples. In this section we will present some numerical examples calculated with the algorithm developed in the previous sections. All examples were computed on an IBM6000 Workstation.

The first example is a bilinear control system in $\mathbb{R}^2$, the two-dimensional linear oscillator given by

$$\ddot{x} + 2b\dot{x} + (1 + u)x = 0$$

or written as a two-dimensional system by $x_1 = x$, $x_2 = \dot{x}$:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ -1-u & -2b \end{pmatrix}}_{=:A(u)} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{5.1}$$

The projection of the system to $\mathbb{S}^1$ by $s = \frac{x}{\|x\|}$ reads

$$\dot{s} = \begin{pmatrix} s_2(1 + us_1^2 + 2bs_1s_2) \\ -(1+u)s_1 - 2bs_2 + s_2^2(us_1 + 2bs_2) \end{pmatrix}. \tag{5.2}$$

For the one-dimensional sphere we may use the parametrization via polar coordinates $\Psi(\varphi) = (\cos\varphi, \sin\varphi)$ where $\Psi^{-1}(s) = \arcsin(s_2)$, $s_2 \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, $\Psi^{-1}(s) = \arcsin(\pi - s_2)$, $s_2 \in [\frac{\pi}{2}, \frac{3\pi}{2}]$. In polar coordinates the cost function reads $g(\varphi, u) = -\sin\varphi(u\cos\varphi + 2b\sin\varphi)$, and we can choose $\Omega = (0 - \varepsilon, \pi + \varepsilon)$ to cover the whole projective space (identified with one half of the sphere).

Tables 5.1–5.3 show the number of iterations in the increasing coordinate algorithm (ops_1) and the number of evaluations of the operator T_h^k in the accelerated algorithm (ops_2) depending

TABLE 5.1
Dependence on the time step h ($k = 0.032$, $\delta = 1.0$).

h	ops_1	ops_2
1.0	13	11477
0.1	42	11477
0.01	51	11477

TABLE 5.2
Dependence on the space discretization k ($h = 0.1$, $\rho = 1.0$).

k	ops_1	ops_2
0.063	28	5918
0.032	42	11477
0.0063	233	49625

TABLE 5.3
Dependence on the discount rate δ ($h = 0.1$, $k = 0.032$).

δ	V_0	ops_1	ops_2
5.0	−0.66	16	2001
2.0	−1.66	35	5543
1.0	−3.32	42	11477
0.1	−33.23	194	121187
0.01	−332.27	1707	–
0.001	−3322.72	16836	–

TABLE 5.4
Lyapunov spectrum for system (5.1) *with $b = 1.5$.*

ρ	$\min(D_1)$	$\max(D_1)$	$\min(D_2)$	$\max(D_2)$
0.0	−2.61	−2.61	−0.38	−0.38
0.1	−2.65	−2.58	−0.42	−0.35
0.2	−2.69	−2.52	−0.47	−0.31
0.3	−2.73	−2.47	−0.52	−0.25
0.4	−2.77	−2.42	−0.57	−0.22
0.5	−2.81	−2.37	−0.63	−0.19
0.6	−2.85	−2.31	−0.69	−0.14
0.7	−2.89	−2.24	−0.75	−0.11
0.8	−2.91	−2.18	−0.82	−0.07
0.9	−2.96	−2.09	−0.90	−0.06
1.0	−2.99	−2.00	−0.99	0.00
1.1	−3.00	−1.90	−1.10	0.03
1.2	−3.03	−1.74	−1.27	0.06
1.3	−3.03	–	–	0.10

on certain parameters with damping parameter $b = 1.5$. Remember that one iteration in the increasing coordinate algorithm corresponds to one evaluation of T_h^k. The used set of control values was ρU with $U = \{-1, 1\}$ and $\rho = 0.5$.

Using the techniques described in Remark 3.2 the whole Lyapunov spectrum for this system was computed for $\rho = \{0.1, 0.2, \ldots, 1.3\}$ with parameters $h = 0.01$, $k = 0.006$, and $\delta = 0.01$ and locally refined grid with $k = 0.0016$ around the variant control set for $\rho \leq 0.5$. (For $\rho = 0.0$ the exponents are just the eigenvalues of A.) The calculated intervals are shown in Table 5.4 and Figure 5.1. For $\rho \leq 1.2$ there exist two control sets D_1 and D_2 and therefore two intervals of Lyapunov exponents. For $\rho = 1.3$ there is only one control set and thus only

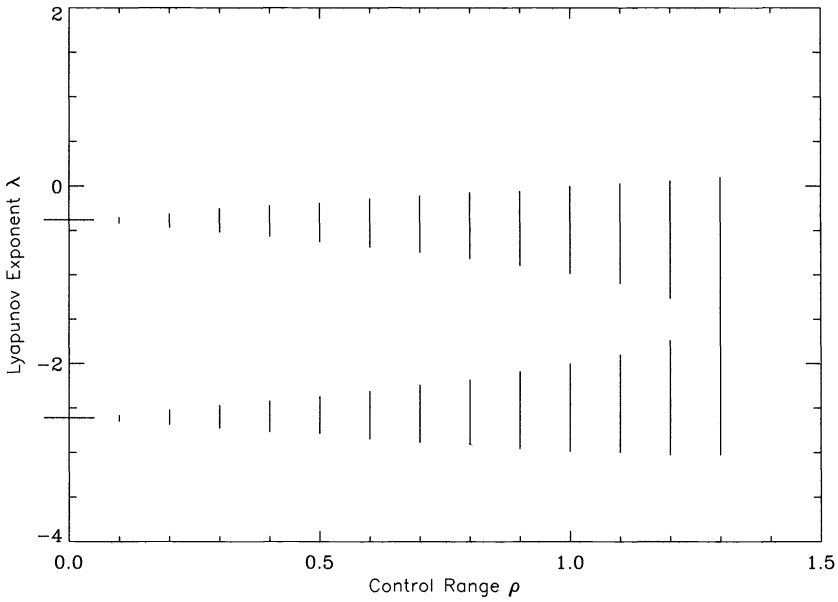


FIG. 5.1. *Lyapunov spectrum of system* (5.1) *with* $b = 1.5$.

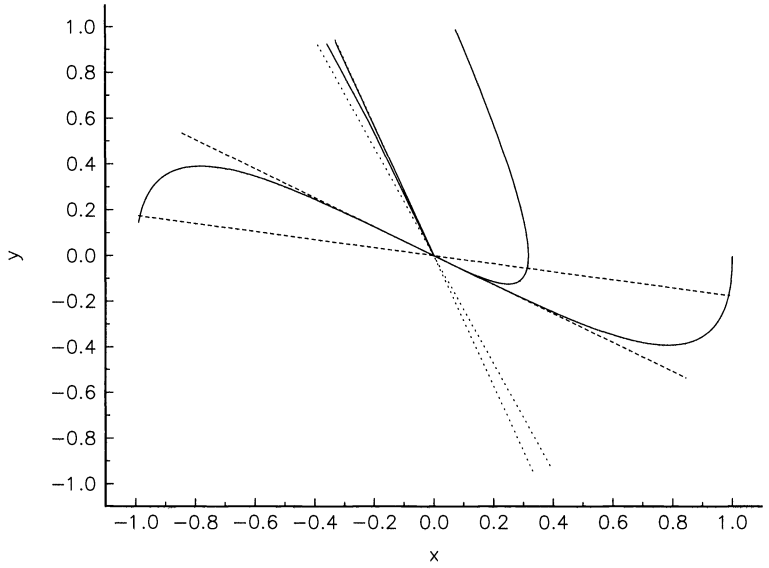


FIG. 5.2. *Trajectories for* $b = 1.5$, $\rho = 0.5$.

one interval. For this system a finer discretization of U does not yield different values for v_h^k; it is sufficient to minimize over the extremal control values.

For $\rho = 0.5$ the system is asymptotically stable for all control functions since the maximal Lyapunov exponent is negative. But as the Lyapunov exponents corresponding to D_1 are much smaller than those of the control set D_2 it can be expected that the optimal trajectories with initial value inside D_1 tend to zero much faster.

Figure 5.2 shows that this is exactly what happens. In this figure the dotted lines correspond to the boundaries of D_1, the dashed lines to the boundary of D_2.

All trajectories in this section were computed using the extrapolation method for ordinary differential equations by Stoer and Bulirsch [19, §7.2.14]. The parameter a from Definition 4 was chosen as $a = h$ (see Remark 4.9).

The second example is the three-dimensional linear oscillator given by

$$\dddot{y} + a\ddot{y} + b\dot{y} + (c + u)y = 0 \tag{5.3}$$

or written as a three-dimensional system by

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -(c+u) & -b & -a \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \tag{5.4}$$

with $a, b, c \in \mathbb{R}$, and $u \in U$. The projected system on $\mathbb{S}^2$ reads

$$\dot{s} = \begin{pmatrix} s_2 - s_1(-\tilde{u}s_1s_3 + s_1s_2 + (1-b)s_2s_3 - as_3^2) \\ s_3 - s_2(-\tilde{u}s_1s_3 + s_1s_2 + (1-b)s_2s_3 - as_3^2) \\ -\tilde{u}s_1 - bs_2 - as_3 - s_3(-\tilde{u}s_1s_3 + s_1s_2 + (1-b)s_2s_3 - as_3^2) \end{pmatrix} \tag{5.5}$$

with $\tilde{u} := c + u$.

For $\mathbb{S}^2$ the parametrization by spherical coordinates is not suitable since this parametrization maps two opposite points to a line and hence it is not invertible on one half of the sphere. Thus the stereographic projection is used instead; it is given by

$$\Psi(x) = \left(\frac{2x_1}{1 + \|x\|^2}, \frac{2x_2}{1 + \|x\|^2}, \frac{2}{1 + \|x\|^2} - 1 \right)$$

and

$$\Psi^{-1}(s) = \left(\frac{1}{1+s_3} s_1, \frac{1}{1+s_3} s_2 \right).$$

The cost function reads

$$g(x, u) = -(c+u)\Psi_1(x)\Psi_3(x) + \Psi_1(x)\Psi_2(x) + (1-b)\Psi_2(x)\Psi_3(x) - a\Psi_3(x)^2$$

with $\Psi = (\Psi_1, \Psi_2, \Psi_3)$. The set Ω was chosen as $\Omega = (-1-\varepsilon, 1+\varepsilon) \times (-1-\varepsilon, 1+\varepsilon)$ to cover the whole $\mathbb{P}^2$ (identified with the upper half of $\mathbb{S}^2$).

All values given have been checked according to Remark 3.2 (ii); in all cases it was possible to find trajectories that realized the values as Lyapunov exponents. Hence the calculated values at least give an approximation of the minimal Lyapunov exponents over the interior of the control sets. To apply the results of Remark 3.2 (iii), i.e., to make sure that this is indeed the Lyapunov spectrum, we have to check the $\rho - \rho'$ inner pair condition described in §3. Unfortunately as of now it is not known how to check this condition analytically. However, the program CS2DIM from Häckl [15] has been used to calculate reachable sets for the system for different ρ-parameters numerically. Since they turned out to be strictly increasing in this example there is strong evidence that the condition is fulfilled.

For Figures 5.3–5.8 spherical coordinates ($s_1 = \sin\theta\cos\varphi$, $s_2 = \sin\theta\sin\varphi$, $s_3 = \cos\theta$)$x = \theta$, $y = \varphi$ were used and the system was transformed by $z(t) := e^{\frac{1}{3}at}y(t)$.

The first parameters considered for this system were $a = 1$, $b = 0$, $c = 0.5$, and $U = \{-0.3, -0.25, \ldots, 0.25, 0.3\}$. Figure 5.3 shows the two control sets of this system. The control sets were computed again using the program CS2DIM [15].

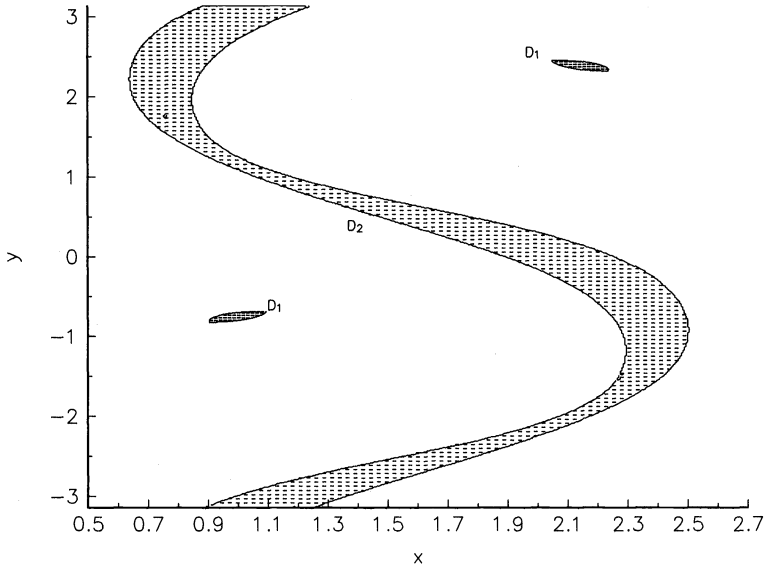


FIG. 5.3. *Control sets of system* (5.5) *with* $a = 1$, $b = 0$, $c = 0.5$.

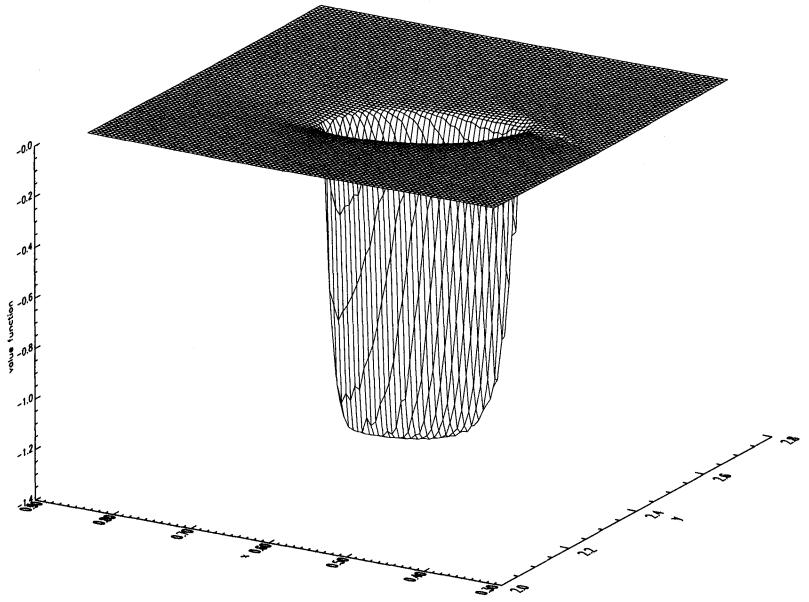


FIG. 5.4. *Value function around* D_1.

The numerical parameters used for this example are $k = 0.003$ around D_1, $k = 0.09$ elsewhere, $h = 0.05$, and $\delta = 0.01$. The discounted value function of this system around D_1 is shown in Figure 5.4. The calculated minimal Lyapunov exponent over D_1 is -1.25, the maximal exponent is -1.15. The calculated minimal and maximal exponents over D_2 are 0.019 and 0.24 and the value function is constant outside D_1. Figure 5.5 shows two

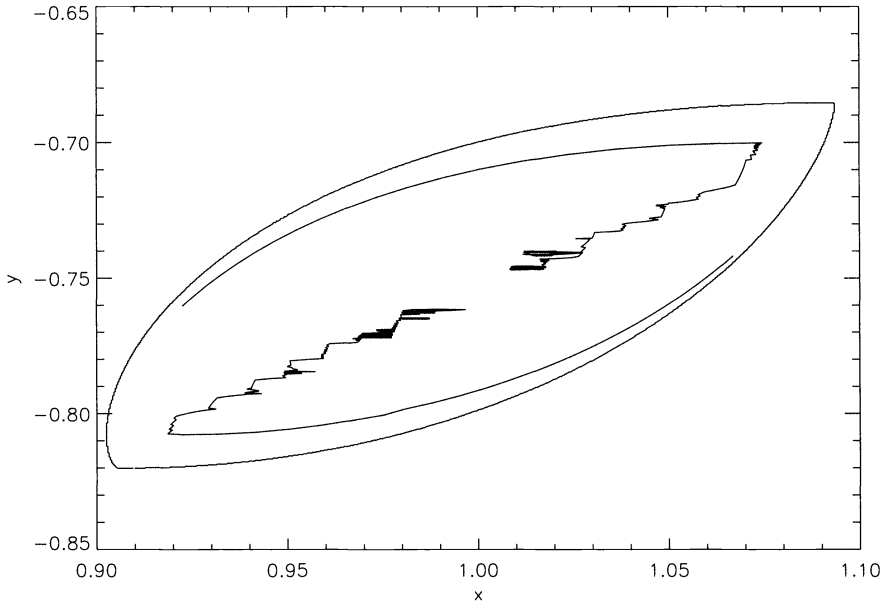


FIG. 5.5. *Optimal trajectories in* D_1.

TABLE 5.5
Stabilized trajectory for system (5.4) *with* $a = 1$, $b = 0$, $c = 0.5$.

t	x_1	x_2	x_3
1	0.124609	−0.169914	0.219449
2	0.031318	−0.043096	0.060307
3	0.008062	−0.010800	0.014665
4	0.002129	−0.002818	0.003757
5	0.000569	−0.000750	0.000986
6	0.000153	−0.000201	0.000264
7	0.000041	−0.000054	0.000071
8	0.000011	−0.000014	0.000019
9	0.000003	−0.000004	0.000005
10	0.000000	−0.000001	0.000001
11	0.000000	0.000000	0.000000

trajectories of the projected system with initial values inside D_1. Table 5.5 shows the values of one corresponding trajectory in $\mathbb{R}^3$.

The second set of parameters considered for this system is $a = -1$, $b = -3$, $c = 0.5$, and $U = \{-1.0, -0.9, \ldots, 0.9, 1.0\}$. Figure 5.6 shows the three control sets of the projected system, the domain of attraction of D_2 (denoted by $A^-(D_2)$), and the domain of attraction of D_2 of the time-reversed system (denoted by $A^+(D_2)$).

Here the numerical parameters were $k = 0.002$ around D_1, $k = 0.045$ elsewhere, $h = 0.05$, and $\delta = 0.01$. Figure 5.7 shows the discounted optimal value function around D_1.

The calculated spectrum for this example is $\lambda(D_1) = [-1.47, -1.17]$, $\lambda(D_2) = [-0.10, 0.43]$, and $\lambda(D_3) = [2.07, 2.36]$.

Figure 5.8 shows an optimal trajectory in $\mathbb{P}^2$, starting in the domain of attraction of D_2. Table 5.6 shows the corresponding trajectory (x_1, x_2, x_3) in $\mathbb{R}^3$ and another trajectory

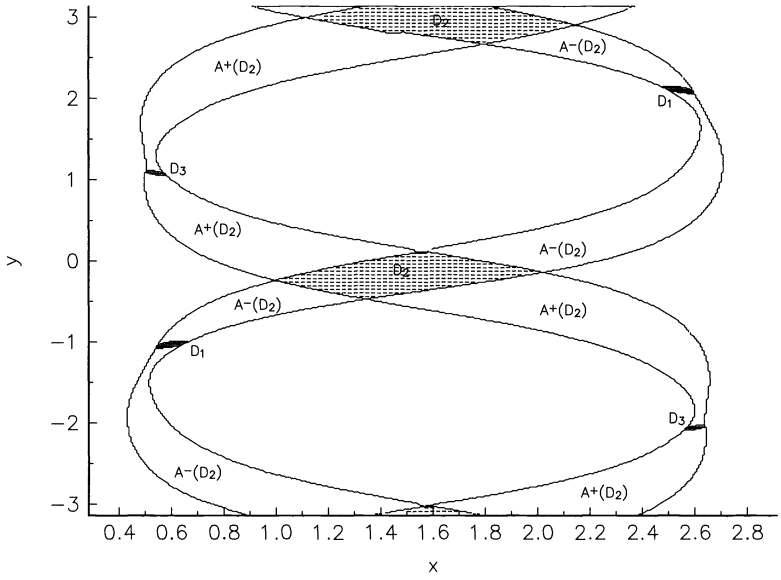


FIG. 5.6. *Control sets of system* (5.5) *with* $a = -1.0$, $b = -3.0$, $c = 0.5$.

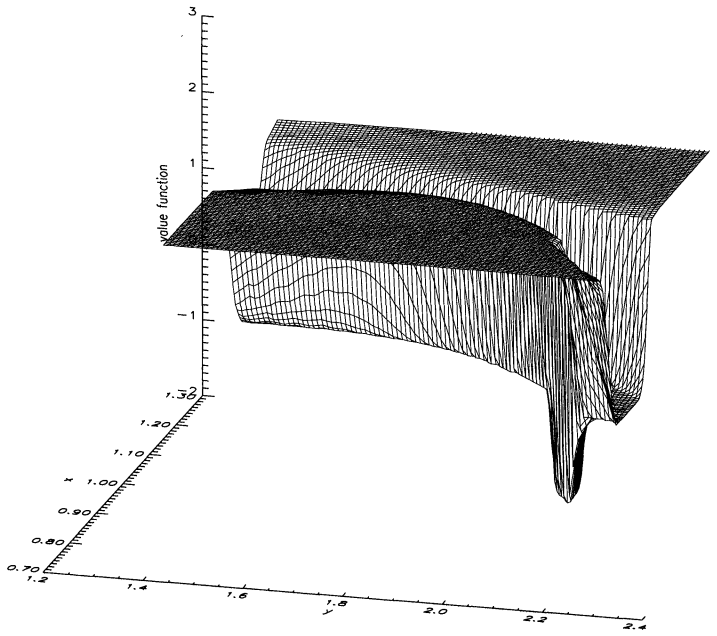


FIG. 5.7. *Value function of the system.*

(y_1, y_2, y_3) in $\mathbb{R}^3$ with projected initial value in D_1. This trajectory tends to zero much faster, which is exactly what one would expect since the minimal Lyapunov exponent inside D_1 is much smaller.

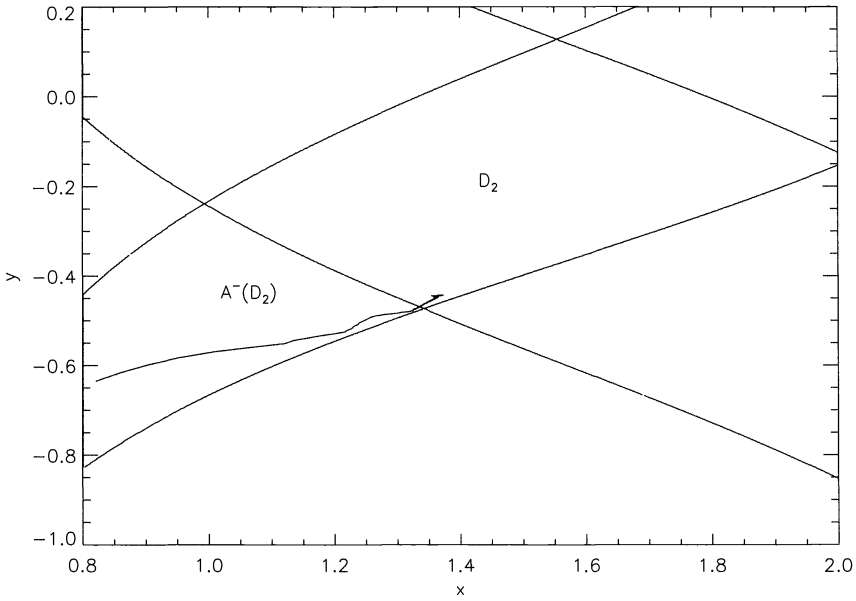


FIG. 5.8. *Optimal trajectory starting in* $A^-(D_2)$.

TABLE 5.6
Stabilized trajectories for system (5.4) *with* $a = -1$, $b = -3$, $c = 0.5$.

t	x_1	x_2	x_3	y_1	y_2	y_3
1	0.576395	−0.119011	0.071718	0.096972	−0.141621	0.200267
5	0.293857	−0.044627	0.007731	0.000260	−0.000384	0.000562
10	0.142984	−0.020156	0.003083	0.000000	−0.000000	0.000000
15	0.070691	−0.009962	0.001172	0.000000	−0.000000	0.000000
20	0.034949	−0.004924	0.000754	0.000000	−0.000000	0.000000
25	0.017279	−0.002434	0.000373	0.000000	−0.000000	0.000000
30	0.008543	−0.001204	0.000184	0.000000	−0.000000	0.000000
35	0.004224	−0.000595	0.000091	0.000000	−0.000000	0.000000
40	0.002088	−0.000294	0.000045	0.000000	−0.000000	0.000000
45	0.001032	−0.000146	0.000022	0.000000	−0.000000	0.000000
50	0.000510	−0.000072	0.000008	0.000000	−0.000000	0.000000

Acknowledgment. I would like to thank Fritz Colonius for his constant help and many useful discussions as well as a number of anonymous referees for their detailed advice.

REFERENCES

[1] M. BARDI AND M. FALCONE, *An approximation scheme for the minimum time function*, SIAM J. Control Optim., 28 (1990), pp. 950–965.

[2] I. CAPUZZO DOLCETTA, *On a discrete approximation of the Hamilton-Jacobi equation of dynamic programming*, Appl. Math. Optim., 10 (1983), pp. 367–377.

[3] I. CAPUZZO DOLCETTA AND M. FALCONE, *Discrete dynamic programming and viscosity solutions of the Bellman equation*, Ann. Inst. H. Poincaré, Anal. Non Linéaire, 6 (supplement) (1989), pp. 161–184.

[4] I. CAPUZZO DOLCETTA AND H. ISHII, *Approximate solutions of the Bellman equation of deterministic control theory*, Appl. Math. Optim., 11 (1984), pp. 161–181.

[5] R. CHABOUR, G. SALLET, AND J. VIVALDA, *Stabilization of nonlinear systems: a bilinear approach*, Math. Control Signals Systems, 6 (1993), pp. 224–246.
[6] F. COLONIUS, *Asymptotic behaviour of optimal control systems with low discount rates*, Math. Oper. Res., 14 (1989), pp. 309–316.
[7] F. COLONIUS AND W. KLIEMANN, *Infinite time optimal control and periodicity*, Appl. Math. Optim., 20 (1989), pp. 113–130.
[8] ———, *Linear control semigroups acting on projective space*, J. Dynamics Differential Equations, 5 (1993), pp. 495–528.
[9] ———, *Maximal and minimal Lyapunov exponents of bilinear control systems*, J. Differential Equations, 101 (1993), pp. 232–275.
[10] ———, *Asymptotic null controllability of bilinear systems*, in Geometry in Nonlinear Control and Differential Inclusions, Vol. 32, B. Jakubczyk and W. Respondek, eds., Banach Center Publications, Warsaw, 1995, pp. 139–148.
[11] ———, *The Lyapunov spectrum of families of time varying matrices*, Trans. Amer. Math. Soc., (1996), to appear.
[12] M. FALCONE, *Numerical solution of deterministic control problems*, in Proc. International Symposium on Numerical Analysis, Madrid, 1985.
[13] ———, *A numerical approach to the infinite horizon problem of deterministic control theory*, Appl. Math. Optim., 15 (1987), pp. 1–13; *corrigenda*, Appl. Math. Optim., 23 (1991), pp. 213–214.
[14] R. L. V. GONZÁLES AND M. M. TIDBALL, *On the rates of convergence of fully discrete solutions of Hamilton-Jacobi equations*, INRIA Rapports de Recherche Nr. 1379, 1991.
[15] G. HÄCKL, *Numerical approximation of reachable sets and control sets*, Random Comput. Dynamics, 1 (1992–1993), pp. 371–394.
[16] W. KLIEMANN, *Recurrence and invariant measures for degenerate diffusions*, Ann. Probab., 15 (1987), pp. 690–707.
[17] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman Publishing, London, 1982.
[18] R. MOHLER, *Bilinear Control Processes*, Academic Press, New York, 1973.
[19] J. STOER AND R. BULIRSCH, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.
[20] F. WIRTH, *Convergence of the value functions of discounted infinite horizon optimal control problems with low discount rates*, Math. Oper. Res., 18 (1993), pp. 1006–1019.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 2051–2063, November 1996

010

CONVERGENCE OF THE BFGS METHOD FOR LC^1 CONVEX CONSTRAINED OPTIMIZATION*

XIAOJUN CHEN†

Abstract. This paper proposes a BFGS-SQP method for linearly constrained optimization where the objective function f is required only to have a Lipschitz gradient. The Karush–Kuhn–Tucker system of the problem is equivalent to a system of nonsmooth equations $F(v) = 0$. At every step a quasi-Newton matrix is updated if $\|F(v_k)\|$ satisfies a rule. This method converges globally, and the rate of convergence is superlinear when f is twice strongly differentiable at a solution of the optimization problem. No assumptions on the constraints are required. This generalizes the classical convergence theory of the BFGS method, which requires a twice continuous differentiability assumption on the objective function. Applications to stochastic programs with recourse on a CM5 parallel computer are discussed.

Key words. quasi-Newton methods, convex programming, nonsmooth equations

AMS subject classifications. 90C30, 90C25

1. Introduction. The Broyden–Fletcher–Goldfarb–Shanno (BFGS) method is the most successful quasi-Newton method for solving convex minimization problems [11], [14]. In this paper we consider the BFGS method for solving the following constrained minimization problem:

$$\begin{aligned} &\min \ f(x) \\ &\text{subject to (s.t.)} \ Ax \le b, \end{aligned} \tag{1.1}$$

where $A \in R^{m\times n}$, $b \in R^m$, and $f : R^n \to R$ is a convex LC^1 function.

The LC^1 property of f means that f is Fréchet differentiable at all points in an open convex set $\Omega \subseteq R^n$ containing $X = \{x \in R^n : Ax \le b\}$, and the gradient function $g := \nabla f : \Omega \to R^n$ is locally Lipschitz in Ω. If f is LC^1 and X is nonempty, then (1.1) is called an LC^1 minimization problem. LC^1 optimization problems arise from nonlinear minimax problems, stochastic programs, augmented Lagrangians, semi-infinite programs, and some differentiable penalty function methods for constrained optimization problems. See [2], [7], [8], [13], [27], [28], [31], [32], [33], [37].

The Karush–Kuhn–Tucker (KKT) system for (1.1) is

$$\nabla f(x) + A^T u = 0,$$

$$u \ge 0, \ b - Ax \ge 0, \ u^T(b - Ax) = 0,$$

where $x \in R^n$ and $u \in R^m$. Let $N = n + m$ and $v = (x^T, u^T)^T$. Then the KKT system is equivalent to a system of nonsmooth equations [25]:

$$F(v) := \begin{pmatrix} \nabla f(x) + A^T u, \\ \min(u, b - Ax) \end{pmatrix} = 0, \tag{1.2}$$

where the "min" operator denotes the componentwise minimum of two vectors.

The local convergence theory of quasi-Newton methods for smooth equations and smooth unconstrained minimization problems is well developed. See [3], [4], [11], [12], [14], [22]. Quasi-Newton methods have been applied to nonsmooth equations in [5], [6], [9], [16], [17],

*Received by the editors September 28, 1994; accepted for publication (in revised form) September 19, 1995.

†School of Mathematics, University of New South Wales, Sydney 2052, Australia (X.Chen@unsw.edu.au). This research was supported by the Australian Research Council.

[19], [21], [23], but local superlinear convergence properties have not been obtained without the differentiability of the function F at a solution of the system of nonsmooth equations. Recently, Bonnans, Gilbert, Lemaréchal, and Sagastizábal [2] presented a conceptual method for LC^1 unconstrained minimization combining the BFGS method with the Moreau–Yosida regularization. Their local superlinear convergence requires that f be twice strongly differentiable at the solution.

Several authors [7], [8], [27], [28], [31], [32] studied the generalized Newton method for solving problem (1.1). Local superlinear convergence properties were established under a "semismoothness" assumption on the gradient function g at a solution [28], [31]. The generalized Newton method utilizes the generalized Hessian instead of $\nabla^2 f$ in a Newton method or a sequential quadratic programming (SQP) method. In many cases, however, calculating the generalized Hessian is very difficult.

Motivated by the fact that the BFGS method combines global convergence, a rate of superlinear convergence, and simple updates for smooth unconstrained minimization problems [4], we present a BFGS-SQP method for solving the LC^1 minimization problem (1.1). This method replaces the Hessian in the SQP method by the updated BFGS matrix and uses an Armijo line search to reduce the objective value. Moreover, this method uses the BFGS formula to update the quasi-Newton matrix if $\|F(v_k)\|$ satisfies a rule at every step.

The goal of this paper is to establish global and superlinear convergence of the BFGS-SQP method for the LC^1 convex optimization problem (1.1). Global convergence of this method only requires Lipschitz continuity of the gradient function and boundedness of the level sets of f in X. Superlinear convergence of this method requires twice strong differentiability of f at the solution of (1.1), but it does not require differentiability of F at the solution of (1.2). Furthermore no assumptions on the constraints are required, for example, the linear independence condition [31]. Note that an LC^1 function f can be twice strongly differentiable at a single point but can fail to be twice differentiable at arbitrarily close neighboring points (cf. [25]). Our results extend the classical convergence theory of the BFGS method for convex, smooth minimization problems which typically requires a twice continuous differentiability assumption on the objective function.

The remainder of the paper is organized as follows. In §2 we review the key analytic properties of the BFGS method. In §3 we give the BFGS-SQP method and study global convergence of this method. In §4 we study superlinear convergence of the BFGS-SQP method. In §5 we discuss applications to stochastic programs with recourse on a CM5 parallel computer.

2. The BFGS method. Quasi-Newton versions of SQP methods for solving linearly constrained minimization problems are iterative methods of the form [14]

$$x_{k+1} = x_k + \alpha_k d_k,$$

where α_k is a steplength and d_k is a solution of the quadratic subproblem

$$\begin{aligned} &\min g_k^T d + \frac{1}{2} d^T B_k d \\ &\text{s.t. } A(x_k + d) \le b. \end{aligned}$$

The matrix B_k is updated at every step by means of a quasi-Newton update formula, and g_k is the gradient of f at x_k. In particular, the BFGS update formula is given by

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}, \tag{2.1}$$

where $y_k = g_{k+1} - g_k$ and $s_k = x_{k+1} - x_k$.

The BFGS formula has an important property that B_{k+1} is positive definite if B_k is positive definite and $y_k^T s_k > 0$.

Powell [29] first proved the global convergence of the BFGS method for unconstrained optimization by measuring the trace

$$\operatorname{tr}(B_{k+1}) = \operatorname{tr}(B_k) - \frac{\|B_k s_k\|^2}{s_k^T B_k s_k} + \frac{\|y_k\|^2}{y_k^T s_k} \tag{2.2}$$

and the determinant

$$\det(B_{k+1}) = \det(B_k)\frac{y_k^T s_k}{s_k^T B_k s_k}.$$

Byrd and Nocedal [4] simplified the proof by using a function

$$\psi(B_k) = \operatorname{tr}(B_k) - \ln(\det(B_k)).$$

Let x_0 be the starting point for the BFGS method. We define the level set

$$\hat{D}_0 = \{x \in R^n : \ f(x) \le f(x_0)\}.$$

Byrd and Nocedal [4] proved the global convergence of the BFGS method under the conditions that f is twice continuously differentiable and there exist positive constants μ and ν such that

$$\mu\|z\|^2 \le z^T \nabla^2 f(x) z \le \nu\|z\|^2 \tag{2.3}$$

for all $z \in R^n$ and all $x \in \hat{D}_0$. Note that (2.3) implies that f has a unique minimizer $x^* \in \hat{D}_0$. They proved the local superlinear convergence under assumption (2.3) and that $\nabla^2 f$ is Lipschitz continuous at the minimizer x^*.

Bonnans, Gilbert, Lemaréchal, and Sagastizábal [2] presented a BFGS proximal method for LC^1 unconstrained optimization problems which combines the Moreau–Yosida regularization and the BFGS method. The BFGS proximal method is an iterative method of the form

$$x_{k+1} = x_k + \alpha_k(x_k^p - x_k),$$

where α_k is a steplength and

$$x_k^p = \arg\min\left\{ f(x) + \frac{1}{2}(x - x_k)^T B_k (x - x_k), \ x \in R^n \right\}. \tag{2.4}$$

The matrix B_k is updated at every step by the BFGS formula (2.1).

Paper [2] gave preliminary results to combine methods for nonsmooth optimization and classical quasi-Newton methods. However, the BFGS proximal method is only a conceptual algorithm because we do not specify how $\{x_k^p\}$ can be calculated from (2.4). The superlinear convergence theorem for the BFGS proximal method requires that f be twice strongly differentiable at the solution. This assumption has been required in superlinear convergence analysis of the Broyden method. See [5] and [19].

3. Global convergence. We will now use ideas from both smooth and nonsmooth optimization [2], [4], [29] to study a new BFGS-SQP method for LC^1 convex constrained optimization (1.1). We show global convergence of this method in this section.

BFGS-SQP method

Given constants $\tau, \sigma, \rho, \eta \in (0, 1)$, $\epsilon_0 > 0$ and an integer $r > 0$, choose $x_0 \in X$, $u_0 \geq 0$, and an $n \times n$ symmetric positive definite matrix B_0. Let $v_0 = (x_0^T, u_0^T)^T$ and let $\delta = \|F(v_0)\|$. For $k \geq 0$

1. Solve the quadratic program

$$\begin{aligned} &\min g_k^T d + \tfrac{1}{2} d^T (B_k + \epsilon_k I) d \\ &\text{s.t. } A(x_k + d) \leq b. \end{aligned} \tag{3.1}$$

Let d_k be the solution of (3.1) and u_{k+1} be the Lagrange multipliers at d_k corresponding to $A(x_k + d) \leq b$.

2. Let t_k be the minimum integer $t \geq 0$ such that

$$f(x_k + \rho^t d_k) - f(x_k) \leq \frac{\sigma}{2} \rho^t g_k^T d_k. \tag{3.2}$$

Let $\alpha_k = \rho^{t_k}$ and let $x_{k+1} = x_k + \alpha_k d_k$.

3. Let $v_{k+1} = (x_{k+1}^T, u_{k+1}^T)^T$.

If $\|F(v_{k+1})\|/\delta \leq \eta$,

let $\delta = \|F(v_{k+1})\|$, $\epsilon_{k+1} = \tau \epsilon_k$.

Otherwise let $\epsilon_{k+1} = \epsilon_k$.

4. If $y_k^T d_k \geq \rho^{r+1} \epsilon_k d_k^T d_k$, update B_k by the BFGS formula (2.1). Otherwise, set $B_{k+1} = B_k$.

5. If x_{k+1} satisfies a prescribed stopping criterion, terminate; otherwise, return to Step 1 with k replaced by $k + 1$.

Without loss of generality, we assume that x_k is feasible but nonoptimal to (1.1). Suppose that B_k is symmetric positive definite. Since $d = 0$ is feasible to (3.1), the optimal objective value of (3.1) must be nonpositive. Moreover, the nonoptimality of x_k to (1.1) implies that (3.1) has a unique optimal solution d_k which satisfies

$$g_k^T d_k + \frac{1}{2} d_k^T (B_k + \epsilon_k I) d_k < 0. \tag{3.3}$$

This implies

$$g_k^T d_k < 0.$$

A standard result in nonlinear programming [1] establishes that the integer t_k is well defined and finite. Therefore, we have

$$y_k^T s_k = \alpha_k y_k^T d_k > 0 \qquad \text{if } y_k^T d_k \geq \rho^{r+1} \epsilon_k d_k^T d_k.$$

By construction, the matrix B_{k+1} is symmetric positive definite. By the convexity of X, the point x_{k+1} is feasible. Hence the BFGS-SQP method is well defined and generates an infinite sequence $\{x_k\} \subseteq X$.

We define the level set $D_0 = \{x \in X : f(x) \leq f(x_0)\}$, where x_0 is the starting point for the BFGS-SQP method.

Assumption 3.1. Assume that the level set D_0 is bounded and there is a positive constant L such that

$$\|g(x) - g(y)\| \leq L\|x - y\| \qquad \text{if } x, y \in D_0.$$

THEOREM 3.1. *Under Assumption* 3.1, *every accumulation point* $\bar{x}$ *of the sequence* $\{x_k\}$ *produced by the BFGS-SQP method is an optimal solution of* (1.1).

Proof. Let $K = \{0, 1, 2, \ldots\}$ and $K_0 = \{k \in K : \epsilon_{k+1} = \tau\epsilon_k \text{ at iteration } k\}$.

If K_0 is infinite, then every accumulation point $\bar{v}$ of the sequence $\{v_k : k \in K_0\}$ is a solution of $F(v) = 0$. By Theorem 9.4.2 in [14], $\bar{x}$ is a global minimum point of (1.1).

The remainder of this proof is for the case where K_0 is finite.

Since D_0 is bounded, the sequence $\{x_k\}$ is bounded. Since K_0 is finite, there is a large $\hat{k}$ such that $\epsilon_k = \epsilon_{\hat{k}}$ for all $k \geq \hat{k}$, and thus $\epsilon_k \geq \epsilon_{\hat{k}}$ for all $k \geq 0$. Let $\hat{\epsilon} = \epsilon_{\hat{k}}$. Since B_k is symmetric positive definite, we have

$$d^T(B_k + \epsilon_k I)d \geq \hat{\epsilon}\|d\|^2 \quad \text{for all } d \in R^n \text{ and all } k \geq 0\,. \tag{3.4}$$

By the variational principle of the quadratic program (3.1), the unique optimal solution d_k satisfies

$$g_k^T d_k + d_k^T(B_k + \epsilon_k I)d_k \leq 0. \tag{3.5}$$

Let $K_1 = \{k \in K, y_k^T d_k \geq \rho^{r+1}\epsilon_k d_k^T d_k\}$. If K_1 is finite, then there is a large $\bar{k}$ such that $B_{k+1} = B_{\bar{k}}$ for $k \geq \bar{k}$. Using the inequality [29]

$$\|y_k\|^2 \leq L y_k^T s_k$$

and the trace relation (2.2), we obtain

$$\operatorname{tr}(B_{k+1}) \leq \operatorname{tr}(B_k) + L \qquad \text{if } k \in K_1.$$

Noticing $B_{k+1} = B_k$ if $k \notin K_1$ we have

$$\operatorname{tr}(B_k) \leq \operatorname{tr}(B_0) + \bar{k}L \qquad \text{for all } k \geq 0.$$

Since the largest eigenvalue λ_{k_n} of B_k is less than the trace, we get

$$d^T(B_k + \epsilon_k I)d \leq (\lambda_{k_n} + \epsilon_k)\|d\|^2 \leq (\operatorname{tr}(B_0) + \bar{k}L + \epsilon_0)\|d\|^2$$

for all $d \in R^n$ and all $k \in K$. Hence we can show that every accumulation point $\bar{x}$ is an optimal solution of (1.1) by a standard proof (see, e.g., [26]). We omit the details.

Now we consider the case where K_1 is infinite. From (3.2), (3.4), and (3.5) and that D_0 is bounded, we have

$$\hat{\epsilon}\sum_{k=0}^{\infty}\alpha_k\|d_k\|^2 \leq \sum_{k=0}^{\infty}\alpha_k d_k^T(B_k+\epsilon_k I)d_k \leq \sum_{k=0}^{\infty} -\alpha_k g_k^T d_k \leq \frac{2}{\sigma}(f(x_0)-f(\bar{x})) < \infty. \tag{3.6}$$

Furthermore, if $k \in K_1$, we have

$$d_k^T d_k \leq y_k^T d_k/(\rho^{r+1}\epsilon_k) \leq L\|s_k\|\|d_k\|/(\rho^{r+1}\hat{\epsilon}) = \alpha_k L d_k^T d_k/(\rho^{r+1}\hat{\epsilon}).$$

This implies

$$\alpha_k \geq \frac{\rho^{r+1}\hat{\epsilon}}{L} \qquad \text{for } k \in K_1.$$

Since K_1 is infinite and $\rho^{r+1}\hat{\epsilon}/L$ is a constant, we obtain

$$\sum_{k=0}^{\infty}\alpha_k \geq \sum_{k\in K_1}\alpha_k \geq \sum_{k\in K_1}\frac{\rho^{r+1}\hat{\epsilon}}{L} = \infty. \tag{3.7}$$

From (3.6) and (3.7), the sequence $\{\|d_k\|\}$ cannot be bounded away from 0. There exists a subset $K_2 \subset K_1$ such that $\lim_{k\in K_2} \|d_k\| = 0$.

Let $\bar{f}$ denote the optimal value of f. Extract from K_2 a further subset, say $K_3 \subset K_2$, such that $\{x_k, k \in K_3\}$ tends to some limit $\bar{x}$ and $\|B_k s_k\| \le \beta \|s_k\|$ for $k \in K_3$, where β is a constant (cf. Theorem 2.1 in [4]). By construction, the point $\bar{x}$ must necessarily belong to X. Let $\{d_k, k \in K_3\}$ be a corresponding sequence of directions. Since $\lim_{k\in K_3} d_k = 0$, $\bar{x}$ is an optimal solution of (1.1) and $f(\bar{x}) = \bar{f}$.

Since $\{f(x_k)\}$ is nonincreasing and has a limit f^*,

$$\begin{aligned} f^* - f(x_k) \le f(x_{k+1}) - f(x_k) &\le \frac{\sigma}{2}\alpha_k g_k^T d_k \\ &\le -\frac{\sigma}{2}\alpha_k d_k^T (B_k + \epsilon_k I) d_k \\ &\le -\frac{\sigma}{2}\hat{\epsilon}\alpha_k \|d_k\|^2. \end{aligned} \tag{3.8}$$

From (3.6), $\alpha_k \|d_k\|^2 \to 0$. Pass to the limit in (3.8), written for $k \in K_3$; we obtain $f^* = \bar{f}$. Then any accumulation point of $\{x_k\}$ is also optimal. □

Analysis in [4] covered a large class of line search strategies. We can generalize the results to LC^1 problems easily. In particular, the Armijo line search in the BFGS-SQP method has the following relationship with two other line search strategies.

LEMMA 3.1. *Under Assumption* 3.1, *there exist positive constants* η_1 *and* η_2 *such that the steplength* α_k *produced by the BFGS-SQP method will satisfy either*

$$f(x_k + \alpha_k d_k) - f(x_k) \le -\eta_1 \frac{(g_k^T d_k)^2}{\|d_k\|^2}$$

or

$$f(x_k + \alpha_k d_k) - f(x_k) \le \eta_2 g_k^T d_k. \qquad \square$$

We can prove Lemma 3.1 by the same technique in [4, Lem. 4.1], since g is Lipschitz continuous and construction of the BFGS-SQP method implies $g_k^T d_k < 0$ and

$$f(x_k + \rho^{t_k - 1} d_k) - f(x_k) > \frac{\sigma}{2}\rho^{t_k-1} g_k^T d_k \qquad \text{if } t_k > 0. \tag{3.9}$$

Lemma 3.1 will be applied to superlinear convergence analysis of the BFGS-SQP method.

4. Superlinear convergence. In this section we first prove superlinear convergence of the BFGS-SQP method under a Dennis–Moré–type condition [10]. Next we show that the condition is satisfied when f is twice strongly differentiable at a solution of (1.1).

Assumption 4.1. The objective function f is twice differentiable at a solution x^*, and $\nabla^2 f(x^*)$ is nonsingular.

Assumption 4.1 and the convexity of f imply that x^* is a unique minimizer of (1.1) [33]. Clearly, if a sequence $\{x_k\}$ generated by the BFGS-SQP method converges to x^*, then $\epsilon_k \to 0$ as $k \to \infty$.

Assumption 4.1 implies that the following limit holds:

$$\lim_{\|d\|\to 0} \frac{\|g(x^* + d) - g(x^*) - \nabla^2 f(x^*) d\|}{\|d\|} = 0. \tag{4.1}$$

THEOREM 4.1. *Under Assumptions* 3.1 *and* 4.1, *the BFGS-SQP method generates a sequence* $\{x_k\}$ *that converges to the unique minimum point* x^* *of* (1.1). *Moreover, if*

$$\lim_{k\to\infty} \frac{\|(\nabla^2 f(x^*) - B_k) d_k\|}{\|d_k\|} = 0 \tag{4.2}$$

and the sequence $\{\|B_k^{-1}\|\}$ *is bounded, then there exists an integer* k_0 *such that* $\alpha_k = 1$ *for all* $k \geq k_0$ *and the sequence* $\{x_k\}$ *converges to* x^* *at least Q-superlinearly; i.e.,*

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0.$$

Proof. From Theorem 3.1, the BFGS-SQP method generates a sequence $\{x_k\}$ that converges to x^*.

Now we show that $\{x_k\}$ converges to x^* at least Q-superlinearly. First we show

$$\|x_k + d_k - x^*\| = o(\|x_k - x^*\|). \tag{4.3}$$

Combining (4.1) and (4.2), we have

$$\begin{aligned} &g_k - g(x^*) - \nabla^2 f(x^*)(x_k - x^*) - (\nabla^2 f(x^*) - B_k)d_k \\ &= g_k - g(x^*) - B_k(x_k - x^*) - (\nabla^2 f(x^*) - B_k)(x_k + d_k - x^*) \\ &= o(\|x_k - x^*\|) + o(\|d_k\|). \end{aligned}$$

Hence we may write

$$\begin{aligned} g(x^*) = g_k + B_k(x^* - x_k) - (\nabla^2 f(x^*) - B_k)(x_k + d_k - x^*) \\ + o(\|x^* - x_k\|) + o(\|d_k\|). \end{aligned}$$

By the variational principle of (1.1), we have

$$\begin{aligned} 0 &\leq (x_k + d_k - x^*)^T g(x^*) \\ &= (x_k + d_k - x^*)^T (g_k + B_k(x^* - x_k) - (\nabla^2 f(x^*) - B_k)(x_k + d_k - x^*) \\ &\quad + o(\|x^* - x_k\|) + o(\|d_k\|)). \end{aligned} \tag{4.4}$$

Since d_k is the solution of (3.1), by the variational principle of (3.1) we have

$$0 \leq (x^* - x_k - d_k)^T (g_k + (B_k + \epsilon_k I)d_k),$$

which implies

$$\begin{aligned} &(x^* - x_k - d_k)^T (B_k + \epsilon_k I)(x^* - x_k - d_k) \\ &\leq (x^* - x_k - d_k)^T (g_k + (B_k + \epsilon_k I)(x^* - x_k)). \end{aligned} \tag{4.5}$$

Adding (4.4) to (4.5), we have

$$\begin{aligned} &(x^* - x_k - d_k)^T (B_k + \epsilon_k I)(x^* - x_k - d_k) \\ &\leq (x^* - x_k - d_k)^T (\epsilon_k (x^* - x_k) + o(\|x^* - x_k\|) + o(\|d_k\|)) \\ &\quad -(x_k + d_k - x^*)^T (\nabla^2 f(x^*) - B_k)(x_k + d_k - x^*). \end{aligned} \tag{4.6}$$

Since f is convex, the nonsingularity of $\nabla^2 f(x^*)$ implies that the matrix $\nabla^2 f(x^*)$ is symmetric positive definite [20]. Hence there exists a constant $\mu > 0$ such that

$$\mu\|z\|^2 \leq z^T \nabla^2 f(x^*) z \qquad \text{for all } z \in R^n.$$

Therefore from (4.6) and that $\epsilon_k \to 0$ as $k \to \infty$ we have

$$\begin{aligned} \mu\|x_k + d_k - x^*\|^2 &\leq (x^* - x_k - d_k)^T (\nabla^2 f(x^*) + \epsilon_k I)(x^* - x_k - d_k) \\ &\leq \|x_k + d_k - x^*\|(o(\|x_k - x^*\|) + o(\|d_k\|)). \end{aligned}$$

This implies

$$\|d_k\| = O(\|x_k - x^*\|) \tag{4.7}$$

and (4.3).

Now we show that there exists an integer $k_0 \geq 0$ such that $\alpha_k = 1$ for all $k \geq k_0$.

From (4.1), we have for all large k,

$$f(x_k) = f(x^*) + g(x^*)^T(x_k - x^*) + \frac{1}{2}(x_k - x^*)^T \nabla^2 f(x^*)(x_k - x^*) + o(\|x_k - x^*\|^2) \tag{4.8}$$

and

$$f(x_k + d_k) = f(x^*) + g(x^*)^T(x_k + d_k - x^*) + \frac{1}{2}(x_k + d_k - x^*)^T \nabla^2 f(x^*)(x_k + d_k - x^*)$$

$$+ o(\|x_k + d_k - x^*\|^2). \tag{4.9}$$

By subtracting (4.8) from (4.9), we obtain

$$f(x_k + d_k) - f(x_k) = \frac{1}{2} g_k^T d_k + (g(x^*) - g_k + \nabla^2 f(x^*)(x_k - x^*))^T d_k$$

$$+ \frac{1}{2} d_k^T (g_k + B_k d_k) + \frac{1}{2} d_k^T (\nabla^2 f(x^*) - B_k) d_k + o(\|x_k - x^*\|^2).$$

Using the relation $d_k^T(g_k + B_k d_k) \leq 0$ and (4.1), (4.2), and (4.7), we have

$$f(x_k + d_k) - f(x_k) - \frac{1}{2} g_k^T d_k \leq o(\|x_k - x^*\|^2). \tag{4.10}$$

Since the sequence $\{\|B_k^{-1}\|\}$ is bounded, there exists a constant $\tilde{c} > 0$ such that for all large k

$$-g_k^T d_k \geq d_k^T (B_k + \epsilon_k I) d_k \geq \tilde{c} \|d_k\|^2.$$

Thus from (4.7) there exists a constant $c > 0$ such that

$$c\|x_k - x^*\|^2 \leq -g_k^T d_k \qquad \text{for all large } k. \tag{4.11}$$

Combining (4.10) and (4.11), we have the existence of k_0 such that for all $k \geq k_0$ there is a positive scalar $\delta < \min\{\frac{1}{2}, \frac{c}{2}(1 - \sigma)\}$ such that

$$f(x_k + d_k) - f(x_k) - \frac{\sigma}{2} g_k^T d_k \leq \frac{1 - \sigma}{2} g_k^T d_k + \delta \|x_k - x^*\|^2$$

$$\leq \left(\frac{1 - \sigma}{2} - \frac{\delta}{c} \right) g_k^T d_k \leq 0.$$

Hence for all $k \geq k_0$ we have $\alpha_k = 1$ and $x_{k+1} = x_k + d_k$. From (4.3) we obtain that $\{x_k\}$ superlinearly converges to x^*. □

The condition (4.2) is a Dennis–Moré–type condition [10] which plays a key role in the superlinear convergence analysis. We give the following sufficient conditions for (4.2) by using Theorem 3.2 in [4].

LEMMA 4.1. *Suppose that f satisfies Assumption* 4.1 *and $y_k^T d_k \geq \rho^{r+1} \epsilon_k d_k^T d_k$ for all large k. Assume that $\{s_k\}$ and $\{y_k\}$ are such that*

$$\frac{\|y_k - \nabla^2 f(x^*) s_k\|}{\|s_k\|} \leq w_k$$

for some sequence $\{w_k\}$ *with the property* $\sum_{k=0}^{\infty} w_k < \infty$. *Then* (4.2) *holds and the sequences* $\{\|B_k\|\}$, $\{\|B_k^{-1}\|\}$ *are bounded.*

The assumption that $y_k^T d_k \geq \rho^{r+1}\epsilon_k d_k^T d_k$ for all large k implies that $s_k^T y_k > 0$ and B_k is updated for all large k in the BFGS-SQP method. See Theorem 3.2 in [4] for a proof of Lemma 4.1.

To prove $y_k^T d_k \geq \rho^{r+1}\epsilon_k d_k^T d_k$ for all large k, we require the strong convexity of f.

Assumption 4.2. f is strongly convex on D_0 with modulus $\frac{\mu}{2} > 0$; i.e., for any $\lambda \in (0, 1)$ there holds

$$\lambda f(x) + (1-\lambda) f(y) - f(\lambda x + (1-\lambda)y) \geq \frac{1}{2}\mu\lambda(1-\lambda)\|x-y\|^2$$

for all $x, y \in D_0$.

There are two equivalent properties to Assumption 4.2 (see Theorems 3.4.4 and 3.4.5 in [24]):

$$f(x) - f(y) \geq g(y)^T(x-y) + \frac{\mu}{2}\|x-y\|^2 \qquad \text{for all } x, y \in D_0 \tag{4.12}$$

and

$$(g(x) - g(y))^T(x-y) \geq \mu\|x-y\|^2 \qquad \text{for all } x, y \in D_0. \tag{4.13}$$

THEOREM 4.2. *Under Assumptions* 3.1, 4.1, *and* 4.2, *if*

$$L \leq \frac{\mu(1+\rho^r)}{2\rho^r}, \tag{4.14}$$

then the following statements hold:

(i) *there exists an integer* $k_0 \geq 0$ *such that for all* $k \geq k_0$ *the sequences* $\{y_k\}$, $\{d_k\}$, $\{\epsilon_k\}$ *generated by the BFGS-SQP method satisfy*

$$y_k^T d_k \geq \rho^{r+1}\epsilon_k d_k^T d_k;$$

(ii) *the sequence* $\{x_k\}$ *converges to* x^* *at least r-linearly; this implies*

$$\sum_{k=0}^{\infty} \|x_k - x^*\| < \infty.$$

Proof. (i) Since

$$y_k^T d_k = y_k^T s_k/\alpha_k \geq \mu\|s_k\|^2/\alpha_k = \mu\alpha_k\|d_k\|^2,$$

it suffices to show that there is a $k_0 \geq 0$ such that

$$\mu\alpha_k \geq \rho^{r+1}\epsilon_k \qquad \text{for } k \geq k_0. \tag{4.15}$$

From (4.12) for any $\alpha \in (0, 1]$ there holds

$$f(x_k) - f(x_k + \alpha d_k) \geq -\alpha g(x_k + \alpha d_k)^T d_k + \frac{\mu}{2}\alpha^2\|d_k\|^2. \tag{4.16}$$

From (4.14), we have

$$\alpha(g(x_k + \alpha d_k) - g_k)^T d_k \leq \alpha\|g(x_k + \alpha d_k) - g_k\|\|d_k\| \leq L\alpha^2\|d_k\|^2 \leq \frac{1+\rho^r}{2\rho^r}\mu\alpha^2\|d_k\|^2. \tag{4.17}$$

Combining (4.16) and (4.17), we have

$$\begin{aligned}
&f(x_k) - f(x_k + \alpha d_k) \geq -\alpha g(x_k + \alpha d_k)^T d_k + \frac{\mu}{2}\alpha^2 \|d_k\|^2 \\
&\geq -\alpha g_k^T d_k - \frac{1+\rho^r}{2\rho^r}\mu\alpha^2\|d_k\|^2 + \frac{\mu}{2}\alpha^2\|d_k\|^2 \\
&= -\frac{\sigma}{2}\alpha g_k^T d_k - \frac{2-\sigma}{2}\alpha g_k^T d_k - \frac{1}{2\rho^r}\mu\alpha^2\|d_k\|^2 \\
&\geq -\frac{\sigma}{2}\alpha g_k^T d_k + \frac{2-\sigma}{2} d_k^T (B_k + \epsilon_k I) d_k - \frac{1}{2\rho^r}\mu\alpha^2\|d_k\|^2 \\
&\geq -\frac{\sigma}{2}\alpha g_k^T d_k + \frac{1}{2}\left((2-\sigma)\epsilon_k - \frac{\mu}{\rho^r}\alpha\right)\alpha\|d_k\|^2 \\
&\geq -\frac{\sigma}{2}\alpha g_k^T d_k + \frac{1}{2}\left(\epsilon_k - \frac{\mu}{\rho^r}\alpha\right)\alpha\|d_k\|^2.
\end{aligned}$$

Hence

$$f(x_k + \alpha d_k) - f(x_k) \leq \frac{\sigma}{2}\alpha g_k^T d_k \qquad \text{if } \alpha \leq \min\left(1, \frac{\rho^r \epsilon_k}{\mu}\right).$$

Since $\epsilon_k \to 0$ as $k \to \infty$, there exists $k_0 \geq 0$ such that $\rho^r \epsilon_k / \mu < 1$ for $k \geq k_0$. Therefore, for $k \geq k_0$ there is an integer $l_k \geq 0$ such that

$$\rho^{l_k+1} \leq \rho^r \frac{\epsilon_k}{\mu} \leq \rho^{l_k}.$$

Consequently, we have

$$f(x_k + \rho^{l_k+1} d_k) - f(x_k) \leq \frac{\sigma}{2}\rho^{l_k+1} g_k^T d_k.$$

By construction of the BFGS-SQP method, we have

$$\alpha_k = \rho^{t_k} \geq \rho^{l_k+1} \geq \frac{\epsilon_k}{\mu}\rho^{r+1};$$

thus (4.15) holds.

(ii) The Lipschitz property of g ensures [29]

$$L y_k^T s_k \geq \|y_k\|^2 \qquad \text{for } k \geq k_0. \tag{4.18}$$

From (4.13), we have

$$y_k^T s_k \geq \mu \|s_k\|^2 \qquad \text{for } k \geq 0. \tag{4.19}$$

From Theorem 2.1 in [4] and Lemma 3.1, (4.18) and (4.19) imply that there is a constant $\beta > 0$ such that

$$f(x_j) - f(x_j + \alpha_j d_j) \geq \beta \|g_j\|^2 \tag{4.20}$$

holds for at least $[p(k - k_0 + 1)](p \in (0, 1))$ values of $j \in [k_0, k]$ where k_0 is defined in (i) of this theorem.

Now we show

$$\frac{1}{2}\mu\|x_k - x^*\|^2 \leq f(x_k) - f(x^*) \leq \frac{1}{\mu}\|g_k\|^2. \tag{4.21}$$

The lower bound follows from (4.12) and

$$(x_k - x^*)^T g(x^*) \geq 0. \tag{4.22}$$

The upper bound follows from

$$f(x_k) - f(x^*) \le g_k^T(x_k - x^*) \le \|g_k\|\|x_k - x^*\|$$

and

$$\|g_k\| \ge \frac{g_k^T(x_k - x^*)}{\|x_k - x^*\|} \ge \frac{(g_k - g(x^*))^T(x_k - x^*)}{\|x_k - x^*\|} \ge \mu\|x_k - x^*\|, \tag{4.23}$$

where (4.23) is derived by (4.13) and (4.22).

From (4.20) and (4.21), the r-linear convergence of $\{x_k\}$ can be obtained by a simple manipulation (see, e.g., [4, p. 733]). We omit the details. □

We are now ready to give the superlinear convergence result.

THEOREM 4.3. *Under the assumptions of Theorem* 4.2, *if there exist a positive constant* $\tilde{L}$ *and a neighborhood* $\mathcal{N}_*$ *of* x^* *such that for any* $x, y \in \mathcal{N}_*$,

$$\frac{\|g(x) - g(y) - \nabla^2 f(x^*)(x - y)\|}{\|x - y\|} \le \tilde{L}\max(\|x - x^*\|, \|y - x^*\|), \tag{4.24}$$

then the sequence $\{x_k\}$ *generated by the BFGS-SQP method converges to the unique solution* x^* *of* (1.1) *Q-superlinearly.*

From Theorems 4.1 and 4.2, there exists $k_0 > 0$ such that for $k \ge k_0$, $x_k, x_{k+1} \in \mathcal{N}_*$, and $y_k^T d_k \ge \rho^{r+1}\epsilon_k d_k^T d_k$. Hence for $k \ge k_0$,

$$\frac{\|y_k - \nabla^2 f(x^*)s_k\|}{\|s_k\|} \le \tilde{L}\max(\|x_k - x^*\|, \|x_{k+1} - x^*\|) =: w_k$$

where $\sum_{k=k_0}^{\infty} w_k < \infty$. Let $w_k = \|y_k - \nabla^2 f(x^*)s_k\|/\|s_k\|$ for $k < k_0$. Since $\|s_k\| > 0$ and the sequences $\|y_k\|$ and $\|s_k\|$ are bounded, $\sum_{k=0}^{k_0-1} w_k < \infty$. Hence from Theorem 4.1 and Lemma 4.1, we conclude the rate of convergence is superlinear.

5. Applications. Some source problems for LC^1 convex optimization have been discussed in [2], [7], [8], [13], [27], [28], [31], [32], [33], [37]. In this section we use the BFGS-SQP method to solve quadratic stochastic programming problems on a 16-node partition of a CM5 parallel computer using data parallel constructs expressed by Fortran 90 style matrix operations.

The quadratic stochastic programming model was introduced by Rockafellar and Wets [34], [35], [36]. A version of a 2-stage quadratic stochastic program with fixed recourse is

$$\begin{aligned} \min \quad & \frac{1}{2}x^T Px + c^T x + \sum_{i=0}^{\gamma} \psi(x, \omega_i)\xi_i \\ \text{s.t.} \quad & Ax \le b \end{aligned} \tag{5.1}$$

where

$$\begin{aligned} \psi(x, \omega) = \max \quad & -\frac{1}{2}z^T Hz + z^T(\omega - Tx) \\ \text{s.t.} \quad & Wz \le q. \end{aligned}$$

Here $P \in R^{n\times n}$ and $H \in R^{n_1\times n_1}$ are symmetric positive definite; $c \in R^n$, $T \in R^{n_1\times n}$, $W \in R^{m_1\times n_1}$, $q \in R^{m_1}$, $\omega_i \in R^{n_1}$, $i = 0, 1, \ldots, \gamma$, and $\xi_i \ge 0$, $i = 0, 1, \ldots, \gamma$ are scalars satisfying $\sum_{i=0}^{\gamma} \xi_i = 1$.

The objective function of (5.1) is strongly convex and has a Lipschitz continuous gradient at all points in X but it is not twice differentiable in X. Hence this problem is an LC^1 convex constrained optimization problem. The gradient of the objective function is

$$Px + c - T^T H^{-\frac{1}{2}} \sum_{i=0}^{\gamma} z_i^*(x)\xi_i,$$

TABLE 1

γ	$k(t_{k-2}, t_{k-1}, t_k)$	$\|x^* - x_k\|$	$\|F(v_k)\|$	CPU(sec)
2^5	6 (6, 6, 1)	3.6×10^{-11}	6.7×10^{-12}	236.3
5^5	5 (6, 6, 1)	2.1×10^{-11}	2.1×10^{-12}	246.0
8^5	4 (1, 4, 2)	1.1×10^{-10}	9.3×10^{-12}	1062.8

where

$$z_i^*(x) = \operatorname{argmax}\left\{-\frac{1}{2}z^T Hz + z^T(\omega_i - Tx),\ Wz \le q\right\} = \Pi_S(H^{-\frac{1}{2}}(\omega_i - Tx)).$$

Here $\Pi_S(u)$ is the projection of $u \in R^{n_1}$ into the set $S = \{s \in R^{n_1}, WH^{-\frac{1}{2}}s \le q\}$. Since the projection operator is nonexpansive, we have a Lipschitz constant $L = \|P\| + \|T^T H^{-\frac{1}{2}}\|\|H^{-\frac{1}{2}}T\|$ for Assumption 3.1. Furthermore, we can choose a positive integer r and positive scalars $\rho < 1$, $\mu \le \lambda_{\min}(P)$ (the smallest eigenvalue of P) such that condition (4.14) holds.

Calculating the objective and its gradient involves a large number of quadratic programs. We tested the BFGS-SQP method for solving (5.1) on a CM5 parallel computer. At each step, $f(x_k)$ and g_k are calculated in parallel. The test problems are randomly generated but with known solution characteristics so different features of the algorithm can be tested.

We chose $n = 100, m = 60, n_1 = 5, m_1 = 3, \kappa(P) = \kappa(H) = 10^2, \kappa(A) = \kappa(W) = 2.5^2, \mu_0 = 30, \nu_0 = 15$. Here $\kappa(G) = \tau^2$ is the condition number of a matrix G whose nonzero eigenvalues are distributed on $[1/\tau, \tau]$, μ_0 is the number of active constraints at the solution x^* with positive multipliers, and ν_0 is the number of active constraints at the solution x^* with zero multipliers. We generate $\omega_i = (\omega_{i,1}, \omega_{i,2}, \omega_{i,3}, \omega_{i,4}, \omega_{i,5}) \in [0, 1]^5$, $i = 1, 2, \ldots, \gamma$ with $\omega_{i,j} \in \{0, \frac{1}{l-1}, \ldots, 1\}$, $\gamma = l^5$. In the BFGS-SQP method, we chose $B_0 = P, \tau = \sigma = \rho = \eta = \epsilon_0 = 0.75$, and $r = 20$. We choose the starting point $x_0 = \Pi_X(\tilde{x}), u_0 = 0$, where $\tilde{x} \in R^n$ was randomly generated. The algorithm terminated when $v_k = (x_k^T, u_k^T)^T$ satisfied $\|F(v_k)\| \le 10^{-11}$. Numerical results are reported in Table 1, where t_k is the number of iterations on the Armijo line search at the kth iteration.

Acknowledgments. The author is grateful to Liqun Qi, Robert S. Womersley, and Tetsuro Yamamoto for their valuable comments.

REFERENCES

[1] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[2] J. BONNANS, J. GILBERT, C. LEMARÉCHAL, AND C. SAGASTIZÁBAL, *A family of variable metric proximal methods*, Math. Programming, 68 (1995), pp. 15–47.

[3] C. G. BROYDEN, J. E. DENNIS, AND J. J. MORÉ, *On the local and superlinear convergence of quasi-Newton methods*, J. Inst. Math. Appl., 12 (1973), pp. 223–246.

[4] R. H. BYRD AND J. NOCEDAL, *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM J. Numer. Anal., 26 (1989), pp. 727–739.

[5] X. CHEN, *On the convergence of Broyden-like methods for nonlinear equations with nondifferentiable terms*, Ann. Inst. Statist. Math., 42 (1990), pp. 387–401.

[6] X. CHEN AND L. QI, *A parameterized Newton method and a quasi-Newton method for nonsmooth equations*, J. Comp. Optim. Appl., 3 (1994), pp. 157–179.

[7] X. CHEN, L. QI, AND R. S. WOMERSLEY, *Newton's method for quadratic stochastic programs with recourse*, J. Comput. Appl. Math., 60 (1995), pp. 29–46.

[8] X. CHEN AND R. S. WOMERSLEY, *A parallel inexact Newton method for stochastic programs with recourse*, Ann. Oper. Res., (1996), to appear.

[9] X. CHEN AND T. YAMAMOTO, *On the convergence of some quasi-Newton methods for nonlinear equations with nondifferentiable operators*, Computing, 48 (1992), pp. 87–94.

[10] J. E. DENNIS AND J. J. MORÉ, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549–560.

[11] J. E. DENNIS, JR., AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1983.

[12] J. E. DENNIS AND H. F. WALKER, *Convergence theorems for least-change secant update methods*, SIAM J. Numer. Anal., 18 (1981), pp. 949–987.

[13] F. FACCHINEI, *Minimization of SC^1 functions and the Maratos effect*, Oper. Res. Lett., 17 (1995), pp. 131–137.

[14] R. FLETCHER, *Practical Methods of Optimization*, 2nd ed., John Wiley & Sons Ltd., Chichester, New York, 1987.

[15] J. C. GILBERT AND C. LEMARÉCHAL, *Some numerical experiments with variable-storage quasi-Newton algorithms*, Math. Programming, 45 (1989), pp. 407–435.

[16] M. A. GOMES-RUGGIERO, J. M. MARTÍNEZ, AND S. A. SANTOS, *Solving nonsmooth equations by means of quasi-Newton methods with globalization*, in Recent Advances in Nonsmooth Optimization, D. Du, L. Qi, and R. S. Womersley, eds., World Scientific, River Edge, NJ, 1995, pp. 121–140.

[17] M. HEINKENSCHLOSS, C. T. KELLEY, AND H. T. TRAN, *Fast algorithms for nonsmooth compact fixed point problems*, SIAM J. Numer. Anal., 29 (1992), pp. 1769–1792.

[18] J. B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, New York, 1993.

[19] C. M. IP AND J. KYPARISIS, *Local convergence of quasi-Newton methods for B-differentiable equations*, Math. Programming, 56 (1992), pp. 71–89.

[20] H. JIANG AND L. QI, *Local uniqueness and convergence of iterative methods for nonsmooth variational inequalities*, J. Math. Anal. Appl., 196 (1995), pp. 314–331.

[21] M. KOJIMA AND S. SHINDO, *Extensions of Newton and quasi-Newton methods to systems of PC^1 equations*, J. Oper. Res. Soc. Japan, 29 (1986), pp. 352–374.

[22] J. M. MARTÍNEZ, *On the relation between two local convergence theories of least change secant update methods*, Math. Comp., 59 (1992), pp. 457–481.

[23] J. M. MARTÍNEZ AND M. C. ZAMBALDI, *Least change update methods for nonlinear systems with nondifferentiable terms*, Numer. Funct. Anal. Optim., 14 (1993), pp. 405–415.

[24] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[25] J. S. PANG, *Newton methods for B-differentiable equations*, Math. Oper. Res., 15 (1990), pp. 311–341.

[26] J. S. PANG, S. P. HAN, AND N. RANGARAJ, *Minimization of locally Lipschitzian functions*, SIAM J. Optim., 1 (1991), pp. 57–82.

[27] J. S. PANG AND L. QI, *Nonsmooth equations: motivation and algorithms*, SIAM J. Optim., 3 (1993), pp. 443–465.

[28] ———, *A globally convergent Newton method for convex SC^1 minimization problems*, J. Optim. Theory Appl., 85 (1995), pp. 633–648.

[29] M. J. D. POWELL, *Some global convergence properties of a variable metric algorithm for minimization without exact line searches*, in Nonlinear Programming, SIAM-AMS Proceedings, Vol. IX, R. W. Cottle and C. E. Lemke, eds., American Mathematical Society, Providence, RI, 1976.

[30] L. QI, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.

[31] ———, *Superlinear convergent approximate Newton methods for LC^1 optimization problems*, Math. Programming, 64 (1994), pp. 277–294.

[32] L. QI AND J. SUN, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.

[33] L. QI AND R. WOMERSLEY, *An SQP algorithm for extended linear-quadratic problems in stochastic programming*, Ann. Oper. Res., 56 (1995), pp. 251–285.

[34] R. T. ROCKAFELLAR AND R. J.-B. WETS, *A dual solution procedure for quadratic stochastic programs with simple recourse*, in Numerical Methods, Lecture Notes in Mathematics 1005, V. Pereyra and A. Reinoza, eds., Springer-Verlag, Berlin, 1983, pp. 252–265.

[35] ———, *Linear-quadratic problems with stochastic penalties: The finite generation algorithm*, in Stochastic Optimization, Lecture Notes in Control and Information Sciences 81, V. I. Arkin, A. Shiraev, and R. J.-B. Wets, eds., Springer-Verlag, Berlin, 1987, pp. 545–560.

[36] ———, *Generalized linear-quadratic problems of deterministic and stochastic optimal control in discrete time*, SIAM J. Control Optim., 28 (1990), pp. 810–822.

[37] C. ZHU AND R. T. ROCKAFELLAR, *Primal-dual projected gradient algorithms for extended linear-quadratic programming*, SIAM J. Optim., 3 (1993), pp. 751–783.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 2064–2076, November 1996

011

EXISTENCE RESULTS FOR NONCOERCIVE VARIATIONAL PROBLEMS*

GRAZIANO CRASTA† AND ANNALISA MALUSA‡

Abstract. The aim of this paper is to give an existence result for a class of one-dimensional, nonconvex, noncoercive problems in the calculus of variations. The main tools for the proof are an existence theorem in the convex case and the closure of the convex hull of the epigraph of functions strictly convex at infinity.

Key words. existence theory, nonconvex problems, noncoercive problems

AMS subject classification. 49J05

1. Introduction. It is well known that if L is a continuous function, such that $\xi \mapsto L(t, x, \xi)$ is convex and superlinear, then the variational problem

$$\min\left\{\int_0^T L(t, u, u')dt \;\Big|\; u \in W^{1,1}([0, T], \mathbb{R}^m),\ u(0) = a,\ u(T) = b\right\} \tag{1.1}$$

has a solution (see, for instance, [7]).

In recent years, the possibility of avoiding the convexity or the superlinearity assumption was investigated by many authors.

Some existence results for nonconvex coercive problems were obtained in the case $L(t, x, \xi) = g(t, x) + f(t, \xi)$ (see, for instance, [5], [14], [16], and the references therein). In particular, in [5] it was proved that the convexity assumption on $f(t, \cdot)$ can be replaced by the condition of concavity of $g(t, \cdot)$.

More recently, some techniques were developed in order to treat convex but noncoercive problems. In this case, even if the functionals considered are lower semicontinuous in the weak topology of $W^{1,1}([0, T], \mathbb{R}^m)$, the direct method of the calculus of variations cannot be applied due to the lack of compactness of the minimizing sequences.

In [10] problem (1.1) was studied with L continuous, bounded from below and convex with respect to ξ, the superlinearity being replaced by a weaker condition which permits construction of a relatively compact minimizing sequence, obtained by considering the minima of suitable coercive approximating problems. The main step in the proof of the existence result in [10] was to show that every minimum point of the approximating problems solves a generalized DuBois–Reymond condition, which implies that the minimizing sequence is bounded in the space $W^{1,\infty}([0, T], \mathbb{R}^m)$.

A similar approach was used in [6] for the autonomous problem $L(t, x, \xi) = g(x) + f(\xi)$, where g is a nonnegative continuous function and $f \in C^1(\mathbb{R}^m, \mathbb{R})$ is a strictly convex function bounded from below, such that

$$\lim_{|\xi| \to +\infty} [f(\xi) - \langle \nabla f(\xi),\ \xi \rangle] = -\infty. \tag{1.2}$$

In that paper, it was proved that for every rectifiable curve C in $\mathbb{R}^m$ joining a to b there exists a unique solution to the problem (1.1) restricted to the class of all absolutely continuous parameterizations $u: I \to \mathbb{R}^m$ of C. Thus, every element u_n of a minimizing sequence can be replaced by the minimum corresponding to the curve parameterized by u_n. It can be shown, still using a DuBois–Reymond condition satisfied by those minima and by (1.2), that this new

*Received by the editors December 5, 1994; accepted for publication September 19, 1995.

†Dipartimento di Matematica Pura ed Applicata, Via Campi 213/B, 41100 Modena, Italy, (crasta@c220.unimo.it).

‡Istituto di Matematica, Facoltà di Architettura, Via Monteoliveto 3, 80134 Napoli, Italy (malusa@ds.cised.unina.it).

sequence is bounded in $W^{1,\infty}([0, T], \mathbb{R}^m)$, so that there exists a minimum point for (1.1) in this space.

In [12] both the superlinearity and the convexity assumptions were dropped for Lagrangians of the form $L(t, x, \xi) = \langle a(t), x\rangle + f(\xi)$ where f is a lower semicontinuous function whose convexification f^{**} satisfies (1.2) for every diverging sequence of points of differentiability of the Lipschitz continuous function f^{**}. The existence of a minimum is proved by a technique relying only on a Lyapunov-type theorem due to Olech (see [15]).

For other results concerning noncoercive problems we mention [1], [2], and [3].

In this paper we consider nonautonomous problems of the form

$$\min\left\{\int_0^T \left[g(t, u) + f(t, u')\right] dt \;\middle|\; u \in W^{1,1}([0, T], \mathbb{R}^m),\ u(0) = a,\ u(T) = b\right\} \tag{1.3}$$

with neither coercivity nor convexity assumptions. More precisely, we introduce the class $\mathcal{E}$ of all functions $\psi : [0, T] \times \mathbb{R}^m \to \mathbb{R}$, bounded from below, such that $\psi(\cdot, \xi)$ is Lipschitz continuous for every fixed $\xi \in \mathbb{R}^m$ and $\psi(t, \cdot)$ is lower semicontinuous and satisfies

$$\lim_{n\to+\infty} \left[\psi^{**}(t^n, \xi^n) - \langle \nabla\psi^{**}(t^n, \xi^n), \xi^n\rangle\right] = -\infty$$

for every sequence $\{t^n\} \in [0, T]$ and for every choice of points ξ^n of differentiability of $\psi^{**}(t^n, \cdot)$ such that $\lim_n |\xi^n| = +\infty$. We show that if $f \in \mathcal{E}$ and there exist two constants A and B, $B > 0$ such that $f(t, \xi) \geq -A + B|\xi|$ for every $(t, \xi) \in [0, T] \times \mathbb{R}^m$ and $g(t, x)$ is a continuous function, Lipschitz continuous with respect to t, concave with respect to x, satisfying $g(t, x) \geq -\alpha - \beta|x|$ for every $(t, x) \in [0, T] \times \mathbb{R}^m$ and for suitable constants α and $0 \leq \beta \leq B/T$, then the problem (1.3) has a solution in the space $W^{1,\infty}([0, T], \mathbb{R}^m)$. This result is the analogue for a class of noncoercive functionals of the one in [5], but it is not a generalization of that result due to the additional requirement of the Lipschitz continuity of the Lagrangian with respect to the variable t. However this extra regularity allows us to obtain the necessary conditions that, used at an intermediate step, also give a regularity result, which is interesting.

As a first step we prove an existence result for (1.3) requiring that f be convex with respect to ξ and dropping the concavity assumption on g. This can be done following [10] and making suitable changes due to the the fact that the Lagrangian is not bounded from below. The second step, linking the convex to the nonconvex case, is based on a result concerning the closure of the convex hull of the epigraph of functions whose convexification is strictly convex at infinity (that is, the graph of the convexification contains no rays). This result is an extension of the classical theorem that holds for superlinear functions (see [13]). We want to remark that the notion of strict convexity at infinity was still used in [11] in order to study noncoercive problems of the type (1.1) with the additional state constraint $\|u\|_{L^\infty} < R$. We shall prove that every function in the class $\mathcal{E}$ is strictly convex at infinity for every fixed t, so that by using the previous results and the Lyapunov theorem on the range of nonatomic measures the existence result for the nonconvex problems follows. The regularity of the solution of (1.3) is a consequence of the regularity of the solution to the relaxed problem.

2. Preliminaries. We shall denote by $\langle x, y\rangle$ the standard scalar product of two vectors $x, y \in \mathbb{R}^m$. For every $1 \leq p \leq +\infty$ we shall denote by $L^p(I, \mathbb{R}^m)$ and $W^{1,p}(I, \mathbb{R}^m)$, respectively, the usual Lebesgue and Sobolev spaces of functions from the interval $I \doteq [0, T]$ into $\mathbb{R}^m$. We shall use the symbol $\|\cdot\|_{L^p}$ to denote the norm in $L^p(I, \mathbb{R}^m)$.

If $A \subset \mathbb{R}^m$, we shall denote by $\operatorname{int} A$ the interior of A and by $\operatorname{co} A$ the convex hull of a A, that is, the smallest convex set which contains A. It is well known that, by Carathéodory's

theorem, the convex hull of A can be characterized by

$$\text{(2.1)}\qquad \operatorname{co} A = \left\{ x \in \mathbb{R}^m \;\middle|\; x = \sum_{i=1}^{m+1} \lambda_i x_i\,,\ \tilde{\lambda} \in E_{m+1}\,,\ x_i \in A,\ i = 1, \dots, m+1 \right\},$$

where $\tilde{\lambda} \doteq (\lambda_1, \dots, \lambda_{m+1})$ and E_{m+1} denotes the standard simplex

$$E_{m+1} \doteq \left\{ (\lambda_1, \dots, \lambda_{m+1}) \in \mathbb{R}^{m+1} \;\middle|\; \lambda_i \geq 0\ \forall i = 1, \dots, m+1,\ \sum_{i=1}^{m+1} \lambda_i = 1 \right\}.$$

Given a function $\psi : \mathbb{R}^m \to \overline{\mathbb{R}}$, we shall denote by $\operatorname{dom}(\psi)$ its effective domain, defined as the subset of $\mathbb{R}^m$ $\{\xi \mid \psi(\xi) < +\infty\}$, and by $\operatorname{epi} \psi$ its epigraph, that is, the set

$$\operatorname{epi} \psi \doteq \{(x, a) \in \mathbb{R}^m \times \mathbb{R} \mid \psi(x) \leq a\}.$$

If $\psi : \mathbb{R}^m \to \overline{\mathbb{R}}$ is Lipschitz continuous in a neighborhood of a point ξ, we shall denote by $\partial\psi(\xi)$ the generalized gradient of ψ at ξ, defined by

$$\text{(2.2)}\qquad \partial\psi(\xi) \doteq \operatorname{co} \left\{ \lim_{i \to +\infty} \nabla\psi(\xi_i) \;\middle|\; \xi_i \to \xi,\ \xi_i \in \mathcal{D}(\psi) \right\},$$

where $\mathcal{D}(\psi)$ denotes the set of points of differentiability of ψ. We recall that a Lipschitz continuous function ψ is almost everywhere differentiable in $\operatorname{int}(\operatorname{dom}(\psi))$.

A function $\psi : \mathbb{R}^m \to (-\infty, +\infty]$ is convex if for every $\xi, \eta \in \mathbb{R}^m$ and for every $\lambda \in [0, 1]$, we have $\psi(\lambda\xi + (1-\lambda)\eta) \leq \lambda\psi(\xi) + (1-\lambda)\psi(\eta)$. We say that ψ is concave if $-\psi$ is convex.

Given a function $\psi : \mathbb{R}^m \to (-\infty, +\infty]$, we shall denote by ψ^* its dual function, defined for every $p \in \mathbb{R}^m$ by

$$\psi^*(p) \doteq \sup_{\xi \in \mathbb{R}^m} \{\langle p\,,\, \xi\rangle - \psi(\xi)\}.$$

It is well known that the bipolar function ψ^{**} coincides with the convexification of ψ, which is the largest convex function φ satisfying $\varphi \leq \psi$.

If $\psi : \mathbb{R}^m \to (-\infty, +\infty]$ is convex, then the generalized gradient of ψ coincides in $\operatorname{int}(\operatorname{dom}(\psi))$ with the subgradient of ψ in the sense of convex analysis, defined at every point $\xi \in \operatorname{dom}(\psi)$ by

$$\text{(2.3)}\qquad \partial\psi(\xi) \doteq \left\{ p \in \mathbb{R}^m \;\middle|\; \psi(\eta) \geq \psi(\xi) + \langle p\,,\, \eta - \xi\rangle \text{ for every } \eta \in \mathbb{R}^m \right\}$$

(see [8, Prop. 2.2.7]). By definition, we set $\partial\psi(\xi) \doteq \emptyset$ for every $\xi \notin \operatorname{dom}(\psi)$. We recall that, if ψ is differentiable at ξ, then $\partial\psi(\xi) = \{\nabla\psi(\xi)\}$.

In the following proposition we collect some well-known properties of the subgradient (see [8] and [13]).

PROPOSITION 2.1. *Let $\psi : \mathbb{R}^m \to (-\infty, +\infty]$ be a convex function. Then the following properties hold:*

(i) *if ψ is bounded from above in a nonempty open set A, then ψ is locally Lipschitz continuous in A;*

(ii) *for every $\xi \in \mathbb{R}^m$ the set $\partial\psi(\xi)$ (possibly empty) is convex and closed in $\mathbb{R}^m$;*

(iii) *if $\xi \in \operatorname{int}(\operatorname{dom}(\psi))$, then $\partial\psi(\xi)$ is a nonempty compact set.*

3. The closure result. In this section we shall prove a result concerning the closure of the convex hull of the epigraph of functions possibly without superlinear growth.

We recall the notion of strict convexity at infinity introduced by Clarke and Loewen in [11].

DEFINITION 3.1. *A convex function $\psi:\mathbb{R}^m \to \mathbb{R}$ is said to be strictly convex at infinity if its graph contains no rays; that is, for every $\nu \in \mathbb{R}^m$, $\nu \neq 0$, and for every $\xi \in \mathbb{R}^m$, the function $\psi_{\nu,\xi}(s) \doteq \psi(s\nu+\xi)$ has the following property: for every $s_0 \in \mathcal{D}(\psi_{\nu,\xi})$ there exists $s_1 \in \mathcal{D}(\psi_{\nu,\xi})$, $s_1 > s_0$, such that $\psi'_{\nu,\xi}(s_1) > \psi'_{\nu,\xi}(s_0)$.*

Remark 3.2. It is easy to see that if $\psi:\mathbb{R}^m \to \mathbb{R}$ is convex then ψ is strictly convex at infinity if and only if $\partial\psi^*(p)$ is either empty or bounded for every $p \in \mathbb{R}^m$.

DEFINITION 3.3. *We shall denote by $\mathcal{G}$ the family of all lower semicontinuous functions $\psi : \mathbb{R}^m \to \mathbb{R}$ such that $\psi^{**} \not\equiv -\infty$ and ψ^{**} is strictly convex at infinity.*

Remark 3.4. Clearly every strictly convex function is strictly convex at infinity. Moreover, every lower semicontinuous superlinear function $\psi:\mathbb{R}^m \to \mathbb{R}$ belongs to $\mathcal{G}$. Indeed, denoting by φ the convexification ψ^{**}, for every fixed $\nu, \xi \in \mathbb{R}^m$, $\nu \neq 0$, by (2.3) it follows that the inequality $\langle \nabla\varphi(s\nu+\xi)\,,\, s\nu\rangle \geq \varphi(s\nu+\xi) - \varphi(\xi)$ holds for every $s \in \mathcal{D}(\varphi_{\nu,\xi})$. This implies that

$$\varphi'_{\nu,\xi}(s) = \langle \nabla\varphi(s\nu+\xi)\,,\, \nu\rangle \geq \frac{\varphi(s\nu+\xi)-\varphi(\xi)}{s} \quad \text{for every } s \in \mathcal{D}(\varphi_{\nu,\xi}),\ s > 0.$$

Since ψ is superlinear, the last term tends to $+\infty$ as s goes to $+\infty$.

LEMMA 3.5. *For every function $\psi \in \mathcal{G}$ satisfying $\psi \geq 0$ and $\psi(0) = 0$ there exist two positive constants C, ρ such that $\psi(\xi) \geq C|\xi|$ for every $|\xi| > \rho$.*

Proof. We can certainly assume that ψ is convex; if not, we replace ψ by ψ^{**}. We start by proving that ψ is coercive; that is, $\psi(\xi) \to +\infty$ as $|\xi| \to +\infty$. Since ψ is convex, the sets $\psi^a \doteq \{\xi \in \mathbb{R}^m \mid \psi(\xi) < a\}$ are convex subsets of $\mathbb{R}^m$ for every $a \geq 0$. By contradiction, suppose that there exists $a > 0$ such that ψ^a is unbounded. Since ψ^a is convex, it contains at least one half-line $\{s\nu \mid s \geq 0\}$ for some $\nu \in \mathbb{R}^m$, $\nu \neq 0$. This means that $\psi_{\nu,0}(s) < a$ for every $s \geq 0$. Since $\psi_{\nu,0}$ is an absolutely continuous function, then for every $\tau > 0$ we have

$$0 \leq \psi_{\nu,0}(\tau) - \psi_{\nu,0}(0) = \int_0^\tau \psi'_{\nu,0}(\sigma)\, d\sigma.$$

Hence there exists $s_0 \in \mathcal{D}(\psi_{\nu,0}) \cap [0,\tau]$ such that $\psi'_{\nu,0}(s_0) \geq 0$. Since ψ is strictly convex at infinity, there exists $s_1 \in \mathcal{D}(\psi_{\nu,0})$, $s_1 > s_0$ such that $\psi'_{\nu,0}(s_1) > 0$. By the convexity of $\psi_{\nu,0}$ it follows that

$$\psi_{\nu,0}(s) \geq \psi_{\nu,0}(s_1) + (s-s_1)\psi'_{\nu,0}(s_1) \quad \text{for every } s \geq 0,$$

and this implies that $\lim_{s\to+\infty} \psi_{\nu,0}(s) = +\infty$ in contradiction with $\psi_{\nu,0} < a$.

Since ψ is coercive, there exist two positive constants ρ, δ such that

$$\psi(\eta) \geq \delta \quad \text{for all } |\eta| = \rho.$$

If $|\xi| > \rho$, let us define $\lambda \doteq \rho/|\xi|$ and $\eta \doteq \lambda\xi$. By the convexity of ψ, and recalling that $\psi(0) = 0$, we get

$$\psi(\xi) \geq \frac{1}{\lambda}\psi(\eta) = \frac{\psi(\eta)}{\rho}|\xi| \geq \frac{\delta}{\rho}|\xi|,$$

so that we conclude by choosing $C = \delta/\rho$. □

We are now in a position to prove the closure result. The proof is based on the fact that if f belongs to the class $\mathcal{G}$ then for every support hyperplane r of f^{**} the function $f - r$ belongs to $\mathcal{G}$. Applying the estimate of Lemma 3.5 to this function, we can follow the lines of the proof of Lemma IX.3.3 in [13].

THEOREM 3.6. *For every $f \in \mathcal{G}$ the set* co epi f *is closed.*

Proof. Let $(\xi, a) \in \partial(\text{co epi } f)$, where ∂S denotes the boundary of the set S, and let $r(\eta) \doteq \langle c\,,\ \eta\rangle + d$ be an affine function such that the hyperplane $H \doteq \{(\eta, r(\eta))\}$ weakly separates co epi f and the point (ξ, a). Let us define the function

$$\phi(\eta) \doteq f(\eta + \xi) - r(\eta + \xi).$$

We have $\phi^{**}(\eta) = f^{**}(\eta + \xi) - r(\eta + \xi)$, $\phi^{**} \geq 0$, $\phi^{**}(0) = 0$. Moreover, for every $\nu \in \mathbb{R}^m$, $\nu \neq 0$, for every $\eta \in \mathbb{R}^m$, and for every $s \in \mathcal{D}(f^{**}_{\nu,\xi+\eta})$ we have $(\phi^{**}_{\nu,\eta})'(s) = (f^{**}_{\nu,\xi+\eta})'(s) - \langle c\,,\ \nu\rangle$. Since f^{**} is strictly convex at infinity, then so is ϕ^{**}. By Lemma 3.5, there exist two positive constants C, ρ such that

$$\phi^{**}(\eta) \geq C|\eta| \quad \text{for every } |\eta| \geq \rho. \tag{3.1}$$

Notice that $(\xi, a) \in \text{co epi } f$ if and only if $(0, 0) \in \text{co epi } \phi$. Moreover, $(\xi, a) \in \partial(\text{co epi } f)$ if and only if $(0, 0) \in \partial(\text{co epi } \phi)$. Hence, to prove the proposition, it suffices to show that $(0, 0) \in \text{co epi } \phi$.

Let $(\xi^n, a^n) \in \text{co epi } \phi$ be such that $\lim_n (\xi^n, a^n) = (0, 0)$. By the characterization (2.1) of the convex hull, for every n there exist $\tilde{\lambda}^n \in E_{m+2}$ and $(\xi^n_j, a^n_j) \in \text{epi } \phi$, $j = 1, \ldots, m+2$, such that

$$\sum_{j=1}^{m+2} \lambda^n_j (\xi^n_j, a^n_j) = (\xi^n, a^n).$$

By the very definition of epigraph it follows that

$$a^n = \sum_{j=1}^{m+2} \lambda^n_j a^n_j \geq \sum_{j=1}^{m+2} \lambda^n_j \phi(\xi^n_j). \tag{3.2}$$

Moreover, (3.2) and the fact that $\phi \geq \phi^{**}$ imply that $a^n \geq \sum_{j=1}^{m+2} \lambda^n_j \phi^{**}(\xi^n_j)$. Since $\phi^{**} \geq 0$, the inequality

$$a^n \geq \lambda^n_j \phi^{**}(\xi^n_j) \tag{3.3}$$

holds for every $j = 1, \ldots, m+2$. Let $J \subset \{1, \ldots, m+2\}$ be the set of all j such that $\{|\xi^n_j|\}_n$ is unbounded, and let $I \doteq \{1, \ldots, m+2\} \backslash J$. By passing to a subsequence, we can assume that there exist $\bar{\xi}_j$, $j \in I$, and $\tilde{\lambda} \in E_{m+2}$, such that

$$\begin{aligned}
&\lim_{n\to+\infty} |\xi^n_j| = +\infty, \quad j \in J,\\
&\lim_{n\to+\infty} \xi^n_j = \bar{\xi}_j, \quad j \in I,\\
&\lim_{n\to+\infty} \lambda^n_j = \lambda_j, \quad j \in \{1, \ldots, m+2\}.
\end{aligned}$$

For every $j \in J$ we have $|\xi^n_j| > \rho$ for n large enough, and then from (3.1) and (3.3) it follows that $a^n \geq C\lambda^n_j |\xi^n_j|$. Since $\lim_n a^n = 0$, we get

$$\lim_{n\to+\infty} \lambda^n_j |\xi^n_j| = 0, \quad j \in J. \tag{3.4}$$

From (3.4) and recalling that $\lim_n \xi^n = 0$, we deduce that

$$\begin{aligned}\sum_{j\in I}\lambda_j\bar{\xi}_j &= \lim_{n\to+\infty}\sum_{j\in I}\lambda_j^n\xi_j^n \\ &= \lim_{n\to+\infty}\left(\sum_{j=1}^{m+2}\lambda_j^n\xi_j^n - \sum_{j\in J}\lambda_j^n\xi_j^n\right) = \lim_{n\to+\infty}\left(\xi^n - \sum_{j\in J}\lambda_j^n\xi_j^n\right) = 0.\end{aligned} \tag{3.5}$$

Moreover, since $\lim_n \lambda_j^n = 0$ for every $j \in J$, we obtain

$$\sum_{j\in I}\lambda_j = \lim_{n\to+\infty}\sum_{j\in I}\lambda_j^n = 1. \tag{3.6}$$

Since ϕ is a nonnegative lower semicontinuous function, we get

$$0 \le \sum_{j\in I}\lambda_j\phi(\bar{\xi}_j) \le \liminf_{n\to+\infty}\sum_{j\in I}\lambda_j^n\phi(\xi_j^n) \le \liminf_{n\to+\infty} a^n = 0. \tag{3.7}$$

There is no loss of generality in assuming that $\lambda_j > 0$ for every $j \in I$; hence (3.7) implies that $\phi(\bar{\xi}_j) = 0$ for every $j \in I$; that is, $(\bar{\xi}_j, 0) \in \operatorname{epi}\phi$ for every $j \in I$. Thus by (3.5) and (3.6) we can conclude that $(0, 0)$ belongs to $\operatorname{co}\operatorname{epi}\phi$. □

Now we state two direct consequences of Theorem 3.6.

COROLLARY 3.7. *If $f \in \mathcal{G}$, then*

$$f^{**}(\xi) = \min\left\{\sum_{j=1}^{m+1}\lambda_j f(\xi_j) \;\Big|\; \sum_{j=1}^{m+1}\lambda_j\xi_j = \xi,\ \tilde{\lambda} \in E_{m+1}\right\}$$

for every $\xi \in \mathbb{R}^m$.

Proof. See [13, Lem. IX.3.3]. □

We recall that a function $f: I \times \mathbb{R}^m \to \mathbb{R}$ is said to be a normal integrand (see [13]) if $f(t, \cdot)$ is lower semicontinuous for almost every (a.e.) $t \in I$ and there exists a Borel function $\tilde{f}: I \times \mathbb{R}^m \to \mathbb{R}$ such that $\tilde{f}(t, \cdot) = f(t, \cdot)$ for a.e. $t \in I$.

COROLLARY 3.8. *Let $f: I \times \mathbb{R}^m \to \mathbb{R}$ be a normal integrand and suppose that $f(t, \cdot) \in \mathcal{G}$ for every $t \in I$. Then for any measurable mapping $p: [0, T] \to \mathbb{R}^m$, there exist a measurable mapping $\tilde{\lambda}: [0, T] \to E_{m+1}$ and $m + 1$ measurable mappings $q_j: [0, T] \to \mathbb{R}^m$ such that*

$$\sum_{j=1}^{m+1}\lambda_j(t)q_j(t) = p(t), \quad \sum_{j=1}^{m+1}\lambda_j(t)f(t, q_j(t)) = f^{**}(t, p(t))$$

for almost all $t \in [0, T]$.

Proof. See [13, Prop. IX.3.1]. □

4. Existence results for variational problems. In this section we shall show that the existence result proved by Cellina and Colombo in [5] holds even for functions of the class $\mathcal{E}$ defined below. In the following, the convexification and the gradient of a function $\psi(t, \xi)$ are understood with respect to ξ.

DEFINITION 4.1. *We shall denote by $\mathcal{E}$ the family of all functions $\psi: I \times \mathbb{R}^m \to \mathbb{R}$, bounded from below, such that $\psi(\cdot, \xi)$ is Lipschitz continuous for every fixed $\xi \in \mathbb{R}^m$, $\psi(t, \cdot)$ is lower semicontinuous for every fixed $t \in I$, and*

$$\lim_{R\to+\infty}\sup_{\substack{t\in I\\ |\xi|>R}}\sup\left\{\psi^{**}(t, \xi) - \langle p\,,\,\xi\rangle \;\middle|\; p \in \partial_\xi\psi^{**}(t, \xi)\right\} = -\infty. \tag{4.1}$$

The following proposition gives a characterization of the family $\mathcal{E}$. The proof is similar to the one of Proposition 3.2 in [12].

PROPOSITION 4.2. *The condition* (4.1) *in Definition* 4.1 *is equivalent to*

$$\lim_{n\to+\infty}\left[\psi^{**}(t^n,\xi^n)-\langle\nabla\psi^{**}(t^n,\xi^n)\,,\,\xi^n\rangle\right]=-\infty \tag{4.2}$$

for every sequence $(t^n,\xi^n)\in I\times\mathbb{R}^m$ *such that* $\xi^n\in\mathcal{D}(\psi^{**}(t^n,\cdot))$, $\lim_n|\xi^n|=+\infty$.

Proof. We have to prove that (4.2) implies (4.1), the other implication being trivial. Let us denote by $\chi(R)$ the argument of the limit in (4.1), and let $\{R_n\}$ be a diverging sequence. For every fixed $n\in\mathbb{N}$, by definition of supremum, there exists $(t^n,\xi^n,p^n)\in I\times\mathbb{R}^m\times\mathbb{R}^m$, with $p^n\in\partial_\xi\psi^{**}(t^n,\xi^n)$ and $|\xi^n|>R_n$, such that

$$\chi(R_n)\le\psi^{**}(t^n,\xi^n)-\langle p^n\,,\,\xi^n\rangle+1. \tag{4.3}$$

From (2.2) and (2.1) there exist $p_j^n\in\partial_\xi\psi^{**}(t^n,\xi^n)$, $\xi_j^n\in\mathcal{D}(\psi^{**}(t^n,\cdot))$ with $|\xi_j^n-\xi^n|<1$, $j\in J\doteq\{1,\dots,m+1\}$, and $\tilde\lambda^n\in E_{m+1}$ such that

$$p^n=\sum_{j=1}^{m+1}\lambda_j^n p_j^n,\qquad\left|\nabla\psi^{**}(t^n,\xi_j^n)-p_j^n\right|<\frac{1}{|\xi^n|+1}\quad\text{for every } j\in J.$$

For every $j\in J$ the last inequality and the fact that $|\xi_j^n-\xi^n|<1$ imply that

$$\left|\langle\nabla\psi^{**}(t^n,\xi_j^n)-p_j^n\,,\,\xi_j^n\rangle\right|<\frac{|\xi_j^n|}{|\xi^n|+1}<1\,. \tag{4.4}$$

By the convexity of $\psi^{**}(t^n,\cdot)$ we have

$$\psi^{**}(t^n,\xi^n)-\psi^{**}(t^n,\xi_j^n)\le\langle p_j^n\,,\,\xi^n-\xi_j^n\rangle\quad\text{for every } j\in J. \tag{4.5}$$

Using (4.4) and (4.5) we obtain

$$\psi^{**}(t^n,\xi^n)-\langle p_j^n\,,\,\xi^n\rangle\le\psi^{**}(t^n,\xi_j^n)-\langle\nabla\psi^{**}(t^n,\xi_j^n)\,,\,\xi_j^n\rangle+1\,. \tag{4.6}$$

Multiplying (4.6) by λ_j^n and summing over j it follows that $\psi^{**}(t^n,\xi^n)-\langle p^n\,,\,\xi^n\rangle\le\mu^n$, where $\mu^n\doteq 1+\max_j\{\psi^{**}(t^n,\xi_j^n)-\langle\nabla\psi^{**}(t^n,\xi_j^n)\,,\,\xi_j^n\rangle\}$.

Since $\lim_n|\xi_j^n|=+\infty$ for every $j\in J$, (4.2) implies that $\lim_n\mu^n=-\infty$. Hence by (4.3) it follows that

$$\lim_{n\to+\infty}\chi(R_n)\le\lim_{n\to+\infty}(\mu^n+1)=-\infty\,.$$

Since χ is a monotone nonincreasing function, (4.1) holds. □

Remark 4.3. Definition 4.1 agrees with the one given in [6] and [12], respectively, in the case of convex time-independent smooth functions and nonconvex time-independent functions.

LEMMA 4.4. *If* $\psi\in\mathcal{E}$, *then* $\psi(t,\cdot)\in\mathcal{G}$ *for every* $t\in I$.

Proof. Let us fix $t\in I$ and denote by φ the convexification with respect to ξ of $\psi(t,\xi)$. By Lemma 3.3 in [12], the effective domain $\mathrm{dom}(\varphi^*)$ of φ^* is an open subset of $\mathbb{R}^m$. Hence by Proposition 2.1(iii), $\partial\varphi^*(p)$ is either bounded, if $p\in\mathrm{dom}(\varphi^*)$, or empty, if $p\notin\mathrm{dom}(\varphi^*)$. By Remark 3.2, the result is thus proved. □

LEMMA 4.5. *Let* $\varphi\colon I\times\mathbb{R}^m\times\mathbb{R}^m\to\mathbb{R}$ *be a lower semicontinuous function, Lipschitz continuous with respect to the first variable. Assume that* $\varphi(t,x,\cdot)$ *is convex for a.e.* $t\in I$ *and for every* $x\in\mathbb{R}^m$ *and that there exist three constants* C_i, $i=0,1,2$, *such that*

$$|v|\le C_0|\varphi(t,x,\xi)|+C_1|x|+C_2 \tag{4.7}$$

for every $(t, x, \xi) \in I \times \mathbb{R}^m \times \mathbb{R}^m$ *and for every* $v \in \partial_t \varphi(t, x, \xi)$, *where* $\partial_t \varphi$ *denotes the generalized gradient of* φ *with respect to* t.

Let $u \in W^{1,1}(I, \mathbb{R}^m)$ *and assume that the function* $t \mapsto \varphi(t, u(t), u'(t))$ *belongs to* $L^1(I)$. *Then there exists* $k_0 \in L^1(I)$ *such that*

$$\left|\varphi(s_2, u(t), u'(t)) - \varphi(s_1, u(t), u'(t))\right| \leq k_0(t)|s_2 - s_1|$$

for every $t, s_1, s_2 \in I$.

Proof. For every fixed $t_1, t_2 \in I$, let us define the function

$$g(s) \doteq |\varphi(t_1 + sd, x, \xi) - \varphi(t_1, x, \xi)|, \quad s \in [0, 1],$$

where $d \doteq t_2 - t_1$. By (4.7), it follows that for a.e. $s \in [0, 1]$

$$g'(s) \leq |d||\partial_t \varphi(t_1 + sd, x, \xi)| \leq |d|\,(C_0 g(s) + C_0|\varphi(t_1, x, \xi)| + C_1|x| + C_2)\,.$$

We can apply Gronwall's inequality to the nonnegative absolutely continuous function g, obtaining

$$|\varphi(t_2, x, \xi) - \varphi(t_1, x, \xi)| = g(1) \leq |t_2 - t_1| e^{C_0 T}\,(C_0|\varphi(t_1, x, \xi)| + C_1|x| + C_2)\,. \tag{4.8}$$

This inequality, with $t_1 = t$ and $t_2 = s_1$, implies that

$$|\varphi(s_1, x, \xi)| \leq |\varphi(t, x, \xi)| + T e^{C_0 T}\,(C_0|\varphi(t, x, \xi)| + C_1|x| + C_2)\,. \tag{4.9}$$

Again by (4.8), with $t_1 = s_1$, $t_2 = s_2$, and by (4.9), it follows that

$$|\varphi(s_2, x, \xi) - \varphi(s_1, x, \xi)| \leq |s_2 - s_1|(\tilde{C}_0|\varphi(t, x, \xi)| + \tilde{C}_1|x| + \tilde{C}_2),$$

where $\tilde{C}_i \doteq C_i e^{C_0 T}(1 + T C_0 e^{C_0 T})$, $i = 0, 1, 2$. Finally, by hypothesis, the function

$$k_0(t) \doteq \tilde{C}_0|\varphi(t, u(t), u'(t))| + \tilde{C}_1|u(t)| + \tilde{C}_2$$

belongs to $L^1(I)$, completing the proof. □

DEFINITION 4.6. *We shall say that* $\theta \in C^1((0, +\infty), \mathbb{R})$ *is a Nagumo function if* θ *is convex and increasing and it satisfies* $\lim_{r \to +\infty} \theta(r)/r = +\infty$.

We begin the study of minimization problems, starting with an existence result for convex functionals. We collect here the basic hypotheses on the integrand.

(H_0) $f \in \mathcal{E}$ and $f(t, \cdot)$ is a convex function for every $t \in I$.

(H_1) There exist two constants A and B, with $B > 0$, such that $f(t, \xi) \geq -A + B|\xi|$ for every $(t, \xi) \in I \times \mathbb{R}^m$.

(H_2) $g: I \times \mathbb{R}^m \to \mathbb{R}$ is Lipschitz continuous with respect to the first variable and continuous with respect to the second, and there exist two constants α and β, with $0 \leq \beta < B/T$, such that $g(t, x) \geq -\alpha - \beta|x|$ for every $(t, x) \in I \times \mathbb{R}^m$.

(H_3) There exist three constants C_i, $i = 0, 1, 2$, such that the condition (4.7) holds with $\varphi(t, x, \xi) \doteq g(t, x) + f(t, \xi)$.

Remark 4.7. If $f \in \mathcal{E}$ is independent of t, then it is easily seen that Lemmas 3.5 and 4.4 imply that condition (H_1) is always satisfied for suitable constants A and B, with $B > 0$.

THEOREM 4.8. *Let* f *and* g *satisfy the hypotheses* (H_0), (H_1), (H_2), (H_3). *Then there exists a solution to the problem*

$$\min\left\{F(u) \;\middle|\; u \in W^{1,1}(I, \mathbb{R}^m),\ u(0) = a,\ u(T) = b\right\}, \tag{4.10}$$

where

$$F(u) \doteq \int_I [f(t, u'(t)) + g(t, u(t))]\, dt\,.$$

Moreover every solution $\tilde{u}$ belongs to $W^{1,\infty}(I, \mathbb{R}^m)$ and satisfies for a.e. $t \in I$

$$f(t, \tilde{u}'(t)) - \langle p(t)\,,\, \tilde{u}'(t)\rangle + g(t, \tilde{u}(t)) = c + \int_0^t v(\tau)\, d\tau\,, \tag{4.11}$$

where c is a constant and $(v(t),\, p(t)) \in (\partial_t f(t, \tilde{u}'(t)) + \partial_t g(t, \tilde{u}(t)),\, \partial_\xi f(t, \tilde{u}'(t)))$ for almost every $t \in I$.

Proof. The proof follows the lines of the proof of Theorem 3 in [10], with some changes due to the fact that in this case the Lagrangian is not bounded from below. As in [10] one can prove, using the De Giorgi semicontinuity result (see [4]) and the Dunford–Pettis criterion of weak compactness in $L^1(I, \mathbb{R}^m)$, that for every Nagumo function θ and for every $l > 0$ there exists a solution u_l to the problem

$$\min\left\{F(u) \;\middle|\; u \in AC^l_\theta(I, \mathbb{R}^m),\ u(0) = a,\ u(T) = b\right\}\,,$$

where $AC^l_\theta(I, \mathbb{R}^m)$ denotes the class of all function $u \in W^{1,1}(I, \mathbb{R}^m)$ such that $\Theta(u) \le l$, with $\Theta(u) \doteq \int_I \theta(|u'(t)|)\, dt$. Let us set $V_\theta(l) \doteq F(u_l)$.

One can easily check that if $V_\theta(l) = V_\theta(l_0)$ for every $l \ge l_0$ then u_{l_0} is a solution to the problem

$$\min\left\{F(u) \;\middle|\; u \in W^{1,1}(I, \mathbb{R}^m),\ \Theta(u) < +\infty,\ u(0) = a,\ u(T) = b\right\}\,. \tag{4.12}$$

Finally, as in [10], if we are able to prove that u_{l_0} belongs to $W^{1,\infty}(I, \mathbb{R}^m)$ then we can conclude that such a function is a solution to (4.10). Furthermore, any other solution $\tilde{u}$ of (4.10) would solve (4.12) for some Nagumo function θ and hence would belong to $W^{1,\infty}(I, \mathbb{R}^m)$ and satisfy (4.11).

Thus it remains to prove that V_θ is eventually constant and that, for l large enough, u_l belongs to $W^{1,\infty}(I, \mathbb{R}^m)$ and satisfies (4.11). Since V_θ is lower semicontinuous, for every $l > 0$ there exists a proximal subgradient (see [9]) of V_θ at l and, since V_θ is nonincreasing, it is nonpositive. If V_θ is not eventually constant, by Proposition 6.1 in [10], there exists a diverging sequence $\{l_k\}$ such that the proximal subgradient of V_θ at l_k takes the form $-r_k$, with $r_k > 0$. Moreover, it is easy to check that, if we set $u_k \equiv u_{l_k}$, then $\Theta(u_k) = l_k$, so that

$$\lim_{k\to+\infty} \|u_k'\|_{L^\infty} \ge \lim_{k\to+\infty} \theta^{-1}(l_k/T) = +\infty\,. \tag{4.13}$$

By definition of r_k and the fact that $\Theta(u_k) = l_k$, it follows that for every $k \in \mathbb{N}$ there exists a positive constant σ_k such that, if we define

$$G(u) \doteq F(u) + r_k\Theta(u) + \sigma_k|\Theta(u) - \Theta(u_k)|^2\,,$$

then we get that $G(u_k) \le G(u)$ for every u admissible for (4.12) and such that $\Theta(u)$ is sufficiently near to $\Theta(u_k)$ (see [10]). By (H_3) and Lemma 4.5, it follows that there exists $k_0 \in L^1(I)$ such that for every $s_1, s_2, t \in I$

$$|f(s_1, u_k'(t)) + g(s_1, u_k(t)) - f(s_2, u_k'(t)) - g(s_2, u_k(t))| \le k_0(t)|s_1 - s_2|\,,$$

so that we can apply Theorem 5 of [10]. Thus we obtain that u_k satisfies

$$E_f(t, u_k'(t)) + g(t, u_k(t)) + r_k E_\theta(|u_k'(t)|) = c_k + \int_0^t v_k(\tau)\, d\tau\,, \tag{4.14}$$

where c_k is a constant, $E_f(t, u_k'(t)) \doteq f(t, u_k'(t)) - \langle p_k(t)\,,\, u_k'(t)\rangle$, with $(v_k(t),\, p_k(t)) \in (\partial_t f(t, u_k'(t)) + \partial_t g(t, u_k(t)),\, \partial_\xi f(t, u_k'(t)))$ for a.e. $t \in I$, and $E_\theta(s) \doteq \theta(s) - s\theta'(s)$.

Moreover there exists $M_1 > 0$ such that $\|u_k\|_{L^\infty} \le M_1$ for every $k \in \mathbb{N}$. Actually, if there exists $t_k \in I$ such that $\limsup_k |u_k(t_k)| = +\infty$, then

$$\limsup_{k\to+\infty} \int_I |u_k'(t)|\,dt \ge \limsup_{k\to+\infty} \left| \int_0^{t_k} u_k'(t)\,dt \right| = \limsup_{k\to+\infty} |u_k(t_k) - a| = +\infty\,,$$

whereas, if we define $u_0(t) \doteq a + \xi t$, with $\xi \doteq (b-a)/T$, then u_0 is admissible for (4.12), $F(u_0) < +\infty$, and

$$F(u_0) \ge F(u_k) \ge (-A-\alpha)T + B\|u_k'\|_{L^1} - \beta\|u_k\|_{L^1} \ge \tilde{A} + (B - \beta T)\|u_k'\|_{L^1}\,, \tag{4.15}$$

so that, by (H_2), $\{u_k'\}$ must be bounded in $L^1(I, \mathbb{R}^m)$.

The boundedness of $\{u_k\}$ in $L^\infty(I, \mathbb{R}^m)$ and the continuity of g guarantee that there exists M_2 such that

$$|g(t, u_k(t)| \le M_2 \tag{4.16}$$

for a.e. $t \in I$ and for every k. Moreover, by (H_3) we obtain

$$\begin{aligned} \left| \int_0^t v_k(s)\,ds \right| &\le \int_I \left[C_0 \left| f(s, u_k'(s)) + g(s, u_k(s)) \right| + C_1 |u_k(s)| + C_2 \right] ds \\ &\le \int_I \left[C_0 \left| \alpha + \beta |u_k(s)| + f(s, u_k'(s)) + g(s, u_k(s)) \right| + \tilde{C}_1 |u_k(s)| + \tilde{C}_2 \right] ds\,, \end{aligned} \tag{4.17}$$

where $\tilde{C}_1 \doteq C_0\beta + C_1$ and $\tilde{C}_2 \doteq C_0|\alpha| + C_2$. Without loss of generality we can assume that f is positive, so that, thanks to (H_2), it follows that for every $k \in \mathbb{N}$

$$f(s, u_k'(s)) + g(s, u_k(s)) + \alpha + \beta|u_k(s)| \ge 0 \quad \text{a.e. } s \in I\,. \tag{4.18}$$

By (4.15), (4.17), and (4.18) there exist $M_3 > 0$ and two constants $\hat{C}_1$, $\hat{C}_2$ such that

$$\left| \int_0^t v_k(s)\,ds \right| \le C_0 F(u_k) + \hat{C}_1 \|u_k\|_{L^1} + \hat{C}_2 \le M_3 \quad \text{for every } t \in I. \tag{4.19}$$

By (4.14), (4.16), and (4.19) we obtain

$$E_f(t, u_k'(t)) + r_k E_\theta(|u_k'(t)|) \le c_k + M_2 + M_3$$

for every $t \in I$ and for every $k \in \mathbb{N}$.

We claim that it is not possible that there exists a subsequence of $\{c_k\}$, still denoted by $\{c_k\}$, such that $\lim_k c_k = -\infty$. Indeed, if this is the case, then for every $t \in I$ we should have

$$\lim_{k\to+\infty} E_f(t, u_k'(t)) + r_k E_\theta(|u_k'(t)|) = -\infty\,. \tag{4.20}$$

Since $f \in \mathcal{E}$ and θ is superlinear, (4.20) implies that $\lim_k |u_k'(t)| = +\infty$ for every $t \in I$, which by Fatou's lemma contradicts the boundedness of u_k' in $L^1(I, \mathbb{R}^m)$.

Thus there exists c^* such that $c_k \ge c^*$ for every k. From (4.14) we obtain, for every $t \in I$,

$$E_f(t, u_k'(t)) + r_k E_\theta(|u_k'(t)|) \ge c^* - M_2 - M_3\,. \tag{4.21}$$

Now let us suppose that for every k there exists $t_k \in I$ such that $\limsup_k |\xi_k| = +\infty$, where $\xi_k \doteq u_k'(t_k)$. Since f and θ belong to $\mathcal{E}$, we have

$$\liminf_{k\to+\infty} \left[E_f(t_k, \xi_k) + r_k E_\theta(|\xi_k|)\right] \le \liminf_{k\to+\infty} \sup_{t\in I} \left\{E_f(t, \xi_k) + r_k E_\theta(|\xi_k|)\right\} = -\infty\,,$$

in contradiction with (4.21). This implies that $\|u_k'\|_{L^\infty}$ is bounded, which contradicts (4.13).

So we can conclude that V_θ is eventually constant. Hence for k sufficiently large $u_k \in W^{1,\infty}(I, \mathbb{R}^m)$ is a solution of (4.12). Moreover $r_k = 0$, so that u_k satisfies (4.11). Then the proof is complete. □

The last part of this section is devoted to the study of the nonconvex case. The hypotheses (H_0) and (H_3) will be replaced, respectively, by

(H_0') $f \in \mathcal{E}$;

(H_3') there exist three constants C_i, $i = 0, 1, 2$, such that the condition (4.7) holds with $\varphi(t, x, \xi) \doteq g(t, x) + f^{**}(t, \xi)$.

Notice that (H_3') requires the Lipschitz continuity of f^{**} with respect to t. The following two lemmas show that this conclusion follows from (H_0') and

(H_4) for every $R > 0$ there exists a constant L such that

$$|f(t, \xi) - f(s, \xi)| \le L|t - s| \quad \text{for every } t, s \in I, \text{ and } \xi \in \overline{B}_R,$$

where $\overline{B}_R$ denotes the closed ball centered at the origin and with radius R.

LEMMA 4.9. *Let $\psi \in \mathcal{E}$ and let us define, for every $(t, p) \in I \times \mathbb{R}^m$, the set*

$$W(t, p) \doteq \{\xi \in \mathbb{R}^m \mid p \in \partial_\xi \psi^{**}(t, \xi)\}\,.$$

Then for every $r > 0$ there exists $R > 0$ such that for every $(t, p) \in I \times \mathbb{R}^m$ the condition $W(t, p) \cap \overline{B}_r \neq \emptyset$ implies $W(t, p) \subset \overline{B}_R$.

Proof. Suppose, by contradiction, that there exist sequences $(t_n, p_n) \subset I \times \mathbb{R}^m$, $(\eta_n) \subset \overline{B}_r$, $(\xi_n) \subset \mathbb{R}^m$, with $\lim_n |\xi_n| = +\infty$, such that for every $n \in \mathbb{N}$

$$p_n \in \partial_\xi \psi^{**}(t_n, \eta_n), \quad p_n \in \partial_\xi \psi^{**}(t_n, \xi_n)\,. \tag{4.22}$$

From (4.22) it follows that for every $n \in \mathbb{N}$

$$\psi^{**}(t_n, \eta_n) - \langle p_n\,,\, \eta_n\rangle = \psi^{**}(t_n, \xi_n) - \langle p_n\,,\, \xi_n\rangle\,. \tag{4.23}$$

Since (η_n) is a bounded sequence, there exists a constant C such that the left-hand side of (4.23) is bounded from below by C. Thus

$$C \le \psi^{**}(t_n, \xi_n) - \langle p_n\,,\, \xi_n\rangle \le \chi(|\xi_n|) \quad \text{for every } n \in \mathbb{N}, \tag{4.24}$$

where $\chi(R)$ is the argument of the limit in (4.1). Since $\lim_n |\xi_n| = +\infty$, from (4.1) we have that $\lim_n \chi(|\xi_n|) = -\infty$, which contradicts (4.24). □

Remark 4.10. Let us fix $\xi \in \mathbb{R}^m$. Let $t \in I$, $\tilde{\lambda} \in E_{m+1}$, $\xi_j \in \mathbb{R}^m$, $j = 1, \dots, m+1$ satisfy

$$f^{**}(t, \xi) = \sum_{j=1}^{m+1} \lambda_j f(t, \xi_j), \quad \xi = \sum_{j=1}^{m+1} \lambda_j \xi_j.$$

Since for every j there exists $p_j \in \partial_\xi f^{**}(t, \xi)$ such that $\xi_j \in W(t, p_j)$, by Lemma 4.9 we obtain that there exists $R > 0$, depending only on $|\xi|$, such that $\xi_j \in \overline{B}_R$ for every $j = 1, \dots, m+1$.

LEMMA 4.11. *If $f \in \mathcal{E}$ satisfies (H_4), then $f^{**}(\cdot, \xi)$ is Lipschitz continuous for every $\xi \in \mathbb{R}^m$.*

Proof. Let us fix $\xi \in \mathbb{R}^m$ and consider $t, s \in I$. By Corollary 3.7, there exist $\tilde{\lambda}, \tilde{\mu} \in E_{m+1}$, $\xi_j, \eta_j \in \mathbb{R}^m$, $j = 1, \dots, m+1$ such that

$$f^{**}(t, \xi) = \sum_{j=1}^{m+1} \lambda_j f(t, \xi_j), \quad f^{**}(s, \xi) = \sum_{j=1}^{m+1} \mu_j f(s, \eta_j),$$

and $\xi = \sum_j \lambda_j \xi_j = \sum_j \mu_j \eta_j$. Moreover, one has

$$f^{**}(t, \xi) \le \sum_{j=1}^{m+1} \mu_j f(t, \eta_j), \quad f^{**}(s, \xi) \le \sum_{j=1}^{m+1} \lambda_j f(s, \xi_j).$$

Then, by Remark 4.10 and (H_4), there exists $L > 0$, depending only on $|\xi|$, such that

$$f^{**}(s, \xi) - f^{**}(t, \xi) \le \sum_{j=1}^{m+1} \lambda_j [f(s, \xi_j) - f(t, \xi_j)] \le \sum_{j=1}^{m+1} \lambda_j L|t - s| = L|t - s|.$$

In the same way one obtains

$$f^{**}(t, \xi) - f^{**}(s, \xi) \le \sum_{j=1}^{m+1} \mu_j [f(t, \eta_j) - f(s, \eta_j)] \le L|t - s|,$$

completing the proof. □

We are now in a position to prove the existence result for the nonconvex case.

THEOREM 4.12. *Let g and f satisfy the basic hypotheses (H_0'), (H_1), (H_2), (H_3'), (H_4) and assume that $g(t, \cdot)$ is concave for every $t \in I$. Then the problem* (4.10) *has a solution $u \in W^{1,\infty}([0, T], \mathbb{R}^m)$.*

Proof. The proof follows the same lines of the one of Theorem 1 in [5]. It is enough to use Theorem 4.8 to obtain a solution $\tilde{u} \in W^{1,\infty}([0, T], \mathbb{R}^m)$ of the relaxed problem and to replace Lemma IX.3.3 and Proposition IX.3.1 of [13] with Corollaries 3.7 and 3.8. Since $\tilde{u}' \in L^\infty([0, T], \mathbb{R}^m)$, it is easily seen, using Lemma 4.9, that we obtain a solution $u \in W^{1,\infty}([0, T], \mathbb{R}^m)$. □

Acknowledgments. The authors wish to thank Arrigo Cellina for kindly suggesting the problem.

REFERENCES

[1] L. AMBROSIO, O. ASCENZI, AND G. BUTTAZZO, *Lipschitz regularity for minimizers of integral functionals with highly discontinuous integrands*, J. Math. Anal. Appl., 142 (1989), pp. 301–316.

[2] B. BOTTERON AND B. DACOROGNA, *Existence and non-existence results for non-coercive variational problems and applications in ecology*, J. Differential Equations, 85 (1990), pp. 214–235.

[3] B. BOTTERON AND P. MARCELLINI, *A general approach to the existence of minimizers of one dimensional non-coercive integrals of the calculus of variations*, Ann. Inst. H. Poincaré Anal. Non Lineaire, 8 (1991), pp. 197–223.

[4] G. BUTTAZZO, *Semicontinuity, Relaxation and Integral Representation in the Calculus of Variations*, Pitman Res. Notes Math. Ser. 207, Longman Scientific and Technical, Harlow, U.K., 1989.

[5] A. CELLINA AND G. COLOMBO, *On a classical problem of the calculus of variations without convexity assumptions*, Ann. Inst. H. Poincaré Anal. Non Lineaire, 7 (1990), pp. 97–106.

[6] A. CELLINA, G. TREU, AND S. ZAGATTI, *On the minimum problem for a class of non-coercive functionals*, J. Differential Equations, to appear.

[7] L. CESARI, *Optimization – Theory and Applications*, Springer-Verlag, New York, 1983.
[8] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
[9] ———, *Methods of dynamic and nonsmooth optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math., Vol. 57, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1989.
[10] ———, *An indirect method in the calculus of variations*, Trans. Amer. Math. Soc., 336 (1993), pp. 655–673.
[11] F. H. CLARKE AND P. D. LOEWEN, *An intermediate existence theory in the calculus of variations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 16 (1989), pp. 487–526.
[12] G. CRASTA, *An existence result for non-coercive non-convex problems in the calculus of variations*, Nonlinear Anal., 26 (1996), pp. 1527–1533.
[13] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1977.
[14] P. MARCELLINI, *Alcune osservazioni sull'esistenza del minimo di integrali del calcolo delle variazioni senza ipotesi di convessità*, Rend. Mat., 13 (1980), pp. 271–281.
[15] C. OLECH, *The Lyapunov theorem: Its extensions and applications*, in Methods of Non-convex Analysis, A. Cellina, ed., Springer-Verlag, New York, 1990.
[16] J. P. RAYMOND, *Existence theorems in optimal control problems without convexity assumptions*, J. Optim. Theory Appl., 67 (1990), pp. 109–132.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 2077–2091, November 1996

012

RELAXATION OF CONSTRAINED CONTROL PROBLEMS*

E. N. BARRON† AND R. JENSEN†

Abstract. The problem of relaxation of optimal control problems with state and control constraints is formulated in this paper. We determine that if the original problem consists of minimizing, over control functions $\zeta(\cdot)$, $g(\xi(T))$ subject to $d\xi/ds = f(s,\xi,\zeta)$, $t < s \leq T$, and $h(s,\xi(s),\zeta(s)) \leq 0$ for a.e. $t \leq s \leq T$, then the relaxed problem consists of minimizing, over measure-valued control functions $\mu(\cdot)$, $g(\hat{\xi}(T))$, subject to $d\hat{\xi}/ds = \int_Z f(s,\hat{\xi}(s),z)\mu(s,dz)$ and $\mu(s) - \operatorname{ess\,sup}_z h(s,\hat{\xi}(s),z) \leq 0$ for a.e. $t \leq s \leq T$. For each s this is the essential supremum of h in z with respect to the measure $\mu(s)$.

Key words. relaxed controls, state and control constraints

AMS subject classifications. 49A40, 49A10, 49C20

1. Introduction.

In this paper we will determine the relaxed formulation of the following constrained optimal control problem of Mayer type:

$$\text{minimize } P_{t,x}(\zeta) = g(\xi(T)) \text{ over controls } \zeta(\cdot)$$

subject to $d\xi(\tau)/d\tau = f(\tau,\xi(\tau),\zeta(\tau)), 0 \leq t < \tau \leq T$, $\xi(t) = x \in R^n$, and the state and control constraints $h(\tau,\xi(\tau),\zeta(\tau)) \leq 0$ for a.e. $t \leq \tau \leq T$. The control functions $\zeta(\cdot)$ typically take values in a compact control set Z. Relaxing this problem consists of enlarging the space of control functions to include control functions $\mu(\cdot)$ which, for each fixed $\tau \in [0,T]$ is a probability measure on the control set Z. Then we have the problem of how to define $f(\cdot,\cdot,\mu)$ as well as $h(\cdot,\cdot,\mu)$.

The reason for relaxing a control problem is twofold. First, it is always pleasant to have an optimal control, but this optimal control is usually found only in the class of relaxed controls, which is compact in the weak topology. We do not, in general nonconvex problems, expect an optimal control to exist outside of the class of measure-valued controls. Of course, optimal controls can be found for special problems with additional hypotheses. For example, even in nonconvex problems, an optimal control will exist in the class of uniformly bounded and uniformly Lipschitz continuous control functions with a fixed Lipschitz constant. Many of these results can be found in books by Cesari [10], Berkovitz [7] and Ekeland and Temam [13]. Second, numerical approximations of a control problem are known (for example, [8]) to converge to the relaxed version of the original problem. For these reasons, when one speaks of relaxing an optimal control problem, the relaxation chosen, since there is often more than one way to do this, must satisfy the following properties:

(i) an optimal relaxed control must exist for all problems with reasonable hypotheses;

(ii) the relaxation should have the same minimum as the infimum of the original problem.

A relaxation satisfying these two properties can be said to be correct.

Relaxation theory has been important ever since the origin of optimal control theory, especially when L. C. Young emphasized the important role of convexity. The standard reference for relaxation theory is [22]. For some more recent results see [21], and for relaxation applied to some state contrained problems (but without mixed control and state constraints) see [20].

The theory has led to a major tool in variational problems, appropriately named Young measures = relaxed controls. It is known that the straightforward idea of defining $f(t,x,\mu)$ by $\int_Z f(t,x,z)\mu(dz)$ and $h(t,x,\mu)$ by $\int_Z h(t,x,z)\mu(dz)$ will *not* work since simple examples

*Received by the editors September 19, 1994; accepted for publication (in revised form) September 28, 1995. The research of the first author was supported in part by grant DMS-9300805 from the National Science Foundation, and the research of the second author was supported by grant DMS-9101799 from the National Science Foundation.

†Department of Mathematical Sciences, Chicago, IL 60626 (enb@math.luc.edu and rrj@math.luc.edu).

show that (ii) will not be satisfied. That is, enlarging to relaxed control functions with this definition of f and h may decrease the minimum. We will prove that the way to relax this problem is to define $f(t, x, \mu)$ in the usual way by $\int_Z f(t, x, z)\mu(dz)$, *but* $h(t, x, \mu)$ should be defined by $\mu - \operatorname{ess\,sup}_{z\in Z} h(t, x, z)$, the essential supremum in z of h, with respect to the measure μ. We will prove that this relaxation is the correct relaxation. In particular, (ii) will be proven by using the uniqueness results of viscosity solutions to first-order PDEs, initiated in the seminal papers of Crandall and Lions [11] and Crandall, Evans, and Lions [12].

A major role in this paper is played by the constraint condition (for one constraint function) that when $h(t, x, z) = 0$, $D_z h(t, x, z) \neq 0$. This condition is very stringent and should be weakened. Nevertheless, it is consistent with the assumption for necessary conditions in [15] and is intuitive as well. That is, in practical problems one expects to be able to control the system back into the desired region when one reaches the boundary. The condition says that the control variable can still be used when the constraint is about to be violated. Helene Frankowska has informed us that under this constraint assumption, the theory of differential inclusions already proves the relaxation theorem in this paper. The specific formulation of the relaxed problem and the proof of the relaxation theorem using viscosity solutions appears to be new. In any case, our goal in this paper is not to find the most general result regarding the state constraint problem but to determine the correct formulation of the relaxed state constrained problem in an explicit form.

The problem of this paper was brought to our attention by an anonymous referee of [3], where we determined the relaxation of optimal control problems with L^∞ cost functionals. The referee pointed out that the result obtained there was a special case of the problem considered in the present paper.

2. Statement of the problem. Consider the controlled system of ordinary differential equations

$$d\xi(\tau)/d\tau = f(\tau, \xi(\tau), \zeta(\tau)), \qquad 0 \leq t < \tau \leq T, \tag{2.1}$$

$$\xi(t) = x \in R^n. \tag{2.2}$$

In this section we will begin by imposing only one constraint on the control functions ζ. Consequently, the controls will be allowed to take values in R^q. Specifically, the control functions $\zeta(\cdot)$ are chosen from the class of functions

$$\mathcal{Z}[t, T] = \{\zeta : [t, T] \to R^q : \zeta \text{ is Lebesgue measurable}\}, \qquad q \geq 1.$$

Throughout this paper we will assume that the following condition holds. The letter K will be a symbol perhaps denoting a different constant at different times.

(A) $f : [0, T] \times R^n \times R^q \to R^n$ is jointly continuous and is Lipschitz in x and z. That is, there is a constant K such that

$$|f(t, x, z) - f(t, x', z')| \leq K(|x - x'| + |z - z'|) \quad \forall x, x' \in R^n,\ z, z' \in R^q. \tag{2.3}$$

In addition, $|f(t, x, z)| \leq K(1 + |x|)$. The given function $g : R^n \to R^1$ is Lipschitz and uniformly bounded below.

We turn now to the single constraint function $h : [0, T] \times R^n \times R^q \to R^1$. Later we will indicate the hypotheses needed when we have more than one constraint. In particular, we will need more than one constraint function in order to incorporate constraints such as $|\zeta| \leq M$, which gives us a compact control set.

(B) $h(t, x, z)$ is continuously differentiable and Lipschitz in all variables and satisfies the two conditions

$$D_z h(t, x, z) \neq 0 \text{ if } h(t, x, z) = 0, \tag{2.4}$$

and

$$\{z \in R^q : h(t,x,z) \le r\} \text{ is compact } \forall\, r \in R^1. \tag{2.5}$$

The goal is to choose a control function $\zeta \in \mathcal{Z}$ which minimizes the cost functional

$$P_{t,x}(\zeta) = g(\xi(T))$$

subject to the constraints that

$$h(\tau, \xi(\tau), \zeta(\tau)) \le 0 \quad \text{for a.e. } t \le \tau \le T.$$

Define the set of feasible control functions

$$\mathcal{Z}_h[t,T] = \{\zeta \in \mathcal{Z}[t,T] : h(\tau, \xi(\tau), \zeta(\tau)) \le 0 \text{ a.e. } t \le \tau \le T\}.$$

The constrained value function $V : [0,T] \times R^n \to R^1$ is defined by

$$V(t,x) = \inf_{\zeta \in \mathcal{Z}_h[t,T]} P_{t,x}(\zeta) \tag{2.6}$$

By the usual convention, $V(t,x) = +\infty$ if $\mathcal{Z}_h[t,T] = \emptyset$. We shall have use of the domain

$$\Omega = \left\{(t,x) \in [0,T] \times R^n : \min_{z \in R^q} h(t,x,z) < 0\right\}$$

and the function

$$\gamma(t,x) = \min_{z \in R^q} h(t,x,z).$$

LEMMA 2.1. *When* (B) *holds, the minimum in the definition of* γ *is achieved for each* (t,x). *Also,* $\Omega = [0,T] \times R^n$.

Proof. Let z_k satisfy $\gamma(t,x) \ge h(t,x,z_k) - 1/k$, $k = 1, 2, \ldots$. Then $z_k \in \{z : h(t,x,z) \le \gamma + 1\}$, which is assumed compact. So we may assume $z_k \to z^*$, and then it is easy to see that z^* provides the minimum.

Under the assumption (2.5) when $\gamma(t,x) = 0$—say, $h(t,x,z_0) = 0$—the fact that $D_z h(t,x,z_0) \ne 0$ implies that $\gamma(t,x) < 0$. Thus, $\{\gamma < 0\}$ everywhere, which says that $\Omega = [0,T] \times R^n$. □

We assume throughout this paper that

$$V : \Omega \to R^1 \text{ is finite for each } (t,x) \in \Omega. \tag{2.7}$$

This is equivalent to assuming that $\mathcal{Z}_h[t,T] \ne \emptyset$ for $(t,x) \in \Omega$.

Remark 2.1. If we assume that there is a constant $C > 0$ such that $\min_z h(t,x,z) \le -C$, it will follow that $V < +\infty$ for all (t,x). Indeed, we set

$$u(t,x) = \inf_{\zeta \in \mathcal{Z}[t,T]} \operatorname{ess\,sup} h(s, \xi(s), \zeta(s)),$$

which is the value function for the L^∞ control problem with cost function h (see [2]–[4]). Then u is the unique viscosity solution of

$$\max\left\{u_t + \min_{\{z : h(t,x,z) \le u\}} D_x u \cdot f(t,x,z),\ \min_z h(t,x,z) - u\right\} = 0$$

in Ω, with $u(T,x) = \min_z h(T,x,z)$. The assumption (2.5) replaces the need for a compact control set used in [4]. Under the assumption $\min_z h(t,x,z) \le -C$, we verify that $u_0(t,x) \equiv -C$ is a viscosity supersolution of the L^∞ problem and therefore $u \le u_0 = -C$ everywhere. But then one can always find a control $\zeta \in \mathcal{Z}[t,T]$ so that $h(s,\xi(s),\zeta(s)) \le -C/2 \le 0$, i.e., $\zeta \in \mathcal{Z}_h$.

Define the Hamiltonian function

$$H(t,x,r,p) = \min_{\{z \in R^q : h(t,x,z) \le r\}} p \cdot f(t,x,z) \tag{2.8}$$

for $r \in R^1$ and $p \in R^n$.

Observe that the minimum is over the r-level set $\{z \in R^q : h(t,x,z) \le r\}$, which is assumed to be compact in R^q $\forall r \in R^1$ by (2.5). If this set is empty, H is defined as $+\infty$. It is known that $H(\cdot,\cdot,r,\cdot)$ is, in general, discontinuous, in which case we must calculate the upper and lower semicontinuous envelopes of H. This was done in [4] for H given in (2.8):

$$H^*(t,x,r,p) \equiv \limsup_{(s,y,\rho,q)\to(t,x,r,p)} H(s,y,\rho,q) = H(t,x,r-0,p)$$

and

$$H_*(t,x,r,p) \equiv \liminf_{(s,y,\rho,q)\to(t,x,r,p)} H(s,y,\rho,q) = H(t,x,r+0,p).$$

In general, upper (lower) * denotes the upper (lower) semicontinuous envelope. Using the continuity of h we see quickly that

$$H_*(t,x,r,p) = H(t,x,r+0,p) = \min_{\{z \in R^q : h(t,x,z) \le 0\}} p \cdot f(t,x,z). \tag{2.9}$$

Using condition (B) we can say more.

LEMMA 2.2. *Assuming* (A) *and* (B), $H^*(t,x,0,p) = H_*(t,x,0,p) = H(t,x,0,p)$.

Proof. We know that

$$H^*(t,x,0,p) = H(t,x,0-0,p) \ge H(t,x,0,p) = H(t,x,0+0,p) = H_*(t,x,0,p),$$

and so it is sufficient to prove that $H(t,x,0,p) \ge H(t,x,0-0,p)$. Fix $(t,x) \in \Omega$ and let z_0 satisfy $h(t,x,z_0) \le 0$ and $\min_{\{z \in R^q : h(t,x,z) \le 0\}} p \cdot f(t,x,z) = p \cdot f(t,x,z_0)$. If $h(t,x,z_0) < 0$, we are done. If $h(t,x,z_0) = 0$, then (B) says that for each $\varepsilon > 0$, we can find $\delta(\varepsilon) > 0$ and $\widehat{z}$ so that $|z_0 - \widehat{z}| < \delta$ and $h(t,x,\widehat{z}) \le -\varepsilon$. Using (A) we get, with K_f the Lipschitz constant for f,

$$\begin{aligned} H(t,x,0,p) = p \cdot f(t,x,z_0) &\ge p \cdot f(t,x,\widehat{z}) - K_f |z_0 - \widehat{z}| \\ &\ge \min_{\{z \in R^q : h(t,x,z) \le -\varepsilon\}} p \cdot f(t,x,z) - K\delta \\ &= H(t,x,-\varepsilon,p) - K\delta. \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, we reach the result. □

Now we will characterize the constrained value function.

THEOREM 2.3. *Assume* (A), (B), *and* (2.7). *The value function* $V : \Omega \to R^1$ *is the unique continuous viscosity solution of the problem*

$$V_t + \min_{\{z \in R^q : h(t,x,z) \le 0\}} D_x V \cdot f(t,x,z) = 0, \quad (t,x) \in \Omega, \tag{2.10}$$

satisfying the terminal condition

$$V(T, x) = g(x), \qquad x \in R^n. \tag{2.11}$$

We recall here the definition due to Ishii (see [16] and [1]) of a viscosity solution for problems with discontinuous Hamiltonians. Even though we have proved the continuity of H, the formulation of the definition using upper and lower semicontinuous envelopes is convenient for use in some of the proofs below.

DEFINITION 2.1. *A locally bounded function u is a viscosity solution of $u_t + H(t, x, 0, D_x u) = 0$ if*

(i) *u is a viscosity subsolution on Ω, i.e., for any $(t_0, x_0) \in \Omega$ for which $u^* - \varphi$ has a maximum, for a smooth function φ, it follows that*

$$\varphi_t + H^*(t_0, x_0, 0, D_x\varphi(t_0, x_0)) \geq 0;$$

and

(ii) *u is a viscosity supersolution on Ω, i.e., for any $(t_0, x_0) \in \Omega$ for which $u_* - \varphi$ has a minimum, for a smooth function φ, it follows that*

$$\varphi_t + H_*(t_0, x_0, 0, D_x\varphi(t_0, x_0)) \leq 0.$$

Since Ω is all of $[0, T] \times R^n$ we do not have to work with *constrained viscosity solutions* (see [9], [17]–[19]) in this paper. These are functions which are subsolutions in the interior of a constraint set and supersolutions on the entire constraint set, including the boundary. In order to relax assumption (B) constrained viscosity solutions will have to be used.

Proof of Theorem 2.3. The fact that V is locally bounded on Ω follows from the classical results of Hestenes [15, Chap. 7, §5] and is a consequence of the implicit function theorem using assumption (B). In fact, it is proven in [15] that if there is a control $\zeta \in \mathcal{Z}_h$ for a starting point $(s, y) \in \Omega$, then the same is true in a neighborhood of (s, y). Let $\varepsilon > 0$ and $(s, y) \in \Omega$. Let $\zeta_\varepsilon \in \mathcal{Z}_h[s, T]$ be near optimal, i.e., $V(s, y) + \varepsilon \geq g(\xi_\varepsilon(T; s, y)) - \varepsilon$, where $\xi_\varepsilon(\cdot\,; s, y)$ is the trajectory associated with ζ_ε. Then, using assumption (A) and standard results in ODEs, it is easy to verify that $V(t, x) \leq V(s, y) + K\delta + \varepsilon$ for all $(t, x) \in B_\delta(s, y)$. It follows that V is locally bounded above. V is bounded below also by assumption (A).

The fact that V satisfies (2.10) follows in a standard way (see [19]) from the dynamic programming principle:

$$V(t, x) = \inf_{\zeta \in \mathcal{Z}_h[t,s]} V(s, \xi(s)), \tag{2.12}$$

where $\mathcal{Z}_h[t, s] = \{\zeta \in \mathcal{Z}[t, s] : h(\tau, \xi(\tau), \zeta(\tau)) \leq 0, \text{ a.e. } t \leq \tau \leq s\}$.

We will only prove that V is a subsolution, the proof that V is a supersolution is similar. Suppose that $V^* - \varphi$ achieves a strict zero maximum at $(s, y) \in \Omega$. If V is not a subsolution, then there is $\delta > 0$ such that

$$\varphi_t + H(s, y, -\delta, D_x\varphi) \leq -\delta$$

at (s, y). By the definition of H and (2.5), there is a $z_\delta \in R^q$ so that $h(s, y, z_\delta) \leq -\delta$ and

$$\varphi_t + D_x\varphi \cdot f(s, y, z_\delta) \leq -\delta. \tag{2.13}$$

Set $\zeta(\tau) \equiv z_\delta$, and let $\xi(\cdot; t, x)$ be the trajectory associated with ζ on $[s, T]$ with $\xi(t) = x$. Set

$$\mathcal{A}(z, \rho) = \{(t, x) : h(t, x, z) \leq \rho\}.$$

There is an $r > 0$ such that $(\tau, \xi(\tau; t, x)) \in \mathcal{A}(\zeta(\tau), -\delta/2)$ for all $(t, x) \in B_r(s, y)$ and for all $s \le t \le \tau \le s + r$. Let $\varepsilon > 0$ such that $\varepsilon < r$ and select $(s_\varepsilon, y_\varepsilon) \in B_r(s, y) \cap \mathcal{A}(z_\delta, -\delta/2)$ such that

$$V(s_\varepsilon, y_\varepsilon) > \varphi(s_\varepsilon, y_\varepsilon) - \varepsilon^2.$$

Using the dynamic programming principle (2.12) we have

$$\varphi(s_\varepsilon, y_\varepsilon) - \varepsilon^2 < V(s_\varepsilon, y_\varepsilon) \le V(s_\varepsilon + \varepsilon, \xi(s_\varepsilon + \varepsilon; s_\varepsilon, y_\varepsilon)) \le \varphi(s_\varepsilon + \varepsilon, \xi(s_\varepsilon + \varepsilon; s_\varepsilon, y_\varepsilon)).$$

From this and (2.13) we get, for small $\varepsilon > 0$, that

$$\begin{aligned} -\varepsilon^2 &< \varphi(s_\varepsilon + \varepsilon, \xi(s_\varepsilon + \varepsilon; s_\varepsilon, y_\varepsilon)) - \varphi(s_\varepsilon, y_\varepsilon) \\ &= \int_{s_\varepsilon}^{s_\varepsilon + \varepsilon} \varphi_t(\tau, \xi(\tau)) + D_x\varphi(\tau, \xi(\tau)) \cdot f(\tau, \xi(\tau), \zeta(\tau))\, d\tau \\ &\le -\delta/2\, \varepsilon. \end{aligned}$$

This leads to a contradiction if we let $\varepsilon \to 0$ and so we conclude that V is indeed a subsolution.

To prove uniqueness and continuity we will use the following comparison principle.

PROPOSITION 2.4. *If $u : \Omega \to R^1$ is an upper semicontinuous subsolution and $v : \Omega \to R^1$ is a lower semicontinuous supersolution of* (2.10), *both of which satisfy the terminal condition* (2.11), *then $u \le v$ on Ω.*

Proof. We only sketch the proof since it is similar to the proof of uniqueness in [4] until we get near the end.

By assumption, in the viscosity sense, u solves

$$u_t + H^*(t, x, 0, D_x u) \ge 0, \qquad (t, x) \in \Omega,$$

and v solves

$$v_t + H_*(t, x, 0, D_x v) \le 0, \qquad (t, x) \in \Omega.$$

If $\beta > 0, M > 0$ and we set $v'(t, x) = v(t, x) + \frac{\beta}{t} + M\beta(T - t) + \beta g_R(|x|)$, where $g_R \in C^1(R^1)$ satisfies $0 \le dg_R(r)/dr \le 1$, $g_R(r) = 0$ if $r \le R$, and $g_R(r) \to +\infty$ as $r \to \infty$. Then $v' \ge v$ and it is easy to check that v' is a solution of

$$v'_t + H_*(t, x, 0, D_x v') + \frac{\beta}{t^2} \le 0, \qquad (t, x) \in \Omega.$$

We refer to [4, Thm. 4.2, p. 1086] for similar details. Set $w(t, x, y) = u(t, x) - v'(t, y)$. Since u is a subsolution and v' is a supersolution, w is a viscosity solution of

$$w_t + H^*(t, x, 0, D_x w) - H_*(t, y, 0, -D_y w) - \frac{\beta}{t^2} \ge 0. \tag{2.14}$$

We want to prove that $u \le v'$ on Ω. For the sake of contradiction, suppose that

$$u(t', x') - v'(t', x') = \sup_{(t,x)\in\Omega} (u - v') > 0.$$

The presence of g_R in v' guarantees that the supremum is achieved even though $\Omega = [0, T] \times R^n$ is unbounded. In addition the terms $M\beta(T - t)$ and β/t allow us to assume that $0 < t' < T$. Consider

$$M_\alpha = \sup_{(t,x,y)} \left(w(t, x, y) - \frac{\alpha}{2}|x - y|^2 \right)$$

for $\alpha > 0$. Then $M_\alpha < +\infty$ for large α, and if $(t_\alpha, x_\alpha, y_\alpha)$ satisfies

$$\lim_{\alpha\to\infty}\left(M_\alpha - w(t_\alpha, x_\alpha, y_\alpha) + \frac{\alpha}{2}|x_\alpha - y_\alpha|^2\right) = 0,$$

then (i) $\alpha|x_\alpha - y_\alpha|^2 \to 0$ and (ii) $M_\alpha \to u(t', x') - v'(t', x')$. Using the definition of subsolution with the test function $\frac{\alpha}{2}|x - y|^2$, at $(t, x, y) = (t_\alpha, x_\alpha, y_\alpha)$ we have from (2.14) that

$$H^*(t, x, 0, \alpha(x - y)) - H_*(t, y, 0, \alpha(x - y)) - \frac{\beta}{t^2} \geq 0.$$

Using the fact that

$$H^*(t, x, 0, \alpha(x - y)) = H(t, x, 0 - 0, \alpha(x - y)) = H(t, x, 0, \alpha(x - y)),$$
$$H_*(t, y, 0, \alpha(x - y)) = H(t, y, 0 + 0, \alpha(x - y)) = H(t, y, 0, \alpha(x - y)),$$

we obtain

$$\begin{aligned} 0 &\leq H(t, x, 0, \alpha(x - y)) - H(t, y, 0, \alpha(x - y)) - \frac{\beta}{t^2} \\ &\leq H(t, x, 0, \alpha(x - y)) - \alpha(x - y) \cdot f(t, y, \overline{z}) - \frac{\beta}{t^2} \\ &\leq H(t, x, 0, \alpha(x - y)) - \alpha(x - y) \cdot f(t, x, \overline{z}) + K_f\alpha|x - y|^2 - \frac{\beta}{t^2}, \end{aligned}$$

where $\overline{z} \in \{z : h(t, y, z) \leq 0\}$ achieves the minimum in $H(t, y, 0, \alpha(x - y))$. We have used the Lipschitz continuity of f in the last line.

Now suppose that $h(t, y, \overline{z}) = 0$; the case when $h(t, y, \overline{z}) < 0$ is similar but easier. Using assumption (B), there exists $\widehat{z}$ so that

$$|\widehat{z} - \overline{z}| \leq K_f|x - y| \text{ and } h(t, y, \widehat{z}) \leq -2K_h|x - y|. \tag{2.15}$$

Then

$$h(t, x, \widehat{z}) = h(t, y, \widehat{z}) + h(t, x, \widehat{z}) - h(t, y, \widehat{z}) \leq -2K_h|x-y| + K_h|x-y| = -K_h|x-y| < 0.$$

Consequently, $\widehat{z}$ is a member of the set over which the minimum is taken in the Hamiltonian $H(t, x, 0, \alpha(x - y))$. Therefore, using (2.15),

$$\begin{aligned} 0 &\leq H(t, x, 0, \alpha(x - y)) - \alpha(x - y) \cdot f(t, x, \overline{z}) + K_f\alpha|x - y|^2 - \frac{\beta}{t^2} \\ &\leq \alpha(x - y) \cdot f(t, x, \widehat{z}) - \alpha(x - y) \cdot f(t, x, \overline{z}) + K_f\alpha|x - y|^2 - \frac{\beta}{t^2} \\ &\leq K_f\alpha|x - y|\,|\widehat{z} - \overline{z}| + K_f\alpha|x - y|^2 - \frac{\beta}{t^2} \\ &\leq K\alpha|x - y|^2 - \frac{\beta}{t^2}. \end{aligned}$$

Since $\alpha|x - y|^2 \to 0$, we have reached a contradiction. This contradiction tells us that $u \leq v' = v + o_\beta(1)$, and since $\beta > 0$ was arbitrary, $u \leq v$. □

Completing the proof of the theorem, we have shown that V^* is an upper semicontinuous subsolution and V_* is a lower semicontinuous supersolution, both satisfying the terminal

condition (2.11). By the proposition, we conclude that $V^* \leq V_*$ and therefore that V is continuous. Uniqueness is also an immediate consequence of the proposition. □

Remark 2.2. If we have $m > 1$ constraint functions—say, $h_1(t, x, z), \ldots, h_m(t, x, z)$—then we generalize assumption (B) to the condition that

(B′) The matrix

$$\left(\frac{\partial h_i}{\partial z_j}\right), \qquad i = i_1, \ldots, i_r, \ j = 1, 2, \ldots, q, \tag{2.16}$$

has rank r, where $i_1, \ldots, i_r$ are the indices i from $1, 2, \ldots, m$ such that $h_i(t, x, z) = 0$.

In the remainder of this paper we will assume that we have $m + 1$ constraint functions with $h_i = h_i(z)$ for $i = 1, 2, \ldots, m$ and $h_{m+1} = h(t, x, z)$. We will set

$$Z = \{z \in R^q : h_i(z) \leq 0, \ i = 1, 2, \ldots, m\}, \tag{2.17}$$

which is assumed to be compact in R^q. The set Z is the usual compact control set of control theory, often taking the form $Z = \{z \in R^q : |z| \leq M\}$. A set of this form is easily expressed using smooth functions $h_1, \ldots, h_m$.

Assuming the rank condition (2.16) for the continuously differentiable functions $h_1(z)$, $\ldots, h_m(z)$, $h(t, x, z)$, we prove, using the method of this section, that the value function with these constraints on the controls is the unique continuous viscosity solution of

$$V_t + H(t, x, 0, D_x V) = 0, \qquad (t, x) \in \Omega, \tag{2.18}$$

and terminal condition (2.11), but now since

$$\{z \in R^q : h_1(z) \leq 0, \ldots, h_m(z) \leq 0, h(t, x, z) \leq 0\} = \{z \in Z : h(t, x, z) \leq 0\},$$

$$H(t, x, r, p) = \min_{\{z \in Z : h(t,x,z) \leq r\}} p \cdot f(t, x, z).$$

The value function is given by $V(t, x) = \inf_{\zeta \in \mathcal{Z}_h[t,T]} g(\xi(T))$, but now we take

$$\mathcal{Z}_h[t, T] = \{\zeta : [t, T] \to Z : h(\tau, \xi(\tau), \zeta(\tau)) \leq 0, \text{ a.e. } t \leq \tau \leq T\}.$$

In other words, we have the admissible controls taking values in the compact set $Z \subset R^q$. We no longer need to assume that the level sets of $h(t, x, z)$ in the z variable, i.e., $\{z \in Z : h(t, x, z) \leq r\}$, are compact because this will follow from the assumption that Z is compact.

3. The relaxed problem. Since our aim in this paper is to determine the relaxed version of the problem just formulated, we need to introduce the space of relaxed controls and the relaxed dynamics.

We assume throughout the remainder of this paper that condition (A) and the assumption (B′) on the constraint functions given in Remark 2.2 hold, as does (2.7).

Let $M(Z)$ denote the space of bounded measures on the compact control set $Z = \{z \in R^q : h_1(z) \leq 0, \ldots, h_m(z) \leq 0\}$ and $\mathcal{PM}(Z)$ be the set of probability measures on Z. Viewing $M(Z)$ as the dual space of $C(Z) =$ continuous functions on Z, we endow $M(Z)$ and $\mathcal{PM}(Z)$ with the weak star topology of $C(Z)^*$. For any $\mu \in \mathcal{PM}(Z)$ define the functions

$$\widehat{f}(t, x, \mu) = \int_Z f(t, x, z) \, \mu(dz), \tag{3.1}$$

and

$$\widehat{h}(t, x, \mu) = \mu - \operatorname*{ess\,sup}_{z \in Z} h(t, x, z). \tag{3.2}$$

This means that we take the essential supremum in $z \in Z$ of $h(\cdot, \cdot, z)$ with respect to the measure μ.

Let $\widehat{\mathcal{Z}}[t, T] = L^\infty([t, T]; \mathcal{PM}(Z))$, and define the space of feasible relaxed controls by

$$\widehat{\mathcal{Z}}_h[t, T] = \{\mu \in \widehat{\mathcal{Z}}[t, T] : \widehat{h}(s, \widehat{\xi}(s), \mu(s)) \leq 0, \text{ a.e. } t \leq s \leq T\},$$

where $\widehat{\xi}(\cdot)$, for any control $\mu \in \widehat{\mathcal{Z}}[t, T]$, is the unique relaxed trajectory given by

$$\widehat{\xi}(\tau) = x + \int_t^\tau \int_Z f(s, \widehat{\xi}(s), z)\, \mu(s, dz)ds. \tag{3.3}$$

Remark 3.1. We proved in [3] that, for each (t, x) fixed, the mapping

$$\mu \in \mathcal{PM}(Z) \mapsto \widehat{h}(t, x, \mu)$$

is weakly sequentially lower semicontinuous. The continuity properties of $\widehat{h}$ in the (t, x) variables are the same as those of h. In general, however, $\widehat{h}$ is not continuously differentiable.

DEFINITION 3.1. *The relaxed value function associated with the constrained problem is*

$$\widehat{V}(t, x) = \inf_{\mu(\cdot) \in \widehat{\mathcal{Z}}_h[t,T]} g(\widehat{\xi}(T)). \tag{3.4}$$

Consider the set

$$\widehat{\Omega} = \left\{(t, x) \in [0, T] \times R^n : \min_{\mu \in \mathcal{PM}(Z)} \widehat{h}(t, x, \mu) < 0\right\}, \tag{3.5}$$

and set

$$\widehat{\gamma}(t, x) = \min_{\mu \in \mathcal{PM}(Z)} \widehat{h}(t, x, \mu). \tag{3.6}$$

PROPOSITION 3.1. $\Omega = \widehat{\Omega} = [0, T] \times R^n$.

Proof. The proof follows from the fact that

$$\widehat{\gamma}(t, x) = \gamma(t, x) = \min_{z \in Z} h(t, x, z) \qquad \forall (t, x) \in [0, T] \times R^n. \tag{3.7}$$

To see that this is true, we have that, for any $\mu \in \mathcal{PM}(Z)$,

$$\widehat{h}(t, x, \mu) = \mu - \operatorname{ess\,sup}_{z \in Z} h(t, x, z) \geq \min_{z \in Z} h(t, x, z),$$

and so $\widehat{\gamma}(t, x) \geq \gamma(t, x)$. If z_0 satisfies $\gamma(t, x) = h(t, x, z_0)$, let $\mu_0 \in \mathcal{PM}(Z)$ be the Dirac measure on Z concentrated at z_0, $\mu_0 = \delta_{z_0}$. Then

$$\widehat{\gamma}(t, x) \leq \widehat{h}(t, x, \mu_0) = h(t, x, z_0) = \gamma(t, x),$$

and (3.7) is proven. □

THEOREM 3.2. *Assume* (A), *the conditions of Remark* 2.2, *and* (2.7). $\widehat{V}$ *is a continuous viscosity solution of the Bellman equation* (2.18) *satisfying the terminal condition* (2.11).

Proof. Since any ordinary control $\zeta \in \mathcal{Z}_h[t, T]$ can be viewed as a relaxed control (by taking the relaxed control to be the Dirac measure concentrated on ζ, $\mu(\tau) = \delta_{\zeta(\tau)}$), it is immediate that $\widehat{V}(t, x) \leq V(t, x)$ $\forall (t, x) \in \Omega$. Assumption (A) implies that $\widehat{V}$ is bounded below. Thus, $\widehat{V}$ is locally bounded. Also, it is clear that $\widehat{V}(T, x) = g(x)$.

Define the relaxed Hamiltonian

$$\widehat{H}(t,x,r,p) = \min_{\{\mu\in\mathcal{PM}(Z):\widehat{h}(t,x,\mu)\le r\}} p\cdot\widehat{f}(t,x,\mu). \tag{3.8}$$

Then, just as in the unrelaxed case,

$$\widehat{H}^*(t,x,r,p) = \widehat{H}(t,x,r-0,p) \text{ and } \widehat{H}_*(t,x,r,p) = \widehat{H}(t,x,r+0,p).$$

We may not yet conclude continuity of the Hamiltonians because we have not imposed condition (B′) on $\widehat{h}$. To obtain continuity we will need the following lemma relating the relaxed Hamiltonian with the ordinary Hamiltonian.

LEMMA 3.3. *$\widehat{H}(t,x,r,p) = H(t,x,r,p)$, i.e.,*

$$\min_{\{\mu\in\mathcal{PM}(Z):\widehat{h}(t,x,\mu)\le r\}} p\cdot\widehat{f}(t,x,\mu) \;=\; \min_{\{z\in Z:h(t,x,z)\le r\}} p\cdot f(t,x,z). \tag{3.9}$$

Proof of Lemma 3.3. First, if the set $\{\mu\in\mathcal{PM}(Z) : \widehat{h}(t,x,\mu)\le r\}$ is nonempty, then it contains a measure μ such that for μ a.e. $z\in Z$, $h(t,x,z)\le r$. Therefore, the set $\{z\in Z : h(t,x,z)\le r\}$ also is nonempty. The converse statement is also clearly true by using Dirac measures. Without loss of generality we may therefore assume the sets are nonempty since otherwise the Hamiltonians are both $+\infty$. Now $\widehat{H}(t,x,r,p)\le H(t,x,r,p)$ since, for each z_0 such that $h(t,x,z_0)\le r$, we can let $\mu=\delta_{z_0}$. To see the opposite inequality, let $\varepsilon>0$ be given and $\mu^*\in\mathcal{PM}(Z)$ satisfy

$$\widehat{H}(t,x,r,p)\ge p\cdot\widehat{f}(t,x,\mu^*)-\varepsilon = \int_Z p\cdot f(t,x,z)\mu^*(dz)-\varepsilon$$

and $\mu^*-\operatorname{ess\,sup}_z h(t,x,z)\le r$. Let $B=\{z\in Z : h(t,x,z)\le r\}$. Then $\mu^*(B)=1$, and

$$\begin{aligned}\widehat{H}(t,x,r,p) &\ge \int_Z p\cdot f(t,x,z)\mu^*(dz)-\varepsilon\\ &= \int_B p\cdot f(t,x,z)\mu^*(dz)-\varepsilon\\ &\ge \min_{z\in B} p\cdot f(t,x,z)-\varepsilon = H(t,x,r,p)-\varepsilon.\end{aligned}$$

Therefore, (3.9) is true. □

One consequence of the lemma is that even though $\widehat{h}(t,x,\mu)$ is not a continuously differentiable function, at least for the purpose of calculating the Hamiltonian we may use the continuously differentiable function h from which $\widehat{h}$ is defined.

Returning to the proof of the theorem, we begin the proof that $\widehat{V}$ is a subsolution of (2.18). Assume that $\widehat{V}^*-\varphi$ achieves a zero maximum at $(s,y)\in\Omega=\widehat{\Omega}$. If $\widehat{V}$ is not a subsolution, then there is $\delta>0$ so that

$$\varphi_t(s,y)+H(s,y,-\delta,D_x\varphi(s,y)) = \varphi_t(s,y)+\widehat{H}(s,y,-\delta,D_x\varphi(s,y))\le -\delta.$$

Observe that we have used the fact that H and $\widehat{H}$ are identical. Now we may use the dynamic programming principle for the relaxed value, i.e.,

$$\widehat{V}(s,y) = \inf_{\{\mu\in\widehat{\mathcal{Z}}[s,\sigma]\}} \widehat{V}(\sigma,\widehat{\xi}(\sigma;s,y)), \qquad s\le\sigma\le T,$$

and we complete the proof that $\widehat{V}$ is a subsolution of (2.18) in a manner entirely similar to that of Theorem 2.3. In this proof we use a control $\zeta(\tau)=z_\delta$ and the relaxed control $\mu(\tau)=\delta_{\zeta(\tau)}$.

The assumption (B′) is used exactly as in the proof of Theorem 2.3 to achieve a contradiction. The continuity of $\widehat{V}$ also follows from the comparison principle once we know that $\widehat{V}$ is a viscosity solution. □

Once we know that $\widehat{V}$ is a viscosity solution of (2.18) and satisfies the terminal condition (2.11), the uniqueness of viscosity solutions is used to yield the following relaxation theorem.

COROLLARY 3.4. *The relaxed value function $\widehat{V}$ and the ordinary value function V are identical on Ω.*

The reader may well ask whence comes our definition of the relaxed problem and, in particular, the use of the function $\widehat{h}$. The following theorem provides the answer to that question.

Denote by $a^+ = \max\{a, 0\}$.

THEOREM 3.5. *Assume* (A), *the conditions of Remark* 2.2, *and* (2.7). *Define, for each* $n = 1, 2, \ldots$, *the value function for the unconstrained problem*

$$V_n(t,x) = \inf_{\zeta \in \mathcal{Z}[t,T]} \left\{ g(\xi(T)) + n \cdot \int_t^T h^+(s, \xi(s), \zeta(s))\, ds \right\},$$

where ξ is the trajectory starting from $x \in R^n$ at time t. Here $\mathcal{Z}[t,T] = \{\zeta : [t,T] \to Z\}$, and recall that $Z = \{z \in R^q : h_i(z) \le 0, 1 \le i \le m\}$. Let $\widehat{V}_n(t,x)$ be the classical unconstrained relaxed value:

$$\widehat{V}_n(t,x) = \inf_{\mu \in \widehat{\mathcal{Z}}[t,T]} \left\{ g(\widehat{\xi}(T)) + n \cdot \int_Z \int_t^T h^+(s, \widehat{\xi}(s), z)\, \mu(s, dz)\, ds \right\}.$$

Then, for $(t,x) \in [0,T] \times R^n$,

$$\lim_{n \to \infty} V_n(t,x) = V(t,x) \tag{3.10}$$

and

$$\lim_{n \to \infty} \widehat{V}_n(t,x) = \widehat{V}(t,x). \tag{3.11}$$

Proof. To prove (3.10), we know (see [19]) that V_n is the unique continuous viscosity solution of

$$\frac{\partial V_n}{\partial t} + \min_{z \in Z} \left(D_x V_n \cdot f(t,x,z) + n \cdot h^+(t,x,z) \right) = 0, \qquad (t,x) \in [0,T) \times R^n,$$

with terminal condition $V_n(T,x) = g(x)$. Set

$$H_n(t,x,p) = \min_{z \in Z} \left(p \cdot f(t,x,z) + n \cdot h^+(t,x,z) \right).$$

Then, by classical penalization arguments (cf. [4] and [6]), assuming that all of the Hamiltonians are finite,

$$\limsup_{(n,s,y,q) \to (\infty,t,x,p)} H_n(s,y,q) = H(t,x,0-0,p)$$

and

$$\liminf_{(n,s,y,q) \to (\infty,t,x,p)} H_n(s,y,q) = H(t,x,0+0,p).$$

Recall that under our assumptions, V is locally uniformly bounded. The stability result of Barles and Perthame (see Fleming and Soner [19, p. 288]) then gives us (3.10).

To prove (3.11), we know that $\widehat{V}_n$ is the continuous viscosity solution of

$$\frac{\partial V_n}{\partial t} + \min_{\mu \in \mathcal{PM}(Z)} \left(D_x \widehat{V}_n \cdot \widehat{f}(t, x, \mu) + n \cdot \int_Z h^+(t, x, z) \mu(dz) \right) = 0, \ (t, x) \in [0, T) \times R^n,$$

with $\widehat{V}_n(T, x) = g(x)$. Now, setting

$$\widehat{H}_n(t, x, r, p) = \min_{\mu \in \mathcal{PM}(Z)} p.\widehat{f}(t, x, \mu) + n \cdot \int_Z (h(t, x, z) - r)^+ \mu(dz),$$

we verify easily that

$$\limsup_{(n,s,y,b,q) \to (\infty,t,x,r,p)} \widehat{H}_n(s, y, b, q) = \min_{\{\mu \in \mathcal{PM}(Z): \int_Z (h(t,x,z)-r+0)^+ \mu(dz) \le 0\}} p.\widehat{f}(t, x, \mu)$$

and

$$\liminf_{(n,s,y,b,q) \to (\infty,t,x,r,p)} \widehat{H}_n(s, y, b, q) = \min_{\{\mu \in \mathcal{PM}(Z): \int_Z (h(t,x,z)-r-0)^+ \mu(dz) \le 0\}} p.\widehat{f}(t, x, \mu).$$

Indeed to verify the first claim, we have, for all $(s, y, b, q) \in B_\varepsilon(t, x, r, p)$,

$$\min_{\{\mu \in \mathcal{PM}(Z): \int_Z (h(s,y,z)-b)^+ \mu(dz) \le 0\}} q \cdot \widehat{f}(s, y, \mu)$$

$$= \lim_{n \to \infty} \min_{\mu \in \mathcal{PM}(Z)} \left(q.\widehat{f}(s, y, \mu) + n \cdot \int_Z (h(s, y, z) - b)^+ \mu(dz) \right)$$

$$\le \lim_{n \to \infty} \min_{\mu \in \mathcal{PM}(Z)} \left(p.\widehat{f}(t, x, \mu) + n \cdot \int_Z (h(t, x, z) - r + (K_h + 1)\varepsilon)^+ \mu(dz) + K_f \varepsilon + |p|\varepsilon \right)$$

$$= \min_{\{\mu \in \mathcal{PM}(Z): \int_Z (h(t,x,z)-r+(K_h+1)\varepsilon)^+ \mu(dz) \le 0\}} \left(p \cdot \widehat{f}(t, x, \mu) + K_f \varepsilon + |p|\varepsilon \right).$$

Since $\varepsilon > 0$ was arbitrary we have shown that

$$\limsup_{(n,s,y,b,q) \to (\infty,t,x,r,p)} \widehat{H}_n(s, y, b, q) \le \min_{\{\mu \in \mathcal{PM}(Z): \int_Z (h(t,x,z)-r+0)^+ \mu(dz) \le 0\}} p.\widehat{f}(t, x, \mu).$$

The reverse inequality is immediate, so the claim is verified.

Now $\int_Z (h(t, x, z) - r + 0)^+ \mu(dz) \le 0$, if and only if $\mu - \operatorname{ess\,sup}_z h(t, x, z) < r$ and so

$$\limsup_{(n,s,y,b,q) \to (\infty,t,x,r,p)} \widehat{H}_n(s, y, b, q)$$

$$= \min_{\{\mu \in \mathcal{PM}(Z) \,:\, \mu - \operatorname{ess\,sup}_z h(t,x,z) \le r-0\}} p.\widehat{f}(t, x, \mu)$$

$$= \widehat{H}^*(t, x, r, p).$$

Similarly,

$$\liminf_{(n,s,y,b,q)\to(\infty,t,x,r,p)} \widehat{H}_n(s, y, b, q)$$

$$= \min_{\{\mu\in\mathcal{PM}(Z)\,:\,\mu-\operatorname{ess\,sup}_z h(t,x,z)\le r+0\}} p.\widehat{f}(t, x, \mu)$$

$$= \widehat{H}_*(t, x, r, p).$$

Once again the Hamiltonian for $\widehat{V}_n$ converges correctly to the Hamiltonian for $\widehat{V}$ and the theorem of Barles and Perthame allows us to conclude. Without loss of generality we are assuming that all of the Hamiltonians in the proof are finite to avoid trivial cases. □

Remark 3.2. It is apparent from Theorem 3.5 that the natural approach to relaxing the constrained problem, i.e., convexifying h by taking $\int_Z h(t, x, z)\mu(dz)$, is *not* correct. The theorem shows that we must relax h^+, not h. The distinction comes from the fact that $(\int_Z h(t, x, z)\mu(dz))^+ \neq \int_Z h^+(t, x, z)\mu(dz)$.

We have proved so far that the relaxed value defined in this paper and the original value function will coincide. Thus, even though the class of controls is enlarged to guarantee the existence of an optimal control, this enlargement will not decrease the value. Furthermore, the fact that the values are the same means that it will be possible to approximate the relaxed value function with ordinary controls and trajectories. This is the first part of establishing that the formulation of the relaxed problem in this paper is the correct one.

The second part now consists of showing that an optimal relaxed control will always exist in our formulation. If we do that, then our formulation will satisfy the properties required in enlarging the class of controls to give us existence of an optimal control.

We recall here the definition of a quasi-convex function. A quasi-convex function—say, $g : X \to R^1$, X a convex set—satisfies

$$g(\lambda x + (1-\lambda)y) \le \max\{g(x), g(y)\}, \qquad 0 < \lambda < 1, \quad x, y \in X.$$

Equivalently, g is quasi-convex if the r-level set of g, $\{x \in X : g(x) \le r\}$ is convex for all $r \in R^1$.

It was proven in [3] that $\mu \mapsto \widehat{h}(t, x, \mu)$ is a quasi-convex function. In fact,

$$\widehat{h}(t, x, \lambda\mu_1 + (1-\lambda)\mu_2) = \max\{\widehat{h}(t, x, \mu_1), \widehat{h}(t, x, \mu_2)\} \tag{3.12}$$

for $0 < \lambda < 1$. Furthermore, it is obvious that $\mu \mapsto \widehat{f}(t, x, \mu)$ is linear and weakly continuous. The continuity properties of $\widehat{f}$ and $\widehat{h}$ in the (t, x) variables are inherited from f and h.

THEOREM 3.6. *For each fixed* $(t, x) \in \Omega$, *there exists an optimal relaxed constrained control* $\mu^*(\cdot) \in \widehat{\mathcal{Z}}_h[t, T]$.

Proof. Let $\{\mu_n\} \subset \widehat{\mathcal{Z}}_h[t, T]$ be a minimizing sequence and $\widehat{\xi}_n$ the associated relaxed trajectory for each $n = 1, 2, \ldots$, starting from $x \in R^n$ at time $t \ge 0$. Then, we know that on a subsequence $\mu_n \rightharpoonup \mu^*$, weak-* for some $\mu^* \in \widehat{\mathcal{Z}}[t, T]$, as well as $\widehat{\xi}_n \to \widehat{\xi}^*$, uniformly on $[t, T]$, with $\widehat{\xi}^*$ the trajectory associated with μ^*. We first need to show that $\mu^* \in \widehat{\mathcal{Z}}_h[t, T]$, i.e., that the constraint $\widehat{h}(s, \widehat{\xi}^*(s), \mu^*(s)) \le 0$ for a.e. $t \le s \le T$, is satisfied. This constraint is equivalently expressed as

$$\lambda - \operatorname*{ess\,sup}_{t\le s\le T} \widehat{h}(s, \widehat{\xi}^*(s), \mu^*(s)) \le 0, \tag{3.13}$$

where λ denotes Lebesgue measure on $[0, T]$. It was proved in [3] and [5] (see Remark 3.3) that the functional $F : \widehat{\mathcal{Z}}[t, T] \to R^1$ defined by

$$F(\mu) = \lambda - \operatorname*{ess\,sup}_{t\le s\le T} \widehat{h}(s, \widehat{\xi}(s), \mu(s))$$

is weakly lower semicontinuous if $\widehat{h}(\cdot, \cdot, \mu)$ is quasi-convex. Since (3.12) shows that this function is indeed quasi-convex, and since $\mu_n \in \widehat{\mathcal{Z}}_h[t, T]$, we have that

$$F(\mu^*) \leq \liminf_{n\to\infty} F(\mu_n) \leq 0,$$

and (3.13) is proven.

Finally, since $\widehat{\xi}_n(T) \to \widehat{\xi}^*(T)$,

$$\widehat{V}(t, x) = \lim_{n\to\infty} g(\widehat{\xi}_n(T)) = g(\widehat{\xi}^*(T)) \geq \widehat{V}(t, x),$$

which says that (μ^*, ξ^*) is optimal. □

Remark 3.3. We will sketch a proof of the fact that $u \in L^\infty([0, T]; R^n) \mapsto \operatorname{ess\,sup} h(u(\tau))$ is weakly lower semicontinuous when h is quasi-convex. The proof here is modeled after a similar statement for integrals of convex functions in [14]. The proof given in [3] is based on Mazur's lemma. Set $G(u) = \operatorname{ess\,sup}_{0\leq\tau\leq T} h(u(\tau))$. Suppose that $u_k \to u$ weak-* in $L^\infty([0, T], R^n)$. We assume that h is represented as

$$h(p) = \max_{1\leq j\leq m} (q_j \cdot p + d_j) \wedge c_j, \qquad q_j \in R^n,\ c_j \in R^1,\ 1 \leq j \leq m < \infty.$$

Indeed, it is proven in [5] that any quasi-convex function can be represented as $h(q) = \sup_{p,c} ((q \cdot p - h^*(p, c)) \wedge c)$, where $h^*(p, c) = \left(\sup_q (p \cdot q - h(q)) \wedge c\right)$. Set $E_j = \{\tau \in [0, T] : h(u(\tau)) = (q_j \cdot u(\tau) + d_j) \wedge c_j\}$. Then $[0, T] = \cup_j E_j$, and we assume that the sets are disjoint. Then, since $u_k \to u$ weak-*,

$$\begin{aligned} G(u) = \operatorname*{ess\,sup}_{0\leq\tau\leq T} h(u(\tau)) &= \max_j \operatorname*{ess\,sup}_{\tau\in E_j} h(u(\tau)) \\ &= \max_j \operatorname*{ess\,sup}_{\tau\in E_j} (q_j \cdot u(\tau) + d_j) \wedge c_j \\ &\leq \liminf_{k\to\infty} \max_j \operatorname*{ess\,sup}_{\tau\in E_j} (q_j \cdot u_k(\tau) + d_j) \wedge c_j \\ &= \liminf_{k\to\infty} G(u_k). \end{aligned}$$

A limiting argument on $m \to \infty$ then completes the proof.

Remark 3.4. Recall that $Z = \{z \in R^q : h_1(z) \leq 0, \ldots, h_m(z) \leq 0\}$. The class $\mathcal{PM}(Z)$ convexifies Z. Also,

$$\{\mu \in \mathcal{PM}(R^q) : \widehat{h}_1(\mu) \leq 0, \ldots, \widehat{h}_m(\mu) \leq 0\} = \{\mu \in \mathcal{PM}(Z)\}.$$

An equivalent formulation of the relaxed problem can be stated, using the greatest quasi-convex minorant of h. It was shown in [5] that the greatest quasi-convex minorant of a function h can be written

$$h^{**}(z) = \min \left\{ \max_{1\leq i\leq q+1} h(z_i) : z = \sum_{i=1}^{q+1} \lambda_i z_i, \lambda_i \geq 0, \sum_i \lambda_i = 1 \right\}.$$

Once we have $h^{**}(t, x, z)$, we may state the relaxed optimal control problem using chattering controls as in [7].

Acknowledgment. We are grateful to the referees for pointing out several simplifications in the use of the set Ω.

REFERENCES

[1] G. BARLES AND B. PERTHAME, *Discontinuous solutions of deterministic optimal stopping time problems*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 557–579.
[2] E. N. BARRON, *Optimal control and calculus of variations in L^∞*, in Optimal Control of Differential Equations, N. H. Pavel, ed., Marcel Dekker, New York, 1994.
[3] E. N. BARRON AND R. JENSEN, *Relaxed minimax control*, SIAM J. Control Optim., 33 (1995), pp. 1028–1039.
[4] E. N. BARRON AND H. ISHII, *The Bellman equation for minimizing the maximum cost*, Nonlinear Anal., 13 (1989), pp. 1067–1090.
[5] E. N. BARRON AND W. LIU, *Calculus of variations in L^∞*, Appl. Math. Optim., to appear.
[6] E. N. BARRON, R. JENSEN, AND J. L. MENALDI, *Optimal control and differential games with measures*, Nonlinear Anal., 21 (1993), pp. 241–268.
[7] L. J. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.
[8] I. CAPUZZO-DOLCETTA, *On a discrete approximation of the Hamilton Jacobi equation of dynamic programming*, Appl. Math. Optim., 10 (1983), pp. 367–377.
[9] I. CAPUZZO-DOLCETTA AND P.-L. LIONS, *Viscosity solutions of Hamilton-Jacobi-Bellman equations and state constraints*, Trans. Amer. Math. Soc., 318 (1990), pp. 643–683.
[10] L. CESARI, *Optimization Theory and Application*, Springer-Verlag, New York, 1983.
[11] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
[12] M. G. CRANDALL, L. C. EVANS, AND P.-L. LIONS, *Some properties of viscosity solutions of Hamilton Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.
[13] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, New York, 1976.
[14] L. C. EVANS, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, CBMS Regional Conf. Ser. in Math. 74, American Mathematical Society, Providence, RI, 1990.
[15] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley & Sons, New York, 1966.
[16] H. ISHII, *Hamilton Jacobi equations with discontinuous hamiltonians on arbitrary open sets*, Bull. Fac. Sci. Engrg. Chuo Univ. Ser. I Math, 28 (1985), pp. 33–77.
[17] H. ISHII AND S. KOIKE, *A new formulation of state constraint problems for first-order PDEs*, SIAM J. Control Optim., 34 (1996), pp. 554–571.
[18] H. M. SONER, *Optimal control with state space constraint* I, SIAM J. Control Optim., 24 (1986), pp. 552–562.
[19] W. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[20] P. LORETI, *Some properties of constrained viscosity solutions of Hamilton Jacobi Bellman equations*, SIAM J. Control Optim., 25 (1987), pp. 1244–1252.
[21] E. MASCOLO AND L. MIGLIACCIO, *Relaxation in control theory*, Appl. Math. Optim., 20 (1989), pp. 97–103.
[22] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 2092–2115, November 1996

013

SOLVABILITY AND RIGHT-INVERSION OF IMPLICIT NONLINEAR DISCRETE-TIME SYSTEMS*

T. FLIEGNER†, Ü. KOTTA‡, AND H. NIJMEIJER§

Abstract. In this paper the problems of solvability and right-invertibility for implicit nonlinear discrete-time control systems are investigated. The concept "solvability" is defined in such a way that consistency of the implicit system equations is locally guaranteed for all input sequences, and an algorithm is introduced to verify the solvability of an implicit system in that sense. It is demonstrated how this mechanism may be used to decide on the right-invertibility or functional reproducibility of a given system. In contrast to previous work on right-invertibility for special classes of implicit nonlinear systems, the approach is not restricted to the characterization of right-invertibility, but it is shown in addition how an inverse system can actually be obtained. The theory is illustrated by a realistic economic example in which the inversion procedure is applied using formula manipulation.

Key words. implicit nonlinear systems, solvability, right-invertibility

AMS subject classifications. 93C10, 93C55, 93B40

1. Introduction.

Controllability questions play a prominent role in the study of dynamical control systems. Among them the problem of right-inversion has, starting with the paper [4], received a lot of attention. Roughly speaking, a control system is right-invertible (also referred to as functionally reproducible or dynamic path controllable) if for all time paths in the output space, one can find an input sequence and an initial state $x(0)$ such that, applied to the system, the system generates precisely the given time path as its output. Practically, the problem consists of two parts: the decision whether or not a given system is right-invertible and, if the answer is affirmative, construction of a feedback/feedforward mechanism producing the desired inputs.

Concerning linear standard state-space systems, the problem of system inversion is satisfactorily solved. Some of the major references are [4], [18], and [19]. There also exist a number of articles dealing with the inversion of standard explicit nonlinear systems in continuous time, such as [16], [17], and [20], and similar work in discrete time (see, e.g., [11], [15]).

In this paper we are concerned with the right-inversion of implicit nonlinear discrete-time systems, a problem which, to the best of our knowledge, has not been considered before in this form. However, note that in [6], [7] left- and right-invertibility are characterized for (implicit) rational systems in a differential/difference algebraic language.

We formulate necessary and sufficient conditions for right-invertibility, and an inversion algorithm is introduced which can be used to obtain a right-inverse system in case such a system exists.

The results of this paper are inspired by a series of articles [12]–[14] by Luenberger in which he studied problems of the existence and uniqueness of solutions for both linear and nonlinear implicit discrete-time systems. As suggested by Luenberger (but not further pursued), we tackle the inversion problem as the solvability problem of the system with respect to the state variables and control variables; that is, we ask whether there exists for every given output sequence a sequence of states and inputs, respectively, such that the system equations are satisfied. For this purpose we generalize the shuffle algorithm given in [14] for linear

*Received by the editors January 1, 1995; accepted for publication (in revised form) October 3, 1995.

†Department of Engineering, University of Cambridge, Cambridge CB21PZ, UK (tf3@eng.cam.ac.uk).

‡Institute of Cybernetics, Estonian Academy of Sciences, Akadeemia tee 21, EE0026 Tallinn, Estonia (Kotta@ioc.ee).

§Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands (h.nijmeijer@math.utwente.nl).

systems to the nonlinear setting. The outcome of this algorithm decides upon local solvability of the given implicit system and rearranges the system equations in a way that solutions can be obtained easily if they exist.

Note that the inversion algorithm introduced in this paper applies to implicit systems as well as to standard state-space systems. In the latter case it just reduces to the traditional inversion algorithm suggested in [11]. The order of the obtained inverse system is the same as the order of the original system.

To our knowledge, the approach suggested by Luenberger and used in this paper for constructing a right-inverse system has not been used before, even for linear systems. All the papers [2], [8], [21] seem to follow the traditional semiimplicit case where the dynamic part and the algebraic part are treated separately. This either causes a lot of trouble or leads to a very complicated inversion procedure. In [2], for instance, the author conjectures that his inversion procedure finds an inverse system whenever one exists, but a formal proof is not available. The procedure in [8] consists of three steps. At a first step, the consistency of the system equations with respect to states and control variables is checked. The second step deals with the construction of a so-called "candidate inverse system" whose order is in general greater than that of the original system. Finally, a finite iterative method is applied to the "candidate inverse" to reduce its dimension, either yielding an inverse system or providing evidence that there is no such system.

In our approach, we try to circumvent some of these disadvantages by directly attacking the problem of obtaining a right-inverse system without a preliminary test of the invertibility conditions. The shuffle algorithm provides both a simple criterion for checking invertibility and a systematic procedure for constructing a right-inverse system. Moreover, our method—in contrast to, for example, [8]—does not require state transformations.

A main part of the paper is dedicated to show that, in spite of the fact that the shuffle algorithm extensively uses the theorem on the functional dependence of functions, which, in turn, is based on the implicit function theorem, the procedure lends itself to a numerical treatment and may even be performed using formula manipulation in case the nonlinearities are not too serious.

The organization of this paper is as follows. In §2 we consider the solvability of nonlinear implicit systems over a finite time interval. Section 3 is concerned with the extension of the shuffle algorithm to nonlinear systems, and §4 connects the results of the previous sections. The problem of right-invertibility is addressed in §5. The purpose of §6 is to show in some detail how the methods introduced so far can be used to treat problems arising in economic policy making. Final remarks conclude the paper.

2. Preliminaries. In the following discussion we consider implicit nonlinear discrete-time systems described by equations of the form

$$\Sigma : \begin{cases} f(x(k+1), x(k), u(k)) &= 0, \\ h(x(k), u(k), y(k)) &= 0, \end{cases} \tag{1}$$

where, for all k, $x(k)$ belongs to some open part $\mathcal{X}$ of $\mathbb{R}^n$, the inputs $u(k)$ are in an open part $\mathcal{U}$ of $\mathbb{R}^m$, and the outputs $y(k)$ belong to some open set $\mathcal{Y}$ of $\mathbb{R}^p$. With regard to right-inversion we moreover assume $p \leq m$.

The mappings $f : \mathcal{X} \times \mathcal{X} \times \mathcal{U} \longrightarrow \mathbb{R}^s$, $s \leq n$, and $h : \mathcal{X} \times \mathcal{U} \times \mathcal{Y} \longrightarrow \mathbb{R}^p$ are supposed to be smooth; that is, all partial derivatives of the functions f and h exist and are continuous.

Remark 2.1. In the terminology of [23], systems of this type are state-space systems. Therefore, we do not hesitate to call the x variables states. In [12] and many other papers they are referred to as *descriptor variables* (or *generalized states* or *semi-states*), expressing their role in the modelling process. In order to differentiate among systems given in implicit

(respectively, explicit) state-space form, we call the latter systems in standard state-space form. Implicit systems may not be necessarily input-output systems; that is, u and y may not possess additional properties which qualify them as inputs and outputs, respectively. Despite of this we often refer to the u and y variables as inputs and outputs even if there is no justification to do so.

We work on a finite time interval $k = 0, 1, \ldots, k_F < \infty$, where k_F is supposed to be greater than n. Furthermore, we adopt a local point of view by assuming the existence of an equilibrium point for the system Σ; that means a point $(x_e, u_e, y_e) \in \mathcal{X} \times \mathcal{U} \times \mathcal{Y}$ satisfying $f(x_e, x_e, u_e) = 0$ and $h(x_e, u_e, y_e) = 0$. Throughout the paper, we work in a neighbourhood $\mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{Y}^0$ of this equilibrium point not further specified, although everything works in the same way in case Σ is given in a neighbourhood of a reference trajectory of the system (cf. §6).

Denote by $\mathcal{X}^0_{k_F}, \mathcal{U}^0_{k_F}$, and $\mathcal{Y}^0_{k_F}$ the sets of state variable sequences

$$\mathcal{X}^0_{k_F} := \{\{x(k)\}_{k=0}^{k=k_F} | x(k) \in \mathcal{X}^0, 0 \le k \le k_F\},$$

control sequences

$$\mathcal{U}^0_{k_F} := \{\{u(k)\}_{k=0}^{k=k_F-1} | u(k) \in \mathcal{U}^0, 0 \le k \le k_F - 1\},$$

and output sequences

$$\mathcal{Y}^0_{k_F} := \{\{y(k)\}_{k=0}^{k=k_F-1} | y(k) \in \mathcal{Y}^0, 0 \le k \le k_F - 1\}.$$

Individual members of these sets will be denoted $\mathbf{x}$, $\mathbf{u}$, and $\mathbf{y}$, respectively.

DEFINITION 2.2. *A triple* $(\mathbf{x}, \mathbf{u}, \mathbf{y}) \in \mathcal{X}^0_{k_F} \times \mathcal{U}^0_{k_F} \times \mathcal{Y}^0_{k_F}$ *is said to be admissible for* Σ *if it satisfies equations* (1) *on the time horizon considered.*

Remark 2.3. Since we are working on a finite time horizon, Σ may equivalently be considered, for a fixed $\mathbf{u}$, as a set of $k_F \times (s + p)$ nonlinear equations in the variables $x(0), y(0), x(1), \ldots, y(k_F - 1), x(k_F)$ (abbreviated $(\mathbf{x}, \mathbf{y})$).

Observe that in case of implicit system equations, the initial state $x(0)$ cannot be arbitrarily selected in general. This peculiarity arises because implicit system equations usually consist of a dynamic and an algebraic part of implicit equations. In the most general case, this algebraic part constitutes implicit relations which, at $k = 0$, relate the components of $x(0)$ and those of $u(0)$, including special cases such as relations between the components of $x(0)$ and $u(0)$, respectively, only. A free choice of $x(0)$ can therefore lead to inconsistencies in the system equations or impose constraints on the inputs. The latter is not desirable because the inputs are then not what is called "free" in system-theoretic terms. This unwanted situation especially occurs when there exist algebraic implicit equations relating only components of u. At any rate, performing the shuffle algorithm propagates these relations between $x(0)$ and $u(0)$, leading to equations of the form

$$L(x(0), u(0), u(1), \ldots, u(l)) = 0 \tag{2}$$

for some $l \le n$ (see Remark 3.1), and if these relations fail to hold for $x(0)$, the system equations are not satisfied. We call the initial variable $x(0)$ admissible if all the implicit algebraic relations which can be derived from Σ and which relate the components of $x(0)$ and $u(k)$, $0 \le k \le l$, hold for $x(0)$. We specify these implicit equations in §4 after extending the shuffle algorithm to nonlinear systems in §3. At the moment, we do not fix $x(0)$.

Moreover, again in contrast to systems in standard state-space form, in general neither existence nor uniqueness of solutions $(\mathbf{x},\mathbf{y})$ for Σ can be guaranteed for arbitrary control sequences $\mathbf{u} \in \mathcal{U}^0_{k_F}$ without further assumptions.

Luenberger [13] introduced the notion of local solvability for systems of the form

$$F_k(x(k+1), x(k)) = 0, \tag{3}$$

where the number $s = n$ of scalar equations of (3) equals dim x. If we interpret both x and y as descriptor variables, Σ may be considered, for a given input sequence $\mathbf{u}$, as a special case of (3), and the notion of solvability may be easily specified to our case. An input sequence $\mathbf{u}$ then merely specifies a certain set of functions F_k. Note, however, that we allow the number of scalar equations in (3) to be less than dim x.

Denote for $\alpha \in \mathbb{Z}$ the time shift operator by σ^α, that is, for instance $\sigma^\alpha x(k) = x(k+\alpha)$. In order to economize notations, we frequently drop time dependence if there is no danger of confusion.

For a given input sequence $\mathbf{u} \in \mathcal{U}^0_{k_F}$, let $\mathcal{M}_\mathbf{u}$ denote the set of pairs $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^0_{k_F} \times \mathcal{Y}^0_{k_F}$ such that (**x**,**u**,**y**) becomes an admissible triple of system Σ. Following Luenberger [13], we define the matrices

$$\begin{aligned}
A^i_{(x,u,y)} &= \frac{\partial f(\sigma x, x, u)}{\partial x}|_{x(i+1),x(i),u(i)}, \qquad 0 \le i \le k_F - 1,\\
E^i_{(x,u,y)} &= \frac{\partial f(\sigma x, x, u)}{\partial \sigma x}|_{x(i),x(i-1),u(i-1)}, \qquad 1 \le i \le k_F,\\
C^i_{(x,u,y)} &= \frac{\partial h(x, u, y)}{\partial x}|_{x(i),u(i),y(i)}, \qquad 0 \le i \le k_F - 1,\\
G^i_{(x,u,y)} &= \frac{\partial h(x, u, y)}{\partial y}|_{x(i),u(i),y(i)}, \qquad 0 \le i \le k_F - 1,
\end{aligned}$$

which are computed for every admissible (**x**,**u**,**y**). These matrices usually differ for different admissible triples. The same holds for the matrix $F_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F)$ defined by

$$F_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F) = \begin{pmatrix}
A^0 & 0 & E^1 & & & & & & \\
C^0 & G^0 & 0 & & & & & & \\
 & & A^1 & 0 & E^2 & & 0 & & \\
 & & C^1 & G^1 & 0 & & & & \\
 & & & & & \ddots & & & \\
 & & 0 & & & & A^{k_F-1} & 0 & E^{k_F} \\
 & & & & & & C^{k_F-1} & G^{k_F-1} & 0
\end{pmatrix}.$$

$F_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F)$ is of size $k_F(s+p) \times [(k_F+1)n + k_F p]$ and is referred to as *solvability matrix*. Essentially, it represents the collection of coefficient matrices of the linearization of Σ along the admissible path (**x**,**u**,**y**).

DEFINITION 2.4 (see [13]). *The system Σ is said to be locally solvable about the equilibrium point (x_e, u_e, y_e) if there exist neighbourhoods $\mathcal{X}^0$, $\mathcal{U}^0$, and $\mathcal{Y}^0$ of x_e, u_e, and y_e, respectively, such that for every control sequence $\mathbf{u} \in \mathcal{U}^0_{k_F}$ the solvability matrix $F_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F)$ has full row rank for every admissible triple.*

Remark 2.5. This definition essentially entails that if the conditions are satisfied, the solution set $\mathcal{M}_\mathbf{u}$ forms a manifold of dimension $(n-s)k_F + n$ for every $\mathbf{u} \in \mathcal{U}^0_{k_F}$ and thus displays a nice structure.

Remark 2.6. In connection with the question under which conditions nonlinear implicit state-space systems can be converted to standard state-space form, some sufficient conditions for the existence of solutions for nonlinear implicit systems that are considered over an infinite time horizon are derived in [9].

If system Σ is solvable on $0 \leq k \leq k_F - 1$, its solution is not unique. This follows from the fact that one has $(k_F+1)n+k_F p$ unknowns but only $k_F(s+p)$ equations. Thus in order to define a unique solution, $k_F(n-s)+n$ additional conditions have to be imposed to make the sequences of state vectors $\mathbf{x}$ and outputs $\mathbf{y}$ corresponding to a prescribed sequence of control vectors unique. Special cases of such additional conditions are initial ($x_{cond} = x(0)$) and final conditions ($x_{cond} = x(k_F)$) (both together also termed end-point conditions) or combinations of them. Evidently, in case of an explicitly given system in standard state-space form, where especially $n = s$, the necessary n additional conditions can always be obtained by fixing the initial value $x(0)$. For an arbitrary system Σ, an end-point conditioning is not always possible, and the degrees of freedom in the solution require one to restrict certain components of the x-vector also at intermediate points of time. Again, we specify these additional conditions in §4 after having extended the shuffle algorithm.

3. The shuffle algorithm. Strictly speaking, Definition 2.4 throws light upon two aspects of the solutions of Σ. If satisfied, it guarantees the existence of admissible triples for arbitrary $\mathbf{u} \in \mathcal{U}^0_{k_F}$, and it excludes pathological solution sets $\mathcal{M}_{\mathbf{u}}$. But it does not provide a computational means to obtain solutions.

In this section we generalize the shuffle algorithm given in [14] for linear systems to nonlinear implicit systems. We maintain the name "shuffle algorithm," although it does not become clear in the nonlinear setting where the name originates. In the linear case this name is derived from the fact that certain matrix blocks are "shuffled" from the right side to the left while performing the algorithm. As already mentioned, the result of this algorithm indicates whether or not Σ is solvable in the sense of Definition 2.4. Moreover it converts the system to a form which allows an easy—which here should be understood as recursive—computation of the solutions. The main operations of the shuffle algorithm are

1. separation of the functionally independent components in the system equations;
2. expression of the functionally dependent components by the independent ones;
3. restriction of the considered time horizon in order to avoid variables at time instants greater than k_F when time-shifting the arguments of the dependent components.

3.1. The shuffle algorithm. Before presenting the shuffle algorithm we want to draw attention to a part of the algorithm which may be difficult to grasp at first sight. It concerns the part immediately after expressing functionally dependent components by the independent ones in each step $l \geq 1$ of the algorithm. Having permuted independent and dependent components to obtain the system

$$[\Phi^{l^T}(x(k), u(k), \ldots, \sigma^{l-1}u(k)), \tilde{F}^{l^T}(\sigma x(k), x(k), u(k), \ldots, \sigma^{l-1}u(k), y(k))]^T = 0,$$

which is considered for $0 \leq k \leq k_F - l$, one temporarily interprets the system as collection of simultaneous equations in the variables $x(0), y(0), x(1), \ldots$ (cf. Remark 2.3). Then the system $F^{l+1} = 0$ is formed by deleting the equations

$$\Phi^l(x(0), u(0), \ldots, u(l-1)) = 0,$$

$$\tilde{F}^l(x(k_F - l + 1), x(k_F - l), u(k_F - l), \ldots, u(k_F - 1), y(k_F - l)) = 0.$$

This comes down to applying the time-shift operator σ to the dependent components Φ^l in the unpermuted system but avoids the occurrence of variables at time instants greater than k_F. The new system is considered on the restricted time horizon $0 \leq k \leq k_F - (l+1)$.

We now present the shuffle algorithm. In performing the algorithm one should keep in mind that most of the derived equations hold only along trajectories of Σ.

Let $F^1 := (f^T, h^T)^T$.

STEP 1

Define

$$\rho_1 := \operatorname{rank} \frac{\partial}{\partial(\sigma x, y)} F^1(\sigma x, x, u, y) \quad \left(= \operatorname{rank} \frac{\partial}{\partial(\sigma x, y)} \begin{bmatrix} f(\sigma x, x, u) \\ h(x, u, y) \end{bmatrix} \right),$$

and assume that $\rho_1 =$ constant in a neighbourhood of (x_e, u_e, y_e). If $\rho_1 < s + p$, then the components of the functions f and h are not independent but functionally related. Choose ρ_1 components of $(f^T, h^T)^T$, denoted by $(\tilde{f}^T, \tilde{h}^T)^T$, which are functionally independent, and define

$$\tilde{F}^1 := (\tilde{f}^T, \tilde{h}^T)^T.$$

The remaining $s + p - \rho_1$ components, denoted by $\hat{F}^1$, are functionally dependent and can therefore be expressed as a function of $\tilde{F}^1$, x, and u, that is,

$$\hat{F}^1(\sigma x, x, u, y) = \xi^1(\tilde{F}^1(\sigma x, x, u, y), x, u).$$

Since $\tilde{F}^1 = 0$, we have

$$\hat{F}^1(\sigma x, x, u, y) = \Phi^1(x, u) \tag{4}$$

for some function Φ^1. Write the system in the form $[\Phi^{1^T}(x, u), \tilde{F}^{1^T}(\sigma x, x, u, y)]^T = 0$, and define

$$F^2(\sigma x, x, u, \sigma u, y) := \begin{pmatrix} \tilde{F}^1(\sigma x, x, u, y) \\ \Phi^1(\sigma x, \sigma u) \end{pmatrix},$$

considered on the time horizon $0 \le k \le k_F - 2$.

Go to the next step.

If $\rho_1 = s + p$, then define

$$\tilde{F}^1 := F^1, \quad \Phi^1 := \hat{F}^1 = 0,$$

and the equation

$$\tilde{F}^1(\sigma x, x, u, y) = 0$$

can be solved locally for $s + p$ components of σx and y. In this case the algorithm stops.

STEP (l+1)

Suppose that in steps 1 through l the functions $F^{l+1} := (\tilde{F}^{l^T}, \Phi^{l^T})^T$, where $\tilde{F}^l$ and Φ^l are ρ_l- and $(s+p-\rho_l)$-dimensional, respectively, considered on the time interval $0 \le k \le k_F - (l+1)$, have been defined in such a way that

$$\tilde{F}^l(\sigma x, x, u, \sigma u, \ldots, \sigma^{l-1} u, y) = 0,$$

$$\Phi^l(\sigma x, \sigma u, \ldots, \sigma^l u) = 0,$$

and $\partial \tilde{F}^l(\cdot)/\partial(\sigma x, y)$ has full row rank equal to ρ_l in some neighbourhood of (x_e, u_e, y_e). Next define

$$\rho_{l+1} := \operatorname{rank} \frac{\partial}{\partial(\sigma x, y)} F^{l+1}(\sigma x, x, u, \ldots, \sigma^l u, y),$$

and assume that $\rho_{l+1} =$ constant in a neighbourhood of (x_e, u_e, y_e). If $\rho_{l+1} < s + p$, then the components of the functions $\tilde{F}^l$ and Φ^l are functionally dependent about (x_e, u_e, y_e). Choose $\rho_{l+1} - \rho_l$ components of Φ^l, denoted by $\tilde{\Phi}^l$, which together with $\tilde{F}^l$ are functionally independent, and define

$$\tilde{F}^{l+1} := (\tilde{F}^{l^T}, \tilde{\Phi}^{l^T})^T.$$

If $\rho_{l+1} = \rho_l$, then

$$\tilde{F}^{l+1} := \tilde{F}^l \text{ and } \hat{\Phi}^l := \Phi^l.$$

The remaining $s + p - \rho_{l+1}$ components, denoted by $\hat{\Phi}^l$, can be expressed as

$$\hat{\Phi}^l(\sigma x, \sigma u, \ldots, \sigma^l u) = \Phi^{l+1}(x, u, \ldots, \sigma^l u). \tag{5}$$

Write the system in the form $[\Phi^{(l+1)^T}(x, u, \ldots, \sigma^l u), \tilde{F}^{(l+1)^T}(\sigma x, x, u, \ldots, \sigma^l u, y)]^T = 0$, and define

$$F^{l+2}(\sigma x, x, u, \ldots, \sigma^{l+1} u, y) := \begin{pmatrix} \tilde{F}^{l+1}(\sigma x, x, u, \ldots, \sigma^l u, y) \\ \Phi^{l+1}(\sigma x, \sigma u, \ldots, \sigma^{l+1} u) \end{pmatrix},$$

considered on the time horizon $0 \le k \le k_F - (l + 2)$.

Go to the next step.

If $\rho_{l+1} = s + p$, define

$$\tilde{F}^{l+1} := F^{l+1}, \quad \Phi^{l+1} := \hat{\Phi}^l = 0,$$

and the equation

$$\tilde{F}^{l+1}(\sigma x, x, y, u, \ldots, \sigma^l u) = 0$$

can be solved locally for $s + p$ components of σx and y. Then the algorithm stops.

Remark 3.1. Before introducing the shuffle algorithm we have explained how the functions F^{l+1} in each step l of the algorithm are obtained by deleting certain equations. Some of these deleted equations, namely,

$$\Phi^1(x(0), u(0)) = 0, \quad \Phi^2(x(0), u(0), u(1)) = 0, \ldots, \quad \Phi^{l+1}(x(0), u(0), \ldots, u(l)) = 0,$$

are just the algebraic restrictions connecting $x(0)$ and successive inputs and which have already been mentioned in the preliminaries. They are indispensable for a consistent initialization of the system. The other part consists of such equations of system Σ in the interpretation of Remark 2.3 which already exhibit a maximal possible rank with respect to $x(k + 1)$ and $y(k)$ for k sufficiently large.

Remark 3.2. The algebraic constraints occur in the first step of the shuffle algorithm in form of the equation $\Phi^1(x, u) = 0$. It can happen that parts of these constraints involve only components of the inputs u. A consequence of this situation is that the inputs cannot be chosen

freely. Moreover it turns out that in this case the system is never solvable in the sense that we have defined above. If these particular constraints can be solved for a number of input components, we can take remedial measures by eliminating them. In case these constraints are inconsistent, the system is not solvable in any sense whatever. In the subsequent discussion we therefore exclude systems which exhibit constraints that only involve input components.

Remark 3.3. The following linear example is intended to show that $\rho_{l+1} = \rho_l$ is not sufficient for the termination of the algorithm, and hence a separate stopping criterion has to be introduced. Consider the system

$$\begin{aligned} x_1(k+1) + x_2(k) + x_3(k) &= 0, & x_2(k+1) + u(k) &= 0, \\ x_1(k) + x_2(k) &= 0, & y(k) + x_1(k) &= 0. \end{aligned}$$

Evidently, $\rho_1 = 3$ for this system, with $x_1(k) + x_2(k) = 0$ the functionally dependent component. Time-shifting this component does not increase the rank of the Jacobian with respect to $x(k+1)$ and $y(k)$, which results in $\rho_2 = 3$. It is not hard to see that in the third step we obtain $\rho_3 = 4$.

3.2. The stopping criterion. We now give the stopping criterion for the shuffle algorithm.

Denote $W^0 = 0$ and

$$W^l(x, u, \sigma u, \dots, \sigma^{l-1}u) = \begin{bmatrix} W^{l-1}(x, u, \sigma u, \dots, \sigma^{l-2}u) \\ \Phi^l(x, u, \sigma u, \dots, \sigma^{l-1}u) \end{bmatrix}, \qquad l \geq 1.$$

Define

$$r_l := \operatorname{rank} \frac{\partial}{\partial x} W^l(x, u, \sigma u, \dots, \sigma^{l-1}u),$$

and assume that for all $l \geq 1$, $r_l =$ constant in a neighbourhood of (x_e, u_e, y_e). Stop if

$$r_l = r_{l-1}. \tag{6}$$

LEMMA 3.1. *The stopping criterion* (6) *of the shuffle algorithm is always reached for some* $l \leq n$.

Proof. Note that the sequence $\{r_l : l \geq 1\}$ is nondecreasing by its definition. Hence, by the finite dimensionality of x, (6) must be reached for some $l \leq n$. □

Remark 3.4. Observe that the functions W^l are just the collection of the functionally dependent components up to step l and, for $k = 0$, contain all information about the algebraic restrictions on $x(0)$.

The next lemma shows that when condition (6) is satisfied, the sequence $\{\rho_l : l \geq 1\}$ defined by the shuffle algorithm has converged so that it can indeed be concluded that the algorithm has terminated.

LEMMA 3.2. *Denote by* α *the first integer* l *such that* (6) *is satisfied. Then* ρ_l *does not increase by further iterations of the algorithm, that is,* $\rho_l = \rho_\alpha$ *for all* $l > \alpha$.

Proof. The stopping criterion (6) or equivalently

$$\operatorname{rank} \frac{\partial}{\partial x} \begin{bmatrix} \Phi^1(\cdot) \\ \vdots \\ \Phi^{\alpha-1}(\cdot) \\ \Phi^\alpha(\cdot) \end{bmatrix} = \operatorname{rank} \frac{\partial}{\partial x} \begin{bmatrix} \Phi^1(\cdot) \\ \vdots \\ \Phi^{\alpha-1}(\cdot) \end{bmatrix}$$

implies that about the point (x_e, u_e)

$$\Phi^\alpha(x, u, \sigma u, \dots, \sigma^{\alpha-1}u) = \mu(\Phi^l, \dots, \Phi^{\alpha-1}, u, \sigma u, \dots, \sigma^{\alpha-1}u).$$

Since $\Phi^l(\cdot) = 0$ for $1 \leq l \leq \alpha - 1$, we have

$$\Phi^\alpha(x, u, \sigma u, \ldots, \sigma^{\alpha-1}u) = \lambda(u, \sigma u, \ldots, \sigma^{\alpha-1}u). \tag{7}$$

According to the shuffle algorithm define

$$\rho_{\alpha+1} := \operatorname{rank} \frac{\partial}{\partial(\sigma x, y)} \begin{bmatrix} \tilde{F}^\alpha(\sigma x, x, y, u, \sigma u, \ldots, \sigma^{\alpha-1}u) \\ \lambda(\sigma u, \ldots, \sigma^\alpha u) \end{bmatrix} = \operatorname{rank} \frac{\partial}{\partial(\sigma x, y)} \tilde{F}^\alpha(\cdot) = \rho_\alpha.$$

This means $\tilde{F}^{\alpha+1} = \tilde{F}^\alpha$ and $\hat{\Phi}^\alpha = \Phi^\alpha$, which by (7) is equal to λ.

Thus, it is clear that for every $l \geq 1$, $\Phi^{\alpha+l}$ does not depend on σx and y anymore, which completes the proof. □

Note that the application of the shuffle algorithm is not unique. In general there exist different selections of $\tilde{\Phi}^l$ in each step $l + 1$, $l \geq 1$, so that the matrix

$$\frac{\partial}{\partial(\sigma x, y)} \tilde{F}^{l+1}(\cdot) = \frac{\partial}{\partial(\sigma x, y)} \begin{bmatrix} \tilde{F}^l(\cdot) \\ \tilde{\Phi}^l(\cdot) \end{bmatrix}$$

has full row rank equal to ρ_{l+1} in a neighbourhood of (x_e, u_e, y_e). Note moreover that different choices of $\tilde{\Phi}^l$ result in different functions $\tilde{F}^{l+1}$ and Φ^{l+1}.

In the shuffle algorithm certain constant rank conditions have been imposed to ensure that the algorithm can be applied about a given equilibrium point. We summarize these conditions in the definition of regularity of an equilibrium point associated with the shuffle algorithm.

DEFINITION 3.5. *We call the equilibrium point (x_e, u_e, y_e) of the implicit system Σ regular with respect to the shuffle algorithm if for some specific application of the shuffle algorithm, the constant rank assumptions of the algorithm with respect to the sequence $\{\rho_l\}$ are satisfied. We call (x_e, u_e, y_e) strongly regular if this holds for each application of the algorithm.*

Although the result of the application of the shuffle algorithm apparently depends on the actual choice of $\tilde{\Phi}^l$ at each step of the algorithm, the sequences of quantities ρ_l and r_l are unique. We especially have the following lemma.

LEMMA 3.3. *About a strongly regular equilibrium point, the integers ρ_l, $l \geq 1$, do not depend on the specific application of the algorithm; that is, for any two applications of the algorithm we have*

$$\rho_l^1 = \rho_l^2, \qquad l \geq 1,$$

where superscripts refer to the application under consideration.

Proof. The proof is given in the appendix. □

From the proof of Lemma 3.3 one concludes that any "successful" application of the algorithm leads to the same sequence $\{\rho_l\}$ and thus to the same α. In particular, applying the shuffle algorithm about a strongly regular equilibrium point, one obtains a uniquely defined sequence of integers $\rho_1 \leq \rho_2 \leq \cdots \leq \rho_l \leq \cdots \leq s + p$. Define $\rho^* := \max\{\rho_l : l \geq 1\}$, and let α be defined as the smallest $l \in \mathbb{N}$ such that $\rho_\alpha = \rho^*$.

4. Necessary and sufficient conditions for local solvability.

4.1. Local solvability. It is clear that about a strongly regular equilibrium point the shuffle algorithm can terminate in either of the following ways:

(i) The rank of $\partial \tilde{F}^\alpha(\cdot)/\partial(\sigma x, y)$ is equal to $s + p$.

(ii) At least one function appears in Φ^α which does not depend on σx.

It will turn out that there exists a close connection between the solvability of the given implicit system and the outcome of the shuffle algorithm. In order to prove this relation, we have to provide an additional result.

In each step of the shuffle algorithm, functionally dependent components of certain functions are expressed as a function of the independent components and parameters. The purpose of the next lemma is to show that these manipulations have no influence on the row ranks of the row blocks of the solvability matrix $F_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F)$.

To this end, consider the equation

$$F(z, x) = (F_1(z, x), \ldots, F_s(z, x))^T = 0,$$

where $\dim x = \dim z = n \geq s$. Suppose that in a neighbourhood of a point (z_0, x_0) with $F(z_0, x_0) = 0$ we have

$$\operatorname{rank} \frac{\partial F(z, x)}{\partial z} = \rho < s,$$

and assume moreover without loss of generality that

$$\operatorname{rank} \frac{\partial (F_1(z, x), \ldots, F_\rho(z, x))}{\partial z} = \rho. \tag{8}$$

Denote $\tilde{F} = (F_1, \ldots, F_\rho)^T$, $\hat{F} = (F_{\rho+1}, \ldots, F_s)^T$. Then we know that we can write

$$\hat{F}(z, x) = \xi(\tilde{F}, x),$$

which becomes for $\tilde{F}(z, x) = 0$ a function $\Phi(x)$ of x only. Now consider the matrices

$$A(z, x) = \begin{pmatrix} \frac{\partial \tilde{F}(z,x)}{\partial x} & \frac{\partial \tilde{F}(z,x)}{\partial z} \\ \frac{\partial \hat{F}(z,x)}{\partial x} & \frac{\partial \hat{F}(z,x)}{\partial z} \end{pmatrix} \text{ and } \overline{A}(z, x) = \begin{pmatrix} \frac{\partial \tilde{F}(z,x)}{\partial x} & \frac{\partial \tilde{F}(z,x)}{\partial z} \\ \frac{\partial \Phi(x)}{\partial x} & 0 \end{pmatrix}.$$

We then have the following lemma.

LEMMA 4.1. $\operatorname{rank} A(z, x) = \operatorname{rank} \overline{A}(z, x)$.

Proof. We show the existence of an invertible matrix,

$$M(z, x) = \begin{pmatrix} M_1(z, x) & M_2(z, x) \\ M_3(z, x) & M_4(z, x) \end{pmatrix},$$

such that $A(z, x) = M(z, x)\overline{A}(z, x)$. Set $M_1(z, x) = I_{\rho\times\rho}$, $M_2(z, x) = 0_{\rho\times(s-\rho)}$, and $M_4(z, x) = I_{(s-\rho)\times(s-\rho)}$. It remains to show the existence of a $((s - \rho) \times \rho)$ matrix $M_3(z, x)$ satisfying

$$M_3(z, x)\frac{\partial \tilde{F}(z, x)}{\partial x} + \frac{\partial \Phi(x)}{\partial x} = \frac{\partial \hat{F}(z, x)}{\partial x} \tag{9}$$

and

$$M_3(z, x)\frac{\partial \tilde{F}(z, x)}{\partial z} = \frac{\partial \hat{F}(z, x)}{\partial z}. \tag{10}$$

From $\hat{F}(z, x) = \xi(\tilde{F}, x)$ it follows that

$$\frac{\partial \hat{F}(z, x)}{\partial z} = \frac{\partial \xi(\tilde{F}, x)}{\partial \tilde{F}}\frac{\partial \tilde{F}(z, x)}{\partial z} \quad \text{and} \quad \frac{\partial \hat{F}(z, x)}{\partial x} = \frac{\partial \xi(\tilde{F}, x)}{\partial \tilde{F}}\frac{\partial \tilde{F}(z, x)}{\partial x} + \left.\frac{\partial \xi(\tilde{F}, x)}{\partial x}\right|_{(\tilde{F},x)}.$$

Observing that about solutions $\xi(0, x) = \Phi(x)$, it immediately follows that $M_3 = \frac{\partial \xi}{\partial F}$ satisfies (9) and (10). This means

$$A(z, x) = M(z, x)\overline{A}(z, x)$$

with a nonsingular matrix $M(z, x)$, and therefore rank $A(z, x) = \text{rank } \overline{A}(z, x)$. □

The main result of this subsection consists of the following theorem.

THEOREM 4.1. *About a strongly regular equilibrium point, the following statements are equivalent:*

1. $F_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F)$ *has full row rank for every* $(\mathbf{x},\mathbf{y},\mathbf{u}) \in \mathcal{M}_{\mathbf{u}} \times \mathbf{u},\ \mathbf{u} \in \mathcal{U}^0_{k_F}$.
2. *The application of the shuffle algorithm with respect to* σx *and* y *terminates with* $\rho^* = s + p$.

Proof. Consider the solvability matrix $F_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F)$ of the original system. According to Lemma 4.1, the solvability matrix $F^1_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F)$ of the system obtained after the first step of the shuffle algorithm can be computed by multiplying the original one from the left by a permutation matrix Π^1 representing the permutations of the components of $F^1(\sigma x, x, u, y)$ performed in the first step and a nonsingular block diagonal matrix, the blocks of which consist of the inverses of matrices as obtained in Lemma 4.1, that is,

$$F^1_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F) = \begin{bmatrix} M^1(x(1), x(0), u(0), y(0)) & & \\ & \ddots & \\ & & M^1(x(k_F), x(k_F-1), u(k_F-1), y(k_F-1)) \end{bmatrix} \Pi^1 F_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F).$$

Now suppose that the solvability matrix $F^l_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F)$ of the system obtained after the lth step of the shuffle algorithm (including the deleted parts) has been obtained by left-multiplying $F^{l-1}_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F)$ by a nonsingular matrix. After the lth step, we are left with the modified system

$$\begin{aligned} \Phi^1(x(0), u(0)) &= 0, \\ &\vdots \\ \Phi^l(x(0), u(0), \ldots, u(l-1)) &= 0, \\ F^{l+1}(\sigma x, x, u, \sigma u, \ldots, \sigma^l u, y) &= 0, \\ \tilde{F}^l(\sigma x(k_F-l), x(k_F-l), u(k_F-l), \ldots, u(k_F-1), y(k_F-l)) &= 0, \\ &\vdots \\ \tilde{F}^1(x(k_F), x(k_F-1), u(k_F-1), y(k_F-1)) &= 0. \end{aligned}$$

The $(l+1)$th step modifies F^{l+1} only. If the solvability matrix of this restricted system is denoted by $F^l_{r_{(\mathbf{x},\mathbf{u},\mathbf{y})}}$ and that for the system after performing the $(l+1)$th step is denoted by $F^{l+1}_{r_{(\mathbf{x},\mathbf{u},\mathbf{y})}}$, we know that

$$F^{l+1}_{r_{(\mathbf{x},\mathbf{u},\mathbf{y})}} = \Gamma F^l_{r_{(\mathbf{x},\mathbf{u},\mathbf{y})}},$$

where Γ is some invertible matrix computed as described above. It is then clear that

$$F^{l+1}_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F) = \text{diag}\,[I, \Gamma, I] F^l_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F), \tag{11}$$

where the identity matrices in (11) represent the unaltered equations of the system above. Thus, applying the shuffle algorithm does not change the rank of the solvability matrices connected with the systems obtained in each step of the algorithm. Now consider the function

$$F^{(\alpha+1)}(\sigma x, x, u, \ldots, \sigma^\alpha u, y) := \begin{pmatrix} \tilde{F}^\alpha(\sigma x, x, u, \ldots, \sigma^{(\alpha-1)} u, y) \\ \Phi^\alpha(\sigma x, \sigma u, \ldots, \sigma^\alpha u) \end{pmatrix}$$

obtained in the final step of the shuffle algorithm. From the proof up to now and the special structure of $F^{(\alpha+1)}$ (Φ^α does not depend on x!), it immediately follows that $F_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F)$ has full row rank if and only if $\rho^* = s + p$. □

From now on we assume that the shuffle algorithm terminates with $\rho_\alpha = s + p$. In this case, the implicit system Σ may be expressed locally in the equivalent form

$$\left.\begin{aligned} \tilde{f}(x(k+1), x(k), u(k)) &= 0, && 0 \le k \le k_F - 1, \\ \tilde{h}(x(k), u(k), y(k)) &= 0, && 0 \le k \le k_F - 1, \\ \tilde{\Phi}^1(x(k+1), u(k+1)) &= 0, && 0 \le k \le k_F - 2, \\ &\vdots \\ \tilde{\Phi}^{\alpha-1}(x(k+1), u(k+1), \ldots, u(k+\alpha-1)) &= 0, && 0 \le k \le k_F - \alpha, \end{aligned}\right\} \tag{12}$$

$$W^{\alpha-1}(x(0), u(0), \ldots, u(\alpha-2)) = 0, \tag{13}$$

where, according to the shuffle algorithm, (12) can be solved for $s + p$ components of $(x(k+1), y(k))$. Assume furthermore that these components contain all components of $y(k)$.

Remark 4.2. This full rank assumption with respect to the output components is very natural in view of right-inversion. In case it is not satisfied, there exist functional relations between the output components, making right-invertibility impossible. Those situations are often a result of overparameterization in the modelling process. A consequence of this assumption is $\tilde{h} = h$ in the first step of the shuffle algorithm.

Hence, possibly after reordering the components of x, (12) is solvable for

$$x^1(k+1) := [x_1(k+1), \ldots, x_s(k+1)] \quad \text{and} \quad y(k)$$

in terms of

$$x(k), u(k), \ldots, u(k+\alpha-1) \quad \text{and} \quad x^2(k+1) := [x_{s+1}(k+1), \ldots, x_n(k+1)],$$

that is,

$$\begin{aligned} x^1(k+1) &= \psi(x(k), u(k), \ldots, u(k+\alpha-1), x^2(k+1)), \\ y(k) &= \phi(x(k), u(k), \ldots, u(k+\alpha-1), x^2(k+1)), \end{aligned} \tag{14}$$

where these equations hold for $0 \le k \le k_F - \alpha$. For the role of (13) see Remark 3.4.

4.2. Uniqueness of solutions. The rest of this section is concerned with the question of how to impose further restrictions in order to make the solution of (12), (13) unique. Introduce the following notation:

$$\tilde{\Phi}^{(1,\alpha-1)} := [\tilde{\Phi}^{1^T}, \ldots, \tilde{\Phi}^{\alpha-1^T}]^T.$$

Consider (12), (13), and specify unique initial and final conditions via equations of the form

$$f^{init}(x^1(0)) = 0, \; f_1^{fin}(x^1(k_F - \alpha + 2)) = 0, \ldots, f_{\alpha-1}^{fin}(x^1(k_F)) = 0. \tag{15}$$

Here, $f^{init}, f_1^{fin}, \ldots, f_{\alpha-1}^{fin}$ are arbitrary functions with dimensions $s - r_{\alpha-1}$, $s + p - \rho_{\alpha-1}, \ldots, s + p - \rho_1$, respectively, such that

$$\operatorname{rank} \frac{\partial}{\partial x^1(0)} \begin{bmatrix} f^{init}(x^1(0)) \\ W^{\alpha-1}(x(0), u(0), \ldots, u(\alpha-2)) \end{bmatrix} = s, \tag{16}$$

$$\operatorname{rank} \frac{\partial}{\partial x^1(k_F-\alpha+2)} \begin{bmatrix} \tilde{f}(x(k_F-\alpha+2), x(k_F-\alpha+1), u(k_F-\alpha+1)) \\ \tilde{\Phi}^{(1,\alpha-2)}(x(k_F-\alpha+2), u(k_F-\alpha+2), \ldots, u(k_F-1)) \\ f_1^{fin}(x^1(k_F-\alpha+2)) \end{bmatrix} = s, \tag{17}$$

$$\vdots$$

$$\operatorname{rank} \frac{\partial}{\partial x^1(k_F-1)} \begin{bmatrix} \tilde{f}(x(k_F-1), x(k_F-2), u(k_F-2)) \\ \tilde{\Phi}^{(1,1)}(x(k_F-1), u(k_F-1)) \\ f_{\alpha-2}^{fin}(x^1(k_F-1)) \end{bmatrix} = s, \tag{18}$$

$$\operatorname{rank} \frac{\partial}{\partial x^1(k_F)} \begin{bmatrix} \tilde{f}(x(k_F), x(k_F-1), u(k_F-1)) \\ f_{\alpha-1}^{fin}(x^1(k_F)) \end{bmatrix} = s. \tag{19}$$

In order to make the solution unique, one has to specify x^2 for the whole time interval as well:

$$x^2(k) = x^{20}(k), \quad 0 \le k \le k_F. \tag{20}$$

After determining initial and final conditions and specifying $x^2(k)$ over the whole time interval, the number of equations and unknowns in equations (12), (13), (15), (20) is the same, and the system is still solvable.

A feature of the system equations as obtained after applying the shuffle algorithm is that they can be solved recursively if they are solvable at all.

1. From

$$f^{init}(x^1(0)) = 0,$$
$$W^{\alpha-1}(x^1(0), u(0), \ldots, u(\alpha-2)) = 0,$$

we compute $x^1(0)$.

2. Then, using $x^1(0)$ and the given control sequence $\mathbf{u} \in \mathcal{U}_{k_F}^0$, we obtain $x^1(k)$, $1 \le k \le k_F - \alpha + 1$, and $y(k)$, $0 \le k \le k_F - \alpha$, from (14).

3. Finally, by means of the set of equations

$$\begin{aligned}
\tilde{f}(x(k_F-\alpha+2), x(k_F-\alpha+1), u(k_F-\alpha+1)) &= 0, \\
h(x(k_F-\alpha+1), u(k_F-\alpha+1), y(k_F-\alpha+1)) &= 0, \\
\tilde{\Phi}^{(1,\alpha-2)}(x(k_F-\alpha+2), u(k_F-\alpha+2), \ldots, u(k_F-1)) &= 0, \\
f_1^{fin}(x^1(k_F-\alpha+2)) &= 0, \\
\vdots \qquad\qquad & \qquad \vdots \\
\tilde{f}(x(k_F-1), x(k_F-2), u(k_F-2)) &= 0, \\
h(x(k_F-2), u(k_F-2), y(k_F-2)) &= 0, \\
\tilde{\Phi}^{(1,1)}(x(k_F-1), u(k_F-1)) &= 0, \\
f_{\alpha-2}^{fin}(x^1(k_F-1)) &= 0, \\
\tilde{f}(x(k_F), x(k_F-1), u(k_F-1)) &= 0, \\
h(x(k_F-1), u(k_F-1), y(k_F-1)) &= 0, \\
f_{\alpha-1}^{fin}(x^1(k_F)) &= 0,
\end{aligned}$$

we obtain $x^1(k_F-\alpha+2)$, $y(k_F-\alpha+1), \ldots, x^1(k_F-1)$, $y(k_F-2)$ and $x^1(k_F)$,$y(k_F-1)$, respectively.

5. Right-invertibility. If the solvability conditions stated in Theorem 4.1 are satisfied, we know that there is for every locally given input sequence $\mathbf{u} \in \mathcal{U}^0_{k_F}$ at least one pair $(\mathbf{x},\mathbf{y})$ satisfying the system equations for $0 \le k \le k_F - 1$ or, stated equivalently, for every such $\mathbf{u}$ there exists $(\mathbf{x},\mathbf{y})$ such that $(\mathbf{x},\mathbf{u},\mathbf{y})$ is an admissible triple. Of course, one can ask as well whether or not the same holds true if, instead of $\mathbf{u}$, a sequence of outputs $\mathbf{y} \in \mathcal{Y}^0_{k_F}$ is considered to be given; that is, does there exist for arbitrary $\mathbf{y}$ a pair $(\mathbf{x},\mathbf{u})$ such that the system equations are satisfied? Obviously, this question can be decided by performing the shuffle algorithm with respect to σx and u. But this is basically what right-invertibility means: given a desired output trajectory, one wishes to determine a control sequence (not necessarily unique) that enforces the prespecified output trajectory. Thus, the problem of right-invertibility appears as the solvability problem of the system with respect to σx and u. To make things precise we use the following definition.

DEFINITION 5.1. *System Σ is said to be locally right-invertible in a neighbourhood of its equilibrium point (x_e, u_e, y_e) if for any sequence $\mathbf{y}^{ref} \in \mathcal{Y}^0_{k_F}$, it is possible to find $\mathbf{x}^{ref} \in \mathcal{X}^0_{k_F}$ and a control sequence $\mathbf{u}^{ref} \in \mathcal{U}^0_{k_F}$ such that $(\mathbf{x}^{ref}, \mathbf{u}^{ref}, \mathbf{y}^{ref})$ is admissible.*

As already mentioned, the right-inversion problem can be solved quite routinely via the shuffle algorithm. The algorithm can also be used to determine explicit recursive equations for the inverse system.

In order to differentiate between the two versions of the shuffle algorithm, that is, between the cases where we check the solvability with respect to σx and y or σx and u, we will refer to the second case as the inversion algorithm.

Introduce in addition to the matrices defined in §2 the following matrices:

$$B^i_{(x,u,y)} = \frac{\partial f(\sigma x, x, u)}{\partial u}|_{x(i+1),x(i),u(i)}, \qquad 0 \le i \le k_F - 1,$$
$$D^i_{(x,u,y)} = \frac{\partial h(x, u, y)}{\partial u}|_{x(i),u(i),y(i)}, \qquad 0 \le i \le k_F - 1,$$

and

$$G_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F) = \begin{pmatrix} A^0 & B^0 & E^1 & & & & \\ C^0 & D^0 & 0 & & & & \\ & & A^1 & B^1 & E^2 & & \\ & & C^1 & D^1 & 0 & & \\ & & & & & \ddots & \\ & & 0 & & & A^{k_F-1} & B^{k_F-1} & E^{k_F} \\ & & & & & C^{k_F-1} & D^{k_F-1} & 0 \end{pmatrix},$$

where $G_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F)$ is of size $k_F(s + p) \times [(k_F + 1)n + k_F m]$. Again, these matrices are defined along every admissible path $(\mathbf{x},\mathbf{u},\mathbf{y})$ of system Σ .

Similarly to Theorem 4.1 one now proves the following theorem.

THEOREM 5.2. *Consider system Σ about—with respect to the inversion algorithm—the strongly regular equilibrium point (x_e, u_e, y_e). Then the following statements are equivalent:*

1. *System Σ is locally right-invertible about (x_e, u_e, y_e).*
2. *$G_{(\mathbf{x},\mathbf{u},\mathbf{y})}(0, k_F)$ has full row rank for every $(\mathbf{x},\mathbf{u},\mathbf{y}) \in \mathcal{M}_{\mathbf{y}} \times \mathbf{y}$, $\mathbf{y} \in \mathcal{Y}^0_{k_F}$.*
3. *The application of the inversion algorithm terminates with $\rho^* = s + p$.*

6. Example: Economic policy making by right-inversion. The inversion algorithm (and the shuffle algorithm) essentially uses the implicit function theorem, which is frequently considered a restriction on the applicability of an algorithm. The aim of this section is to show by means of a realistic model that for many problems a numerical treatment is still possible.

Moreover if the nonlinearities are such that they include only, for instance, rational expressions (as is the case in most macroeconomic models), then an inverse system can even be obtained by symbolic computations.

The model we want to use for our considerations was created by Klein and Goldberger (cf. [10]) and adapted by a number of economists over the course of the years. It aimed at modelling the U.S. economy of the years 1929–1952 and is probably the first macroeconomic model used for economic forecasting (cf. [3]). This was done in a way still in use nowadays. The approach simply consists of predetermining a policy scenario and investigating the corresponding behaviour of the state variables by solving the system equations.

In this section we are going to demonstrate how the right-inversion approach of §5 can be employed to actively enforce a desired evolution of a number of (important) state variables depending on the number of control variables available to a policy maker.

6.1. The Klein–Goldberger model. In the sequel we use the Klein–Goldberger model as adapted by Adelman and Adelman [1]. The model consists of 22 state and 10 exogenous variables, the latter containing 4 control variables, which have the following meaning.

Exogenous control variables:

u_1 = government employee compensation,
u_2 = government expenditures for goods and services,
u_3 = government payment to farmers,
u_4 = number of government employees.

Exogenous variables with no interpretation as control variables:

z_1 = index of agricultural exports,
z_2 = number of persons in the United States,
z_3 = number of persons in the labour force,
z_4 = number of nonfarm entrepreneurs,
z_5 = number of farm operators,
z_6 = time in years; $z_6(0) = 0$ corresponds to the year 1929.

State variables:

x_1 = consumer expenditures in 1939 dollars,
x_2 = gross private domestic capital formation in 1939 dollars,
x_3 = corporate savings,
x_4 = corporate profits,
x_5 = capital consumption charges,
x_6 = private employee compensation,
x_7 = number of wage-and-salary earners,
x_8 = index of hourly wages,
x_9 = farm income,
x_{10} = index of agricultural prices,
x_{11} = end-of-year liquid assets held by persons,
x_{12} = end-of-year liquid assets held by businesses,
x_{13} = gross national product,

x_{14} = nonwage nonfarm income,
x_{15} = price index of gross national product,
x_{16} = end-of-year stock of private capital,
x_{17} = end-of-year corporate surplus,
x_{18} = indirect taxes less subsidies,
x_{19} = personal and payroll taxes less transfers,
x_{20} = corporate income tax,
x_{21} = personal and corporate taxes less transfers,
x_{22} = taxes less transfers associated with farm income,

where states and exogenous variables are combined by the following set of simultaneous implicit nonlinear equations:

$$x_1(k) = 0.55[x_6(k) + u_1(k) - x_{19}(k)] + 0.41[x_{14}(k) - x_{21}(k) - x_3(k)] + 0.34[x_9(k) + u_3(k) - x_{22}(k)] + 0.26x_1(k-1) + 0.072x_{11}(k-1) + 0.26z_2(k) - 22.26, \tag{21}$$

$$x_2(k) = 0.78[x_{14}(k-1) - x_{21}(k-1) + x_9(k-1) + u_3(k-1) - x_{22}(k-1) + x_5(k-1)] - 0.073x_{16}(k-1) + 0.14x_{12}(k-1) - 16.71, \tag{22}$$

$$x_3(k) = -3.53 + 0.72[x_4(k) - x_{20}(k)] - 0.027x_{17}(k-1), \tag{23}$$

$$x_4(k) = -7.60 + 0.68x_{14}(k), \tag{24}$$

$$x_5(k) = 7.25 + 0.05[x_{16}(k) + x_{16}(k-1)] + 0.044[x_{13}(k) - u_1(k)], \tag{25}$$

$$x_6(k) = -1.40 + 0.24[x_{13}(k) - u_1(k)] + 0.24[x_{13}(k-1) - u_1(k-1)] + z_6(k), \tag{26}$$

$$x_7(k) = \frac{[26.08 + x_{13}(k) - u_1(k) - 0.08x_{16}(k) - 0.08x_{16}(k-1) - 2.05z_6(k)]}{2.17 \times 1.062} + u_4(k) - [z_4(k) + z_5(k)]/1.062, \tag{27}$$

$$x_8(k) = x_8(k-1) + 4.11 - 0.74[z_3(k) - x_7(k) - z_4(k) - z_5(k)] + 0.52[x_{15}(k-1) - x_{23}(k-1)] + 0.54z_6(k), \tag{28}$$

$$x_9(k) = 0.054[x_6(k) + u_1(k) - x_{19}(k) + x_{14}(k) - x_{21}(k) - x_3(k)] + \frac{0.012z_1(k)x_{10}(k)}{x_{15}(k)}, \tag{29}$$

$$x_{10}(k) = 1.39x_{15}(k) + 32.0, \tag{30}$$

$$x_{11}(k) = 0.14[x_6(k) + u_1(k) - x_{19}(k) + x_{14}(k) - x_{21}(k) - x_3(k) + x_9(k) + u_3(k) - x_{22}(k)] + 76.03(1.5)^{-0.84}, \tag{31}$$

$$x_{12}(k) = 0.26x_6(k) - 2.55 - 0.26[x_{15}(k) - x_{15}(k-1)] + 0.61x_{12}(k-1), \tag{32}$$

$$x_{13}(k) = x_1(k) + x_2(k) + u_2(k), \tag{33}$$

$$x_{14}(k) = x_{13}(k) - x_{18}(k) - x_5(k) - x_6(k) - u_1(k) - x_9(k) - u_3(k), \tag{34}$$

$$x_{15}(k) = \frac{1.062x_7(k)x_8(k)}{(x_6(k) + u_1(k))}, \tag{35}$$

$$x_{16}(k) = x_{16}(k-1) + x_2(k) - x_5(k), \tag{36}$$

$$x_{17}(k) = x_{17}(k-1) + x_3(k), \tag{37}$$

$$x_{18}(k) = 0.0924x_{13}(k) - 1.3607, \tag{38}$$

$$x_{19}(k) = 0.1549x_6(k) + 0.131u_1(k) - 6.9076, \tag{39}$$

$$x_{20}(k) = 0.4497x_4(k) + 2.7085, \tag{40}$$

$$x_{21}(k) = \frac{0.2695x_{15}(k-1)}{x_{15}(k)}[x_{14}(k-1) - x_{20}(k-1) - x_3(k-1)] + 0.248[x_{14}(k) - x_{20}(k) - x_3(k)] + 0.4497x_4(k) - 5.7416, \tag{41}$$

$$x_{22}(k) = 0.0512[x_9(k) + u_3(k)], \tag{42}$$

$$x_{23}(k) = x_{15}(k-1) \tag{43}$$

with $x = (x_1, \ldots, x_{22})^T$ the vector of state variables, $u = (u_1, \ldots, u_4)$ the vector of exogenous control variables, and $z = (z_1, \ldots, z_6)$ the vector of exogenous variables without control interpretation. Considering the system equations above, one recognizes a number of deviations from the model form as given by (1). This mainly concerns the occurrence of lagged exogenous variables such as for instance $u_1(k)$ and $u_1(k-1)$ in (26). A second point is the explicit time dependence of the system via the variable z_6. We are therefore left with structural equations of the form

$$x(k) - f[x(k), x(k-1), u(k), u(k-1), z(k)] = 0, \tag{44}$$

where (43) is introduced only to avoid lag-2 variables.

The flexibility of the algorithm appears among other things in its robustness with respect to the model form. This is due to the fact that all variables which do not enter into the computation of Jacobians are treated on an equal footing by simply regarding them as parameters. Hence, even though it is possible to bring (44) to the form (1) by redefining variables if required, there is actually no need to do so. On the contrary, it would unnecessarily complicate the model!

The incorporation of time-varying exogenous variables without an interpretation as control variables in the model equations leads to the nonexistence of equilibrium points. The role of the equilibrium point is now taken over by the points of a reference trajectory about which the inversion algorithm may be performed. Eventually, the results can be patched together, provided that the applications of the inversion algorithm about each single reference point led to the same number of independent functions in each step of the algorithm, and moreover independent components could be chosen to be the same.

In the next subsection we demonstrate the right-invertibility of the considered model along the lines of §5. As it is common practice in economics, the goal will be the guidance of selected state variables as to enforce a desired behaviour.

6.2. Right-invertibility of the Klein–Goldberger model. The first part of the following investigations is dedicated to the generation of a reference trajectory for the considered model beyond the year 1952. This trajectory should reflect a reasonable—meaning economically possible—behaviour of the underlying system. Such a trajectory can be generated by determining the evolution of the state variables under the influence of exogenous variables obtained by extrapolating their historical trends and with initial conditions as actually observed. Information about initial values and extrapolations for the exogenous variables are taken from [1].

In order to perform the inversion algorithm we have to fix output equations. In view of the fact that we have only four control variables at our disposal, their number will naturally be bounded above by four if one wants to have an outlook for a positive result with respect to right-invertibility. This follows from well-known necessary right-invertibility conditions which generalize Tinbergen's counting rule for the inversion of static systems (cf. [22]).

As already mentioned, the most common objective in economics is the control of selected state variables rather than the control of certain combinations of them. The output equations are therefore chosen in the following form:

$$y_i(k) - x_{j_i}(k) = 0, \qquad i = 1, \ldots, 4; \qquad j_i \neq j_l \text{ for } i \neq l;$$

that is, the outputs consist of mutually different components of the vector of state variables. We now perform the inversion algorithm with an arbitrary selection of four outputs of the form above.

The Jacobian (with respect to $x(k)$ and $u(k)$) obtained in the first step of the algorithm is a simply structured rational matrix with only six nonconstant elements resulting from the nonlinearities in equations (29), (35), and (41). According to the occurrence of x_{15} and $(x_6 + u_1)$ as divisors in matrix elements, it can be computed for all combinations of variables where these terms do not vanish.

To determine its rank, one has to distinguish two cases. In the first case the selection of outputs is such that $x_{j_i} \neq x_2$, $i = 1, \ldots, 4$. Then the rank is maximal—that is, 27—independent (with exception of the restriction made above) of the chosen values of $x(k)$, $u(k)$, $y(k)$, and $z(k)$ and thus also along the considered reference trajectory resulting in a termination of the inversion algorithm. By Theorem 5.2 (global) right-invertibility of the system is ensured.

The remaining case is concerned with the situation in which $x_{j_i} = x_2$ for some i. Obviously, the rank of the Jacobian matrix obtained now is one less than maximal. Bearing in mind that the rank is computed with respect to $x(k)$ and $u(k)$, this is easily seen by comparing equations (22) and $y_i(k) - x_2(k) = 0$. Since (22) does not contain variables at time instant k others than $x_2(k)$, the corresponding rows of the Jacobian matrix will be dependent. The observed rank defect indicates the functional dependence of this output component which makes a representation

$$\begin{aligned} y_i(k) - x_2(k) = &- 0.78[x_{14}(k-1) - x_{21}(k-1) + x_9(k-1) + u_3(k-1) \\ &- x_{22}(k-1) + x_5(k-1)] + 0.073x_{16}(k-1) \\ &- 0.14x_{12}(k-1) + 16.71 + y_i(k) = 0 \end{aligned} \tag{45}$$

independent of $x(k)$ and $u(k)$ possible. Applying the shift operator σ to (45) we get a new equation in the variables $x(k)$ and $u(k)$. Adding this equation to the system of equations already recognized as functionally independent (seen as functions of $x(k)$ and $u(k)$) and computing the rank of the resulting Jacobian give full rank. Again the system is right-invertible. Hence, the inversion algorithm terminates after at most two steps irrespective of the selection of output functions.

In the next subsection we consider both cases with the help of concrete examples.

6.3. Some simulation results. As a first example we want to retrace results obtained by Chow in [5]. In this article similar objectives—namely, the guidance of certain state variables along desired paths—are pursued. However, the employed methods differ essentially from ours. Chow formulates the intended goal as an optimal control problem which is solved with the method of dynamic programming minimizing a quadratic loss function with respect to the control variables. The remarkable point is that dynamic programming is not applied directly to the nonlinear problem but instead relies on the linearized system. This involves an iterative computation of the control variables until convergence occurs.

As an illustration of his method, Chow used the Klein–Goldberger model. In a first control experiment he considered the number of wage-and-salary earners x_7, the gross national product x_{13}, the nonwage income x_{14}, and the price index of the gross national product x_{15} as targets and tried to steer them to grow at 2%, 5%, 5%, and 1% per year, respectively, from their initial values at 1952. This will also be the objective in our first simulation.

We have already seen in the previous subsection that with this choice of targets, the inversion algorithm terminates in the first step with a favourable outcome. Stated differently, the unmodified system can be solved for the controls and the state variables, which enables the computation of control values enforcing the intended growth of the targets. For the considered

example this leads for instance to the following time-varying nonlinear feedback generating the unique control sequence which achieves our goal. In the equations below most of the coefficients occur rounded off to four-digit accuracy.

$$
\begin{aligned}
u_1(k) = [&5.743y_1(k) + 1.034y_1^2(k) + 1.842y_4(k) + 0.316u_1(k-1)y_4(k)\\
&- 0.316y_2(k)y_4(k) - 1.034y_1(k)z_3(k) + 1.034y_1(k)z_4(k) + 1.034y_1(k)z_5(k)\\
&+ 0.755y_1(k)z_6(k) - 0.382y_4(k)z_6(k) + 1.397y_1(k)x_8(k-1)\\
&- 0.316y_4(k)x_{13}(k-1) + 0.727(y_1(k)x_{15}(k-1) - y_1(k)x_{23}(k-1))]/y_4(k),\\
u_2(k) = [&-1.149 - 0.768y_1(k) - 0.138y_1^2(k) + 30.160y_4(k) - 0.009u_1(k-1)y_4(k)\\
&- 0.768u_3(k-1)y_4(k) + 0.729y_2(k)y_4(k) + 0.191y_3(k)y_4(k)\\
&- 0.26y_4(k)z_2(k) + 0.138y_1(k)(z_3(k) - z_4(k) - z_5(k)) - 0.101y_1(k)z_6(k)\\
&+ y_4(k)(0.010z_6(k) - 0.26x_1(k-1) - 0.768x_5(k)) - 0.187y_1(k)x_8(k-1)\\
&- 0.768y_4(k)x_9(k-1) - 0.072y_4(k)x_{11}(k) - 0.138y_4(k)x_{12}(k-1)\\
&+ 0.009y_4(k)x_{13}(k-1) - 0.768y_4(k)x_{14}(k-1) - 0.097y_1(k)x_{15}(k-1)\\
&- 0.121x_{15}(k-1)(x_3(k-1) - x_{14}(k-1) + x_{20}(k-1))\\
&+ 0.103y_4(k)x_{16}(k-1) - 0.009y_4(k)x_{17}(k-1) + 0.768y_4(k)x_{21}(k-1)\\
&+ 0.768y_4(k)x_{22}(k-1) + 0.097y_1(k)x_{23}(k-1)]/y_4(k),\\
u_3(k) = [&-65.994 - 4.331y_1(k) - 0.780y_1^2(k) - 8.743y_4(k) + 0.013u_1(k-1)y_4(k)\\
&- y_4(k)(0.037u_3(k-1) - 0.853y_2(k) + 1.017y_3(k)) + 0.780y_1(k)z_3(k)\\
&- 0.780y_1(k)z_4(k) - 0.780y_1(k)z_5(k) - 0.569y_1(k)z_6(k) - 0.015y_4(k)z_6(k)\\
&- 0.037y_4(k)x_5(k-1) - 1.054y_1(k)x_8(k-1) - 0.037y_4(k)x_9(k-1)\\
&- 0.007y_4(k)x_{12}(k-1) - 0.013y_4(k)x_{13}(k-1) - 0.037y_4(k)x_{14}(k-1)\\
&- 0.548x_{15}(k-1)(y_1(k) + 0.016x_3(k-1) - 0.016x_{14}(k-1))\\
&- y_4(k)(0.092x_{16}(k-1) + 0.001x_{17}(k-1)) - 0.016x_{15}(k-1)x_{20}(k-1)\\
&+ 0.037y_4(k)(x_{21}(k-1)x_{22}(k-1)) + 0.548y_1(k)x_{23}(k-1)]/y_4(k),\\
u_4(k) = [&2.500y_1(k) + 0.450y_1^2(k) - 11.307y_4(k) + 0.137u_1(k-1)y_4(k)\\
&+ y_4(k)(0.026u_3(k-1) + 1.000y_1(k) - 0.573y_2(k)) - 0.450y_1(k)z_3(k)\\
&+ 0.450y_1(k)z_4(k) + 0.942y_4(k)z_4(k) + 0.450y_1(k)z_5(k) + 0.942y_4(k)z_5(k)\\
&+ 0.329y_1(k)z_6(k) + 0.723y_4(k)z_6(k) + 0.026y_4(k)x_5(k-1)\\
&+ 0.608y_1(k)x_8(k-1) + 0.026y_4(k)x_9(k-1) + 0.005y_4(k)x_{12}(k-1)\\
&- 0.137y_4(k)x_{13}(k-1) + 0.026y_4(k)x_{14}(k-1) + 0.316y_1(k)x_{15}(k-1)\\
&+ 0.064y_4(k)x_{16}(k-1) - 0.026y_4(k)x_{21}(k-1) - 0.026y_4(k)x_{22}(k-1)\\
&- 0.316y_1(k)x_{23}(k-1)]/y_4(k).
\end{aligned}
$$

This feedback has been obtained by means of formula manipulation packages and leads to the controls depicted in Figure 1 (computed with the exact coefficients). Incidentally, they coincide with those computed in [5] as far as they are available. The resulting closed-loop system generates the desired target paths (Figure 2). As an example for the second case we are going to steer the gross private domestic capital formation x_2 by means of the control variable u_2 to grow at a rate of, say, 5% per year, with extrapolations used for the other controls. It should be noticed that since x_2 is exclusively dependent on lagged variables and in view of fixed initial conditions for all involved quantities, we have no access to x_2 in the first year of the planning period. Afterwards, exact tracking can be achieved as above.

FIG. 1. *Time paths of the control variables.*

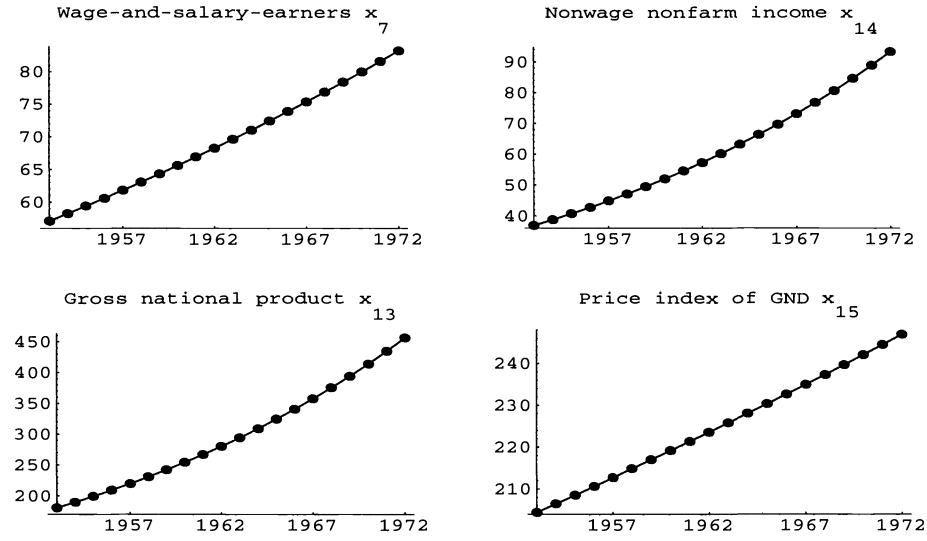


FIG. 2. *Computed target values.*

The results of the simulations can be seen in Figure 3. Clockwise, this figure shows the evolution of the gross private domestic capital formation x_2 as obtained by using extrapolated values for all exogenous variables, the desired time path of x_2, the computed sequence of controls attaining the control objective, and, finally, the resulting evolution of x_2. Again, exact matching is achieved starting with the year 1954.

7. Final remarks. In this article we focused on the problems of solvability and right-invertibility of implicit nonlinear discrete-time systems. The concept "solvability" has been defined for this class of systems in such a way that the existence of solutions for any locally given input sequence is ensured. Moreover we showed how this property can be checked algorithmically and how the algorithmically modified system can be equipped with additional conditions in order to obtain unique solutions which can be calculated recursively.

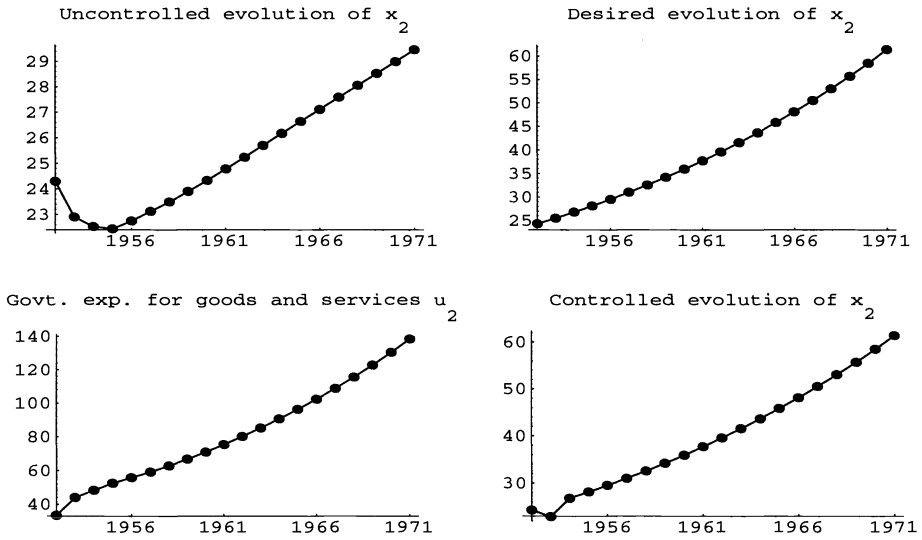


FIG. 3. *Time paths of the second control experiment.*

A second question was that of right-inversion of implicit systems. It turned out that the problem of right-inversion appears as a special solvability problem which can be solved quite routinely with the methods developed so far, lending itself to a numerical treatment.

We would like to mention that the approach used here to investigate the solvability of a difference-algebraic equation shows at least some similarity with the procedure which is used to determine what is called the *index* of a differential-algebraic equation, a natural number which is considered an indicator for the numerical complexity of such an equation. This similarity opens a wide field of research on the mutual connections between the solvability theories of constraint systems in discrete and continuous time.

Apart from a theoretical characterization of solvability and right-invertibility for implicit nonlinear discrete-time systems, a main objective was to demonstrate that important classes of such systems may be conveniently approached within the developed framework also from a computational point of view. One should not forget, however, that right-invertibility merely reflects an ideal with respect to a complete access to the outputs and does not say anything whatever about the possibility of a practical realization of the inputs required to enforce a desired output behaviour. Moreover, even if the necessary inputs can be generated, their values may not be acceptable from a practical point of view. A way out of this situation could consist in fixing an acceptance region for the inputs and finding, within this region, those inputs generating the output which is "nearest" to the desired one. We believe that the right-inversion approach as introduced above could be helpful in this direction.

Appendix.

Proof of Lemma 3.3. Consider the system of equations

$$f(z, x, u) = 0,$$
$$h(x, u, y) = 0,$$

and define $F := (f^T, h^T)^T$. Observe that we use slightly different notations to keep notations compact. In the first step of the shuffle algorithm let

$$\operatorname{rank} \partial F(z, x, u, y)/\partial(z, y) = \rho_1.$$

We show that ρ_2 does not depend on the selection of ρ_1 functionally independent (with respect to (z, y)) components of F in the first step. For the other steps, the proof is analogous. Now let $(\tilde{F}_1, \dots, \tilde{F}_{\rho_1}, \hat{F}_1, \dots, \hat{F}_{s+p-\rho_1})^T$ and $(\tilde{F}_1^*, \dots, \tilde{F}_{\rho_1}^*, \hat{F}_1^*, \dots, \hat{F}_{s+p-\rho_1}^*)^T$ denote two arbitrary permutations of components of F such that

$$\operatorname{rank} \frac{\partial F}{\partial(z, y)} = \operatorname{rank} \frac{\partial(\tilde{F}_1, \dots, \tilde{F}_{\rho_1})}{\partial(z, y)} = \operatorname{rank} \frac{\partial(\tilde{F}_1^*, \dots, \tilde{F}_{\rho_1}^*)}{\partial(z, y)} = \rho_1.$$

It then follows that

$$\hat{F}_k(z, x, u, y) = \xi_k(\tilde{F}_1, \dots, \tilde{F}_{\rho_1}, x, u), \qquad 1 \le k \le s + p - \rho_1,$$

and along trajectories of the system

$$\hat{F}_k(z, x, u, y) = \Phi_k(x, u).$$

Analogous equations hold for the starred functions. If $\{\tilde{F}_1, \dots, \tilde{F}_{\rho_1}\} = \{\tilde{F}_1^*, \dots, \tilde{F}_{\rho_1}^*\}$, nothing has to be shown. Assume therefore that this is not the case. Then there exists a number d with $1 \le d \le \rho_1 \le s + p$ such that the following hold:

$$\operatorname{card}(\{\tilde{F}_1^*, \dots, \tilde{F}_{\rho_1}^*\} \cap \{\hat{F}_1, \dots, \hat{F}_{s+p-\rho_1}\}) = d,$$

$$\operatorname{card}(\{\tilde{F}_1^*, \dots, \tilde{F}_{\rho_1}^*\} \cap \{\tilde{F}_1, \dots, \tilde{F}_{\rho_1}\}) = \rho_1 - d,$$

$$\operatorname{card}(\{\hat{F}_1^*, \dots, \hat{F}_{s+p-\rho_1}^*\} \cap \{\tilde{F}_1, \dots, \tilde{F}_{\rho_1}\}) = d,$$

$$\operatorname{card}(\{\hat{F}_1^*, \dots, \hat{F}_{s+p-\rho_1}^*\} \cap \{\hat{F}_1, \dots, \hat{F}_{s+p-\rho_1}\}) = s + p - \rho_1 - d.$$

Rename the components $(\tilde{F}_1, \dots, \tilde{F}_{\rho_1})$ and $(\hat{F}_1, \dots, \hat{F}_{s+p-\rho_1})$ on the one hand and the components $(\tilde{F}_1^*, \dots, \tilde{F}_{\rho_1}^*)$ and $(\hat{F}_1^*, \dots, \hat{F}_{s+p-\rho_1}^*)$ on the other hand to obtain

$$\begin{aligned}&(\tilde{F}_1^*, \dots, \tilde{F}_d^*, \tilde{F}_{d+1}^*, \dots, \tilde{F}_{\rho_1}^*, \hat{F}_1^*, \dots, \hat{F}_d^*, \hat{F}_{d+1}^*, \dots, \hat{F}_{s+p-\rho_1}^*)^T \\ &\quad = (\hat{F}_1, \dots, \hat{F}_d, \tilde{F}_1, \dots, \tilde{F}_{\rho_1-d}, \tilde{F}_{\rho_1-d+1}, \dots, \tilde{F}_{\rho_1}, \hat{F}_{d+1}, \dots, \hat{F}_{s+p-\rho_1})^T\end{aligned}$$

Observe that we only renamed components rather than altering groups of components that had been selected to be independent and dependent, respectively.

(i) For components $\hat{F}_1^*, \dots, \hat{F}_d^*$ we obtain

$$\begin{aligned}\hat{F}_k^* &= \xi_k^*(\tilde{F}_1^*, \dots, \tilde{F}_d^*, \tilde{F}_{d+1}^*, \dots, \tilde{F}_{\rho_1}^*, x, u) \\ &= \xi_k^*(\xi_1(\tilde{F}_1, \dots, \tilde{F}_{\rho_1}, x, u), \dots, \xi_d(\tilde{F}_1, \dots, \tilde{F}_{\rho_1}, x, u), \tilde{F}_{d+1}^*, \dots, \tilde{F}_{\rho_1}^*, x, u).\end{aligned}$$

Restricting ourselves to $\tilde{F}_1 = \cdots = \tilde{F}_{\rho_1} = 0$ we get, after applying the shift operator σ,

$$\xi_k^*(\Phi_1(z, u), \dots, \Phi_d(z, u), 0, \dots, 0, z, u) \equiv 0$$

$$\Longrightarrow \frac{\partial \xi_k^*}{\partial(z, y)} = \frac{\partial \xi_k^*}{\partial \tilde{F}_1^*} \frac{\partial \Phi_1}{\partial(z, y)} + \cdots + \frac{\partial \xi_k^*}{\partial \tilde{F}_d^*} \frac{\partial \Phi_d}{\partial(z, y)} + \frac{\partial \Phi_k^*}{\partial(z, y)} \equiv 0$$

$$\Longrightarrow \frac{\partial \Phi_k^*}{\partial(z, y)} = -\left(\frac{\partial \xi_k^*}{\partial \tilde{F}_1^*} \frac{\partial \Phi_1}{\partial(z, y)} + \cdots + \frac{\partial \xi_k^*}{\partial \tilde{F}_d^*} \frac{\partial \Phi_d}{\partial(z, y)} \right).$$

(ii) For components $\hat{F}^*_{d+1}, \ldots, \hat{F}^*_{s+p-\rho_1}$ we get by similar computations

$$\frac{\partial \Phi^*_k}{\partial(z, y)} = \frac{\partial \Phi_k}{\partial(z, y)} - \left(\frac{\partial \xi^*_k}{\partial \tilde{F}^*_1} \frac{\partial \Phi_1}{\partial(z, y)} + \cdots + \frac{\partial \xi^*_k}{\partial \tilde{F}^*_d} \frac{\partial \Phi_d}{\partial(z, y)} \right).$$

Partial derivatives with respect to y occurring in (i) and (ii) are of course equal to zero and appear only for formal reasons. $\partial \Phi^*/\partial(z, y)$ and $\partial \Phi/\partial(z, y)$ are therefore connected via

$$\text{(46)} \quad \left(\begin{array}{c} \frac{\partial \Phi^*_1}{\partial(z,y)} \\ \vdots \\ \frac{\partial \Phi^*_d}{\partial(z,y)} \\ \hline \frac{\partial \Phi^*_{d+1}}{\partial(z,y)} \\ \vdots \\ \frac{\partial \Phi^*_{s+p-\rho_1}}{\partial(z,y)} \end{array} \right) = \left(\begin{array}{ccc|ccc} -\frac{\partial \xi^*_1}{\partial \tilde{F}^*_1} & \cdots & -\frac{\partial \xi^*_1}{\partial \tilde{F}^*_d} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial \xi^*_d}{\partial \tilde{F}^*_1} & \cdots & -\frac{\partial \xi^*_d}{\partial \tilde{F}^*_d} & 0 & \cdots & 0 \\ \hline -\frac{\partial \xi^*_{d+1}}{\partial \tilde{F}^*_1} & \cdots & -\frac{\partial \xi^*_{d+1}}{\partial \tilde{F}^*_d} & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial \xi^*_{s+p-\rho_1}}{\partial \tilde{F}^*_1} & \cdots & -\frac{\partial \xi^*_{s+p-\rho_1}}{\partial \tilde{F}^*_d} & 0 & \cdots & 1 \end{array} \right) \left(\begin{array}{c} \frac{\partial \Phi_1}{\partial(z,y)} \\ \vdots \\ \frac{\partial \Phi_d}{\partial(z,y)} \\ \hline \frac{\partial \Phi_{d+1}}{\partial(z,y)} \\ \vdots \\ \frac{\partial \Phi_{s+p-\rho_1}}{\partial(z,y)} \end{array} \right).$$

Denote the coefficient matrix of (46) by A. A is invertible if and only if the upper left corner is nonsingular. It will turn out later that this is indeed the case. Now consider the set of components

$$\begin{aligned} (\tilde{F}_1, \ldots, \tilde{F}_{\rho_1}) &= (\tilde{F}^*_{d+1}, \ldots, \tilde{F}^*_{\rho_1}, \hat{F}^*_1, \ldots, \hat{F}^*_d) \\ &= (\tilde{F}^*_{d+1}, \ldots, \tilde{F}^*_{\rho_1}, \xi^*_1(\tilde{F}^*_1, \ldots, \tilde{F}^*_{\rho_1}, x, u), \ldots, \xi^*_d(\tilde{F}^*_1, \ldots, \tilde{F}^*_{\rho_1}, x, u)). \end{aligned}$$

This provides the following relation between

$$\frac{\partial(\tilde{F}_1, \ldots, \tilde{F}_{\rho_1})}{\partial(z, y)} \quad \text{and} \quad \frac{\partial(\tilde{F}^*_1, \ldots, \tilde{F}^*_{\rho_1})}{\partial(z, y)}:$$

$$\text{(47)} \quad \left(\begin{array}{c} \frac{\partial \tilde{F}_1}{\partial(z,y)} \\ \vdots \\ \frac{\partial \tilde{F}_{\rho_1-d}}{\partial(z,y)} \\ \hline \frac{\partial \tilde{F}_{\rho_1-d+1}}{\partial(z,y)} \\ \vdots \\ \frac{\partial \tilde{F}_{\rho_1}}{\partial(z,y)} \end{array} \right) = \left(\begin{array}{ccc|ccc} 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \\ \hline \frac{\partial \xi^*_1}{\partial \tilde{F}^*_1} & \cdots & \frac{\partial \xi^*_1}{\partial \tilde{F}^*_d} & \frac{\partial \xi^*_1}{\partial \tilde{F}^*_{d+1}} & \cdots & \frac{\partial \xi^*_1}{\partial \tilde{F}^*_{\rho_1}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \xi^*_d}{\partial \tilde{F}^*_1} & \cdots & \frac{\partial \xi^*_d}{\partial \tilde{F}^*_d} & \frac{\partial \xi^*_d}{\partial \tilde{F}^*_{d+1}} & \cdots & \frac{\partial \xi^*_d}{\partial \tilde{F}^*_{\rho_1}} \end{array} \right) \left(\begin{array}{c} \frac{\partial \tilde{F}^*_1}{\partial(z,y)} \\ \vdots \\ \frac{\partial \tilde{F}^*_d}{\partial(z,y)} \\ \hline \frac{\partial \tilde{F}^*_{d+1}}{\partial(z,y)} \\ \vdots \\ \frac{\partial \tilde{F}^*_{\rho_1}}{\partial(z,y)} \end{array} \right).$$

Denote the coefficient matrix of (47) by B. Observe that the lower left corner of B is, up to the sign, equal to the upper left corner of A. Since

$$\frac{\partial(\tilde{F}_1, \ldots, \tilde{F}_{\rho_1})}{\partial(z, y)} \quad \text{and} \quad \frac{\partial(\tilde{F}^*_1, \ldots, \tilde{F}^*_{\rho_1})}{\partial(z, y)}$$

have full row rank by assumption, B and consequently its lower left corner must be invertible in a neighbourhood of $(x_e, u_e, y_e) \implies A$ is invertible.

It finally follows that

$$\begin{pmatrix} \frac{\partial \tilde{F}_1^*}{\partial(z,y)} \\ \vdots \\ \frac{\partial \tilde{F}_{\rho_1}^*}{\partial(z,y)} \\ \frac{\partial \Phi_1^*}{\partial(z,y)} \\ \vdots \\ \frac{\partial \Phi_{s+p-\rho_1}^*}{\partial(z,y)} \end{pmatrix} = \begin{pmatrix} B^{-1} & 0 \\ 0 & A \end{pmatrix} \begin{pmatrix} \frac{\partial \tilde{F}_1}{\partial(z,y)} \\ \vdots \\ \frac{\partial \tilde{F}_{\rho_1}}{\partial(z,y)} \\ \frac{\partial \Phi_1}{\partial(z,y)} \\ \vdots \\ \frac{\partial \Phi_{s+p-\rho_1}}{\partial(z,y)} \end{pmatrix},$$

which proves that ρ_2 will not depend on the choice of independent components in Step 1.

REFERENCES

[1] I. ADELMAN AND F. L. ADELMAN, *The dynamic properties of the Klein-Goldberger model*, Econometrica, 27 (1959), pp. 596–625.

[2] G. BEAUCHAMP, *Algorithms for Singular Systems*, Ph.D. thesis, School of Electrical Engineering, Georgia Institute of Technology, Atlanta, 1990.

[3] R. G. BODKIN, L. R. KLEIN, AND K. MARWAH, *A History of Macroeconometric Model-Building*, reprinted, Elgar, Aldershot, UK, 1993.

[4] R. W. BROCKETT AND M. D. MESAROVIC, *The reproducibility of multivariable systems*, J. Math. Anal. Appl., 11 (1965), pp. 548–563.

[5] G. C. CHOW, *An approach to the feedback control of nonlinear economic systems*, in Frontiers of Quantitative Economics, Vol. IIIa, Papers Invited for Presentation at the Econometric Society Third World Congress, Toronto, 1975, M. D. Intriligator, ed., 1977, pp. 263–279.

[6] S. EL ASMI AND M. FLIESS, *Formule d'inversion*, in Analysis of Controlled Dynamical Systems, B. Bonnard, B. Bride, J. P. Gauthier, and I. Kupka, eds., Birkhäuser, Boston, 1991, pp. 201–210.

[7] ———, *Invertibility of discrete-time systems*, in Proc. 2nd IFAC Symposium on Nonlinear Control Systems Design, Bordeaux, 1992, pp. 192–196.

[8] M. EL-TOHAMI, V. LOVASS-NAGY, AND D. L. POWERS, *On minimal-order inverses of discrete-time descriptor systems*, Int. J. Control, 41 (1985), pp. 991–1004.

[9] T. FLIEGNER, H. NIJMEIJER, AND Ü. KOTTA, *Some aspects of nonlinear discrete-time descriptor systems in economics*, in Predictability and Nonlinear Modelling in Natural Sciences and Economics, J. Grasman and G. van Straten, eds., Kluwer, Dordrecht, 1994, pp. 582–590.

[10] L. R. KLEIN AND A. S. GOLDBERGER, *An Econometric Model of the United States*, 1929-1952, North-Holland, Amsterdam, 1955.

[11] Ü. KOTTA, *Right inversion of a discrete-time nonlinear system*, Int. J. Control, 51 (1990), pp. 1–9.

[12] D. G. LUENBERGER, *Dynamic equations in descriptor form*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 312–321.

[13] ———, *Nonlinear descriptor systems*, J. Econom. Dynamics Control, 1 (1979), pp. 219–242.

[14] ———, *Time-invariant descriptor systems*, Automatica, 14 (1978), pp. 473–480.

[15] H. NIJMEIJER, *Remarks on the control of discrete-time nonlinear systems*, in Perspectives in Control Theory, Proc. Sielpia Conference, B. Jakubczyk, K. Malanowski, and W. Respondek, eds., Birkhäuser, 1990, pp. 261–276.

[16] W. RESPONDEK AND H. NIJMEIJER, *On local right-invertibility of nonlinear control systems*, Control Theory Adv. Tech., 4 (1988), pp. 325–348.

[17] W. RESPONDEK, *Right and left invertibility of nonlinear control systems*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 133–176.

[18] M. K. SAIN AND J. L. MASSEY, *Invertibility of linear time-invariant dynamical systems*, IEEE Trans. Automat. Control, AC-14 (1969), pp. 141–149.

[19] L. M. SILVERMAN, *Inversion of multivariable linear systems*, IEEE Trans. Automat. Control, AC-14 (1969), pp. 270–276.

[20] S. N. SINGH, *A modified algorithm for invertibility in nonlinear systems*, IEEE Trans. Automat. Control, AC-26 (1980), pp. 595–598.

[21] S. TAN AND J. VANDEWALLE, *Inversion of singular systems*, IEEE Trans. Circuits Systems I Fund Theory Appl., CS-35 (1988), pp. 583–587.

[22] J. TINBERGEN, *On the Theory of Economic Policy*, North-Holland, Amsterdam, 1952.

[23] J. C. WILLEMS, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 259–294.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 2116–2132, November 1996

014

A TARGET RECOGNITION PROBLEM: SEQUENTIAL ANALYSIS AND OPTIMAL CONTROL*

MARK H. A. DAVIS† AND MOHAMMAD FARID‡

Abstract. An iterative computational method for determining the value function of an optimal control problem, related to target tracking, is presented. The target is assumed to be located in a fixed known position in space, but its identity (hostile or friendly) is known only with a prior probability. An observation of the target can be made at any location, and its error has position-dependent probability. The objective is finding the optimal navigation and observation strategy which leads to a final decision (i.e., the target is friendly or hostile). The value function is shown to be the unique viscosity solution of a variational inequality. Furthermore it is the unique fixed point of a nondecreasing concave operator.

Key words. optimal control, viscosity solutions, variational inequality, dynamic programming, Hamilton–Jacobi–Bellman equation, target tracking, hypothesis testing

AMS subject classifications. 49J40, 49L25, 62C10, 62K05, 62L10, 93C15

1. Introduction.

An aircraft locates a target and seeks to classify it as one of a finite number of possible objects. This is done by taking a sequence of observations, each observation (consisting, say, of emission of a radar pulse) being processed to produce a classification which is subject to errors occurring with distance and relative orientation-dependent probabilities. After a certain number of observations a final classification decision is made. There are known penalties for misclassification and costs associated with taking observations. The latter might be actual costs of physically taking an observation, or more indirect penalties associated with, for example, the risk of giving one's own position away by emitting a radar pulse. The problem is to decide how to navigate, when to take the observations, and at what point to make a final classification, in such a way as to minimize overall costs.

This problem incorporates features of several traditional statistical and control-theoretic paradigms. It is *sequential analysis* in that a data-dependent number of observations is taken. There is an element of *experimental design* in that the distribution of the observations is not fixed in advance but depends on some design parameters. It is *optimal control* in that these design parameters are actually the control inputs to a dynamical system. This combination of features does not appear to have been studied before, and our objective is to outline how the problem is solved in a fairly simple setting, described in more detail below.

The problem arose in connection with a study of Bayes-optimal tracking and interception strategies. It has of course been widely appreciated since the pioneering work of Fel'dbaum [F] on "dual control" that control problems in which the control is used with a view to acquiring information as well as steering along a low-cost trajectory are typically extremely hard. Considerable insight into the form of optimal strategies in the case of a binary (friend/foe) classification with penalties for missing a foe or hitting a friend has been gained by exact solution of very simple discrete-time models [H]. It turns out that, typically, an optimal strategy avoids "commitment" as long as possible, steering to a point as close to the target as possible but from which both hitting the target and avoiding it are still feasible. Up to arrival at this point the main function of the control action is information gathering. At the "commitment point"

*Received by the editors August 31, 1994; accepted for publication (in revised form) October 5, 1995. This research was supported by an SERC/EPSRC grant and agreement 2037/393/RAE with the UK Defence Research Agency (DRA).

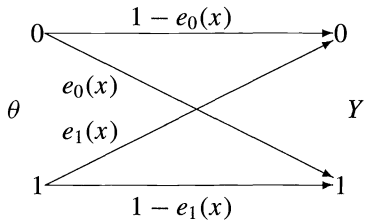
†Center for Process Systems Engineering, Imperial College, London SW7 2BY, UK (mhad@ic.ac.uk).

‡Control Engineering Research Center, City University, London EC1V 0HB, UK (M.Farid@city.ac.uk). This work was carried out while this author was with the Center for Process Systems Engineering, Imperial College, London SW72BY, UK.

a classification decision must be made and the appropriate subsequent action taken. To study tracking in a more general setting it seems worthwhile to isolate the "information-gathering" phase as a separate problem in its own right, and this is what we do in the present paper. Of course, various applications of the problem formulation, other than the one outlined above, are easily envisaged.

The paper is laid out as follows. In §2 below a precise formulation of the problem is given. As will be seen, the target is assumed to be stationary and the "own vehicle" dynamics are deterministic, so the only sources of uncertainty in the problem arise from the classification mechanism. For simplicity we assume a binary (friend/foe) classification, but the theory would be the same for any finite classification. In §3 Bayes' formula is used to make an information-updating procedure after each observation. A dynamic programming formulation of the problem is described, and the variational inequality formally satisfied by the value function is obtained. Later in §4 the value function is proven to be the unique viscosity solution of this variational inequality. We also present a computational algorithm for the value function and prove its convergence. Finally, some simulation results are presented in §5.

2. Problem formulation. A target is located in a certain fixed known position in space but may be hostile or friendly. Let $\theta = \mathbf{1}_{(\text{target is hostile})}$, and suppose that the event $(\theta = 1)$ has prior probability p_0. An observation of the target can be made at any point x in space (or, more generally, at any point in the state space of the vehicle; see below), and the observation device produces an output Y which takes the values 0, 1 and misclassifies the target with position-dependent probabilities $e_0(x)$, $e_1(x)$, i.e., $P[Y = 1|\theta = 0] = e_0(x)$ and $P[Y = 0|\theta = 1] = e_1(x)$ (see below).



For definiteness, we assume that

$$0 < r \leq e_0(x), e_1(x) \leq \frac{1}{2}. \tag{1}$$

The observations are taken from a vehicle whose state $x(s)$ satisfies the dynamical equation

$$\dot{x}(s) = g(x(s), u(s)). \tag{2}$$

Here $x(s) \in \mathbb{R}^n$ and $u(s)$ is a control, taking values in U, a subset of $\mathbb{R}^m$. We make the following assumptions:

(3) U is compact,

(4) $g \in C(\mathbb{R}^n \times U)$,

(5) $\exists K > 0$: $|g(x, u) - g(x', u)| \leq K|x - x'| \quad \forall x, x' \in \mathbb{R}^n$ and $\forall u \in U$,

(6) g is bounded.

Equation (2) then has a global solution for any measurable control function $u(\cdot)$ and starting point $x(0)$.

An *observation strategy* is a collection $S = \{k, \tau_1, \dots, \tau_k, u, d\}$, where k is a nonnegative integer, $\tau_1 \le \tau_2 \le \cdots \le \tau_k$ are observation times (there are none of these if $k = 0$), $u = \{u(t), 0 \le t \le \tau_k\}$ is a control, and d is a final classification of the target. Observations Y_j are taken at each time τ_j, $j = 1, \dots, k$ (when the vehicle's state is $x(\tau_j)$).

Admissible strategies are nonanticipative in the sense that either $k = 0$ and $d \in \{0, 1\}$ is fixed decision or $\tau_1 \ge 0$ and $u_1 = \{u(t), 0 \le t \le \tau_1\}$ are fixed and τ_j and $u_j = \{u(t), \tau_{j-1} < t \le \tau_j\}$ are functions of $(Y_1, \dots, Y_{j-1})$ only, for $j = 2, \dots, k$. It may be the case that $\tau_{j-1} = \tau_j$ with positive probability. The number k is a stopping time of the filtration $\mathcal{Y}_j = \sigma\{Y_1, \dots, Y_j\}$, and the decision d is $\mathcal{Y}_k$-measurable. A penalty or cost $\ell_1 > 0$ is incurred if a hostile target is identified as friendly, and a cost $\ell_2 > 0$ is incurred for the converse error. In addition, a cost of $c(x(\tau_j))$ is paid each time that an observation is taken, and a cost $\ell(x, u)$ per unit time is paid when the vehicle is in state x, control action u is applied, and $t < \tau_k$. We assume $c(x)$, $\ell(x, u) \ge \delta > 0$, and that c, ℓ are uniformly continuous on $\mathbb{R}^n$ and $\mathbb{R}^n \times U$, respectively. The problem is now to choose a strategy S to minimize

$$J(S) = \mathbb{E}\left\{\int_0^{\tau_k} \ell(x(s), u(s))\, ds + \sum_{j=1}^{k} c(x(\tau_j)) + \ell_1(1-d)\theta + \ell_2\, d(1-\theta)\right\}. \tag{7}$$

One possible strategy is to make an immediate decision at time zero without taking any observations, and for this strategy clearly $J(S) \le L = \max(\ell_1, \ell_2)$. We can therefore restrict attention to strategies such that $\mathbb{E}[k]$, $\mathbb{E}[\tau_k] \le L/\delta$, so that in particular k and τ_k are finite with probability 1.

Let $\pi_{\tau_k} = P[\theta = 1 | \mathcal{Y}_k]$. Temporarily writing $\pi = \pi_{\tau_k}$, we have

$$\begin{aligned}\mathbb{E}[\ell_1(1-d)\theta + \ell_2\, d(1-\theta)] &= \mathbb{E}[\ell_1(1-d)\pi + \ell_2\, d(1-\pi)] \\ &= \mathbb{E}[\ell_1\pi + (\ell_2 - (\ell_1+\ell_2)\pi)d].\end{aligned}$$

The minimum cost decision is therefore $d = \mathbf{1}_{(\pi \ge \ell_2/(\ell_1+\ell_2))}$. Thus d can be taken out of the problem and the cost written as

$$J(S) = \mathbb{E}\left\{\int_0^{\tau_k} \ell(x(s), u(s))\, ds + \sum_{j=1}^{k} c(x(\tau_j)) + h(\pi_{\tau_k})\right\}, \tag{8}$$

where

$$h(\pi) := \min\{\pi\ell_1,\ (1-\pi)\ell_2\}. \tag{9}$$

3. Dynamic programming. For a strategy S, define the filtration $\mathcal{F}_t$ as follows:

$$\mathcal{F}_t := \sigma\{Y_j\, \mathbf{1}_{(\tau_j \le t)},\quad j = 1, 2, \dots\}.$$

Clearly $\mathcal{F}_t \bigcap(\tau_j \le t < \tau_{j+1}) = \mathcal{F}_{\tau_j} \bigcap(\tau_j \le t < \tau_{j+1})$. We denote $\pi(t) = P[\theta = 1 | \mathcal{F}_t]$. At time τ_j, random variable Y_j is observed. Denote temporarily $\pi = \pi(\tau_j^-)$ $(= \pi(\tau_{j-1})$ if $\tau_j > \tau_{j-1})$ and $x = x(\tau_j)$. By Bayes' formula, π is updated to π_0' or π_1' when Y_j takes the values 0 or 1, respectively, where π_0' and π_1' are given by

$$\begin{aligned}\pi_0' &= P_\pi[\theta = 1 | Y = 0] \\ &= \frac{P[Y = 0|\theta = 1]\, P_\pi[\theta = 1]}{P[Y = 0|\theta = 1]\, P_\pi[\theta = 1] + P[Y = 0|\theta = 0]\, P_\pi[\theta = 0]} \\ &= \frac{e_1(x)\pi}{e_1(x)\pi + (1 - e_0(x))(1-\pi)}, \qquad (10)\end{aligned}$$

and similarly

$$\pi_1' = \frac{(1 - e_1(x))\pi}{(1 - e_1(x))\pi + e_0(x)(1 - \pi)}. \tag{11}$$

If $V: \mathbb{R}^n \times [0, 1] \longmapsto \mathbb{R}$ is a given bounded measurable function, we therefore have

$$\begin{aligned}\mathbb{E}[V(x(\tau_j), \pi(\tau_j))|\mathcal{F}_{\tau_j^-}] &= V(x, \pi_0')\ P_\pi[Y_{\tau_j} = 0] + V(x, \pi_1')\ P_\pi[Y_{\tau_j} = 1] \\ &= V(x, \pi_0') + (V(x, \pi_1') - V(x, \pi_0'))((1 - e_1(x))\pi \\ &\qquad + e_0(x)(1 - \pi)),\end{aligned}$$

since

$$\begin{aligned}P_\pi[Y_{\tau_j} = 1] &= P_\pi[Y_{\tau_j} = 1|\theta = 1]\ P_\pi[\theta = 1] + P_\pi[Y_{\tau_j} = 1|\theta = 0]\ P_\pi[\theta = 0] \\ &= (1 - e_1(x))\pi + e_0(x)(1 - \pi),\end{aligned} \tag{12}$$

and $P_\pi[Y_{\tau_j} = 0] = 1 - P_\pi[Y_{\tau_j} = 1]$. We define an operator M acting on the space of bounded measurable functions as follows:

$$MV(x, \pi) := c(x) + V(x, \pi_0') + (V(x, \pi_1') - V(x, \pi_0'))((1 - e_1(x))\pi + e_0(x)(1 - \pi)), \tag{13}$$

where π_0', π_1' are defined by (10) and (11). $MV(x, \pi)$ represents the average cost paid if one observation Y is taken, updating the original π to $\pi^+ = \pi_0'\ \mathbf{1}_{(Y=0)} + \pi_1'\ \mathbf{1}_{(Y=1)}$, and then a cost $V(x, \pi^+)$ is paid. For example, let S' be the strategy such that $k = 1$ and $\tau_k = 0$; i.e., one observation is taken at time zero and then a decision is made. Then $J(S') = M\upsilon_0(x, p_0)$, where $\upsilon_0(x, \pi) = h(\pi)$, h is given by (9), and p_0 is the prior probability that $\theta = 1$.

Define the *value function* $W(x, \pi)$ as

$$W(x, \pi) = \inf_{S \in \mathbb{S}_{ad}} \mathbb{E}\left\{\int_0^{\tau_k} \ell(x(s), u(s))\, ds\ + \sum_{j=1}^{k} c(x(\tau_j)) + h(\pi(\tau_k))\right\}, \tag{14}$$

where x and π are, respectively, the starting point and the prior classification probability and $\mathbb{S}_{ad}$ is the set of all admissible strategies as defined in §2. At time zero, there are three possible courses of action:

(i) Take no observations and make an immediate decision; expected cost $= h(\pi)$.

(ii) Take an observation immediately and then continue "optimally"; expected cost $= MW(x, \pi)$.

(iii) Maneuver for a short time t before continuing optimally. In this case we have

$$W(x, \pi) \leq \inf_{u \in U} \left\{\int_0^t \ell(x(s), u(s))\, ds\ + W(x(t), \pi)\right\},$$

where $x(s)$ is the solution of the dynamical equation (2). Assuming that W is C^1 in x, this leads in the standard way to the inequality

$$\sup_{u \in U}[-g(x, u) \cdot D_x W(x, \pi) - \ell(x, u)] \leq 0.$$

Combining this with the inequalities $W \leq h$ and $W \leq MW$ obtained from (i) and (ii) we obtain the basic *variational inequality*

$$\max\left\{\sup_{u \in U}[-g(x, u) \cdot D_x W(x, \pi) - \ell(x, u)],\ W - h,\ W - MW\right\} = 0. \tag{15}$$

The (x, π) space splits into three regions $\mathcal{C}, \mathcal{S}$, and $\mathcal{O}$ in which each of the three expressions in (15) is equal to zero (in the respective order). The conjectured form of the optimal strategy is this: starting at $(x, \pi) \in \mathcal{C}$ (the *continuation region*) the process evolves, with π constant, until region $\mathcal{O}$ (for *observation*) is hit at, say, (x', π). (It cannot be the case that $\mathcal{S}$ is hit first.) An observation is now taken, and the process jumps to (x', π^+), where $\pi^+ = \pi_0'$ or π_1' (see (10) and (11)). If $(x', \pi^+) \in \mathcal{C}$, this process is repeated; if $(x', \pi^+) \in \mathcal{S}$ (*stopping region*), one stops, paying $h(\pi^+)$, while if $(x', \pi^+) \in \mathcal{O}$, one takes another observation. It is thus possible in principle to take a sequence of observations at the same place and time, but as argued above, both k and τ_k are a.s. finite. The variational inequality in (15) may be written as follows:

$$\max\left\{\sup_{u\in U}[-g(x,u)\cdot D_x W(x,\pi) - \ell(x,u)],\ W(x,\pi) - NW(x,\pi)\right\} = 0, \tag{16}$$

where $NW(x, \pi) := h(\pi) \wedge MW(x, \pi)$. The variational inequality given in (16) will be used throughout the following section.

4. Viscosity solution. Since we cannot guarantee in advance that (16) has a C^1 solution, a rigorous theory should be sought in the framework of viscosity solutions [CL] and [FS]. An approach to this is as follows. Let $B(E)_+$ denote the set of all positive bounded measurable functions on E. Define

$$NV(x, \pi) := h(\pi) \wedge MV(x, \pi), \tag{17}$$

and for a given function $\psi \in B(\mathbb{R}^n \times [0, 1])_+$ define an operator $\mathcal{G}$ by

$$\mathcal{G}\psi(x, \pi) := \inf_{(u,t_f)} \int_0^{t_f} \ell(x(s), u(s))\, ds + N\psi(x(t_f), \pi), \tag{18}$$

where the infimum is taken over pairs $t_f \geq 0$ and $u \in L^\infty([0, t_f]; U)$. Thus the $\mathcal{G}$-operator defines a free end-time deterministic optimal control problem.

Let us give the definition of viscosity solutions for a variational inequality [B].

DEFINITION 4.1. *Let* $S_0 = \mathbb{R}^n \times (0, 1)$, $\psi \in BUC(\bar{S}_0)$, *and consider the following variational inequality for all* $(x, \pi) \in S_0$:

$$\max\left\{\sup_{u\in U}[-g(x,u)\cdot D_x V(x,\pi) - \ell(x,u)],\ V(x,\pi) - \psi(x,\pi)\right\} = 0. \tag{19}$$

Assume $V \in BUC(\bar{S}_0)$; *then viscosity solutions are defined as follows.*

(a) *V is a viscosity subsolution of* (19) *in* S_0 *if for each* $w \in C^1(S_0)$,

$$\max\left\{\sup_{u\in U}[-g(\bar{x},u)\cdot D_x w(\bar{x},\bar{\pi}) - \ell(\bar{x},u)],\ V(\bar{x},\bar{\pi}) - \psi(\bar{x},\bar{\pi})\right\} \leq 0$$

at every $(\bar{x}, \bar{\pi}) \in S_0$ *which is a local maximum of* $V - w$ *on* S_0.

(b) *V is a viscosity supersolution of* (19) *in* S_0 *if for each* $w \in C^1(S_0)$,

$$\max\left\{\sup_{u\in U}[-g(\bar{x},u)\cdot D_x w(\bar{x},\bar{\pi}) - \ell(\bar{x},u)],\ V(\bar{x},\bar{\pi}) - \psi(\bar{x},\bar{\pi})\right\} \geq 0$$

at every $(\bar{x}, \bar{\pi}) \in S_0$ *which is a local minimum of* $V - w$ *on* S_0.

(c) *V is a viscosity solution of* (19) *in* S_0 *if it is both a viscosity subsolution and a viscosity supersolution of* (19) *in* S_0. □

THEOREM 4.2. *Assume* (1)–(6). *Moreover assume that* c, ℓ, *and* ψ *are bounded and uniformly continuous on* $\mathbb{R}^n$, $\mathbb{R}^n \times U$, *and* $\bar{S_0}$, *respectively. We also assume that* e_0 *and* e_1 *are uniformly continuous on* $\mathbb{R}^n$ *and* ψ *is positive. Then* $V(x,\pi) = \mathcal{G}\psi(x,\pi)$ *is bounded uniformly continuous in* $\bar{S_0}$ *and is a viscosity solution of the equation*

$$F_\psi(x,\pi,V,D_xV) = 0, \tag{20}$$

where

$$F_\psi(x,\pi,r,p) = \max\left\{\sup_{u\in U}[-g(x,u)\cdot p - \ell(x,u)],\ r - N\psi(x,\pi)\right\}.$$

The following three lemmas prove that $V \in BUC(\bar{S_0})$.

LEMMA 4.3. $\psi \in BUC(\bar{S_0})$ *implies that* $N\psi \in BUC(\bar{S_0})$.

Proof. The following is an outline of the proof. Given the above assumptions, π_0' and π_1' ((10) and (11)) are bounded uniformly continuous functions in $\bar{S_0}$. So $\psi \in BUC(\bar{S_0})$ implies that $M\psi \in BUC(\bar{S_0})$. It is easily verified that h (refer to (9)) is bounded uniformly continuous in $[0,1]$. Finally, $N\psi(x,\pi) = \min\{h(\pi),\ M\psi(x,\pi)\}$, so $N\psi \in BUC(\bar{S_0})$. □

LEMMA 4.4. *For* $\rho \geq 0$ *define*

$$m_\ell(\rho) = \sup\{|\ell(x,u) - \ell(y,u)|\colon\ |x-y| \leq \rho,\ u \in U\},$$

$$m_\psi(\rho) = \sup\{|N\psi(x,\pi) - N\psi(y,\pi')|\colon\ |x-y| + |\pi - \pi'| \leq \rho\}.$$

Then the uniform continuity of ℓ *and* $N\psi$ *implies that* $m_\ell, m_\psi \in C([0,\infty])$ *and* $m_\ell(0) = 0$, $m_\psi(0) = 0$.

Proof. It can easily be verified. For a similar argument refer to the proof of theorem II(10.1), pp. 95–97 in [FS]. □

LEMMA 4.5. $V \leq h(\ell_2/(\ell_1+\ell_2))$ *and*

$$|V(x,\pi) - V(y,\pi')| \leq \tau_{\max} m_\ell(|x-y|e^{K\tau_{\max}}) + m_\psi(|x-y|e^{K\tau_{\max}}) + m_\psi(|\pi-\pi'|),$$

where $\tau_{\max} = h(\ell_2/(\ell_1+\ell_2))/\delta$ *and* K *is the Lipschitz constant. Hence* $V \in BUC(\bar{S_0})$.

Proof. Since h achieves its maximum at $\ell_2/(\ell_1+\ell_2)$ and $V \leq h$ (take $t_f = 0$ in (18)), one can write

$$V(x,\pi) \leq h(\ell_2/(\ell_1+\ell_2)),$$

$$V(x,\pi) = \inf_{\tau\in[0,\tau_{\max}],\ u(\cdot)\in L^\infty([0,\tau_{\max}];U)} \int_0^\tau \ell(x(s),u(s))\,ds + N\psi(x(\tau),\pi).$$

Using $\inf\ldots - \inf\ldots \leq \sup\ldots$ and Gronwall's inequality [PSV], one can obtain the required result. □

Proof of Theorem 4.2. The above lemmas imply that $V \in BUC(\bar{S_0})$. Now let us prove that V is a viscosity solution of (20). The argument is very similar to that in [B]. Let $\varphi \in C^1(\mathbb{R}^n \times (0,1))$, and assume that $(\bar{x},\bar{\pi}) \in \mathbb{R}^n \times (0,1)$ is a local maximum point of $V - \varphi$. Then for a fixed control $\bar{u} \in U$ and all $T > 0$ we have

$$V(\bar{x},\bar{\pi}) \leq \int_0^T \ell(x(s),\bar{u})\,ds + V(x(T),\bar{\pi}).$$

But for small values of T one can write

$$V(x(T),\bar{\pi}) - \varphi(x(T),\bar{\pi}) \leq V(\bar{x},\bar{\pi}) - \varphi(\bar{x},\bar{\pi}),$$

so replacing for $V(x(T), \bar{\pi})$ from the above inequality gives

$$\int_0^T \ell(x(s), \bar{u})\, ds + (\varphi(x(T), \bar{\pi}) - \varphi(\bar{x}, \bar{\pi})) \geq 0.$$

Then dividing both sides of the above inequality by T and taking the limit when $T \downarrow 0$, one finally gets

$$-g(\bar{x}, \bar{u}) \cdot D_x \varphi(\bar{x}, \bar{\pi}) - \ell(\bar{x}, \bar{u}) \leq 0 \quad \forall\, \bar{u} \in U.$$

According to the definition of $V(x, \pi)$ one can immediately say that $V(\bar{x}, \bar{\pi}) \leq N\psi(\bar{x}, \bar{\pi})$, so

$$\max\left\{\sup_{u \in U}[-g(\bar{x}, u) \cdot D_x \varphi(\bar{x}, \bar{\pi}) - \ell(\bar{x}, u)],\ V(\bar{x}, \bar{\pi}) - N\psi(\bar{x}, \bar{\pi})\right\} \leq 0; \tag{21}$$

in other words, V is a viscosity subsolution of (20).

Now let $\varphi \in C^1(\mathbb{R}^n \times (0, 1))$, and assume that $(\bar{x}, \bar{\pi}) \in \mathbb{R}^n \times (0, 1)$ is a local minimum point of $V - \varphi$. We have two cases.

Case 1. If $V(\bar{x}, \bar{\pi}) = N\psi(\bar{x}, \bar{\pi})$, then there is nothing to prove.

Case 2. $V(\bar{x}, \bar{\pi}) < N\psi(\bar{x}, \bar{\pi})$, which means $(\bar{x}, \bar{\pi}) \in \mathcal{C}$ (continuation region). So $\exists \epsilon > 0$ such that $\forall\, T \in (0, \epsilon)$ we have

$$V(\bar{x}, \bar{\pi}) = \inf_{u \in U}\left\{\int_0^T \ell(x(s), u(s))\, ds + V(x(T), \bar{\pi})\right\},$$

but if we choose T small enough, one can write

$$V(x(T), \bar{\pi}) - \varphi(x(T), \bar{\pi}) \geq V(\bar{x}, \bar{\pi}) - \varphi(\bar{x}, \bar{\pi}).$$

Again replacing for $V(x(T), \bar{\pi})$ from the above inequality gives

$$\inf_{u \in U}\left\{\int_0^T \ell(x(s), u(s))\, ds + (\varphi(x(T), \bar{\pi}) - \varphi(\bar{x}, \bar{\pi}))\right\} \leq 0.$$

Then dividing both sides by T and taking the limit when $T \downarrow 0$, one finally obtains

$$\sup_{u \in U}[-g(\bar{x}, u) \cdot D_x \varphi(\bar{x}, \bar{\pi}) - \ell(\bar{x}, u)] \geq 0,$$

which immediately implies

$$\max\left\{\sup_{u \in U}[-g(\bar{x}, u) \cdot D_x \varphi(\bar{x}, \bar{\pi}) - \ell(\bar{x}, u)],\ V(\bar{x}, \bar{\pi}) - N\psi(\bar{x}, \bar{\pi})\right\} \geq 0, \tag{22}$$

so V is a viscosity supersolution of (20). If we use the definition of viscosity solutions and inequalities in (21) and (22), it is easy to see that V is a viscosity solution of (20). □

Now we prove that $V(x, \pi) = \mathcal{G}\psi(x, \pi)$ is the unique viscosity solution of (20).

THEOREM 4.6. *Consider the variational inequality for all* $(x, \pi) \in \mathbb{R}^n \times (0, 1)$,

$$\max\left\{\sup_{u \in U}[-\ell(x, u) - g(x, u) \cdot D_x V(x, \pi)],\ V(x, \pi) - \psi(x, \pi)\right\} = 0, \tag{23}$$

and make all the assumptions of Theorem 4.2; *then if* $\psi(x, \pi) > 0\ \forall\, (x, \pi) \in \mathbb{R}^n \times (0, 1)$, *there is at most one viscosity solution of* (23).

We will need the following lemmas.

LEMMA 4.7. *Define* $H(x, p) := \sup_{u \in U}[-\ell(x, u) - g(x, u) \cdot p]$; *then* $H(x, p)$ *is a convex function with respect to* p.

Proof. This is easily verified. □

LEMMA 4.8. *Define*

$$\tilde{H}(x, r, p) := \max\{H(x, p),\ r - \psi(x)\},$$

where

$$\ell(x, u) \geq \delta > 0 \quad \forall\, (x, u) \in R^n \times U,$$
$$\psi(x) > 0 \quad \forall\, x \in R^n.$$

Then $\tilde{H}(x, V, DV)$ *has a strict subsolution, i.e.,*

$$\exists w \in C^1(\mathbb{R}^n) \cap BUC(\mathbb{R}^n) \text{ s.t. } \tilde{H}(x, w, Dw) < 0 \text{ in } \mathbb{R}^n.$$

Proof. It is easily seen that $w(x) \equiv 0$ is a strict subsolution. □

LEMMA 4.9. *Assume that* $u(x)$ *is a viscosity subsolution of* $\tilde{H}(x, V, DV) = 0$. *Then* $\eta u(x) + (1 - \eta)w(x)$ *is a strict viscosity subsolution of* $\tilde{H}$, *where* $w(x)$ *is a strict subsolution and* $\eta \in (0, 1)$.

Proof. u is a viscosity subsolution of $\tilde{H}(x, V, DV) = 0$, so for all $\varphi \in C^1(\mathbb{R}^n)$, if $u - \varphi$ attains a local maximum at $x_0 \in \mathbb{R}^n$, then

$$\tilde{H}(x_0, u(x_0), D\varphi(x_0)) \leq 0.$$

When $u - \varphi$ attains a local maximum at x_0, so does $\eta u + (1 - \eta)w - (\eta\varphi + (1 - \eta)w)$ for $\eta \in (0, 1)$. Here $w \in C^1(\mathbb{R}^n)$ is a strict subsolution of $\tilde{H}$. Now we show $\eta u + (1 - \eta)w$ is a strict viscosity subsolution of $\tilde{H}$:

$$\begin{aligned}
&\tilde{H}(x_0,\ \eta u(x_0) + (1 - \eta)w(x_0),\ \eta D\varphi(x_0) + (1 - \eta)Dw(x_0)) \\
&\quad = \max\{H(x_0, \eta D\varphi(x_0) + (1 - \eta)Dw(x_0)),\ \eta u(x_0) + (1 - \eta)w(x_0) - \psi(x_0)\} \\
&\quad \leq \max\{\eta H(x_0, D\varphi(x_0)) + (1 - \eta)H(x_0, Dw(x_0)),\ \eta u(x_0) + (1 - \eta)w(x_0) - \psi(x_0)\} \\
&\quad \leq \eta \max\{H(x_0, D\varphi(x_0)),\ u(x_0) - \psi(x_0)\} \\
&\qquad + (1 - \eta)\max\{H(x_0, Dw(x_0)),\ w(x_0) - \psi(x_0)\} \\
&\quad < 0,
\end{aligned}$$

because u is a viscosity subsolution and w is a strict subsolution. □

Proof of Theorem 4.6. Without loss of generality we drop π in all arguments. Consider the following auxiliary test function:

$$\Phi(x, y) = v_1(x) - v_2(y) - \frac{(x - y)^2}{2\epsilon}, \qquad \epsilon > 0,$$

where v_1 and v_2 are a viscosity subsolution and a viscosity supersolution of (23), respectively. In our problem any viscosity subsolution or supersolution must be bounded and uniformly continuous, so $\Phi(x, y)$ has a maximiser over $\mathbb{R}^n \times \mathbb{R}^n$ at $(\bar{x}_\epsilon, \bar{y}_\epsilon) \in \bar{Q} \times \bar{Q}$, where Q is a bounded subset of $\mathbb{R}^n$. Obviously

$$\Phi(x, y) \leq \Phi(\bar{x}_\epsilon, \bar{y}_\epsilon) \quad \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n. \tag{24}$$

For $\rho \geq 0$ define

$$D_\rho = \{(x, y) \in \bar{Q} \times \bar{Q} : |x - y|^2 \leq \rho\},$$
$$m_{v_1}(\rho) = 2\sup\{|v_1(x) - v_1(y)| : (x, y) \in D_\rho\},$$
$$K_1 = \sup\{m_{v_1}(\rho) : \rho \geq 0\}.$$

$\bar{Q}$ is compact and v_1 is uniformly continuous on $\bar{Q}$. Thus $m_{v_1} \in C([0,\infty))$ with $m_{v_1}(0) = 0$. We have

$$\Phi(\bar{y}_\epsilon, \bar{y}_\epsilon) \le \Phi(\bar{x}_\epsilon, \bar{y}_\epsilon),$$

so

$$\frac{(\bar{x}_\epsilon - \bar{y}_\epsilon)^2}{\epsilon} \le 2(v_1(\bar{x}_\epsilon) - v_1(\bar{y}_\epsilon)) \le K_1, \tag{25}$$

which gives

$$|\bar{x}_\epsilon - \bar{y}_\epsilon| \le \sqrt{K_1 \epsilon}. \tag{26}$$

One can observe that $|\bar{x}_\epsilon - \bar{y}_\epsilon| \to 0$ as $\epsilon \downarrow 0$. Using (25) and (26) it is easily verified that $(\bar{x}_\epsilon - \bar{y}_\epsilon)^2/\epsilon \downarrow 0$ as $\epsilon \downarrow 0$. Now let us define $p_\epsilon := (\bar{x}_\epsilon - \bar{y}_\epsilon)/\epsilon$. Then

$$\begin{aligned} |H(\bar{x}_\epsilon, p_\epsilon) - H(\bar{y}_\epsilon, p_\epsilon)| &\le \sup_{u\in U} |(g(\bar{y}_\epsilon, u) - g(\bar{x}_\epsilon, u)) \cdot p_\epsilon + \ell(\bar{y}_\epsilon, u) - \ell(\bar{x}_\epsilon, u)| \\ &\le \sup_{u\in U} |\ell(\bar{x}_\epsilon, u) - \ell(\bar{y}_\epsilon, u)| + \sup_{u\in U} |g(\bar{x}_\epsilon, u) - g(\bar{y}_\epsilon, u)|\, |p_\epsilon| \\ &\le m_\ell(|\bar{x}_\epsilon - \bar{y}_\epsilon|) + K|\bar{x}_\epsilon - \bar{y}_\epsilon||p_\epsilon|, \end{aligned} \tag{27}$$

where K is the Lipschitz constant of g and $m_\ell \in C([0,\infty))$ with $m_\ell(0) = 0$. Using (27) one can show that $|H(\bar{x}_\epsilon, p_\epsilon) - H(\bar{y}_\epsilon, p_\epsilon)| \to 0$ as $\epsilon \downarrow 0$.

Define

$$w_1(x) = v_2(\bar{y}_\epsilon) + \frac{(x - \bar{y}_\epsilon)^2}{2\epsilon}.$$

Obviously $w_1 \in C^\infty(\mathbb{R}^n)$ and $v_1 - w_1$ attains its local maximum at $\bar{x}_\epsilon$. By Lemma 4.8 we know that (23) has a strict subsolution, which is $w(x) \equiv 0$. So according to Lemma 4.9 we have

$$\max\{H(\bar{x}_\epsilon, \eta p_\epsilon),\ \eta v_1(\bar{x}_\epsilon) + (1-\eta)w(\bar{x}_\epsilon) - \psi(\bar{x}_\epsilon)\} < 0. \tag{28}$$

Now define

$$w_2(y) = v_1(\bar{x}_\epsilon) - \frac{(\bar{x}_\epsilon - y)^2}{2\epsilon},$$

where $w_2 \in C^\infty(\mathbb{R}^n)$ and $v_2 - w_2$ attains its local minimum at $\bar{y}_\epsilon$, so

$$\max\{H(\bar{y}_\epsilon, p_\epsilon),\ v_2(\bar{y}_\epsilon) - \psi(\bar{y}_\epsilon)\} \ge 0. \tag{29}$$

The inequality in (28) implies

$$\begin{aligned} &H(\bar{x}_\epsilon, \eta p_\epsilon) < 0 \quad \forall\, \epsilon > 0 \text{ and } \eta \in (0,1), \\ &\eta v_1(\bar{x}_\epsilon) + (1-\eta)w(\bar{x}_\epsilon) - \psi(\bar{x}_\epsilon) < 0. \end{aligned}$$

Now we show that the inequality in (29) implies only that

$$v_2(\bar{y}_\epsilon) \ge \psi(\bar{y}_\epsilon) \quad \forall\, \epsilon < \epsilon_0 \text{ for some } \epsilon_0 > 0. \tag{30}$$

An argument similar to (27) verifies that $H(x, p)$ is uniformly continuous in x and p, so we have

$$\lim_{\epsilon \downarrow 0,\ \eta \uparrow 1} |H(\bar{x}_\epsilon, \eta p_\epsilon) - H(\bar{y}_\epsilon, p_\epsilon)| = 0,$$

but we know that $H(\bar{x}_\epsilon, \eta p_\epsilon) < 0\ \forall\, \epsilon > 0$ and $\eta \in (0, 1)$, so $\exists\, \epsilon_0 > 0$ such that $H(\bar{y}_\epsilon, p_\epsilon) < 0\ \forall\, \epsilon < \epsilon_0$. Together with (29) this implies (30). We also have

$$v_1(\bar{x}_\epsilon) \le \psi(\bar{x}_\epsilon), \tag{31}$$

because v_1 is a viscosity subsolution of (23). Using (24), (30), and (31) we have

$$\begin{aligned} v_1(x) - v_2(x) &\le v_1(\bar{x}_\epsilon) - v_2(\bar{y}_\epsilon) - \frac{(\bar{x}_\epsilon - \bar{y}_\epsilon)^2}{2\epsilon} \quad \forall\, x \in \mathbb{R}^n \\ &= (v_1(\bar{x}_\epsilon) - \psi(\bar{x}_\epsilon)) - (v_2(\bar{y}_\epsilon) - \psi(\bar{y}_\epsilon)) + (\psi(\bar{x}_\epsilon) - \psi(\bar{y}_\epsilon)) - \frac{(\bar{x}_\epsilon - \bar{y}_\epsilon)^2}{2\epsilon} \\ &\le (\psi(\bar{x}_\epsilon) - \psi(\bar{y}_\epsilon)) - (\bar{x}_\epsilon - \bar{y}_\epsilon)^2/2\epsilon, \end{aligned}$$

so when $\epsilon \downarrow 0$ one gets $v_1(x) - v_2(x) \le 0$ or

$$v_1(x) \le v_2(x). \tag{32}$$

The inequality in (32) simply says that any viscosity subsolution is less than or equal to any viscosity supersolution. A viscosity solution is both a viscosity subsolution and a viscosity supersolution, so if V_1 and V_2 are two different viscosity solutions for (23), then by (32) we must have $V_1 \le V_2$ and $V_1 \ge V_2$, which implies that $V_1 = V_2$. So there is at most one viscosity solution of (23). □

COROLLARY 4.10. *$V(x, \pi) = \mathcal{G}\psi(x, \pi)$ is the unique viscosity solution of* (20). This is easily verified by replacing ψ with $N\psi$ in Theorem 4.6 and using Theorem 4.2.

LEMMA 4.11. *Define $v_0(x, \pi) = h(\pi)$ and $v_n(x, \pi) = \mathcal{G}v_{n-1}(x, \pi)$ for $n = 1, 2, \ldots$, where h and $\mathcal{G}$ are defined in* (9) *and* (18), *respectively. Then $v_n(x, \pi)$ is the minimal cost with at most n observations.*

Proof. In view of the definition of N and the fact that $t_f = 0$ is admissible in (18), it is clear that $v_1(x, \pi) = \mathcal{G}v_0(x, \pi)$ is the minimal cost if *at most one* observation is taken. The result follows by induction. □

THEOREM 4.12. *Consider an operator $\mathcal{G}$, defined by* (18). *Then*

$$W(x, \pi) = \inf_{S \in \mathbb{S}_{ad}} \mathbb{E}\left\{ \int_0^{\tau_k} \ell(x(s), u(s))\, ds + \sum_{j=1}^{k} c(x(\tau_j)) + h(\pi(\tau_k)) \right\}$$

is the unique fixed point of $\mathcal{G}$, where $\mathbb{S}_{ad}$ is the set of all admissible strategies as defined in §2.

We will need the following lemmas.

LEMMA 4.13. *W is a fixed point of $\mathcal{G}$, i.e., $W = \mathcal{G}W$.*

Proof. Define

$$\mathbb{S}_{ad}^{ob} = \{S \in \mathbb{S}_{ad}\colon\ \tau_1 \ge 0\},$$

$$\mathbb{S}_{ad}^{st} = \{S = \text{immediate stopping}\},$$

$$\mathbb{S}_{ad}(\tilde{u}, \tilde{\tau}) = \{S \in \mathbb{S}_{ad}^{ob}\colon\ u(t) = \tilde{u}(t), \quad 0 \le t \le \tilde{\tau} \text{ and } \tau_1 = \tilde{\tau}\},$$

where $\mathbb{S}_{ad}^{ob}$ is the set of all admissible strategies with at least one observation. Obviously $\mathbb{S}_{ad} = \mathbb{S}_{ad}^{ob} \cup \mathbb{S}_{ad}^{st}$, so

$$\inf_{S\in\mathbb{S}_{ad}} J(S) = \inf_{S\in\mathbb{S}_{ad}^{ob}\cup\mathbb{S}_{ad}^{st}} J(S)$$

$$= \min\left\{h(\pi),\ \inf_{S\in\mathbb{S}_{ad}^{ob}} J(S)\right\}. \tag{33}$$

We also have

$$\inf_{S\in\mathbb{S}_{ad}^{ob}} J(S) = \inf_{(\tilde{u},\tilde{\tau})} \inf_{S\in\mathbb{S}_{ad}(\tilde{u},\tilde{\tau})} \mathbb{E}_{x,\pi}\left\{\int_0^{\tilde{\tau}} \ell(x(s),\tilde{u}(s))\,ds + c(x(\tilde{\tau}))\right.$$

$$\left. + \int_{\tilde{\tau}}^{\tau_k} \ell(x(s),u(s))\,ds + \sum_{j=2}^{k} c(x(\tau_j)) + h(\pi_{\tau_k})\right\}$$

$$= \inf_{(\tilde{u},\tilde{\tau})}\left\{\int_0^{\tilde{\tau}} \ell(x(s),\tilde{u}(s))\,ds + c(x(\tilde{\tau}))\right.$$

$$\left. + \inf_{S\in\mathbb{S}_{ad}(\tilde{u},\tilde{\tau})} \mathbb{E}_{x(\tilde{\tau}),\pi}\left[\int_{\tilde{\tau}}^{\tau_k} \ell(x(s),u(s))\,ds + \sum_{j=2}^{k} c(x(\tau_j)) + h(\pi_{\tau_k})\right]\right\},$$

and it is easy to observe that

$$\inf_{S\in\mathbb{S}_{ad}^{ob}} J(S) = \inf_{(\tilde{u},\tilde{\tau})}\left\{\int_0^{\tilde{\tau}} \ell(x(s),\tilde{u}(s))\,ds + MW(x(\tilde{\tau}),\pi)\right\}. \tag{34}$$

Combining (33) and (34) yields

$$W(x,\pi) = \inf_{(\tilde{u},\tilde{\tau})}\left\{\int_0^{\tilde{\tau}} \ell(x(s),\tilde{u}(s))\,ds + NW(x(\tilde{\tau}),\pi)\right\},$$

and the result follows. □

LEMMA 4.14. *N, acting on the space of bounded measurable functions, is nondecreasing and concave.*

Proof. Assume that for all $(x,\pi) \in \mathbb{R}^n \times [0,1]$, we have $V_1(x,\pi) \le V_2(x,\pi)$. Then it is easy to show that

$$NV_1(x,\pi) \le NV_2(x,\pi),$$

because $MV_1(x,\pi) \le MV_2(x,\pi)$, so N is nondecreasing. Now assume that $\mu \in [0,1]$. Then

$$\begin{aligned} N(\mu V_1 + (1-\mu)V_2) &= \min\{\mu h + (1-\mu)h,\ M(\mu V_1 + (1-\mu)V_2)\} \\ &= \min\{\mu h + (1-\mu)h,\ \mu MV_1 + (1-\mu)MV_2\} \\ &\ge \min\{\mu h + (1-\mu)h,\ \mu h + (1-\mu)MV_2,\ \mu MV_1 + (1-\mu)h, \\ &\qquad \mu MV_1 + (1-\mu)MV_2\} \\ &= \mu\min\{h,\ MV_1\} + (1-\mu)\min\{h,\ MV_2\}, \end{aligned}$$

so $N(\mu V_1 + (1-\mu)V_2) \ge \mu NV_1 + (1-\mu)NV_2$ and N is concave. □

LEMMA 4.15. *$\mathcal{G}$ is nondecreasing and concave.*

Proof. Let us assume that $V_1 \le V_2\ \forall (x,\pi) \in \mathbb{R}^n \times [0,1]$. Then

$$\begin{aligned}\mathcal{G}V_2(x,\pi) &= \inf_{(u,\tau)} \left\{ \int_0^\tau \ell(x(s),u(s))\,ds + NV_2(x(\tau),\pi) - NV_1(x(\tau),\pi) + NV_1(x(\tau),\pi) \right\} \\ &= \inf_{(u,\tau)} \left\{ \int_0^\tau \ell(x(s),u(s))\,ds + NV_1(x(\tau),\pi) + [NV_2(x(\tau),\pi) - NV_1(x(\tau),\pi)] \right\} \\ &\ge \mathcal{G}V_1(x,\pi) + \inf_{(u,\tau)} [NV_2(x(\tau),\pi) - NV_1(x(\tau),\pi)],\end{aligned}$$

but

$$\inf_{(u,\tau)} [NV_2(x(\tau),\pi) - NV_1(x(\tau),\pi)] \ge 0,$$

because $NV_2(x(\tau),\pi) \ge NV_1(x(\tau),\pi)$, so

$$\mathcal{G}V_2(x,\pi) \ge \mathcal{G}V_1(x,\pi)$$

and $\mathcal{G}$ is nondecreasing. Now assume that $\mu \in [0,1]$. Then

$$\begin{aligned}\mathcal{G}(\mu V_1 + (1-\mu)V_2) &= \inf_{(u,\tau)} \left\{ \mu \int_0^\tau \ell(x(s),u(s))\,ds + (1-\mu)\int_0^\tau \ell(x(s),u(s))\,ds \right. \\ &\qquad \left. + N(\mu V_1 + (1-\mu)V_2) \right\} \\ &\ge \inf_{(u,\tau)} \left\{ \mu \int_0^\tau \ell(x(s),u(s))\,ds + \mu N V_1 \right. \\ &\qquad \left. + (1-\mu)\int_0^\tau \ell(x(s),u(s))\,ds + (1-\mu)NV_2 \right\} \\ &\ge \mu \inf_{(u,\tau)} \left\{ \int_0^\tau \ell(x(s),u(s))\,ds + NV_1 \right\} \\ &\quad + (1-\mu) \inf_{(u,\tau)} \left\{ \int_0^\tau \ell(x(s),u(s))\,ds + NV_2 \right\} \\ &= \mu\mathcal{G}V_1 + (1-\mu)\mathcal{G}V_2,\end{aligned}$$

so

$$\mathcal{G}(\mu V_1 + (1-\mu)V_2) \ge \mu\mathcal{G}V_1 + (1-\mu)\mathcal{G}V_2. \qquad \square$$

LEMMA 4.16. *Let $V = \mathcal{G}\psi$ for some $\psi \in B(\mathbb{R}^n \times [0,1])_+$. Then there exists $\kappa \in \mathbb{R}_+$ such that $V \le \kappa\mathcal{G}0$.*

Proof. For all $\psi \in B(\mathbb{R}^n \times [0,1])_+$ we have

$$V(x,\pi) \le h(\pi). \tag{35}$$

Let $\max_{\pi\in[0,1]} h(\pi) = h_{\max}$. (The maximum is actually achieved at $\pi = \ell_2/(\ell_1+\ell_2)$.) One can verify that $\mathcal{G}0 \ge h(\pi) \wedge \delta$, because $c \ge \delta > 0$. Now two cases can happen.

Case 1. If $h_{\max} > \delta$, then

$$h(\pi) \le \frac{h_{\max}}{\delta}\,(h(\pi) \wedge \delta)$$

$$\le \frac{h_{\max}}{\delta}\,\mathcal{G}0.$$

Case 2. If $h_{\max} \le \delta$, then $\mathcal{G}0 = h(\pi)$.

Finally, using (35) one can write $V \le (1 \vee h_{\max}/\delta)\,\mathcal{G}0$, i.e., $\kappa = (1 \vee h_{\max}/\delta)$. □

LEMMA 4.17. *Let* $\mathcal{P}_{\mathcal{G}} = \{V \in B(\mathbb{R}^n \times [0,1])_+ : V \le \kappa \mathcal{G}0 \text{ for some } \kappa \in \mathbb{R}_+\}$. *Then* $\mathcal{G}$ *has at most one fixed point in* $\mathcal{P}_{\mathcal{G}}$.

Proof. The following proof is very similar to the one on p. 250 of [D]. Let $V_1, V_2 \in \mathcal{P}_{\mathcal{G}}$, and assume that $\mathcal{G}V_1 = V_1$ and $\mathcal{G}V_2 = V_2$. Without loss of generality we can consider

$$V_1 - V_2 \notin B(\mathbb{R}^n \times [0,1])_+.$$

So let $t_0 = \sup\{t \in \mathbb{R}^+ : V_1 \ge tV_2\}$, $t_0 \in [0,1)$ since $V_1 \ge 0$ and $V_1 \not\ge V_2$. We have $V_1 \ge t_0 V_2$, so

$$\begin{aligned} V_1 = \mathcal{G}V_1 &\ge \mathcal{G}(t_0 V_2) \\ &= \mathcal{G}(t_0 V_2 + (1-t_0)0) \\ &\ge t_0 \mathcal{G}V_2 + (1-t_0)\mathcal{G}0 \\ &= t_0 V_2 + (1-t_0)\mathcal{G}0 \\ &\ge t_0 V_2 + \frac{(1-t_0)}{\kappa} V_2 \quad \text{for some } \kappa \in \mathbb{R}_+ \\ &= \left(t_0 + \frac{1-t_0}{\kappa}\right) V_2, \end{aligned} \tag{36}$$

since $V_2 \in \mathcal{P}_{\mathcal{G}}$. Inequality (36) gives $V_1 \ge (t_0 + (1-t_0)/k)V_2$, which contradicts the definition of t_0, so we must have $V_1 \ge V_2$. Now assume that $V_2 - V_1 \notin B(\mathbb{R}^n \times [0,1])_+$ and conclude that $V_2 \ge V_1$, so $V_1 = V_2$. □

Proof of Theorem 4.12. According to Lemma 4.13 $W(x,\pi)$ is a fixed point of $\mathcal{G}$. Then Lemma 4.17 proves that $W(x,\pi)$ is the unique fixed point of $\mathcal{G}$. □

Before proceeding to the following theorem let us remark that the definition of viscosity solutions for $F_V(x,\pi,V,D_xV) = 0$ can be stated exactly in the same way as Definition 4.1, but ψ must be replaced by NV throughout the definition.

THEOREM 4.18. *W is the unique viscosity solution of the equation*

$$F_V(x,\pi,V,D_xV) = 0, \tag{37}$$

where $W(x,\pi)$ *is given by* (14).

The following lemmas are needed.

LEMMA 4.19. *W is a viscosity solution of* (37).

Proof. This is due to the fact that if $V = \mathcal{G}V$, then $V \le NV$, and one can follow the proof of Theorem 4.2. □

LEMMA 4.20. *Any viscosity solution of* (37) *is a fixed point of* $\mathcal{G}$.

Proof. If V is a viscosity solution of (37), then by Corollary 4.10 V is the unique viscosity solution of the equation $F_V(x,\pi,\tilde{V},D_x\tilde{V}) = 0$ (regarded as an equation for unknown $\tilde{V}$ with V "frozen"), and this solution is given by $V = \mathcal{G}V$; i.e., V is a fixed point of $\mathcal{G}$. □

Proof of Theorem 4.18. According to Theorem 4.12, W is the unique fixed point of $\mathcal{G}$. Using the above lemmas it is clear that W is the unique viscosity solution of (37). □

The following theorem describes a computational algorithm for the value function and gives the necessary convergence result.

THEOREM 4.21. *Let* $v_0 = h$ *and* $v_n = \mathcal{G}v_{n-1}$ *for* $n = 1, 2, \ldots$. *Then* $\lim_{n\to\infty} v_n = W$, *where* W *is the value function given by* (14).

Proof. It is obvious that $0 \le v_n \le h$ for $n = 0, 1, \ldots$. Since $v_1 \le v_0$, a simple induction argument proves that $v_n \le v_{n-1}$ for all $n = 1, 2, \ldots$, because $\mathcal{G}$ is a nondecreasing operator (Lemma 4.15). So the limit $V(x, \pi) = \lim_{n\to\infty} v_n(x, \pi)$ certainly exists and $V = \mathcal{G}V$, which implies $V = W$ since W is the unique fixed point of $\mathcal{G}$ (Theorem 4.12). □

5. Simulation results. As an example, consider the following scalar system:

$$\begin{aligned} &\dot{x}(s) = u(s); \quad x(0) = x_0, \\ &U = [-K, K], \\ &\ell(x(s), u(s)) = \ell, \\ &c(x) = 1. \end{aligned}$$

We also assume that the target is located at $x = 0$. This simple one-dimensional example is obviously a minimum-time optimal control problem, so we expect to get bang-bang control. This problem may be solved by the following iterative computational scheme ((18) and (17)):

$$\begin{aligned} &v_n(x, \pi) = \mathcal{G}v_{n-1}(x, \pi), \\ &\mathcal{G}v_{n-1}(x, \pi) = \inf_{(u,t_f)} \int_0^{t_f} \ell \, ds + Nv_{n-1}(x(t_f), \pi), \\ &Nv_{n-1}(x, \pi) = h(\pi) \wedge Mv_{n-1}(x, \pi), \end{aligned} \tag{38}$$

with $v_0(x, \pi) = h(\pi)$. Let $\gamma > 0$ as the time step. Then Σ_0^γ would be the state-space set, where

$$\Sigma_0^\gamma = \{x_j = x_0 + j\gamma K, \ j = 0, \pm 1, \pm 2, \ldots\}.$$

Then (38) can be written as

$$v_n(x_0, \pi) = \inf_{j=0,\pm1,\pm2,\ldots} [|j|\gamma\ell + Nv_{n-1}(x_j, \pi)]. \tag{39}$$

After obtaining $W(x, \pi)$ one can find the optimal observation strategy. In (39), replacing $v_{n-1}(x_j, \pi)$ and $v_n(x_0, \pi)$ with $W(x_j, \pi)$ and $W(x_0, \pi)$, respectively, gives the necessary formula:

$$W(x_0, \pi) = \inf_{j=0,\pm1,\pm2,\ldots} [|j|\gamma\ell + NW(x_j, \pi)].$$

So starting from any initial condition (x_0, π), one can obtain τ_1. Then choosing $(x(\tau_1), \pi(\tau_1))$ as the new starting point, τ_2 can be calculated and one can continue until $(x(\tau_k), \pi(\tau_k)) \in \mathcal{S}$. The following forms are assumed for $e_0(x)$ and $e_1(x)$ throughout the simulation:

$$e_0(x) = \frac{1}{2}(1 - \exp(-b_1 x)), \qquad b_1 > 0,$$

$$e_1(x) = \frac{1}{2}(1 - \exp(-b_2 x)), \qquad b_2 > 0.$$

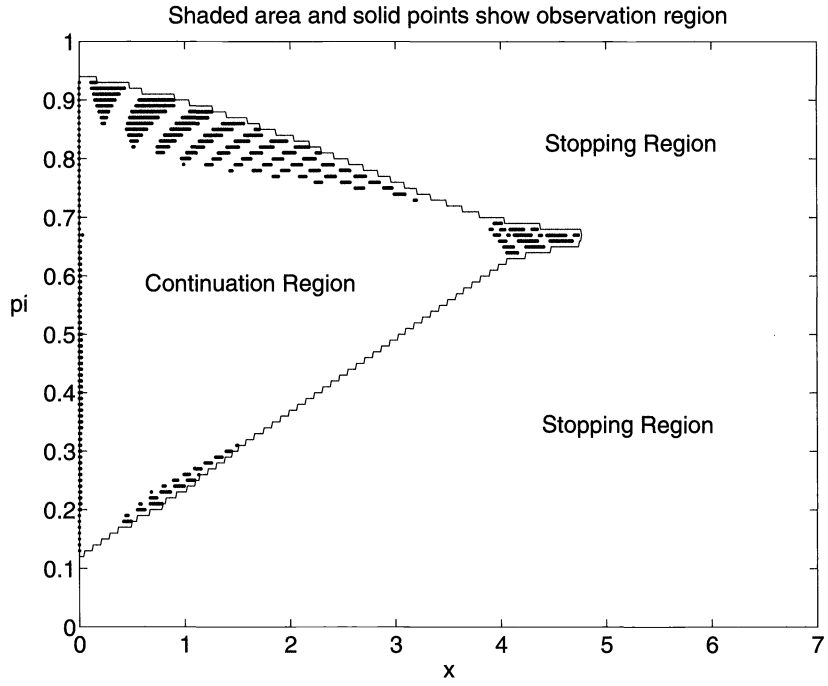


FIG. 1. *Stopping, observation, and continuation regions.*

Although assumption (1) is not satisfied here, one can show that all the results hold when $0 \leq e_0(x), e_1(x) \leq \frac{1}{2}$, $e_0(0) = 0$, $e_1(0) = 0$, and $\forall x \in \mathbb{R}^n \setminus \{0\}$ $e_0(x), e_1(x) \neq 0$. Figures 1–3 show the simulation results for $\gamma = 0.01$, $b_1 = 0.25$, $b_2 = 0.5$, $\ell_1 = 8$, $\ell_2 = 16$, and $\ell = 1$. The discretization step along the π-axis is equal to 0.01. Stopping, observation, and continuation regions are shown in Figure 1. $\ell_2 > \ell_1$, so there are more observation points for $\pi \geq 0.5$. Some continuation points are scattered in the observation region. This arises either due to numerical inaccuracy (discretization) or as a result of position-dependent probabilities ($e_0(x)$ and $e_1(x)$). So at some starting points the optimal strategy is as follows: go a bit forward and then make an observation, because otherwise the cost is slightly more. When $\ell_1 = \ell_2$ and $b_1 = b_2$, the regions are symmetric with respect to $\pi = 0.5$. Figure 2 shows the maximum difference, i.e., $\max |\upsilon_n - \upsilon_{n-1}|$, at each iteration. At the third iteration, υ_n is approximately equal to υ_{n-1}. Thus one can claim that most optimal strategies will have at most two observations. The value function is also depicted in Figure 3. The observation strategy was determined for a friendly target ($\theta = 0$) and starting point ($x_0 = 1, \pi = 0.25$) $\in C$. The results were as follows:

$$\begin{aligned}
&\tau_1 = 0.07, \qquad \tau_2 = 1,\\
&k = 2,\\
&x(\tau_1) = 0.93, \qquad x(\tau_2) = 0,\\
&\pi(\tau_1) = 0.72, \qquad \pi(\tau_2) = 0,\\
&d = 0.
\end{aligned}$$

The sequence of two observations, which was produced by a random generator with the associated distributions, was $Y_1 = 1$ and $Y_2 = 0$.

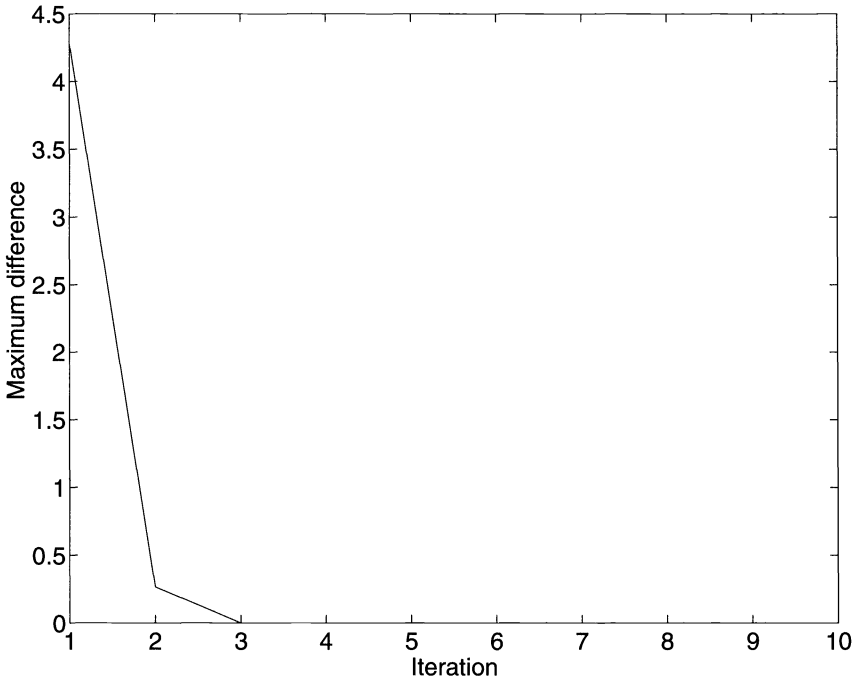


FIG. 2. $\max |v_n - v_{n-1}|$ *versus iteration.*

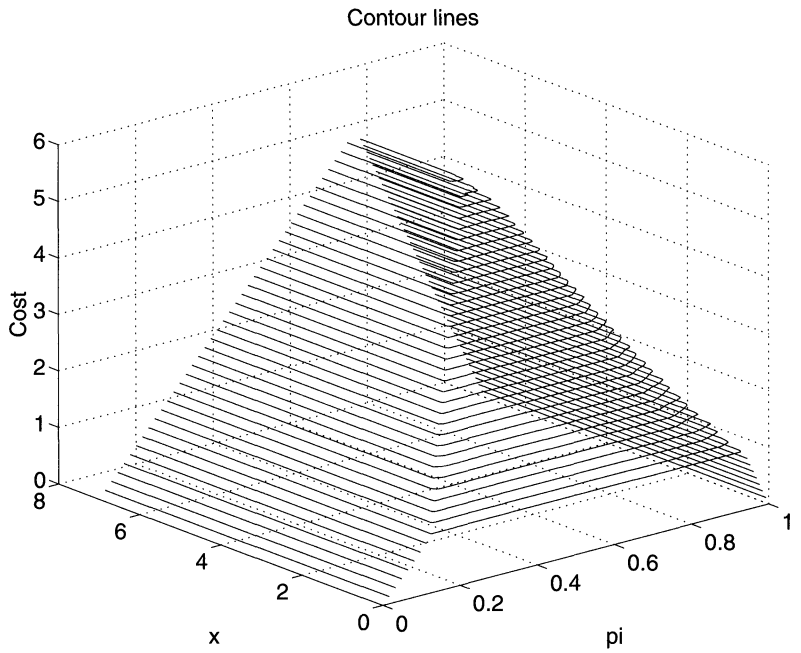


FIG. 3. *Value function.*

As one may expect, any observation error within the vicinity of target can cause a wrong decision. Due to discretization inaccuracies, finding the accurate boundaries between stopping,

observation, and continuation regions is very difficult. Nevertheless, the resulting difference in the value function is very small.

Acknowledgment. We would like to thank Guy Barles, Martin Clark, Jeremy Hodgson, David Salmond, Mete Soner, Gwen Tanner, Mihail Zervos, and the referees of this paper for their help, comments and suggestions. Thanks are due to Dimitris Sarris for doing the simulations.

REFERENCES

[B] G. BARLES, *Deterministic impulse control problems*, SIAM J. Control Optim., 23 (1985), pp. 419–432.

[CL] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[D] M. H. A. DAVIS, *Markov Models and Optimization*, Chapman & Hall, London, 1993.

[F] A. A. FEL'DBAUM, *Optimal Control Systems*, Academic Press, New York, 1965.

[FS] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[H] J. A. HODGSON, *Target Classification and Interception Using Dynamic Programming*, M.Sc. thesis, Electrical and Electronic Engineering, Imperial College of London, 1993.

[PSV] L. C. PICCININI, G. STAMPACCHIA, AND G. VIDOSSICH, *Ordinary Differential Equations in* $\mathbb{R}^n$, Springer-Verlag, New York, 1984.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 2133–2171, November 1996

015

HEAVY TRAFFIC CONVERGENCE OF A CONTROLLED, MULTICLASS QUEUEING SYSTEM*

L. F. MARTINS†, S. E. SHREVE‡, AND H. M. SONER‡

Abstract. This paper provides a rigorous proof of the connection between the optimal sequencing problem for a two-station, two-customer-class queueing network and the problem of control of a multidimensional diffusion process, obtained as a heavy traffic limit of the queueing problem. In particular, the diffusion problem, which is one of "singular control" of a Brownian motion, is used to develop policies which are shown to be asymptotically nearly optimal as the traffic intensity approaches one in the queueing network. The results are proved by a viscosity solution analysis of the related Hamilton–Jacobi–Bellman equations.

Key words. Brownian networks, queueing, heavy traffic, viscosity solutions, stochastic control

AMS subject classifications. 60K25, 93E20, 60J65

1. Introduction.

This paper provides a rigorous proof of the connection between the optimal sequencing problem for a two-station, two-customer-class queueing network and the problem of control of a multidimensional diffusion process, obtained as a heavy traffic limit of the queueing problem. In particular, the diffusion problem, which is one of "singular control" of a Brownian motion (also called "regulated Brownian motion" by Harrison (1985)), is used to develop policies which are shown to be asymptotically optimal as the traffic intensity approaches one in the queueing network.

The diffusion we wish to control here has been given the name *Brownian network* by Harrison (1988), who proposed such models as approximations to multiclass queueing networks. The idea of using diffusion approximations for single-class queueing systems dates back to Inglehart and Whitt (1970), Reiman (1984), and Johnson (1983). More recently, Reiman (1988), Peterson (1990), and Dai and Kurtz (1995) have obtained diffusion approximations for multiclass queues.

The control of Brownian networks for the purpose of obtaining control policies for queueing networks was initiated by Wein (1990a, 1990b, 1992) and Harrison and Wein (1989, 1990). These papers derive rules for sequencing customer services and for controlling input to queueing networks. Laws and Louth (1990) and Laws (1992) use Brownian networks to derive queueing network routing policies as well. All these papers are based on a heuristic understanding, amply supported by simulations, of the connection between the Brownian network control problem and the original queueing problem. Such a connection has been rigorously established in models with a single customer class by Kushner and Ramachandran (1988, 1989), Kushner and Martins (1990, 1991), and Krichagina et al. (1993, 1994). These papers use weak convergence methods. After the completion of this paper, Kushner and Martins (1994) used these methods to obtain the convergence of the value function considered in this paper. For weak convergence methods, the exogenous processes (e.g., arrival and service processes) can be quite general, provided that they have finite first and second moments.

In this paper, we assume that the arrival processes are Poisson and the service times are exponentially distributed. We base our analysis on the Hamilton–Jacobi–Bellman (HJB) equation, which, in turn, is based on the Markov property. In contrast to most other rigorous treatments of convergence, we treat a network with multiple customer classes. Our analysis

*Received by the editors April 8, 1994; accepted for publication (in revised form) October 20, 1995. This research was partially supported by the Army Research Office and the National Science Foundation through the Center for Nonlinear Analysis. The research of the second and third authors was partially supported by Army Research Office grant DAAH04-95-1-0226.

†Department of Mathematics, Cleveland State University, Cleveland, OH 44115.

‡Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15123.

uses the theory of viscosity solutions of HJB equations. Viscosity solutions were first introduced by Crandall and Lions (1984), and equivalent definitions were given by Crandall, Evans, and Lions (1984). For recent developments, we refer the reader to Crandall, Ishii, and Lions (1992) and Fleming and Soner (1993).

The particular example chosen for our study has also been examined by Harrison and Wein (1989) and Chen, Yang, and Yao (1991). The former work derives a plausible asymptotically nearly optimal sequencing policy for the queueing network in one of the parameter cases that we study; we confirm the asymptotic near-optimality of this policy. The latter work, which does not introduce the Brownian network, solves the original queueing problem in some parameter cases; we obtain consistent results in the case where comparison of results is appropriate, and we obtain an asymptotically nearly optimal policy in a parameter case not solved by Chen, Yang, and Yao (1991).

This paper is organized as follows. In §2 we describe enough of the queueing system problem, including the heavy traffic assumptions, to enable us to summarize our results. We complete the problem formulation in §3. Sections 4 and 5 establish elementary results concerning the value function for the queueing system problem. In §6 we define the limit of the value functions for a sequence of queueing systems. Of course, our goals are to represent this limit as the value function for a diffusion control problem and to use this representation to construct asymptotically optimal policies for the queueing systems. In §7 we introduce the associated controlled Brownian network, and in §8 we reduce the Brownian network problem to one of workload control. Section 9 dispatches the easy Case I. Section 10 provides an overview of the harder Case II. The remaining sections are devoted to the technical analysis of a subcase of Case II, which we call Case IIA.

We choose only Case IIA for full treatment because

(i) it includes the common situation of seeking to minimize the sum of the queue lengths when the service time at station one is independent of customer class;

(ii) a closed-form solution to the queueing system problem in this subcase is unknown;

(iii) the convergence result in this subcase requires new methodology; and

(iv) the workload control problem in this subcase has a simple solution.

We believe that the techniques developed here can be extended to the other cases, but this would first require the solution of nontrivial singular stochastic control problems to prove existence of the functions Ψ_1 and Ψ_2, which appear in the discussion of cases IIB, IIC, and IID in §2.

2. Summary of results. We study a family of two-station queueing networks with Poisson arrivals and exponential service times. In the nth network, customers of class 1 and 2 arrive at station 1 with arrival rates $\lambda_1^{(n)}$ and $\lambda_2^{(n)}$, respectively, and are served at respective rates $\mu_1^{(n)}$ and $\mu_2^{(n)}$. Class 1 customers then exit the system, whereas class 2 customers proceed to station 2, where they are redesignated as class 3 customers and served at rate $\mu_3^{(n)}$. See Figure 1.

The cost per unit time of holding a class i customer is $c_i > 0$. The objective is to minimize

$$E \int_0^\infty e^{-\alpha t/n} \sum_{i=1}^{3} c_i Q_i^{(n)}(t) dt, \tag{2.1}$$

where $Q_i^{(n)}$ is the number of class i customers queued or undergoing service at time t, and α is a positive constant.

In order to minimize this objective, we may decide at each time t whether to serve a class 1 or a class 2 customer. Service can be switched away from one class to the other and subsequently switched back, resuming where it left off. We may also decide to idle station 1,

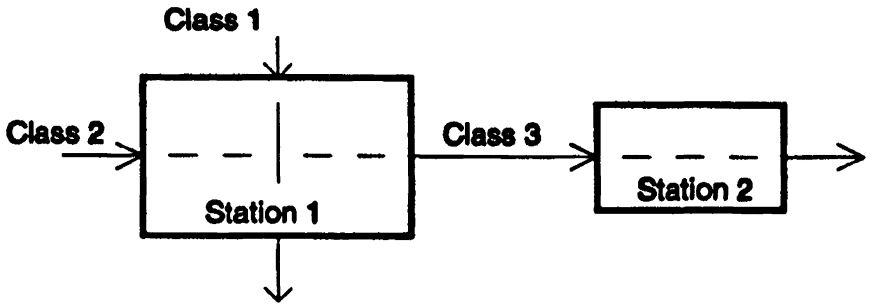


FIG. 1. *Criss-cross network.*

even though there are customers who could be served. This may be desirable if there are no class 1 customers and the cost c_3 is high relative to c_2 so that we prefer not to serve any class 2 customers until a backlog of class 3 customers has been reduced.

We want these networks to approach heavy traffic conditions as $n \to \infty$. Therefore, we define numbers $b_1^{(n)}$ and $b_2^{(n)}$ by the formulas

$$\frac{\lambda_1^{(n)}}{\mu_1^{(n)}} + \frac{\lambda_2^{(n)}}{\mu_2^{(n)}} = 1 - \frac{b_1^{(n)}}{\sqrt{n}}, \qquad \frac{\lambda_2^{(n)}}{\mu_3^{(n)}} = 1 - \frac{b_2^{(n)}}{\sqrt{n}} \tag{2.2}$$

so that $1 - \frac{b_1^{(n)}}{\sqrt{n}}$ is the traffic intensity at station 1 and $1 - \frac{b_2^{(n)}}{\sqrt{n}}$ is the traffic intensity at station 2. The *heavy traffic assumption* is that for $i = 1, 2, 3$ and $j = 1, 2$ the limits

$$\lambda_j = \lim_{n\to\infty} \lambda_j^{(n)}, \quad \mu_i = \lim_{n\to\infty} \mu_i^{(n)}, \quad b_j = \lim_{n\to\infty} b_j^{(n)}$$

are defined and positive and satisfy

$$\sup_n \left[\sqrt{n} \sum_{j=1}^{2} |\lambda_j^{(n)} - \lambda_j| + \sqrt{n} \sum_{i=1}^{3} |\mu_i^{(n)} - \mu_i| + \sum_{j=1}^{2} |b_j^{(n)} - b_j| \right] < \infty. \tag{2.3}$$

Our analysis divides naturally into two main cases, and the second case divides into four subcases. We describe our results in each case.

Case I ($c_1\mu_1 - c_2\mu_2 + c_3\mu_2 \le 0$). As long as customer class 2 is present, it should be served. If all class 2 customers have been served, then class 1 customers should be served.

This result agrees with Theorem 5.2 of Chen, Yang, and Yao (1991). The expected cost reduction per unit of service effort devoted to a class 2 customer is $(c_2 - c_3)\mu_2$, since service turns a class 2 customer into a class 3 customer. In Case I, $(c_2 - c_3)\mu_2$ dominates $c_1\mu_1$, the expected cost reduction per unit of service effort to a class 1 customer. This results in the simple fixed priority rule of serving class 2 customers whenever they are present.

Case II ($c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$). We further divide this case into four subcases. Reasoning behind this subdivision is given in §10 below.

Case IIA ($c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$, $c_2\mu_2 - c_3\mu_2 \ge 0$, $c_2\mu_2 - c_1\mu_1 \ge 0$). Now a unit of service applied to class 1 results in a greater expected cost reduction than a unit of service to class 2. In §12 we prove the asymptotic near-optimality (see the last paragraph of §6 for this concept) of the policy of serving class 1 unless the number of class 3 customers falls below a positive threshold, in which case priority is switched to class 2 so that station 2 is not starved. The switching threshold depends on the queue lengths in the following way. Let

$a, b : [0, \infty) \to [0, \infty)$ be bounded, concave, increasing functions satisfying

$$a(0) = b(0) = 0, \qquad \eta \triangleq b(\infty) < (a(\infty))^2.$$

Define $\gamma(z_1, z_2, z_3) \triangleq a(z_1)a(z_3) - b(z_2)$. The nearly asymptotically optimal policy is given by

$$\text{serve class 1 if} \quad \gamma(Q_1^{(n)}(t)/\sqrt{n}, Q_2^{(n)}(t)/\sqrt{n}, Q_3^{(n)}(t)/\sqrt{n}) \geq 0,$$

$$\text{serve class 2 if} \quad \gamma(Q_1^{(n)}(t)/\sqrt{n}, Q_2^{(n)}(t)/\sqrt{n}, Q_3^{(n)}(t)/\sqrt{n}) < 0,$$

where $Q_i^{(n)}(t)$ denotes the number of class i customers present at time i. As $\eta \downarrow 0$, this policy approaches asymptotic optimality.

Harrison and Wein's (1989) model with $c_1 = c_2 = c_3 = 1$, $\mu_1 = \mu_2 = 2$, $\mu_3 = 1$ falls into this subcase, and their proposed policy is to serve class 1 if and only if $Q_3^{(n)}(t)/\sqrt{n}$ exceeds a positive constant which is independent of n and the other queue lengths. They showed by simulation that with a properly chosen constant, this policy outperforms the rules "first-in, first-out," "longest expected remaining processing time," and "shortest expected remaining processing time." They also found that its performance was within about 5% of a lower bound that they obtained for the optimal cost. We have not done simulation testing of our policy.

The heuristic justification of the policy in Case IIA suggests that the same policy is asymptotically optimal under only the Case II condition $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$. Our proof of the result stated in Case IIA suggests otherwise. Although we have not worked out a full proof for the other three subcases, the proof for Case IIA strongly suggests the following conjectures. A brief motivation of the following conjectures is given in §10 below. Chen, Yang, and Yao (1991) offer a heuristic policy, based only on the length of the queue at the second station, for all subcases of Case II. (Note that one must set $r = 0$ in Chen, Yang, and Yao in order to compare to our result.)

Case IIB ($c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0, c_2\mu_2 - c_3\mu_2 < 0, c_2\mu_2 - c_1\mu_1 \geq 0$). There is a continuous, increasing function $\Psi_2 : [0, \infty) \to [0, \infty)$ satisfying

$$0 \leq \Psi_2(\omega_2) < \mu_3\omega_2/\mu_2 \quad \forall \omega_2 \geq 0, \qquad \lim_{\omega_2 \to \infty} \Psi_2(\omega_2) = \infty.$$

Class 1 should be given priority unless either the queue length $Q_1^{(n)}$ of class 1 customers falls to zero or the queue length $Q_3^{(n)}$ of class 3 customers falls below some positive threshold. While either of these conditions is satisfied, priority should be switched to class 2, except that whenever $Q_1^{(n)} = 0$ and

$$\frac{Q_2^{(n)}}{\sqrt{n}\,\mu_2} < \Psi_2\left(\frac{Q_2^{(n)} + Q_3^{(n)}}{\sqrt{n}\,\mu_3}\right),$$

station 1 should be idled. This idleness can be explained by the fact that it is cheaper to hold class 2 customers at station 1 than to send them on to be held as class 3 customers at station 2; note that in this subcase, $c_2 < c_3$. Also observe that when $Q_1^{(n)} = 0$, the pair $(Q_2^{(n)}/\mu_2, (Q_2^{(n)} + Q_3^{(n)})/\mu_3)$ is equal to the expected impending service time for the two stations embodied in customers anywhere in the network; see §8 below for details. The term $1/\sqrt{n}$ that appears in the above formulas is related to time scaling that will be introduced in the next section.

Case IIC ($c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0, c_2\mu_2 - c_3\mu_2 \geq 0, c_2\mu_2 - c_1\mu_1 < 0$). There exists a continuous, increasing function $\Psi_1 : [0, \infty) \to [0, \infty)$ satisfying

$$0 \leq \Psi_1(\omega_1) < \mu_2\omega_1/\mu_3 \quad \forall \omega_1 \geq 0, \qquad \lim_{\omega_1 \to \infty} \Psi_1(\omega_1) = \infty.$$

Class 1 should be given priority unless either $Q_1^{(n)} = 0$ or $Q_3^{(n)}$ is less than a positive threshold. While either of these conditions is satisfied, priority should be switched to class 2, except that when $Q_1^{(n)} > 0$ and

$$\frac{Q_2^{(n)}}{\sqrt{n}\,\mu_3} < \Psi_1\left(\frac{Q_1^{(n)}}{\sqrt{n}\,\mu_1} + \frac{Q_2^{(n)}}{\sqrt{n}\,\mu_2}\right),$$

priority should be given to class 1, even though this may cause station 2 to starve. Idling station 2 can be explained by the fact that the cost of operating the network can be reduced more quickly by serving class 1 than by serving class 2; note that $c_1\mu_1 > c_2\mu_2$. As in the previous case, the term $1/\sqrt{n}$ is related to time scaling, and when $Q_3^{(n)} = 0$, the pair

$$\left(\frac{Q_1^{(n)}}{\mu_1} + \frac{Q_2^{(n)}}{\mu_2}, \frac{Q_2^{(n)}}{\mu_3}\right)$$

is equal to the expected impending service time for the two stations embodied in customers anywhere in the network.

Case IID ($c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0, c_2\mu_2 - c_3\mu_2 < 0, c_2\mu_2 - c_1\mu_1 < 0$). This case is a combination of Case IIB and Case IIC. We conjecture the existence of functions Ψ_1 and Ψ_2 as described above. Idling can occur at either station 1 or station 2, as described in Case IIB and Case IIC, respectively.

3. The queueing network problem. For the queueing network of the previous section, for $i = 1, 2$, let $\{A_i^{(n)}(t); 0 \le t < \infty)$ be the class i customer *arrival process*, assumed to be Poisson with intensity $\lambda_i^{(n)}$. For $i = 1, 2, 3$, let $\{S_i^{(n)}(t); 0 \le t < \infty\}$ be the class i customer *service process*, assumed to be Poisson with intensity $\mu_i^{(n)}$. We take all these processes to be left-continuous, and we denote by $\{\mathcal{F}^{(n)}(t); 0 \le t < \infty\}$ the filtration generated by these five processes.

A *control law* $\{Y(t), U(t); 0 \le t < \infty\}$ is a pair of left-continuous, $\{\mathcal{F}^{(n)}(t); 0 \le t < \infty\}$-adapted, $\{0, 1\}$-valued processes. The process $Y(t)$ indicates whether station 1 is active ($Y(t) = 1$) or idle ($Y(t) = 0$), and $U(t)$ indicates whether station 1 is serving customer class 1 ($U(t) = 1$) or customer class 2 ($U(t) = 0$). Given nonnegative initial queue lengths $Q_1^{(n)}(0)$, $Q_2^{(n)}(0)$, and $Q_3^{(n)}(0)$ for the three customer classes, and given a control law (Y, U), there is a unique triple of *queue length processes* satisfying

$$Q_1^{(n)}(t) = Q_1^{(n)}(0) + A_1^{(n)}(t) - \int_0^t Y(s)U(s)1_{\{Q_1^{(n)}(s)\ge 1\}} dS_1^{(n)}(s),$$

$$Q_2^{(n)}(t) = Q_2^{(n)}(0) + A_2^{(n)}(t) - \int_0^t Y(s)(1 - U(s))1_{\{Q_2^{(n)}(s)\ge 1\}} dS_2^{(n)}(s),$$

$$Q_3^{(n)}(t) = Q_3^{(n)}(0) + \int_0^t Y(s)(1 - U(s))1_{\{Q_2^{(n)}(s)\ge 1\}} dS_2^{(n)}(s) - \int_0^t 1_{\{Q_3^{(n)}(s)\ge 1\}} dS_3^{(n)}(s),$$

where 1_A is the indicator of the set A. We denote the vector of queue length processes by

$$Q^{(n)}(t) = (Q_1^{(n)}(t), Q_2^{(n)}(t)Q_3^{(n)}(t)).$$

(Note: Because the interservice times are exponentially distributed, the processes

$$\int_0^t Y(s)U(s)1_{\{Q_1^{(n)}(s)\ge 1\}} dS_1^{(n)}(s) \quad \text{and} \quad S_1^{(n)}\left(\int_0^t Y(s)U(s)1_{\{Q_1^{(n)}(s)\ge 1\}} ds\right)$$

have the same law. This permits us to write $Q_1^{(n)}(t)$ in terms of the former, although the latter more nearly reflects the way we interpret the system. If service of a customer is preempted and later resumed, we assume that service begins where it was left off. After resumption of service, the time to completion has the same exponential distribution as the original distribution of the service time. Similar comments apply to $Q_2^{(n)}(t)$ and $Q_3^{(n)}(t)$.)

The vector of *scaled queue length processes* is

$$Z^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} Q^{(n)}(nt).$$

For fixed controls $(y, u) \in \{0, 1\}^2$, this is a Markov chain with lattice state space $L^{(n)} \triangleq \{\frac{k}{\sqrt{n}}; k = 0, 1, \ldots\}^3$, and its infinitesimal generator is (see Chung (1960))

$$\begin{aligned}(\mathcal{L}^{n,y,u}\varphi)(z) \triangleq\ & n\lambda_1^{(n)}\left[\varphi\left(z + \frac{1}{\sqrt{n}}e_1\right) - \varphi(z)\right] + n\lambda_2^{(n)}\left[\varphi\left(z + \frac{1}{\sqrt{n}}e_2\right) - \varphi(z)\right] \\ & + n\mu_1^{(n)} yu\left[\varphi\left(z - \frac{1}{\sqrt{n}}e_1\right) - \varphi(z)\right] 1_{\{z_1>0\}} \\ & + n\mu_2^{(n)} y(1-u)\left[\varphi\left(z - \frac{1}{\sqrt{n}}e_2 + \frac{1}{\sqrt{n}}e_3\right) - \varphi(z)\right] 1_{\{z_2>0\}} \\ & + n\mu_3^{(n)}\left[\varphi\left(z - \frac{1}{\sqrt{n}}e_3\right) - \varphi(z)\right] 1_{\{z_3>0\}},\end{aligned} \tag{3.1}$$

where $z = (z_1, z_2, z_3)$, $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, and $e_3 = (0, 0, 1)$. In particular, given any control law $(Y(\cdot), U(\cdot))$, for any real-valued function φ on $L^{(n)}$, the process

$$e^{-\alpha t}\varphi(Z^{(n)}(t)) + \int_0^t e^{-\alpha s}[\alpha\varphi(Z^{(n)}(s)) - \mathcal{L}^{n,Y(s),U(s)}\varphi(Z^{(n)}(s))]ds \tag{3.2}$$

is a local martingale.

Using the positive holding costs c_1, c_2, c_3, we define the *holding cost function* $h(z) = \sum_{i=1}^3 c_i z_i$. Given an initial condition $Z^{(n)}(0) = z \in L^{(n)}$ and a control law $(Y(\cdot), U(\cdot))$, we define the associated *cost function* at z by

$$J_{Y,U}^{(n)}(z) \triangleq E\int_0^\infty e^{-\alpha t}h(Z^{(n)}(t))dt. \tag{3.3}$$

In terms of the original queue length process, this cost can be written as (cf. (2.1))

$$n^{-\frac{3}{2}}E\int_0^\infty e^{-(\alpha t)/n}h(Q^{(n)}(t))dt.$$

The *value function* at z is

$$J_*^{(n)}(z) \triangleq \inf\{J_{Y,U}^{(n)}(z);\ (Y, U) \text{ is a control law}\}. \tag{3.4}$$

4. Stationary control laws for the queueing network. A *stationary control law* for the nth queueing network is a pair of functions $Y : L^{(n)} \to \{0, 1\}$, $U : L^{(n)} \to \{0, 1\}$. The value of the control at time t is given in feedback form as $(Y(Z^{(n)}(t)), U(Z^{(n)}(t)))$. Because the

queueing network is driven by time-homogeneous Markov arrival and service processes, we have

$$J_*^{(n)}(z) = \inf\{J_{Y,U}^{(n)}(z);\ (Y, U) \text{ is a } \textit{stationary} \text{ control law}\}. \tag{4.1}$$

Let $\mathcal{L}^{n,Y,U}$ denote the infinitesimal generator of the controlled process with stationary controls Y, U. Then $\mathcal{L}^{n,Y,U}$ is given as in (3.1) with the pair (y, u) replaced by $(Y(z), U(z))$.

PROPOSITION 4.1. *For any stationary control law (Y, U), the function $J_{Y,U}^{(n)}$ is the unique, linearly growing solution of the equation*

$$\alpha\varphi - \mathcal{L}^{n,Y,U}\varphi - h = 0 \quad \textit{on} \quad L^{(n)}. \tag{4.2}$$

If φ is a linearly growing subsolution of this equation, i.e.,

$$\alpha\varphi - \mathcal{L}^{n,Y,U}\varphi - h \le 0 \quad \textit{on} \quad L^{(n)},$$

or if φ is a linearly growing supersolution, *i.e.,*

$$\alpha\varphi - \mathcal{L}^{n,Y,U}\varphi - h \ge 0 \quad \textit{on} \quad L^{(n)},$$

then $\varphi \le J_{Y,U}^{(n)}$ or $\varphi \ge J_{Y,U}^{(n)}$, respectively.

Proof. Under any control law, we have

$$EZ_1^{(n)}(t) \le Z_1^{(n)}(0) + \frac{1}{\sqrt{n}} EA_1^{(n)}(nt) = Z_1^{(n)}(0) + \sqrt{n}\,\lambda_1^{(n)} t, \tag{4.3}$$

$$EZ_2^{(n)}(t) \le Z_2^{(n)}(0) + \frac{1}{\sqrt{n}} EA_2^{(n)}(nt) = Z_2^{(n)}(0) + \sqrt{n}\,\lambda_2^{(n)} t, \tag{4.4}$$

$$EZ_3^{(n)}(t) \le Z_3^{(n)}(0) + \frac{1}{\sqrt{n}} ES_2^{(n)}(nt) = Z_3^{(n)}(0) + \sqrt{n}\,\mu_2^{(n)} t, \tag{4.5}$$

so

$$J_{Y,U}^{(n)}(z) \le h(z) + \frac{\sqrt{n}}{\alpha} h(\lambda_1^{(n)}, \lambda_2^{(n)}, \mu_2^{(n)}), \tag{4.6}$$

and $J_*^{(n)}$ has the same upper bound. Using the bounds (4.3)–(4.5) and the dominated convergence theorem, one can show that for any linearly growing φ, the local martingale (3.2) is in fact a martingale. In particular, if (Y, U) is a stationary control law, then

$$\begin{aligned} J_{Y,U}^{(n)}(Z^{(n)}(0)) &= Ee^{-\alpha t} J_{Y,U}^{(n)}(Z^{(n)}(t)) \\ &\quad + E\int_0^t e^{-\alpha s}[\alpha J_{Y,U}^{(n)}(Z^{(n)}(s)) - (\mathcal{L}^{n,Y(Z^{(n)}(s)),U(Z^{(n)}(s))} J_{Y,U}^{(n)})(Z^{(n)}(s))]ds. \end{aligned}$$

But for a stationary control law, the Markov property implies

$$J_{Y,U}^{(n)}(Z^{(n)}(0)) = Ee^{-\alpha t} J_{Y,U}^{(n)}(Z^{(n)}(t)) + E\int_0^t e^{-\alpha s} h(Z^{(n)}(s))ds.$$

Comparing these two equations, we see that

$$\begin{aligned} E\int_0^t e^{-\alpha s}[\alpha J_{Y,U}^{(n)}(Z^{(n)}(s)) - (\mathcal{L}^{n,Y(Z^{(n)}(s)),U(Z^{(n)}(s))} J_{Y,U}^{(n)})(X^{(n)}(s))]ds \\ = E\int_0^t e^{-\alpha s} h(Z^{(n)}(s))ds. \end{aligned}$$

Dividing by t and letting $t \downarrow 0$, we see that $J^{(n)}_{Y,U}$ satisfies (4.2). Uniqueness of this solution will follow from the second part of the proposition.

If φ is a linearly growing subsolution of (4.2), then the martingale property for (3.2) implies

$$\varphi(Z^{(n)}(0)) \leq E\left[e^{-\alpha t}\varphi(Z^{(n)}(t)) + \int_0^t e^{-\alpha s}h(Z^{(n)}(s))ds\right].$$

Letting $t \to \infty$, using (4.3)–(4.5) and the linear growth of φ, we obtain $\varphi \leq J^{(n)}_{Y,U}$. The supersolution claim is proved similarly. □

5. The HJB equation for the queueing network. For $\varphi : L^{(n)} \to \mathcal{R}$, we define the nonlinear operator $\mathcal{L}^{n,*}$ acting on φ by

$$\mathcal{L}^{n,*}\varphi(z) \triangleq \min\{\mathcal{L}^{n,y,u}\varphi(z);\ (y,u) \in \{0,1\}^2\} \quad \forall z \in L^{(n)}. \tag{5.1}$$

The HJB equation for the nth queueing network is

$$\alpha\varphi - \mathcal{L}^{n,*}\varphi - h = 0 \quad \text{on} \quad L^{(n)}. \tag{5.2}$$

PROPOSITION 5.1. *The value function $J^{(n)}_*$ is the unique, linearly growing solution of the HJB equation* (5.2). *If φ is a linearly growing subsolution (respectively, supersolution) of this equation, then $\varphi \leq J^{(n)}_*$ (respectively, $\varphi \geq J^{(n)}_*$). Furthermore, any stationary control law (Y^*, U^*) satisfying*

$$\mathcal{L}^{n,Y^*,U^*}J^{(n)}_* = \mathcal{L}^{n,*}J^{(n)}_* \tag{5.3}$$

is optimal.

Proof. We first prove the comparisons. Let φ be a linearly growing subsolution of (5.2). Then, for any stationary control (Y, U), we have

$$\alpha\varphi - \mathcal{L}^{n,Y,U} - h \leq \alpha\varphi - \mathcal{L}^{n,*}\varphi - h \leq 0.$$

Proposition 4.1 implies $\varphi \leq J^{(n)}_{Y,U}$, and minimization over (Y, U) yields $\varphi \leq J^{(n)}_*$.

Now let φ be a linearly growing supersolution of (5.2), and choose a stationary control (Y, U) satisfying

$$\alpha\varphi - \mathcal{L}^{n,Y,U}\varphi - h = \alpha\varphi - \mathcal{L}^{n,*}\varphi - h \geq 0.$$

Proposition 4.1 implies $\varphi \geq J^{(n)}_{Y,U}$, which dominates $J^{(n)}_*$.

A linearly growing solution of (5.2) can be constructed by the policy iteration algorithm. Let (Y_0, U_0) be any stationary control, and choose (Y_{k+1}, U_{k+1}) recursively so that

$$\mathcal{L}^{n,Y_{k+1},U_{k+1}}J^{(n)}_{Y_k,U_k} = \mathcal{L}^{n,*}J^{(n)}_{Y_k,U_k}.$$

Then $J^{(n)}_\infty \triangleq \lim_{k\to\infty} J^{(n)}_{Y_k,U_k}$ can be shown to be a linearly growing solution of (5.2); we omit the details. By the comparisons already proved, any linearly growing solution of (5.2) must agree with $J^{(n)}_*$, and since (5.2) has a linearly growing solution, $J^{(n)}_*$ is a solution. □

Remark 5.2. In certain situations, we will need to extend the definition of the operator $\mathcal{L}^{n,y,u}$ to allow (y, u) to take values in the square $[0, 1]^2$, rather than just at the corners. Fractional values of u correspond to processor sharing at the first station, and fractional values

of y correspond to partial utilization of this station. The only property that will be needed, however, is that (5.1) can be rewritten as

$$\mathcal{L}^{n,*}\varphi(z) = \min\{\mathcal{L}^{n,y,u}\varphi(z);\ \ (y,u) \in [0,1]^2\}, \tag{5.4}$$

a fact easily verified by noting from (3.1) that the minimum in (5.4) will be obtained at some corner of $[0,1]^2$.

6. The heavy traffic limit of the value function. In order to let $n \to \infty$, we need an upper bound, independent of n, for the nonnegative functions $\{J_*^{(n)}\}_{n=1}^{\infty}$.

PROPOSITION 6.1. *There are constants K_1 and K_2, independent of n, such that*

$$J_*^{(n)}(z) \le K_1 + K_2(z_1 + z_2 + z_3) \quad \forall z \in L^{(n)}.$$

Proof. Define $\varphi(z) = \sum_{i=1}^{3}(z_i + e^{-z_i})$ for all $z \in L^{(n)}$. Set $u = \lambda_1/\mu_1$. We begin by verifying that $(\mathcal{L}^{n,1,u}\varphi)(z)$ is bounded above by a constant independent of n and $z \in L^{(n)}$.

From (2.2) and (2.3) we have $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} = 1$, $\lambda_2 = \mu_3$, $\lambda_i^{(n)} = \lambda_i + O(\frac{1}{\sqrt{n}})$, and $\mu_i^{(n)} = \mu_i + O(\frac{1}{\sqrt{n}})$, where $O(n^p)$ denotes a term whose absolute value is bounded by Kn^p and K is a constant independent of n and $z \in L^{(n)}$. We may thus rewrite (3.1) as

$$\begin{aligned}
(\mathcal{L}^{n,l,u}\varphi)(z) &= n\lambda_1\left[\varphi\left(z + \frac{1}{\sqrt{n}}e_1\right) + \varphi\left(z - \frac{1}{\sqrt{n}}e_1\right) - 2\varphi(z)\right] \\
&\quad + n\lambda_2\left[\varphi\left(z + \frac{1}{\sqrt{n}}e_2\right) + \varphi\left(z - \frac{1}{\sqrt{n}}e_2 + \frac{1}{\sqrt{n}}e_3\right) + \varphi\left(z - \frac{1}{\sqrt{n}}e_3\right) - 3\varphi(z)\right] \\
&\quad - n\lambda_1\left[\varphi\left(z - \frac{1}{\sqrt{n}}e_1\right) - \varphi(z)\right]1_{\{z_1=0\}} \\
&\quad - n\lambda_2\left[\varphi\left(z - \frac{1}{\sqrt{n}}e_2 + \frac{1}{\sqrt{n}}e_3\right) - \varphi(z)\right]1_{\{z_2=0\}} \\
&\quad - n\lambda_2\left[\varphi\left(z - \frac{1}{\sqrt{n}}e_3\right) - \varphi(z)\right]1_{\{z_3=0\}} \\
&\quad + O(\sqrt{n})\left[\sum_{i=1}^{3}\left|\varphi\left(z \pm \frac{1}{\sqrt{n}}e_i\right) - \varphi(z)\right| + \left|\varphi\left(z - \frac{1}{\sqrt{n}}e_2 + \frac{1}{\sqrt{n}}e_3\right) - \varphi(z)\right|\right].
\end{aligned}$$

Since $\varphi(z \pm \frac{1}{\sqrt{n}}e_i) - \varphi(z) = \pm\frac{1}{\sqrt{n}}(1 - e^{-z_i}) + O(\frac{1}{n})$ and $\varphi(z - \frac{1}{\sqrt{n}}e_2 + \frac{1}{\sqrt{n}}e_3) - \varphi(z) = \frac{1}{\sqrt{n}}(e^{-z_2} - e^{-z_3}) + O(\frac{1}{n})$, we have

$$(\mathcal{L}^{n,1,u}\varphi)(z) = -\sqrt{n}\ \lambda_2(1 - e^{-z_3}) + O(1).$$

This implies that $\mathcal{L}^{n,1,u}\varphi$ is bounded above by a constant K_0.

Let $K_2 = \frac{1}{\alpha}\max\{c_1, c_2, c_3\}$, so $\alpha K_2\varphi - h \ge 0$. Put $\Psi = \frac{1}{\alpha}K_2K_0 + K_2\varphi$. Then

$$\alpha\Psi - \mathcal{L}^{n,1,u}\Psi - h = \alpha K_2\varphi - h + K_2(K_0 - \mathcal{L}^{n,1,u}\varphi) \ge 0,$$

which shows that $\alpha\Psi - \mathcal{L}^{n,*}\Psi - h \ge 0$. (Recall Remark 5.2.) The supersolution part of Proposition 5.1 implies

$$J_*^{(n)}(z) \le \Psi(z) \le \frac{1}{\alpha}K_2K_0 + 3K_2 + K_2(z_1 + z_2 + z_3). \qquad \square$$

We wish to consider $\lim_{n\to\infty} J_*^{(n)}$, but since each $J_*^{(n)}$ is defined on a different set $L^{(n)}$, the definition of this limit is not straightforward. Borrowing the technique developed by Barles and Perthame (1988) (see also Fleming and Soner (1993, §7.3)), we define the *upper semicontinuous limit* $J^\#$ of $\{J_*^{(n)}\}_{n=1}^\infty$ by

$$J^\#(z) \triangleq \lim_{\epsilon\downarrow 0} \lim_{n\to\infty} \sup_{\substack{\zeta\in L^{(n)} \\ \|\zeta - z\|<\epsilon}} J_*^{(n)}(\zeta) \quad \forall z \in [0,\infty)^3 \tag{6.1}$$

and the *lower semicontinuous limit* $J_\#$ by

$$J_\#(z) \triangleq \lim_{\epsilon\downarrow 0} \lim_{n\to\infty} \inf_{\substack{\zeta\in L^{(n)} \\ \|\zeta - z\|<\epsilon}} J_*^{(n)}(\zeta) \quad \forall z \in [0,\infty)^3. \tag{6.2}$$

Then $J^\#$ is upper semicontinuous, $J_\#$ is lower semicontinuous, and

$$0 \le J_\#(z) \le J^\#(z) \le K_1 + K_2(z_1 + z_2 + z_3) \quad \forall z \in [0,\infty)^3. \tag{6.3}$$

We shall eventually show that $J_\# = J^\#$, and we shall use a Brownian network problem to suggest, for each $\eta > 0$, a sequence of stationary policies $\{(Y^n, U^n)\}_{n=1}^\infty$ such that

$$\lim_{\epsilon\downarrow 0} \lim_{n\to\infty} \sup_{\substack{\zeta\in L^n \\ \|\zeta - z\|<\epsilon}} |J^{(n)}_{Y^n,U^n}(\zeta) - J_\#(\zeta)| \le \eta \quad \forall z \in [0,\infty)^3. \tag{6.4}$$

We call such a family (parametrized by η) of sequences of policies *asymptotically nearly optimal*.

7. The controlled Brownian network. We first introduce the controlled Brownian network and then explain by an analysis of the infinitesimal generator $\mathcal{L}^{n,y,u}$ why it is relevant. Let M_1, M_2, and M_3 be continuous martingales relative to a filtration $\{\mathcal{F}(t)\}$ satisfying the usual conditions that each $\mathcal{F}(t)$ contains all null sets of $\mathcal{F}(\infty)$ and that $\mathcal{F}(t) = \cap_{s>t}\mathcal{F}(s)$ for all t. Assume that for all t,

$$\langle M_1\rangle(t) = 2\lambda_1 t, \qquad \langle M_2\rangle(t) = \langle M_3\rangle(t) = 2\lambda_2 t, \tag{7.1}$$

$$\langle M_1, M_2\rangle(t) = \langle M_1, M_3\rangle(t) = 0, \qquad \langle M_2, M_3\rangle(t) = -\lambda_2 t. \tag{7.2}$$

Given $z \in [0,\infty)^3$, we will say that the quadruple $(\ell_0, \ell_1, \ell_2, \ell_3)$ of $\{\mathcal{F}(t)\}$-adapted processes is *admissible for initial condition* z, provided that

(i) $(\ell_0, \ell_1, \ell_2, \ell_3)$ are right-continuous with left-hand limits, with the convention that $\ell_i(0^-) = 0$, $i = 1, 2, 3$;

(ii) ℓ_0 is of finite variation on bounded intervals;

(iii) ℓ_1, ℓ_2, and ℓ_3 are nondecreasing,

(iv) the *state process* $Z(t) = (Z_1(t), Z_2(t), Z_3(t))$ is in $[0,\infty)^3$ for all $t \ge 0$, where

$$Z_1(t) \triangleq z_1 + M_1(t) + \mu_1\ell_0(t) + \ell_1(t), \tag{7.3}$$

$$Z_2(t) \triangleq z_2 - b_1\mu_2 t + M_2(t) - \mu_2\ell_0(t) + \ell_2(t), \tag{7.4}$$

$$Z_3(t) \triangleq z_3 + (b_1\mu_2 - b_2\mu_3)t + M_3(t) + \mu_2\ell_0(t) - \ell_2(t) + \ell_3(t). \tag{7.5}$$

The *cost function* associated with $(\ell_0, \ell_1, \ell_2, \ell_3)$, admissible at $z \in [0,\infty)^3$, is

$$V_{\ell_0,\ell_1,\ell_2,\ell_3}(z) \triangleq E\int_0^\infty e^{-\alpha t} h(Z(t))dt.$$

The *value function* for the controlled Brownian network is

$$V(z) \triangleq \inf\{V_{\ell_0,\ell_1,\ell_2,\ell_3}(z);\ (\ell_0, \ell_1, \ell_2, \ell_3) \text{ is admissible at } z\}, \qquad z \in [0, \infty)^3. \tag{7.6}$$

The cross variation formulas (7.1), (7.2) imply that the vector of martingales (M_1, M_2, M_3) is nothing more than a three-dimensional standard Brownian motion multiplied by a nonsingular matrix, so this vector of martingales is also a Markov process. If we set the control processes $\ell_0, \ell_1, \ell_2, \ell_3$ equal to zero, the state process $Z(t)$ given by (7.3)–(7.5) is Markov with infinitesimal generator

$$\mathcal{L}\varphi = -b_1\mu_2\varphi_2 + (b_1\mu_2 - b_2\mu_3)\varphi_3 + \lambda_1\varphi_{11} + \lambda_2\varphi_{22} - \lambda_2\varphi_{23} + \lambda_2\varphi_{33}, \tag{7.7}$$

where φ is any C^2 function from $[0, \infty)^3$ to $\mathbf{R}$ with φ_i denoting partial derivative with respect to the ith variable.

The controlled Brownian network is an intermediate problem between the queueing networks studied thus far and the workload control problem of the next section. Although the value function is well defined by (7.6), the problem does not have an optimal solution. We shall see in the next section that one would like to keep the state $Z(t)$ on a face of the orthant $[0, \infty)^3$, but this is not possible with the bounded variation control processes $\ell_0, \ell_1, \ell_2, \ell_3$. Fortunately, when we pass to the workload formulation, we will obtain a well-posed control problem.

We conclude this section with an asymptotic expansion of the infinitesimal generator $\mathcal{L}^{n,y,u}$ of (3.1) for the controlled queueing network. This expansion is needed for the proofs in the following sections and also explains the origin of the Brownian network problem introduced in this section.

Suppose that $\varphi : [0, \infty)^3 \to \mathcal{R}$ is thrice continuously differentiable, and all derivatives of φ up to order three are bounded uniformly on $[0, \infty)^3$. Fix $(y, u) \in \{0, 1\}^2$, and define

$$\theta^{(n)} = \sqrt{n}\left[\frac{\lambda_1^{(n)}}{\mu_1^{(n)}} - u\right], \quad \sigma_1 = \sqrt{n}\mu_1^{(n)}(1-y)u, \quad \sigma_2 = \sqrt{n}\mu_2^{(n)}(1-y)(1-u). \tag{7.8}$$

Recalling (2.2), we may write

$$u = -\frac{\theta^{(n)}}{\sqrt{n}} + \frac{\lambda_1^{(n)}}{\mu_1^{(n)}}, \qquad (1-u) = \frac{\theta^{(n)}}{\sqrt{n}} + \frac{\lambda_2^{(n)}}{\mu_2^{(n)}} + \frac{b_1^{(n)}}{\sqrt{n}}. \tag{7.9}$$

For $z \in [0, \infty)^3$, we set

$$\mathcal{B}_1^{n,y,u}\varphi(z) \triangleq \sqrt{n}\,\mu_1^{(n)} yu\left[\varphi_1(z) - \frac{1}{2\sqrt{n}}\varphi_{11}(z)\right] 1_{\{z_1=0\}}, \tag{7.10}$$

$$\begin{aligned}\mathcal{B}_2^{n,y,u}\varphi(z) \triangleq \sqrt{n}\,\mu_2^{(n)} y(1-u)\Bigg(&\varphi_2(z) - \varphi_3(z) - \frac{1}{2\sqrt{n}}\varphi_{22}(z) \\ &+ \frac{1}{\sqrt{n}}\varphi_{23}(z) - \frac{1}{2\sqrt{n}}\varphi_{33}(z)\Bigg) 1_{\{z_2=0\}},\end{aligned} \tag{7.11}$$

$$\mathcal{B}_3^{n,y,u}\varphi(z) \triangleq \sqrt{n}\,\mu_3^{(n)}\left[\varphi_3(z) - \frac{1}{2\sqrt{n}}\varphi_{33}(z)\right] 1_{\{z_3=0\}} \tag{7.12}$$

so that (3.1) becomes, by use of a Taylor expansion,

$$\begin{aligned}\mathcal{L}^{n,y,u}\varphi(z) &= \sqrt{n}\,\lambda_1^{(n)}\left[\varphi_1(z)+\frac{1}{2\sqrt{n}}\varphi_{11}(z)\right]+\sqrt{n}\,\lambda_2^{(n)}\left[\varphi_2(z)+\frac{1}{2\sqrt{n}}\varphi_{22}(z)\right]\\ &+(\sqrt{n}\,\mu_1^{(n)}u-\sigma_1)\left[-\varphi_1(z)+\frac{1}{2\sqrt{n}}\varphi_{11}(z)\right]+\mathcal{B}_1^{n,y,u}\varphi(z)\\ &+(\sqrt{n}\,\mu_2^{(n)}(1-u)-\sigma_2)\left[-\varphi_2(z)+\varphi_3(z)+\frac{1}{2\sqrt{n}}\varphi_{22}(z)-\frac{1}{\sqrt{n}}\varphi_{23}(z)+\frac{1}{2\sqrt{n}}\varphi_{33}(z)\right]\\ &+\mathcal{B}_2^{n,y,u}\varphi(z)\\ &+\sqrt{n}\,\mu_3^{(n)}\left[-\varphi_3(z)+\frac{1}{2\sqrt{n}}\varphi_{33}(z)\right]+\mathcal{B}_3^{n,y,u}\varphi(z)+O\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Because the derivatives of φ are bounded, we can conclude from (2.3) and (7.9) that

$$\begin{aligned}\mathcal{L}^{n,y,u}\varphi(z) &= \mathcal{L}\varphi(z)+\theta^{(n)}\left[\nabla\varphi(z)\cdot\xi^{(n)}+\frac{1}{\sqrt{n}}\mathcal{A}^{(n)}\varphi(z)\right]\\ &+\sigma_1\left[\varphi_1(z)-\frac{1}{\sqrt{n}}\varphi_{11}(z)\right]\\ &+\sigma_2\left[\varphi_2(z)-\varphi_3(z)-\frac{1}{2\sqrt{n}}(\varphi_{22}(z)-2\varphi_{23}(z)+\varphi_{33}(z))\right]\\ &+\sum_{i=1}^{3}\mathcal{B}_i^{n,y,u}\varphi(z)+O\left(\frac{1}{\sqrt{n}}\right),\end{aligned}\tag{7.13}$$

where $\mathcal{L}\varphi$ is given by (7.7) and

$$\xi^{(n)} \triangleq (\mu_1^{(n)}, -\mu_2^{(n)}, \mu_2^{(n)}),\tag{7.14}$$

$$\mathcal{A}^{(n)}\varphi \triangleq -\frac{1}{2}\mu_1^{(n)}\varphi_{11}+\frac{1}{2}\mu_2^{(n)}(\varphi_{22}-2\varphi_{23}+\varphi_{33}).\tag{7.15}$$

The expressions in (7.14), (7.15) are bounded uniformly in n. However, $\theta^{(n)}$, σ_1, and σ_2 are of order $\sqrt{n}$, as are the terms $\mathcal{B}_i^{n,y,u}\varphi$. The term $\nabla\varphi\cdot\xi^{(n)}+\frac{1}{\sqrt{n}}\mathcal{A}^{(n)}\varphi$ in (7.13) agrees with $\nabla\varphi\cdot\xi$ up to an error of order $\frac{1}{\sqrt{n}}$, but this term cannot immediately be replaced by $\nabla\varphi\cdot\xi$ because $\theta^{(n)}$ multiplying it is of order $\sqrt{n}$. In §11 we treat this term by adding a corrector to the function which is the argument of $\mathcal{L}^{n,y,u}$. The corrector causes the offending term to vanish.

Equation (7.13) suggests that the controlled Brownian motion $Z(t)$ given by (7.3)–(7.5) approximates the scaled queue length process $Z^{(n)}(t)=\frac{1}{\sqrt{n}}Q^{(n)}(nt)$. The control variable $\theta^{(n)}$ in (7.13), which can be either positive or negative, corresponds to pushing in approximately the direction $\xi \triangleq (\mu_1, -\mu_2, \mu_2)$ or the direction $-\xi$. In (7.3)–(7.5), this pushing is accomplished by the locally finite variation process ℓ_0. The processes ℓ_1, ℓ_2, and ℓ_3 appearing in (7.3)–(7.5) allow us to enforce the condition $Z(t)\in[0,\infty)^3$ for all $t\geq 0$. We have set up the controlled Brownian network to allow ℓ_i to grow even when $Z_i(t)>0$; this corresponds to idling the serving stations.

Remark 7.1. When all derivatives of φ up to order three are bounded uniformly on $[0,\infty)^3$, then $O(\frac{1}{\sqrt{n}})$ in (7.13) is a term whose absolute value is bounded by $K/\sqrt{n}$, where

K is a constant independent of n and $z \in [0, \infty)^3$. If φ is of class C^3, but only with locally bounded derivatives, then the term $O(\frac{1}{\sqrt{n}})$ is bounded by $k(z)/\sqrt{n}$, where $k(\cdot)$ is a locally bounded function of $z \in [0, \infty)^3$. We need the uniform bound in (11.24) when we are obtaining a lower bound on the limit of the queueing system value functions so that we can proceed to (11.25). The nonuniform bound is sufficient for (12.66) when the upper bound is sought.

8. The workload formulation. Following Harrison and Wein (1989), we introduce the *workload transformation*

$$\omega(z_1, z_2, z_3) \triangleq \left(\frac{z_1}{\mu_1} + \frac{z_2}{\mu_2}, \frac{z_2 + z_3}{\mu_3} \right), \tag{8.1}$$

which maps the state space $[0, \infty)^3$ of the controlled Brownian network onto the state space $[0, \infty)^2$ of the *workload control problem* formulated in this section. If (z_1, z_2, z_3) represents the three queue lengths, then $(w_1, w_2) = \omega(z_1, z_2, z_3)$ is the expected impending service time for the two stations embodied in customers anywhere in the network. The workload formulation reduces the dimensionality of the control problem from three (the number of customer classes) to two (the number of stations).

Because we can use the control process ℓ_0 in (7.3)–(7.5) to instantaneously change the state $Z(t)$ in the directions $\pm\xi = \pm(\mu_1, -\mu_2, \mu_2)$ at no cost, the Brownian network value function V of (7.6) will be constant along the direction ξ. This means that $V(z)$ can be written as a function of $\omega(z)$, because $\omega(z)$ does not change along the ξ-direction. It also means that one would want to keep the process $Z(t)$ on the locus of points in $[0, \infty)^3$ which minimize h along line segments parallel to ξ. To find this locus, one considers for each $(w_1, w_2) \in [0, \infty)^2$ the linear program

$$\begin{aligned} \text{minimize} \quad & c_1 z_1 + c_2 z_2 + c_3 z_3 \\ \text{subject to} \quad & \frac{z_1}{\mu_1} + \frac{z_2}{\mu_2} = w_1, \\ & \frac{z_2}{\mu_3} + \frac{z_3}{\mu_3} = w_2, \\ & z_1 \geq 0, \quad z_2 \geq 0, \quad z_3 \geq 0. \end{aligned}$$

Denote by $\hat{h}(w_1, w_2)$ the value of this linear program. We have two major cases.

Case I ($c_1\mu_1 - c_2\mu_2 + c_3\mu_2 \leq 0$). In this case,

$$\hat{h}(w_1, w_2) = c_1\mu_1 w_1 + c_3\mu_3 w_2, \tag{8.2}$$

and the minimizer in the linear program is

$$z_1^* = \mu_1 w_1, \quad z_2^* = 0, \quad z_3^* = \mu_3 w_2. \tag{8.3}$$

Case II ($c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$). Now

$$\hat{h}(w_1, w_2) = \begin{cases} (c_2\mu_2 - c_3\mu_2)w_1 + c_3\mu_3 w_2 & \text{if} \quad \mu_3 w_2 \geq \mu_2 w_1, \\ c_1\mu_1 w_1 + \dfrac{\mu_3}{\mu_2}(c_2\mu_2 - c_1\mu_1)w_2 & \text{if} \quad \mu_3 w_2 \leq \mu_2 w_1. \end{cases} \tag{8.4}$$

The minimizing values are

$$z_1^* = 0, \quad z_2^* = \mu_2 w_1, \quad z_3^* = \mu_3 w_2 - \mu_2 w_1 \quad \text{if} \quad \mu_3 w_2 \geq \mu_2 w_1, \tag{8.5a}$$

$$z_1^* = \frac{\mu_1}{\mu_2}(\mu_2 w_1 - \mu_3 w_2), \quad z_2^* = \mu_3 w_2, \quad z_3^* = 0 \quad \text{if} \quad \mu_3 w_2 \leq \mu_2 w_1. \tag{8.5b}$$

The *workload control problem* has state equations

$$W_1(t) = w_1 - b_1 t + \frac{1}{\mu_1} M_1(t) + \frac{1}{\mu_2} M_2(t) + m_1(t), \tag{8.6}$$

$$W_2(t) = w_2 - b_2 t + \frac{1}{\mu_3} M_2(t) + \frac{1}{\mu_3} M_3(t) + m_2(t), \tag{8.7}$$

where the pair (m_1, m_2) of $\{\mathcal{F}(t)\}$-adapted control processes is *admissible for initial condition* $w = (w_1, w_2) \in [0, \infty)^2$, provided that

(i) m_1 and m_2 are right-continuous with left-hand limits, with the convention that

$$m_i(0^-) = 0, \qquad i = 1, 2;$$

(ii) m_1 and m_2 are nondecreasing;

(iii) the state process $W(t) = (W_1(t), W_2(t))$ is in $[0, \infty)^2$ for all $t \geq 0$.

(We have in mind, of course, that $m_1(t) = \frac{\ell_1(t)}{\mu_1} + \frac{\ell_2(t)}{\mu_2}$, $m_2(t) = \frac{\ell_3(t)}{\mu_3}$, where ℓ_1 and ℓ_3 are part of an admissible quadruple $(\ell_0, \ell_1, \ell_2, \ell_3)$ for the controlled Brownian network.) The *cost function* associated with (m_1, m_2) at $w \in [0, \infty)^2$ is

$$\hat{V}_{m_1, m_2}(w) \triangleq E \int_0^\infty e^{-\alpha t} \hat{h}(W(t)) dt,$$

and the *value function* at w is

$$\hat{V}(w) = \inf\{\hat{V}_{m_1, m_2}(w); \quad (m_1, m_2) \text{ is admissible at } w\}. \tag{8.8}$$

Although we do not need this fact for our analysis, one can show that V of (7.6) and $\hat{V}$ of (8.8) are related by the equation

$$V(z_1, z_2, z_3) = \hat{V}\left(\frac{z_1}{\mu_1} + \frac{z_2}{\mu_2}, \frac{z_2 + z_3}{\mu_3}\right) \quad \forall\, z \in [0, \infty)^3. \tag{8.9}$$

If one had an optimal (m_1^*, m_2^*) for the workload control problem, then as an optimal policy for the Brownian network problem, one would want to take $\ell_1^*(t) = \mu_1 m_1^*(t)$, $\ell_2^*(t) \equiv 0$, $\ell_3^*(t) = \mu_3 m_2^*(t)$ and choose ℓ_0 to ensure that $Z^*(t)$ is always given by (8.3) or (8.5) with $w_i = W_i^*(t)$, $i = 1, 2$, depending on the sign of $c_1\mu_1 - c_2\mu_2 + c_3\mu_2$. However, such an ℓ_0 does not exist, so the Brownian network control problem is ill posed.

9. Solution of Case I. This is the case $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 \leq 0$. Since $\hat{h}$ given by (8.2) is increasing in each variable separately, the optimal control processes m_1 and m_2 act only when $W_1 = 0$ or $W_2 = 0$, respectively. More precisely,

$$m_1(t) \triangleq \max_{0 \leq s \leq t} \left[-w_1 + b_1 s - \frac{1}{\mu_1} M_1(s) - \frac{1}{\mu_2} M_2(s)\right]^+, \tag{9.1}$$

$$m_2(t) \triangleq \max_{0 \leq s \leq t} \left[-w_2 + b_2 s - \frac{1}{\mu_3} M_2(s) - \frac{1}{\mu_3} M_3(s)\right]^+ \tag{9.2}$$

(see, e.g., Harrison (1985)) are the minimal nondecreasing processes which ensure that the associated state processes remain nonnegative almost surely. In particular,

$$m_i(t) = \int_0^t 1_{\{W_1(s) = 0\}} dm_i(s), \qquad 0 \leq t < \infty. \tag{9.3}$$

One can actually compute the value function

$$\begin{aligned}\hat{V}(w_1, w_2) &= \hat{V}_{m_1,m_2}(w_1, w_2) \\ &= A + \gamma_1 B_1 w_1 + \gamma_2 B_2 w_2 + B_1 e^{-\gamma_1 w_1} + B_2 e^{-\gamma_2 w_2},\end{aligned} \tag{9.4}$$

where $\gamma_1 > 0$, $\gamma_2 > 0$ solve the quadratic equations

$$\left(\frac{\lambda_1}{\mu_1^2} + \frac{\lambda_2}{\mu_2^2}\right)\gamma_1^2 + b_1\gamma_1 - 1 = 0, \qquad \frac{\lambda_2}{\mu_3^2}\gamma_2^2 + b_2\gamma_2 - 1 = 0,$$

and $B_1 = c_1\mu_1/\gamma_1$, $B_2 = c_3\mu_3/\gamma_2$, $A = -\gamma_1 b_1 B_1 - \gamma_2 b_2 B_2$.

The formula $z_2^* = 0$ in (8.3) suggests that customer class 2 should always have priority, a fact already established by Chen, Yang, and Yao (1991). Thus, for the queueing networks, we define the stationary control law (independent of n in this case)

$$Y(z) \triangleq \begin{cases} 1 & \text{if} \quad z_1 > 0 \quad \text{or} \quad z_2 > 0, \\ 0 & \text{if} \quad z_1 = z_2 = 0, \end{cases}$$

$$U(z) \triangleq \begin{cases} 0 & \text{if} \quad z_2 > 0, \\ 1 & \text{if} \quad z_2 = 0. \end{cases}$$

One can show that (Y, U) is asymptotically optimal in the sense of (6.4) with $\eta = 0$. We omit the proof, focusing instead on the more complicated Case IIA below.

10. Discussion of Case II. This is the case $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$. We shall complete the analysis only for Case IIA.

Case IIA ($c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0, c_2\mu_2 - c_3\mu_2 \geq 0, c_2\mu_2 - c_1\mu_1 \geq 0$). In this case, the function $\hat{h}$ given by (8.4) is nondecreasing in each variable separately. The optimal control processes for the workload problem are still given by (9.1), (9.2) and satisfy (9.3), but $\hat{V}$ no longer has the simple closed form (9.4). Because

$$\hat{V}(w) = E\int_0^\infty e^{-\alpha t}\hat{h}(W(t))dt \quad \forall\, w \in [0, \infty)^2, \tag{10.1}$$

the Feynman–Kac formula and elliptic regularity imply that $\hat{V}$ is C^2 on the open quadrant $(0, \infty)^2$, $\hat{V}$ is C^1 on the closed quadrant $[0, \infty)^2$, and

$$\hat{V}_1(0, w_2) = \hat{V}_2(w_1, 0) = 0 \quad \forall (w_1, w_2) \in [0, \infty)^2, \tag{10.2}$$

$$\alpha\hat{V} - \hat{\mathcal{L}}\hat{V} - \hat{h} = 0 \quad \text{on} \quad (0, \infty)^2, \tag{10.3}$$

where

$$\hat{\mathcal{L}}\hat{\varphi} \triangleq -b_1\hat{\varphi}_1 - b_2\hat{\varphi}_2 + \left(\frac{\lambda_1}{\mu_1^2} + \frac{\lambda_1}{\mu_2^2}\right)\hat{\varphi}_{11} + \frac{\lambda_2}{\mu_2\mu_3}\hat{\varphi}_{12} + \frac{\lambda_2}{\mu_3^2}\hat{\varphi}_{22} \tag{10.4}$$

for any C^2 function $\hat{\varphi} : (0, \infty)^2 \to \mathbf{R}$. To verify that $\hat{V}$ is of class C^2 on $(0, \infty)^2$, we can let Ω be an arbitrary domain in $(0, \infty)^2$ satisfying an exterior sphere condition and then solve the Dirichlet problem

$$\alpha u - \hat{\mathcal{L}}u - \hat{h} = 0 \quad \text{in} \quad \Omega, \quad u = \hat{V} \quad \text{on} \quad \partial\Omega.$$

This problem has a solution u which is C^2 in Ω (Gilbarg and Trudinger (1977, Thm. 11.5)), and it is an easy exercise using Itô's formula to verify that $u = \hat{V}$. To prove the claimed boundary behavior along the w_1-axis, we can choose a domain Ω in $(0, \infty) \times \mathbf{R}$ which is symmetric about the w_1-axis, and we can extend $\hat{\mathcal{L}}$, $\hat{h}$, and $\hat{V}$ across the w_1-axis by even symmetry so that the extended functions are continuous and piecewise differentiable. The above Dirichlet problem still has a solution u which is C^1 in Ω (Gilbarg and Trudinger (1977, Thm. 8.9)), and because of the even symmetry, $w_2 = 0$ on the intersection of Ω with the w_1-axis. Again, Itô's formula can be used to verify that u and $\hat{V}$ agree in the intersection of Ω with the upper half-plane. The same argument applies to the w_2-axis, and the condition $\hat{V}_1(0, 0) = \hat{V}_2(0, 0) = 0$ is obtained by letting Ω be symmetric with respect to the origin. Note that with $\hat{\varphi}$ as in (10.4) we have

$$(\hat{\mathcal{L}}\hat{\varphi})(\omega(z)) = \mathcal{L}(\hat{\varphi} \circ \omega)(z) \quad \forall\, z \in [0, \infty)^3. \tag{10.5}$$

The principal result of this paper is the following theorem.

THEOREM 10.1. *Assume Case* IIA. *Then*

$$J_\#(z) = J^\#(z) = \hat{V}(\omega(z)) \quad \forall\, z \in [0, \infty)^3,$$

where $J_\#$ and $J^\#$ are the lower and upper semicontinuous limits of the queueing network value functions, defined by (6.1) *and* (6.2), *respectively.*

The proof of Theorem 10.1 is the subject of the next two sections. Since $J_\# \leq J^\#$, it suffices to prove the two inequalities

$$\hat{V}(\omega(z)) \leq J_\#(z), \quad J^\#(z) \leq \hat{V}(\omega(z)) \quad \forall\, z \in [0, \infty)^3.$$

In §11, we prove the first of these inequalities, and in §12, we prove the second. The proof of the second inequality requires the construction of a sequence of asymptotically nearly optimal stationary policies (defined by (12.5a), (12.5b)) which satisfy (6.4). We establish additional properties of $\hat{V}$ in the next section.

Case IIB ($c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$, $c_2\mu_2 - c_3\mu_2 < 0$, $c_2\mu_2 - c_1\mu_1 \geq 0$). Now $\hat{h}$ is strictly decreasing in w_1 for $w_1 \in [0, \frac{\mu_3}{\mu_2} w_2]$, which suggests that w_1 should not be allowed to fall too far below $\frac{\mu_3}{\mu_2} w_2$. Numerical experimentation supports the conjecture that there is a continuous, increasing function $\Psi_2 : [0, \infty) \longrightarrow [0, \infty)$ such that the optimal control process m_1 in the workload control problem acts whenever $W_1(t) = \Psi_2(W_2(t))$ to ensure that the inequality $W_1(t) \geq \Psi_2(W_2(t))$ is always satisfied. The rest of the conjecture was set out in §2.

Cases IIC, IID. The functions Ψ_1 and Ψ_2 appearing in the conjectures in §2 about these cases are the free boundaries on which reflection should occur in the optimal control of the workload processes.

11. The lower bound. Throughout the remainder of the paper, we assume Case IIA. In particular, $\hat{V}$ is given by (10.1), where W is determined by (8.6), (8.7), (9.1), and (9.2), and $\hat{h}$ is given by (8.4). The purpose of this section is to prove the following proposition.

PROPOSITION 11.1. *Assume Case* IIA. *Then*

$$\hat{V}(\omega(z)) \leq J_\#(z) \quad \forall\, z \in [0, \infty)^3, \tag{11.1}$$

where $J_\#$, defined by (6.2), *is the lower semicontinuous limit of the queueing network value functions and ω is given by* (8.1).

The proof of Proposition 11.1 proceeds through several steps.

LEMMA 11.2. *The workload value function* $\hat{V} : [0, \infty)^2 \longrightarrow [0, \infty)$ *is strictly increasing in each variable and is convex. There are positive constants* K_0, K_1, *and* K_2 *such that*

$$K_0\hat{h}(w) \leq \hat{V}(w) \leq K_1 + K_2\hat{h}(w) \quad \forall\, w \in [0, \infty). \tag{11.2}$$

Furthermore, the partial derivatives $\hat{V}_1$ *and* $\hat{V}_2$ *are uniformly bounded on* $[0, \infty)^2$.

Proof. We may rewrite (8.4) as

$$\hat{h}(w) = \max\left\{(c_2\mu_2 - c_3\mu_2)w_1 + c_3\mu_3 w_2,\, c_1\mu_1 w_1 + \frac{\mu_3}{\mu_2}(c_2\mu_2 - c_1\mu_1)w_2\right\} \quad \forall\, w \in [0, \infty)^2,$$

which shows that $\hat{h}$ is convex, and $\hat{V}$, being the value function for a control problem with linear dynamics and a convex state space, inherits the convexity of $\hat{h}$. The representation (10.1) of $\hat{V}$ shows that $\hat{V}$ is strictly increasing in each variable and grows at most linearly. Such a function must also grow at least linearly, and we have (11.2) for suitable positive constants K_0, K_1, and K_2. A linearly growing convex function must have bounded partial derivatives. □

The idea behind the proof of Proposition 11.1 is to alter the function

$$v(z) \triangleq \hat{V}(\omega(z)), \qquad z \in [0, \infty)^3, \tag{11.3}$$

in order to obtain a subsolution of the HJB equation (5.2). Proposition 5.1 will then imply that $J_*^{(n)}$ of (10.4) dominates the altered function. We then let $n \to \infty$ to obtain (11.1).

The construction of the altered version of v requires three steps. First, we mollify v to obtain a smoother function. Next, we compose v with a truncation function so as to restrict attention to a compact subset of the domain of v. Finally, we add a "corrector" to cancel the order $\frac{1}{\sqrt{n}}$ error incurred when $\nabla v(z) \cdot \xi^{(n)} + \frac{1}{\sqrt{n}}\mathcal{A}^{(n)}v(z)$ is replaced by $\nabla v \cdot \xi$ in the expansion of $\mathcal{L}^{n,y,u}v$ (see (7.13)). This error must be cancelled because $\theta^{(n)}$ multiplying the error is of order $\sqrt{n}$.

Step 1. *Mollification.* Let ρ_ϵ be a nonnegative C^∞ function with support contained in the open ball $B_\epsilon(0)$ of radius $\epsilon > 0$ centered at the origin in $\mathbf{R}^2$ and such that $\int_{B_\epsilon(0)} \rho_\epsilon = 1$. We define

$$\hat{V}_0^\epsilon(w) = \int_{B_\epsilon(0)} \hat{V}(x + w)\rho_\epsilon(x)dx = \int_{\mathbf{R}^2} \hat{V}(x)\rho_\epsilon(x - w)dx.$$

(To make this definition possible, we first extend $\hat{V}$ to $\mathbf{R}^2$ so that it remains continuous.) We likewise define

$$\hat{h}^\epsilon(w) = \int_{B_\epsilon(0)} \hat{h}(x + w)\rho_\epsilon(x)dx = \int_{\mathbf{R}^2} \hat{h}(x)\rho_\epsilon(x - w)dx.$$

On the set $(\epsilon, \infty)^2$, (10.3) implies $\alpha\hat{V}_0^\epsilon - \hat{\mathcal{L}}\hat{V}_0^\epsilon - \hat{h}^\epsilon = 0$.

Let us set $\hat{V}^\epsilon(w_1, w_2) = \hat{V}_0^\epsilon(w_1 + 2\epsilon, w_2 + 2\epsilon)$. The Lipschitz continuity of $\hat{h}$ (and hence $\hat{h}^\epsilon$) implies

$$|\alpha\hat{V}^\epsilon - \hat{\mathcal{L}}\hat{V}^\epsilon - \hat{h}^\epsilon| \leq L_0\epsilon \quad \text{on} \quad (-\epsilon, \infty)^2, \tag{11.4}$$

where L_0 is a Lipschitz constant for $\hat{h}$. Like $\hat{V}$, each $\hat{V}^\epsilon$ is strictly increasing in each variable and satisfies (11.2). Furthermore,

$$\frac{\partial}{\partial w_i}\hat{V}_0^\epsilon(w) = \int_{B_\epsilon(0)} \hat{V}_i(x + w)\rho_\epsilon(x)dx = \int_{\mathbf{R}^2} \hat{V}_i(x)\rho_\epsilon(x - w)dx.$$

Computing all higher-order derivatives by differentiating under the second integral, we see that for each N, there is a constant $C_{N,\epsilon}$ such that all partial derivatives up to order N of $\hat{V}^\epsilon$ are bounded in absolute value by $C_{N,\epsilon}$ uniformly on $(-\epsilon, \infty)^2$.

We next define

$$v^\epsilon(z) = \hat{V}^\epsilon(\omega(z)), \qquad z \in (-\delta, \infty)^3, \tag{11.5}$$

where $\delta \triangleq \frac{\epsilon}{2} \min\{\mu_1, \mu_2, \mu_3\}$, and note that v^ϵ converges to v of (11.3) uniformly on $[0, \infty)^3$, i.e.,

$$\lim_{\epsilon \downarrow 0} \sup_{z \in [0,\infty)^3} |v^\epsilon(z) - v(z)| = 0. \tag{11.6}$$

From its definition, we see that v^ϵ satisfies

$$\nabla v^\epsilon(z) \cdot \xi = 0 \quad \forall\, z \in (-\delta, \infty)^3, \tag{11.7}$$

where $\xi = (\mu_1, -\mu_2, \mu_2)$. Inequality (11.4) and the fact that $\hat{h}$ solves the linear program in §8 imply

$$\begin{aligned} \alpha v^\epsilon(z) - \mathcal{L} v^\epsilon(z) &\le \hat{h}^\epsilon\left(\frac{z_1}{\mu_1} + \frac{z_2}{\mu_2}, \frac{z_2}{\mu_3} + \frac{z_3}{\mu_3}\right) + L_0\epsilon \\ &\le c \cdot z + 2L_0\epsilon, \end{aligned} \tag{11.8}$$

where $c \triangleq (c_1, c_2, c_3)$. There is a constant $C_{1,\epsilon}$ such that

$$\|v_i^\epsilon\|_\infty + \|v_{ij}^\epsilon\|_\infty + \|v_{ijk}^\epsilon\|_\infty + \|v_{ijkl}^\epsilon\|_\infty + \|v_{ijklm}^\epsilon\|_\infty \le C_{1,\epsilon} \quad \forall\, i, j, k, l, m \in \{1, 2, 3\}, \tag{11.9}$$

where $\|\cdot\|_\infty$ is the supremum norm on $(-\delta, \infty)^3$.

Finally, because $\hat{v}$ is increasing in each variable separately, we have

$$v_1^\epsilon \ge 0, \quad v_2^\epsilon - v_3^\epsilon \ge 0, \quad v_3^\epsilon \ge 0 \quad \text{on} \quad (-\delta, \infty)^3, \tag{11.10}$$

and there is a constant K_1^ϵ satisfying

$$v^\epsilon(z) \le K_1^\epsilon + \frac{1}{\alpha} c \cdot z \quad \forall\, z \in [0, \infty)^3, \tag{11.11}$$

the last inequality following from (11.8) and the uniform boundedness of $\mathcal{L} v^\epsilon$.

Step 2. *Truncation.* Fix $\beta > 0$, and let $\varphi_\beta : \mathcal{R} \to \mathcal{R}$ be a C^∞ function with the following properties:

(i) $\varphi_\beta(x) = x$ if $x \le \frac{1}{\beta}$;
(ii) $\varphi_\beta'(x) = 0$ if $x \ge \frac{4}{\beta}$;
(iii) $-\beta \le \varphi_\beta''(x) \le 0 \ \forall\, x \ge 0$.

One could, for example, take the C^1 function

$$f(x) = \begin{cases} x & \text{if} \quad x \le \dfrac{2}{\beta}, \\ -\dfrac{1}{2}\beta x^2 + 3x - \dfrac{2}{\beta} & \text{if} \quad \dfrac{2}{\beta} \le x \le \dfrac{3}{\beta}, \\ \dfrac{5}{2\beta} & \text{if} \quad x \ge \dfrac{3}{\beta} \end{cases}$$

and mollify it.

We truncate v^ϵ by defining $u^{\beta,\epsilon}(z) = \varphi_\beta(v^\epsilon(z))$ for $z \in (-\delta, \infty)^3$. Then

$$\alpha u^{\beta,\epsilon}(z) - \mathcal{L}u^{\beta,\epsilon}(z) = \alpha\varphi_\beta(v^\epsilon(z)) - \varphi'_\beta(v^\epsilon(z))\mathcal{L}v^\epsilon(z) + O(\beta),$$

where $|O(\beta)|$ is bounded by β times a constant depending on $C_{1,\epsilon}$ in (11.9). Because $\varphi_\beta(x) = x$ for $0 \le x \le \frac{1}{\beta}$, we have from (11.8) that

$$v^\epsilon(z) \le \frac{1}{\beta} \Rightarrow \alpha u^{\beta,\epsilon}(z) - \mathcal{L}u^{\beta,\epsilon}(z) = \alpha v^\epsilon(z) - \mathcal{L}v^\epsilon(z) \le c \cdot z + 2L_0\epsilon.$$

Conditions (i)–(iii) above imply

(iv) $0 \le \varphi'_\beta(x) \le 1 \; \forall x \ge 0$.

This fact, (11.8), and (11.11) allow us to argue that

$$\begin{aligned}
\alpha u^{\beta,\epsilon}(z) - \mathcal{L}u^{\beta,\epsilon}(z) &\le \alpha\varphi_\beta(v^\epsilon(z)) - \varphi'_\beta(v^\epsilon(z))[\alpha v^\epsilon(z) - c \cdot z - 2L_0\epsilon] + O(\beta) \\
&\le \alpha[\varphi_\beta(v^\epsilon(z)) - \varphi'_\beta(v^\epsilon(z))v^\epsilon(z)] \\
&\quad - [1 - \varphi'_\beta(v^\epsilon(z))]c \cdot z + c \cdot z + 2L_0\epsilon + O(\beta) \\
&\le \alpha[\varphi_\beta(v^\epsilon(z)) - \varphi'_\beta(v^\epsilon(z))v^\epsilon(z) \\
&\qquad - (1 - \varphi'_\beta(v^\epsilon(z))(v^\epsilon(z) - K_1^\epsilon))] \\
&\quad + c \cdot z + 2L_0\epsilon + O(\beta) \\
&= \alpha[\varphi_\beta(v^\epsilon(z)) - v^\epsilon(z) + K_1^\epsilon(1 - \varphi'_\beta(v^\epsilon(z)))] \\
&\quad + c \cdot z + 2L_0\epsilon + O(\beta).
\end{aligned}$$

Since $\varphi_\beta(x) \le x$ for all x,

$$\varphi_\beta(v^\epsilon(z) - K_1^\epsilon) \le v^\epsilon(z) - K_1^\epsilon,$$

from which we conclude that

$$\begin{aligned}
\varphi_\beta(v^\epsilon(z)) - v^\epsilon(z) + K_1^\epsilon(1 - \varphi'_\beta(v^\epsilon(z))) &\le \varphi_\beta(v^\epsilon(z)) - \varphi_\beta(v^\epsilon(z) - K_1^\epsilon) - K_1^\epsilon\varphi'_\beta(v^\epsilon(z)) \\
&\le \int_{v^\epsilon(z)-K_1^\epsilon}^{v^\epsilon(z)} [\varphi'_\beta(r) - \varphi'_\beta(v^\epsilon(z))]dr \\
&= -\int_{v^\epsilon(z)-K_1^\epsilon}^{v^\epsilon(z)} \int_r^{v^\epsilon(z)} \varphi''_\beta(\rho)d\rho\, dr \\
&\le \frac{1}{2}(K_1^\epsilon)^2\beta = O(\beta).
\end{aligned}$$

Thus, regardless of whether $v^\epsilon(z) \le \frac{1}{\beta}$ or $v^\epsilon(z) \ge \frac{1}{\beta}$, we have

$$\alpha u^{\beta,\epsilon}(z) - \mathcal{L}u^{\beta,\epsilon}(z) \le c \cdot z + 2L_0\epsilon + O(\beta), \qquad z \in (-\delta, \infty)^3. \tag{11.12}$$

Also, from (11.7), (11.9), and (11.10), we have

$$\nabla u^{\beta,\epsilon}(z) \cdot \xi = 0 \quad \forall\, z \in (-\delta, \infty)^3, \tag{11.13}$$

$$\begin{aligned}
&\|u_i^{\beta,\epsilon}\|_\infty + \|u_{ij}^{\beta,\epsilon}\|_\infty + \|u_{i,j,k}^{\beta,\epsilon}\|_\infty + \|u_{i,j,k,l}^{\beta,\epsilon}\|_\infty + \|u_{i,j,k,l,m}^{\beta,\epsilon}\|_\infty \le C_{2,\epsilon} \\
&\quad \forall\ \in i, j, k, \ell, m \in \{1, 2, 3\},
\end{aligned} \tag{11.14}$$

$$u_1^{\beta,\epsilon} \ge 0, \quad u_2^{\beta,\epsilon} - u_3^{\beta,\epsilon} \ge 0, u_3^{\beta,\epsilon} \ge 0 \quad \text{on} \quad (-\delta, \infty)^3, \tag{11.15}$$

where $C_{2,\epsilon}$ is a constant depending on ϵ. In place of (11.11), we have now the existence of a constant $R_1^{\beta,\epsilon}$ such that

$$u_i^{\beta,\epsilon}(z_1, z_2, z_3) = 0 \quad \text{if} \quad z_1 \vee z_2 \vee z_3 \geq R_1^{\beta,\epsilon}, \quad i = 1, 2, 3. \tag{11.16}$$

Step 3. *Construction of the corrector.* For each n, define

$$\begin{aligned} \Psi^{n,\beta,\epsilon}(z) &\triangleq -\sqrt{n}\nabla u^{\beta,\epsilon}(z) \cdot \xi^{(n)} - \mathcal{A}^n u^{\beta,\epsilon}(z) \\ &= \sqrt{n}\nabla u^{\beta,\epsilon}(z) \cdot (\xi - \xi^{(n)}) - \mathcal{A}^n u^{\beta,\epsilon}(z), \qquad z \in (-\delta, \infty)^3 \end{aligned} \tag{11.17}$$

($\mathcal{A}^n$ is defined by (7.15)), and note that $\Psi^{n,\beta,\epsilon}$ and its first, second, and third partial derivatives are bounded uniformly in z, n, and β and

$$\Psi^{u,\beta,\epsilon}(z) = 0 \quad \text{if} \quad z_1 \vee z_2 \vee z_3 \geq R_1^{\beta,\epsilon}. \tag{11.18}$$

Define $t(z) \triangleq \min\{\frac{z_1}{\mu_1}, \frac{z_3}{\mu_2}\}$, and let $T : [-\delta, \infty)^3 \to \mathbf{R}$ be a mollification of t which does not depend on the variable z_2 and which satisfies the following:

(v) T is thrice continuously differentiable;
(vi) the derivatives of T, up to second order, are uniformly bounded;
(vii) $\mu_1 T_1 + \mu_2 T_3 = 1$ on $[0, \infty)^3$;
(viii) $t(z) \leq T(z) \leq t(z) + \nu_2 \; \forall\, z \in [-\nu_1, \infty]^3$,

where $\nu_1 > 0$ and $\nu_2 > 0$ are chosen so that

$$z - t(z)\xi^{(n)} \in (-\delta, \infty)^3, \quad z - (t(z) + \nu_2)\xi^{(n)} \in (-\delta, \infty)^3 \quad \forall\, z \in [-\nu_1, \infty)^3, \forall n.$$

We may now define the corrector

$$f^{n,\beta,\epsilon}(z) \triangleq C_{3,\epsilon,\beta}(w_1^{(n)} + w_2^{(n)}) + \int_0^{T(z)} \Psi^{n,\beta,\epsilon}(z - \rho\xi^{(n)})d\rho, \qquad z \in [-\nu_1, \infty)^3, \tag{11.19}$$

where

$$w_1^{(n)} \triangleq \frac{z_1}{\mu_1^{(n)}} + \frac{z_2}{\mu_2^{(n)}}, \qquad w_2^{(n)} = \frac{z_2}{\mu_3^{(n)}} + \frac{z_3}{\mu_3^{(n)}},$$

and the constant $C_{3,\epsilon,\beta}$ is chosen below, independently of n.

Direct computation reveals that

$$\nabla f^{n,\beta,\epsilon} \cdot \xi^{(n)} = \Psi^{n,\beta,\epsilon} + O\left(\frac{1}{\sqrt{n}}\right) \quad \text{on} \quad [-\nu_1, \infty)^3. \tag{11.20}$$

Also, (11.18) shows that the term $\int_0^{T(z)} \Psi^{n,\beta,\epsilon}(z - \rho\xi^{(n)})d\rho$ and its first, second, and third partial derivatives are bounded uniformly in z and n. Consequently, we may choose the constant $C_{3,\epsilon,\beta}$ independently of n so that

$$f_1^{n,\beta,\epsilon} \geq 0, \; f_2^{n,\beta,\epsilon} - f_3^{n,\beta,\epsilon} \geq 0, \; f_3^{n,\beta,\epsilon} \geq 0 \quad \text{on} \quad [-\nu_1, \infty)^3. \tag{11.21}$$

Similarly, we may choose a constant $K_{\epsilon,\beta}$ independently of n so that

$$-K_{\epsilon,\beta} \leq f^{n,\beta,\epsilon}(z) \leq K_{\epsilon,\beta}(1 + c \cdot z) \quad \forall\, z \in [-\nu_1, \infty)^3. \tag{11.22}$$

Finally, we have

$$\|f_i^{n,\beta,\epsilon}\|_\infty + \|f_{i,j}^{n,\beta,\epsilon}\|_\infty + \|f_{i,j,k}^{n,\beta,\epsilon}\|_\infty \leq C_{4,\epsilon,\beta} \quad \forall i, j, k \in \{1, 2, 3\}, \tag{11.23}$$

where $C_{4,\epsilon,\beta}$ is independent of n and we mean $\|\cdot\|_\infty$ to be the supremum norm on $[0, \infty)^3$.

Step 4. *Subsolution confirmation.* We set

$$g^{n,\beta,\epsilon}(z) = u^{\beta,\epsilon}(z) + \frac{1}{\sqrt{n}} f^{n,\beta,\epsilon}(z) \quad \forall z \in [-\nu_1, \infty)^3$$

and check that for $\frac{1}{\sqrt{n}} < \nu_1$, $g^{n,\beta,\epsilon}$ is *nearly* a subsolution of (5.2). Using (7.13), we compute for $z \in [0, \infty)^3$:

$$\begin{aligned}
\alpha g^{n,\beta,\epsilon} - \mathcal{L}^{n,y,u} g^{n,\beta,\epsilon} &= \alpha u^{\beta,\epsilon} - \mathcal{L}^{n,y,u} u^{\beta,\epsilon} + \frac{1}{\sqrt{n}} [\alpha f^{n,\beta,\epsilon} - \mathcal{L}^{n,y,u} f^{n,\beta,\epsilon}] \\
&\quad - \theta^{(n)} \left[\nabla u^{\beta,\epsilon} \cdot \xi^{(n)} + \frac{1}{\sqrt{n}} \mathcal{A}^n u^{\beta,\epsilon} \right] \\
&\quad - \frac{\theta^{(n)}}{\sqrt{n}} \left[\nabla f^{n,\beta,\epsilon} \cdot \xi^{(n)} + \frac{1}{\sqrt{n}} \mathcal{A}^n f^{n,\beta,\epsilon} \right] + \sigma_1 \left[-g_1^{n,\beta,\epsilon} + \frac{1}{2\sqrt{n}} g_{11}^{n,\beta,\epsilon} \right] \\
&\quad + \sigma_2 \left[-g_2^{n,\beta,\epsilon} + g_3^{n,\beta,\epsilon} + \frac{1}{2\sqrt{n}} (g_{22}^{n,\beta,\epsilon} - 2g_{23}^{n,\beta,\epsilon} + g_{33}^{n,\beta,\epsilon}) \right] \\
&\quad - \sum_{i=1}^{3} \mathcal{B}_i^{n,y,u} g^{n,\beta,\epsilon} + O\left(\frac{1}{\sqrt{n}}\right).
\end{aligned} \tag{11.24}$$

The term $O(\frac{1}{\sqrt{n}})$ is independent of z (see Remark 7.1). The terms multiplying σ_1 and σ_2 are respective Taylor series approximations of the nonpositive (see (11.15), (11.21)) differences

$$g^{n,\beta,\epsilon}\left(z - \frac{1}{\sqrt{n}} e_1\right) - g^{n,\beta,\epsilon}(z), \qquad g^{n,\beta,\epsilon}\left(z - \frac{1}{\sqrt{n}} e_2 + \frac{1}{\sqrt{n}} e_3\right) - g^{n,\beta,\epsilon}(z),$$

and the error in these approximations is of order $\frac{1}{n}$. Thus

$$\sigma_1 \left[-g_1^{n,\beta,\epsilon}(z) + \frac{1}{2\sqrt{n}} g_{11}^{n,\beta,\epsilon}(z) \right] \le O\left(\frac{1}{\sqrt{n}}\right),$$

and the term involving σ_2 has the same upper bound. A similar argument shows that

$$\mathcal{B}_i^{n,y,u} g^{n,\beta,\epsilon}(z) \le O\left(\frac{1}{\sqrt{n}}\right), \qquad i \in \{1, 2, 3\}.$$

Dropping these five terms and using (11.12), (11.22), (11.23), (11.17), and (11.20), we see that

$$\alpha g^{n,\beta,\epsilon}(z) - \mathcal{L}^{n,y,u} g^{n,\beta,\epsilon}(z) \le c \cdot z + 2L_0\epsilon + O(\beta) + \frac{\alpha K_{\epsilon,\beta}}{\sqrt{n}} (1 + c \cdot z) + O\left(\frac{1}{\sqrt{n}}\right),$$

where as before $|O(\frac{1}{\sqrt{n}})|$ is bounded by a constant, $K = K(\epsilon, \beta)$ times $\frac{1}{\sqrt{n}}$, and $|O(\beta)|$ is bounded by a constant, $K = K(\epsilon)$ times β. Minimization of $\mathcal{L}^{n,y,u} g^{n,\beta,\epsilon}(z)$ over $(y, u) \in \{0, 1\}^2$ results in the inequality

$$\alpha g^{n,\beta,\epsilon}(z) - \mathcal{L}^{n,*} g^{n,\beta,\epsilon}(z) \le \left(1 + \frac{\alpha K_{\epsilon,\beta}}{\sqrt{n}}\right) c \cdot z + 2L_0\epsilon + O(\beta) + O\left(\frac{1}{\sqrt{n}}\right) \quad \forall z \in [0, \infty)^3.$$

This shows that

$$\left[g^{n,\beta,\epsilon} - \frac{1}{\alpha} \left(2L_0\epsilon + O(\beta) + O\left(\frac{1}{\sqrt{n}}\right) \right) \right] \left[1 + \frac{\alpha K_{\epsilon,\beta}}{\sqrt{n}} \right]^{-1}$$

is a subsolution of (5.2), and Proposition 5.1 implies

$$g^{n,\beta,\epsilon}(z) \le \left(1 + \frac{\alpha K_{\epsilon,\beta}}{\sqrt{n}}\right) J_*^{(n)}(z) + \frac{1}{\alpha}\left(2L_0\epsilon + O(\beta) + O\left(\frac{1}{\sqrt{n}}\right)\right) \quad \forall z \in L^{(n)}. \tag{11.25}$$

Step 5. *Passage to the limit.* As a final step, we let $n \to \infty$, then $\beta \downarrow 0$, and then $\epsilon \downarrow 0$ in (11.25). From the bound (11.22), we see that $\lim_{n\to\infty} g^{n,\beta,\epsilon}(z) = u^{\beta,\epsilon}(z)$ and the convergence is uniform on compact sets. Therefore,

$$u^{\beta,\epsilon}(z) \le J_\#(z) + \frac{1}{\alpha}(2L_0\epsilon + O(\beta)) \quad \forall z \in [0,\infty)^3,$$

where $J_\#$ is given by (6.2). As $\beta \downarrow 0$, φ_β approaches the identity function and $u^{\beta,\epsilon}(z) \to v^\epsilon(z)$. Finally, (11.6) implies $\lim_{\epsilon\downarrow 0} v^\epsilon(z) = v(z)$, given by (11.3). This concludes the proof of Proposition 11.1. □

12. The upper bound. In this section we prove the following proposition.

PROPOSITION 12.1. *Assume Case* IIA. *Then*

$$J^\#(z) \le \hat{V}(\omega(z)) \quad \forall z \in [0,\infty)^3, \tag{12.1}$$

where $J^\#$, *defined in* (6.1), *is the upper semicontinuous limit of the queueing network value functions and* ω *is given by* (8.1).

The proof of Proposition 12.1 depends on the construction of a sequence of asymptotically nearly optimal policies for the queueing networks. We now describe this sequence of policies.

Let a and b be functions in $C_b^4([0,\infty))$ which are strictly increasing, concave, and such that $a(0) = b(0) = 0$ and

$$\delta_1 \triangleq \lim_{x\to\infty} a(x), \qquad \delta_2 \triangleq \lim_{x\to\infty} b(x) \tag{12.2}$$

are finite. We also require

$$\delta_2 < \delta_1^2. \tag{12.3}$$

We define a function $\gamma : [0,\infty)^3 \to \mathcal{R}$ by

$$\gamma(z) = a(z_1)a(z_3) - b(z_2) \tag{12.4}$$

and use γ to define a stationary control law (Y, U) for the nth queueing network by setting

$$Y(z) = 1; \tag{12.5a}$$

$$U(z) = \begin{cases} 1 & \text{if} \quad \gamma(z) \ge 0, \\ 0 & \text{if} \quad \gamma(z) < 0 \end{cases} \tag{12.5b}$$

for $z \in L^{(n)}$.

The control used at $z \in L^{(n)}$ thus depends on which "side" of the surface $\gamma = 0$ the scaled queue length vector z is located. Note that if δ_1 and δ_2 are small, the surface $\gamma = 0$ is "close" to the set $\{z \in [0,\infty)^3 : z_1 z_3 = 0\}$, which, by the discussion in §8, is the set toward which we would like to push our system. The meaning of "close" above is better explained in Lemma 12.2(d) and (e) below.

The policy (12.5a), (12.5b) is also chosen in such a way as to simplify the boundary terms in expansion (7.13). Suppose that $z_2 = 0$. Then $\gamma(z) \ge 0$, so that $U(z) = 1$, and

$\mathcal{B}_2^{n,Y(z),U(z)}\varphi(z) = 0$ for any φ. In the same way, if $z_1 = 0$ and $z_2 > 0$, then $\gamma(z) < 0$ and the $\mathcal{B}_1$ term vanishes. We finally note that (12.5b) and (7.8) imply, for $z \in L^{(n)}$,

$$\theta^n(z)\gamma(z) \le 0, \tag{12.6}$$

with strict inequality if $\gamma(z) \neq 0$.

The next lemma collects the relevant properties of γ.

LEMMA 12.2. (a) *For* $z \in [0,\infty)^3$,

$$\nabla\gamma(z)\cdot\xi > 0 \quad \text{and} \quad \nabla\gamma(z)\cdot\xi^{(n)} > 0, \tag{12.7}$$

where $\xi = (\mu_1, -\mu_2, \mu_3)$ *and* $\xi^{(n)} = (\mu_1^{(n)}, -\mu_2^{(n)}, \mu_3^{(n)})$. *Furthermore, there is a* $\delta_3 > 0$ *such that*

$$\nabla\gamma(z)\cdot\xi \ge \delta_3 \quad \text{if} \quad \gamma(z) = 0. \tag{12.8}$$

(b) *For all* $z \in [0,\infty)^3$, *there is a unique* $r \in \mathcal{R}$ *such that* $\gamma(z+r\xi) = 0$, ($\gamma(z+r\xi^{(n)}) = 0$, *respectively*). *We denote this* r *by* $\rho(z)$ ($\rho^{(n)}(z)$, *respectively*).

(c) *The functions* $\rho : [0,\infty)^3 \to \mathcal{R}$ *and* $\rho^{(n)} : [0,\infty)^3 \to \mathcal{R}$ *defined in* (b) *are four times differentiable and linearly growing, and their derivatives of order up to four are bounded on* $[0,\infty)^3$ *uniformly in* n. *We also have*

$$\gamma(z)\rho(z) \le 0; \quad \gamma(z)\rho^{(n)}(z) \le 0 \quad \forall z \in [0,\infty)^3, \tag{12.9}$$

with strict inequality if $\gamma(z) \neq 0$.

(d) *Set*

$$\delta_4 \triangleq \frac{1}{\mu_1 \wedge \mu_2} a^{-1}(\sqrt{\delta_2}), \qquad \delta_4^{(n)} \triangleq \frac{1}{\mu_1^{(n)} \wedge \mu_2^{(n)}} a^{-1}(\sqrt{\delta_2}).$$

Then for all $z \in [0,\infty)^3$ *we have*

$$|\rho(z)| \le \delta_4; \quad |\rho^{(n)}(z)| \le \delta_4^{(n)} \quad \text{if} \quad z_1 z_3 = 0. \tag{12.10}$$

(e) *The constant* $C_0 \triangleq c_1\mu_1 - c_2\mu_2 + c_3\mu_3$, *which is positive under the Case* II *assumption, satisfies*

$$|c\cdot z - \hat{h}(\omega(z))| \le C_0\delta_4 \tag{12.11}$$

for all $z \in [0,\infty)^3$ *such that* $\gamma(z) = 0$.

Proof. We will give the proof for the vector ξ. The proof for $\xi^{(n)}$ is identical.

(a) We have

$$\nabla\gamma(z)\cdot\xi = \mu_1 a'(z_1)a(z_3) + \mu_3[a(z_1)a'(z_3) + b'(z_2)], \tag{12.12}$$

which is obviously positive for all $z \in [0,\infty)^3$. Now suppose that $\gamma(z) = 0$. Since by (12.3)

$$\left(\frac{\delta_2}{\delta_1}\right)^2 < \delta_2,$$

by (12.2) there is an x_0 such that $b(x_0) > (\frac{\delta_2}{\delta_1})^2$. Therefore x_0 satisfies, $\frac{\delta_2}{\sqrt{b(x_0)}} < \delta_1$. If $z_2 \le x_0$, we then have

$$\nabla\gamma(z)\cdot\xi \ge \mu_2 b'(x_0).$$

Next, assume $z_2 \geq x_0$. Then $a(z_1)a(z_3) = b(z_2) \geq b(x_0)$, so either

$$a(z_1) \geq \sqrt{b(x_0)} \tag{12.13}$$

or

$$a(z_3) \geq \sqrt{b(x_0)}. \tag{12.14}$$

We also have that $a(z_1)a(z_3) < \delta_2$, so (12.13) implies

$$z_3 \leq a^{-1}\left(\frac{\delta_2}{\sqrt{b(x_0)}}\right),$$

and we get

$$\nabla\gamma(z)\cdot\xi \geq \mu_2 a(z_1)a'(z_3) \geq \mu_2\sqrt{b(x_0)}a'\left(a^{-1}\left(\frac{\delta_2}{\sqrt{b(x_0)}}\right)\right).$$

By repeating the same argument following from (12.14), we conclude that we can take

$$\delta_3 = \min\left\{\mu_2 b'(x_0), (\mu_1 \wedge \mu_2)\sqrt{b(x_0)}a'\left(a^{-1}\left(\frac{\delta_2}{\sqrt{b(x_0)}}\right)\right)\right\}.$$

(b) Suppose that $z \in [0,\infty)^3$. Denote by I the internal $[-(\frac{z_1}{\mu_1} \wedge \frac{z_3}{\mu_2}), \frac{z_2}{\mu_2}]$. Note that $z + r\xi \in [0,+\infty)^3$ if and only if $r \in I$. Let $h(r) = \gamma(z + r\xi)$ for $r \in I$. From (12.7), h is strictly increasing in I. Also, $h(-(\frac{z_1}{\mu_1} \wedge \frac{z_3}{\mu_2})) \leq 0$ and $h(\frac{z_2}{\mu_2}) \geq 0$. The conclusion follows.

(c) Since $\gamma \in C^4([0,\infty)^3)$, we have $\rho \in C^4([0,\infty)^3)$, by the implicit function theorem. Also,

$$\rho_i(z) = \frac{-\gamma_i(z + \rho(z)\xi)}{\nabla\gamma(z + \rho(z)\xi)\cdot\xi}, \tag{12.15}$$

which is bounded by (12.8) and boundedness of γ_i. Boundedness of higher-order derivatives is obtained by repeatedly differentiating (12.15). Relation (12.9) follows from (12.7) and the definition of ρ.

(d) Assume that $z_3 = 0$. Then

$$\begin{aligned}\delta_2 &\geq b(z_2 - \mu_2\rho(z)) = a(z_1 + \mu_1\rho(z))a(\mu_2\rho(z))\\ &\geq a((\mu_1 \wedge \mu_2)\rho(z))^2,\end{aligned}$$

so

$$\rho(z) \leq \frac{1}{\mu_1 \wedge \mu_2}a^{-1}(\sqrt{\delta_2}),$$

where the term on the right is well defined by (12.2) and (12.3). We have a symmetrical argument if $z_1 = 0$, and thus we can take

$$\delta_4 = \frac{1}{\mu_1 \wedge \mu_2}a^{-1}(\sqrt{\delta_2}).$$

We note that if we hold the function $a(\cdot)$ fixed and vary $b(\cdot)$ so that $\delta_2 \downarrow 0$, then $\delta_4 \downarrow 0$ also.

(e) Let $z \in [0, \infty)^3$ be such that $\gamma(z) = 0$. There is a unique $\bar{z}$ such that $\bar{z}_1 \bar{z}_3 = 0$ and $z = \bar{z} + \rho(\bar{z})\xi$. Furthermore, by (8.5), $\hat{h}(\omega(z))\hat{h}(\omega(\bar{z})) = c \cdot \bar{z}$. It follows that

$$\begin{aligned} |c \cdot z - \hat{h}(\omega(z)) &= |c \cdot (\bar{z} + \rho(\bar{z})\xi) - \hat{h}(\omega(\bar{z}))| \\ &= (c \cdot \xi)\rho(\bar{z}) \leq (c \cdot \xi)\delta_4. \qquad \square \end{aligned}$$

Let $J^{(n)} = J^{(n)}_{Y,U}$ be the cost associated with policy (12.5a), (12.5b). The next proposition shows that the sequence of costs $J^{(n)}$ grows linearly in z, uniformly in n. In particular, it shows that the sequence $\{J^{(n)}(z)\}_{n=1}^{\infty}$ is bounded for each z.

PROPOSITION 12.3. *There are constants K_1 and K_2 independent of n such that*

$$J^{(n)}(z) \leq K_1|z| + K_2 \quad \forall z \in L^{(n)}, \forall n. \tag{12.16}$$

Proof. We will construct a function $\varphi : [0, \infty)^3 \to \mathcal{R}$ such that

$$k_1|z| - k_2 \leq \varphi(z) \leq k_3|z| + k_4 \quad \forall z \in [0, +\infty)^3 \tag{12.17}$$

and

$$\mathcal{L}^{n,Y(z),U(z)}\varphi(z) \leq k_5 \quad \forall z \in L^{(n)}, \tag{12.18}$$

where $\mathcal{L}^{n,y,u}$ is given by (3.1) and $k_1, \ldots, k_5$ are positive constants independent of n. Once φ has been constructed, we proceed as follows. Let $L_1 = \frac{3}{\alpha k_1} \max\{c_1, c_2, c_3\}$. Then $\alpha L_1 \varphi \geq h - \alpha L_1 k_2$. Put $\Psi = L_1\varphi + \frac{L_1 k_5}{\alpha} + L_1 k_2$. Then

$$\alpha\Psi - \mathcal{L}^{n,Y,U}\Psi - h = \alpha L_1\varphi - (h - \alpha L_1 k_2) + L_1(k_5 - \mathcal{L}^{n,Y,U}\varphi) \geq 0,$$

and, by Proposition 4.1,

$$J^{(n)}(z) \leq \Psi(z) \leq L_1 k_3|z| + L_1 k_4 + \frac{L_1 k_5}{\alpha} + L_1 k_2.$$

It remains to construct φ. We fix a real number $\Delta > \sup_n\{\delta_4^{(n)}\mu_2^{(n)}\}$, and define

$$\zeta \stackrel{\Delta}{=} \Delta - \sup_n\{\delta_4^{(n)}\mu_2^{(n)}\} \tag{12.19}$$

and

$$K = \inf_m \left\{\frac{1}{\mu_2^{(m)}}\right\} a^{-1}\left(\frac{b(\zeta)}{\delta_1}\right) > 0. \tag{12.20}$$

Inequality (12.3) guarantees that K can be defined this way. In order to construct φ, we will need to define a number of auxiliary functions. We let $G \in C^4(\mathcal{R})$ be an even convex function such that

$$G(x) = |x| \quad \text{if} \quad |x| \geq 1. \tag{12.21}$$

We note that $G'(0) = 0$ and

$$xG'(x) \geq 0 \quad \forall x \in \mathcal{R}. \tag{12.22}$$

Let $H \in C^4(\mathcal{R})$ be a nondecreasing convex function such that

$$H'(x) = \begin{cases} 0 & \text{if} \quad x \le \Delta, \\ 1 & \text{if} \quad x \ge 2\Delta, \end{cases} \tag{12.23}$$

and let $\lambda \in C_b^4(\mathcal{R})$ be a bounded nonincreasing function such that

$$\lambda(x) = 0 \quad \text{if} \quad x \le 0, \quad \lambda(x) \le 0 \quad \text{if} \quad x > 0, \quad \lambda(K) \le -\sup_n\{\mu_2^{(n)}\}. \tag{12.24}$$

We then define

$$f(z) = G(\mu_1^{(n)} z_3 - \mu_2^{(n)} z_1) + H(z_2), \quad z \in [0,\infty)^3, \tag{12.25}$$

and, finally, we set

$$\varphi(z) = f(z + \rho^{(n)}(z)\xi^{(n)}) - \int_0^{\rho^{(n)}(z)} \lambda(s)ds, \qquad z \in [0,\infty)^3. \tag{12.26}$$

We first verify (12.18). Because the derivatives of φ up to order four are bounded, we may use the expansion (7.13) to write for any $z \in L^{(n)}$

$$\begin{aligned} \mathcal{L}^{n,Y(z),U(z)}\varphi(z) = \mathcal{L}\varphi(z) + \theta^{(n)}(z)\left[\nabla\varphi(z)\cdot\xi^{(n)} + \frac{1}{\sqrt{n}}\mathcal{A}^n\varphi(z)\right] \\ + \mathcal{B}_1^{n,Y(z),U(z)}\varphi(z) + \mathcal{B}_3^{n,Y(z),U(z)}\varphi(z) + O\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \tag{12.27}$$

Note that, as remarked earlier, there is no $\mathcal{B}_2$ term above. Also, the $\mathcal{B}_1$ term above is nonzero only if $z_1 = z_2 = 0$.

The terms $\mathcal{L}\varphi(z)$, $\frac{g^{(n)}(z)}{\sqrt{n}}\mathcal{A}^n(z)$, and $O(\frac{1}{\sqrt{n}})$ in (12.27) are bounded. Furthermore, $\nabla\varphi(z)\cdot\xi^{(n)} = \lambda(\rho^{(n)}(z))$, so from (12.24) we get $\rho^{(n)}(z)\nabla\varphi(z)\cdot\xi^{(n)} \le 0$, and from (12.6) and Lemma 12.2(c), it follows that

$$\theta^{(n)}(z)\nabla\varphi(z)\cdot\xi^{(n)} \le 0. \tag{12.28}$$

From (7.10) we have

$$\mathcal{B}_1^{n,Y(z),U(z)}\varphi(z) = \sqrt{n}\mu_1^{(n)}\varphi_1(z)1_{\{z_1=z_2=0\}} + O(1),$$

where $O(1)$ is uniformly bounded in n. If $z_1 = z_2 = 0$, then $\gamma(z) = \rho^{(n)}(z) = 0$. Also a simple computation from (12.25) and (12.26) shows that, if $z_1 = z_2 = 0$, then $\varphi_1(z) = -\mu_2^{(n)}G'(\mu_1^{(n)}z_3) \le 0$, which shows that $\mathcal{B}_1^{n,Y(z),U(z)}\varphi(z)$ is uniformly bounded from above in n.

By an analogous reasoning, to show that $\mathcal{B}_3^{n,Y^{(n)}(z),U^{(n)}(z)}\varphi(z)$ is bounded from above, it is enough to show that $\varphi_3(z) \le 0$ if $z_3 = 0$. A direct computation gives, for $z_3 = 0$,

$$\varphi_3(z) = \mu_1^{(n)}G'(-\mu_2^{(n)}z_1) - \rho_3^{(n)}(z)[\mu_2^{(n)}H'(z_2 - \mu_2^{(n)}\rho^{(n)}(z)) + \lambda(\rho^{(n)}(z))].$$

Note first that, from (12.15), $\rho_3^{(n)}(z) \le 0$. Also, from (12.22), $G'(-\mu_2^{(n)}z_1) \le 0$. To show that $\varphi_3(z)$ is nonpositive, it is then enough to show that

$$A = \mu_2^{(n)}H'(z_2 - \mu_2\rho^{(n)}(z)) + \lambda(\rho^{(n)}(z))$$

is nonpositive. Suppose first that $z_2 \le \Delta$. Then (12.23) implies $A = \lambda(\rho^{(n)}(z)) \le 0$. Next assume that $z_2 \ge \Delta$. Then by (12.10) and (12.19) we have $z_2 - \mu_2^{(n)}\rho^{(n)}(z) \ge \zeta$. It follows that

$$b(\zeta) \le b(z_2 - \mu_2^{(n)}\rho^{(n)}(z)) = a(z_1 + \mu_1^{(n)}\rho^{(n)}(z))a(\mu_2^{(n)}\rho^{(n)}(z)) \le \delta_1 a(\mu_2^{(n)}\rho^{(n)}(z))$$

so that

$$\rho^{(n)}(z) \ge \frac{1}{\mu_2^{(n)}} a^{-1}\left(\frac{b(\zeta)}{\delta_1}\right) \ge K$$

by (12.20). Using (12.23) and (12.24), we now get

$$A \le \mu_2^{(n)} + \lambda(\rho^{(n)}(z)) \le \mu_2^{(n)} - \sup_m |\mu_2^{(m)}| \le 0$$

as desired. This concludes the proof of (12.18). We continue with the proof of (12.17). First note that it is clear that φ grows at most linearly, since its derivatives are bounded. To exhibit a linearly growing lower bound for φ, we start by showing that φ grows linearly on the surface $\gamma(z) = 0$. It is obvious that there is a constant m_1 such that

$$H(z_2) \ge z_2 - m_1. \tag{12.29}$$

Now suppose that $\gamma(z) = 0$ and $\mu_1^{(n)}z_3 - \mu_2^{(n)}z_1 \ge 1$. This implies

$$\delta_2 > b(z_2) = a(z_1)a(z_3) \ge a(z_1)a\left(\frac{\mu_2^{(n)}}{\mu_1^{(n)}}z_1\right) \ge a\left(\left(1 \wedge \frac{\mu_2^{(n)}}{\mu_1^{(n)}}\right)z_1\right)^2$$

so that

$$z_1 \le k_1 \triangleq \sup_n \left\{1 \vee \frac{\mu_1^{(n)}}{\mu_2^{(n)}}\right\} a^{-1}(\sqrt{\delta_2})$$

and it follows that, in this case,

$$G(\mu_1^{(n)}z_3 - \mu_2^{(n)}z_1) = \mu_1^{(n)}z_3 - \mu_2^{(n)}z_1 \ge \mu_1^{(n)}z_3 - \mu_2^{(n)}k_1 \ge \mu_1^{(n)}z_3 + z_1 - (1 + \mu_2^{(n)})k_1.$$

Thus, there are constants m_2, m_3, and m_4 such that for all n

$$G(\mu_1^{(n)}z_3 - \mu_2^{(n)}z_1) \ge m_2z_1 + m_3z_3 - m_4 \quad \text{if} \quad \gamma(z) = 0, \mu_1^{(n)}z_3 - \mu_2^{(n)}z_1 \ge 1. \tag{12.30}$$

By analogous reasoning, there are m_5, m_6, and m_7 such that for all n

$$G(\mu_1^{(n)}z_3 - \mu_2^{(n)}z_1) \ge m_5z_1 + m_6z_3 - m_7 \quad \text{if} \quad \gamma(z) = 0, \mu_1^{(n)}z_3 - \mu_2^{(n)}z_1 \le -1. \tag{12.31}$$

Now suppose that $\gamma(z) = 0$ and $|\mu_1^{(n)}z_3 - \mu_2^{(n)}z_1| \le 1$. Then (z_1, z_3) is contained in the set

$$A_n \triangleq \{(x, y) : x \ge 0, y \ge 0, |\mu_1^{(n)}y - \mu_2^{(n)}x| \le 1, \quad a(x) \wedge a(y) \le \sqrt{\delta_2}\}.$$

The set $A = \cup_{n\ge1}A_n$ is clearly bounded. Let k_2 be such that $A \subseteq [0, k_2]^2$. We then have

$$G(\mu_1^{(n)}z_3 - \mu_2^{(n)}z_1) \ge z_1 + z_3 - 2k_2 \quad \text{if} \quad \gamma(z) = 0, |\mu_1^{(n)}z_3 - \mu_2^{(n)}z_1| \le 1. \tag{12.32}$$

Putting (12.29)–(12.32) together, we see that there are constants K_0, K_1 such that

$$f(z) \geq K_0|z| - K_1 \quad \text{if} \quad z \in [0,\infty)^3, \gamma(z) = 0,$$

so

$$f(z + \rho^{(n)}(z)\xi^{(n)}) \geq K_0|z + \rho^{(n)}(z)\xi^{(n)}| - K_1.$$

To prove the lower bound (12.17), it is then enough to show that there is a K_2 such that

$$|z + \rho^{(n)}(z)\xi^{(n)}| \geq K_2|z| \quad \forall z \in [0,\infty)^3.$$

Since all norms on $\mathcal{R}^3$ are equivalent, we may rewrite this as

$$\sum_{i=1}^{3}(z_i + \rho_i^{(n)}(z)\xi_i^{(n)}) \geq K_2 \sum_{i=1}^{3} z_i \quad \forall z \in [0,\infty)^3. \tag{12.33}$$

For $z \in [0,\infty)^3$,

$$\begin{aligned}\sum_{i=1}^{3}(z_i + \rho_i^{(n)}(z)\xi_i^{(n)}) &\geq A \\ &\triangleq \min\left\{\sum_{i=1}^{3} z_i + r\xi_i^{(n)}; -\left(\frac{z_1}{\mu_1^{(n)}} \wedge \frac{z_3}{\mu_2^{(n)}}\right) \leq r \leq \frac{z_2}{\mu_2^{(n)}}\right\} \\ &= z_1 + z_2 + z_3 - \mu_1^{(n)}\left(\frac{z_1}{\mu_1^{(n)}} \wedge \frac{z_3}{\mu_2^{(n)}}\right).\end{aligned}$$

Suppose that

$$\frac{z_1}{\mu_1^{(n)}} \leq 2\frac{z_3}{\mu_2^{(n)}}.$$

Then

$$A \geq z_2 + z_3 \geq \frac{\mu_2^{(n)}}{4\mu_1^{(n)}} z_1 + z_2 + \frac{1}{2} z_3.$$

If, however,

$$\frac{z_1}{\mu_1^{(n)}} > 2\frac{z_3}{\mu_2^{(n)}},$$

then

$$A \geq \frac{1}{2} z_1 + z_2 + z_3.$$

We can therefore take $K_2 = \frac{1}{2} \wedge \inf_n \frac{\mu_1^{(n)}}{4\mu_2^{(n)}}$ in (12.33). □

In view of Proposition 12.3 we can define

$$J^{\infty}(z) = \lim_{\delta \downarrow 0} \overline{\lim_{n\to\infty}} \max_{\substack{\zeta \in L^{(n)} \\ |\zeta - z| \leq \delta}} J^{(n)}(\zeta).$$

We obviously have $J^{\#}(z) \leq J^{\infty}(z)$, so to prove Proposition 12.1 it is enough to show that

$$J^{\infty}(z) \leq \hat{V}(\omega(z)) \quad \forall z \in [0,\infty)^3. \tag{12.34}$$

The remainder of this section is devoted to the proof of (12.34). First observe that because of the special structure of the control process U, the scaled queue length process $Z^{(n)}$ moves very rapidly towards the surface $\gamma = 0$. Therefore in the limit we expect the value of $J^\infty(z)$ to be equal to $J^\infty(z + \rho(z)\xi)$. Indeed we have the following result.

LEMMA 12.4. *Let $z \in [0, \infty)^3$ be such that $z_3 > 0$. Then*

$$J^\infty(z) \le J^\infty(z + \rho(z)\xi). \tag{12.35}$$

Since the proof of this lemma is rather technical, we will first give a brief outline of the proof. Let z^0 be a point in $(0, \infty)^3$, and suppose that there is a smooth function φ such that $J^\infty - \varphi$ has a strict maximum at z_0. Then by an elementary argument, we can construct a sequence z^n converging to z^0 such that z^n maximizes the difference $J^{(n)} - \varphi$. Setting $u = U(z)$, we have from the definition (3.1) of $\mathcal{L}^{n,1,u}$ and from Proposition 4.1 that

$$\mathcal{L}^{n,1,u}\varphi(z^n) \ge \mathcal{L}^{n,1,u}J^{(n)}(z^n) = \alpha J^{(n)}(z^n) - h(z^n). \tag{12.36}$$

For $z \in (0, \infty)^3$, the only unbounded term in the expansion (7.13) of $\mathcal{L}^{n,1,u}\varphi(z)$ is

$$\sqrt{n}\left[\frac{\lambda_1^{(n)}}{\mu_1^{(n)}} - U(z)\right]\nabla\varphi(z) - \xi^{(n)}.$$

By the definition (12.5b) of U, the term in the brackets has the same sign as $-\gamma$. By letting n go to infinity and using the fact that the right-hand side of (12.36) remains bounded, we conclude that

$$\gamma(z^0)\nabla\varphi(z^0) \cdot \xi \le 0.$$

Now if J^∞ is differentiable, then $\nabla J^\infty(z^0) = \nabla\varphi(z^0)$, and we would have

$$\gamma(z^0)\nabla J^\infty(z^0) \cdot \xi \le 0 \quad \forall z^0 \in (0, \infty)^3.$$

This inequality tells us that $\frac{d}{dt}J^\infty(z + t\xi)$ has the sign as $-\gamma(z + t\xi)$, so $J^\infty(z + t\xi)$ is maximized at $t = \rho(z)$.

Proof. Fix $\bar{z} \in [0, \infty)^3$ such that $\bar{z}_3 > 0$. If $\bar{z}_1 = \bar{z}_2 = 0$, then $\rho(\bar{z}) = 0$ and (12.35) holds at $z = \bar{z}$. We thus assume, without loss of generality, that

$$\bar{z}_1 \vee \bar{z}_2 > 0. \tag{12.37}$$

The idea of the proof is to construct, for each sufficiently small positive ϵ, a point z^ϵ satisfying $\gamma(z^\epsilon) = 0$ such that

$$\lim_{\epsilon \downarrow 0} z^\epsilon = \bar{z} + \rho(\bar{z})\xi, \qquad J^\infty(\bar{z}) \le \overline{\lim_{\epsilon \downarrow 0}} J^\infty(z^\epsilon). \tag{12.38}$$

The upper semicontinuity of J^∞, which follows from its definition, will then yield the desired result, (12.35).

Step 1. *Choice of constants.* From Proposition 12.3 it follows that there are positive constants k_1 and k_2 such that

$$J^{(n)}(z) \le k_1|\omega(z)| + k_2 \quad \forall z \in L^{(n)}, \forall n \ge 1. \tag{12.39}$$

We put $k_3 \triangleq 2(1 + k_1 + k_2 + |\omega(\bar{z})|)$, $k_4 \triangleq 2(k_1k_3 + k_2 + 1)$. For $\epsilon > 0$, set

$$\beta(\epsilon) \triangleq \min\{\rho(z);\, z \in [0, \infty)^3, |\omega(z) - \omega(\bar{z})|^2 \le k_4\epsilon, z_3 = 0\}, \tag{12.40}$$

where we use the convention $\min \phi = \infty$. Since $\bar{z}_3 > 0$, there is an $\epsilon_0 > 0$ such that

$$\rho(\bar{z}) < \beta(\epsilon) \quad \forall \epsilon \in (0, \epsilon_0). \tag{12.41}$$

Henceforth, we consider only $\epsilon \in (0, 1 \wedge \epsilon_0)$.

Step 2. *Definition of* φ_ϵ. Let $\psi_\epsilon \in C^4(\mathcal{R})$ be such that

$$\psi_\epsilon(0) = 0 \quad \text{and} \quad \psi_\epsilon(x) \geq 0 \quad \forall x \in \mathcal{R},$$

$$x\psi_\epsilon'(x) > 0 \quad \forall x \in \mathcal{R}, x \neq 0, \tag{12.42}$$

$$\psi_\epsilon(\beta(\epsilon)) = k_1 k_3 + k_2 + 1, \tag{12.43}$$

$$\psi_\epsilon(\rho(\bar{z})) \leq \epsilon. \tag{12.44}$$

Inequality (12.41) guarantees the existence of such a ψ_ϵ. We put

$$\varphi_\epsilon(z) \triangleq \psi_\epsilon(\rho(z)) + \frac{1}{2\epsilon}|\omega(z) - \omega(\bar{z})|^2, \qquad z \in [0, \infty)^3. \tag{12.45}$$

Step 3. *Definition of* z^ϵ. Because φ_ϵ grows quadratically and $J^{(n)}$ grows linearly (Proposition 12.3), the difference $J^{(n)} - \varphi_\epsilon$ attains its maximum at some point $z^n \in L^{(n)}$. Indeed, the linear growth of $J^{(n)}$ is uniform in n, so the sequence $\{z^n\}$ is bounded. We can thus select a convergent subsequence $\{z^{n_k}\}$ such that

$$\begin{aligned} \lim_{k\to\infty} [J^{(n_k)}(z^{n_k}) - \varphi_\epsilon(z^{n_k})] &= \overline{\lim_{n\to\infty}} [J^{(n)}(z^n) - \varphi_\epsilon(z^n)] \\ &\geq J^\infty(z) - \varphi_\epsilon(z) \quad \forall z \in [0, \infty)^3. \end{aligned} \tag{12.46}$$

Let z^ϵ denote the limit of $\{z^n\}$. From (12.46) we see that

$$J^\infty(z^\epsilon) - \varphi_\epsilon(z^\epsilon) \geq J^\infty(z) - \varphi_\epsilon(z) \quad \forall z \in [0, \infty)^3. \tag{12.47}$$

Step 4. *Bounds on* $\{z^n\}$ *and* z^ϵ. Because z^n maximizes $J^{(n)} - \varphi_\epsilon$, we have

$$J^{(n)}(z^n) - \varphi_\epsilon(z^n) \geq J^{(n)}(\bar{z}) - \varphi_\epsilon(\bar{z}) \geq -\epsilon > -1, \tag{12.48}$$

where we have used (12.44). It follows then from (12.39) that

$$\frac{1}{2\epsilon}|\omega(z^n) - \omega(\bar{z})|^2 \leq \varphi_\epsilon(z^n) \leq J^{(n)}(z^n) + 1 \leq k_1|\omega(z^n)| + k_2 + 1. \tag{12.49}$$

This inequality implies

$$\frac{1}{2\epsilon}|\omega(z^n)|^2 \leq k_1|\omega(z^n)| + k_2 + 1 + \frac{1}{\epsilon}|\omega(z^n)|\,|\omega(\bar{z})|,$$

so if $|\omega(z^n)| \geq \epsilon$, we have

$$\frac{1}{2\epsilon}|\omega(z^n)| \leq k_1 + \frac{1}{\epsilon}[k_2 + 1 + |\omega(\bar{z})|] \leq \frac{k_3}{2\epsilon}.$$

Regardless of the value of $|\omega(z^n)|$, we have

$$|\omega(z^n)| \leq \min\{\epsilon, k_3\} \leq k_3. \tag{12.50}$$

Substitution of (12.50) into (12.49) yields

$$|\omega(z^n) - \omega(\bar{z})|^2 \leq k_4\epsilon. \tag{12.51}$$

Taking the limit along the sequence $\{n_k\}$, we obtain from (12.48), (12.50), and (12.51)

$$J^\infty(z^\epsilon) - \varphi_\epsilon(z^\epsilon) \geq -\epsilon > -1, \tag{12.52}$$

$$|\omega(z^\epsilon)| \leq k_3, \tag{12.53}$$

$$|\omega(z^\epsilon) - \omega(\bar{z})|^2 \leq k_4\epsilon. \tag{12.54}$$

Finally, we show that $z_3^\epsilon > 0$. If z_3^ϵ were zero, then (12.40) and (12.54) would imply $\beta(\epsilon) \leq \rho(z^\epsilon)$. But (12.52) implies

$$\psi_\epsilon(\rho(z^\epsilon)) \leq \varphi_\epsilon(z^\epsilon) < J^\infty(z^\epsilon) + 1 \leq k_1|\omega(z^\epsilon)| + k_2 + 1 \leq k_1k_3 + k_2 + 1,$$

where we have used (12.53) and the limit form of (12.39). From (12.42) and (12.43), we conclude that $\rho(z^\epsilon) < \beta(\epsilon)$ and hence $z_3^\epsilon > 0$.

For sufficiently small ϵ, (12.37) and (12.54) imply $z_1^\epsilon \vee z_2^\epsilon > 0$. Thus, we may choose $\epsilon_1 \in (0, 1 \wedge \epsilon_0)$ such that, for each $\epsilon \in (0, \epsilon_1)$, there is a positive integer k^ϵ satisfying

$$z_3^{n_k} > 0, \quad z_1^{n_k} \vee z_2^{n_k} > 0 \quad \forall k \geq k^\epsilon. \tag{12.55}$$

Step 5. $\gamma(z^\epsilon) = 0$. Given $\epsilon \in (0, \epsilon_1)$, let k^ϵ be as in (12.55), and let $k \geq k^\epsilon$ be given. Because z^{n_k} maximizes $J^{(n_k)} - \varphi_\epsilon$, we have $\mathcal{L}^{n_k,y,u}(J^{(n_k)} - \varphi_\epsilon)(z^{n_k}) \leq 0$ (see (3.1)). Setting $y = 1$ and $u = U^{n_k}(z^{n_k})$, we have from Proposition 4.1 that

$$\mathcal{L}^{n_k,1,u}\varphi_\epsilon(z^{n_k}) \geq \mathcal{L}^{n_k,1,u}J^{(n_k)}(z^{n_k}) = \alpha J^{(n_k)}(z^{n_k}) - h(z^{n_k}),$$

which is bounded below uniformly in k because $\{z^{n_k}\}$ is bounded. But with $\theta^{(n_k)} \triangleq \sqrt{n_k}[\frac{\lambda_1^{(n_k)}}{\mu_1^{(n_k)}} - U^{n_k}(z^{n_k})]$, we have from (7.13) that

$$\mathcal{L}^{n_k,1,u}\varphi_\epsilon(z^{n_k}) = \mathcal{L}\varphi(z^{n_k}) + \theta^{(n_k)}\left[\nabla\varphi(z^{n_k})\cdot\xi^{(n_k)} + \frac{1}{\sqrt{n_k}}\mathcal{A}^n\varphi(z^{n_k})\right] + O\left(\frac{1}{\sqrt{n_k}}\right). \tag{12.56}$$

The terms σ_1 and σ_2 in (7.13) vanish because $y = 1$. The terms $\mathcal{B}_i^{n_k,1,u}$, $i = 1, 2, 3$, vanish because $z_1^{n_k} \wedge z_2^{n_k} > 0$, $z_3^{n_k} > 0$, and

1. $z_1^{n_k} = 0 \Rightarrow z_2^{n_k} > 0, \gamma(z_2^{n_k}) < 0, u = 0$;
2. $z_2^{n_k} = 0 \Rightarrow \gamma(z^{n_k}) \geq 0, u = 1$.

All the terms on the right-hand side of (12.56) are bounded uniformly in k, except possibly $\theta^{(n_k)}\nabla\varphi(z^{n_k})\cdot\xi^{(n_k)}$. Therefore, this term must be bounded from below, uniformly in k, i.e.,

$$\inf_{k\geq k^\epsilon}\left\{\sqrt{n_k}\left[\frac{\lambda_1^{(n_k)}}{\mu_1^{(n_k)}} - U^{n_k}(z^{n_k})\right]\nabla\varphi(z^{n_k})\cdot\xi^{(n_k)}\right\} > -\infty. \tag{12.57}$$

We use (12.57) to prove that $\gamma(z^\epsilon) = 0$. If $\gamma(z^\epsilon) > 0$, then $\gamma(z^{n_k}) > 0$ and $U^{n_k}(z^{n_k}) = 1$ for sufficiently large k. Since

$$\lim_{k\to\infty}\left[\frac{\lambda_1^{(n_k)}}{\mu_1^{(n_k)}} - 1\right] = \frac{\lambda_1}{\mu_1} - 1 = -\frac{\lambda_2}{\mu_2} < 0,$$

(12.57) implies that

$$\nabla\varphi(z^\epsilon)\cdot\xi = \lim_{k\to\infty}\nabla\varphi(z^{n_k})\cdot\xi^{(n_k)} \leq 0. \tag{12.58}$$

But

$$\nabla\varphi(z^\epsilon)\cdot\xi = \psi_\epsilon'(\rho(z^\epsilon))\nabla\rho(z^\epsilon)\cdot\xi = -\psi_\epsilon'(\rho(z^\epsilon)), \tag{12.59}$$

where we have used (12.15). From (12.58), (12.59), and (12.42), we conclude that $\rho(z^\epsilon) \geq 0$. Lemma 12.2(c) implies that $\gamma(z^\epsilon) \leq 0$, which contradicts our initial assumption $\gamma(z^\epsilon) > 0$. A similar argument rules out the possibility $\gamma(z^\epsilon) < 0$, and we are left with the conclusion $\gamma(z^\epsilon) = 0$.

Step 6. *Conclusion.* Because $\lim_{\epsilon\downarrow 0}\omega(z^\epsilon) = \omega(\bar z)$ (see (12.54)) and $\gamma(z^\epsilon) = 0$, and we have $\lim_{\epsilon\downarrow 0} z^\epsilon = \bar z + \rho(\bar z)\xi$. We may now take the limit as $\epsilon \downarrow 0$ in (12.47), using (12.44), to conclude

$$J^\infty(\bar z + \rho(\bar z)\xi) \geq \overline{\lim_{\epsilon\downarrow 0}} J^\infty(z^\epsilon) \geq \lim_{\epsilon\downarrow 0}[J^\infty(\bar z) - \varphi_\epsilon(\bar z)] = J^\infty(\bar z).$$

This completes the proof of (12.38). □

The next few lemmas allow us to remove the condition $z_3 > 0$ in Lemma 12.4.

LEMMA 12.5. *Suppose that $\varphi \in C^3([0,\infty)^3)$ is such that $J^\infty - \varphi$ has a local maximum at $\bar z$. Assume that $\bar z_1 \vee \bar z_2 > 0$ and either $\bar z_3 > 0$ or $\bar z_3 = 0$ and $\varphi_3(\bar z) < 0$. Then $\gamma(\bar z)\nabla\varphi(\bar z)\cdot\xi \leq 0$, where $\xi = (\mu_1, -\mu_2, \mu_3)$.*

Proof. Assume for the moment that $J^\infty - \varphi$ has a *strict* local maximum at $\bar z$. Let $\tilde K \subset \mathcal{R}^3$ be a compact set whose interior contains $\bar z$ and is such that $\bar z$ strictly maximizes $J^\infty - \varphi$ over $[0,\infty)^3 \cap \operatorname{int}(\tilde K)$. Define $K = [0,\infty)^3 \cap \tilde K$, and let z^n maximize $J^{(n)} - \varphi$ over the finite set $K \cap L^{(n)}$. Choose a convergent subsequence $\{z^{n_k}\}$ with limit z^∞ such that

$$\lim_{k\to\infty}[J^{(n_k)}(z^{n_k}) - \varphi(z^{n_k})] = \overline{\lim_{n\to\infty}}[J^{(n)}(z^n) - \varphi(z^n)].$$

Then

$$\begin{aligned} J^\infty(z^\infty) - \varphi(z^\infty) &\geq \overline{\lim_{n\to\infty}}[J^{(n)}(z^n) - \varphi(z^n)] \\ &\geq J^\infty(z) - \varphi(z) \quad \forall z \in [0,\infty)^3 \cap \operatorname{int}(\tilde K). \end{aligned}$$

It follows that $z^\infty = \bar z$.

There exists a positive integer k_0 such that $z_1^{n_k} \vee z_2^{n_k} > 0$ for all $k \geq k_0$, and either $z_3^{n_k} > 0$ for all $k \geq k_0$ or else $z_3^{n_k} = 0$ and $\varphi_3(z^{n_k}) < 0$ for all $k \geq k_0$. Consequently, $\mathcal{B}_3^{n,1,U^{n_k}(z^{n_k})}\varphi(z^{n_k})$ given by (7.12) is either zero or else is bounded from above, uniformly in $k \geq k_0$. This observation allows us to use the argument in Step 5 of the proof of Lemma 12.4 to conclude that $\gamma(\bar z)\nabla\varphi(\bar z)\cdot\xi \leq 0$.

If the maximum attained by $J^\infty - \varphi$ at $\bar z$ is not strict, we introduce the function $\varphi_\delta(z) \triangleq \varphi(z) + \delta e^{-|z-\bar z|^2}$. For all $\delta > 0$, the function $J^\infty - \varphi_\delta$ attains a strict maximum at $\bar z$. Furthermore, $\nabla\varphi_\delta(\bar z) = \nabla\varphi(\bar z)$. The preceding argument shows that $\gamma(z)\nabla\varphi(\bar z)\cdot\xi = \gamma(z)\cdot\nabla\varphi_\delta(\bar z)\cdot\xi \leq 0$. □

LEMMA 12.6. *Suppose that $\varphi \in C^3([0,\infty)^3)$ is such that $J^\infty - \varphi$ has a local maximum at $\bar z$. Assume that $\bar z_3 = 0$. Then $J^\infty(\bar z) \leq J^\infty(\bar z + \rho(\bar z)\xi)$.*

Proof. We assume without loss of generality that $\bar z_2 > 0$, since if $\bar z_2 = 0$, then $\rho(\bar z) = 0$ and the result follows. The assumption $\bar z_2 > 0$ is equivalent to $\gamma(\bar z) < 0$.

Suppose the desired result is false, i.e., there is an $\epsilon > 0$ such that $J^\infty(\bar z) \geq J^\infty(\bar z + \rho(\bar z)\xi) + \epsilon$. Continuity of ρ and upper semicontinuity of J^∞ imply the existence of $\delta > 0$ such that

$$\begin{aligned} |z - \bar z| < \delta \Rightarrow J^\infty(z + \rho(z)\xi) &\leq J^\infty(\bar z + \rho(\bar z)\xi) + \frac{\epsilon}{2} \\ &\leq J^\infty(\bar z) - \frac{\epsilon}{2}. \end{aligned}$$

From Lemma 12.4, we have then that

$$|z - \bar{z}| < \delta, z_3 > 0 \Rightarrow J^\infty(z) \leq J^\infty(\bar{z}) - \frac{\epsilon}{2};$$

i.e., J^∞ jumps up at $\bar{z}$ as the boundary $z_3 = 0$ is approached. Consequently, with $K > 0$ and $\tilde{\varphi}(z) \triangleq \varphi(z) - Kz_3$, the function $J^\infty - \tilde{\varphi}$ has a local maximum at $\bar{z}$. For K large enough, we also have $\bar{\varphi}_3(\bar{z}) < 0$. Applying Lemma 12.5 to $\tilde{\varphi}$ and recalling our assumption $\gamma(\bar{z}) < 0$, we see that

$$0 \leq \nabla\tilde{\varphi}(\bar{z}) \cdot \xi = \frac{1}{\mu_1}\varphi_1(\bar{z}) - \frac{1}{\mu_2}\varphi_2(\bar{z}) + \frac{1}{\mu_2}\varphi_3(\bar{z}) - \frac{K}{\mu_2}.$$

This inequality is violated for sufficiently large K. □

PROPOSITION 12.7. *$J^\infty(z) \leq J^\infty(z+\rho(z)\xi)$ for all $z \in [0,\infty)^3$, where $\xi = (\mu_1, -\mu_2, \mu_3)$.*

Proof. In light of Lemma 12.5, it is enough to consider the case $z_3 = 0$. As in that lemma, we assume without loss of generality that $z_1 \vee z_2 > 0$. Consider the so-called *sup-convolution* (Crandall, Ishii, and Lions (1992), Fleming and Soner (1993))

$$J^\epsilon(z) = \sup_{y \in [0,\infty)^3} \left\{ J^\infty(y) - \frac{1}{\epsilon}|y - z|^2 \right\}.$$

By the linear growth of J^∞, there is, for each ϵ, a point $y(\epsilon)$ at which the supremum is attained. We then have

$$J^\infty(y(\epsilon)) - \frac{1}{\epsilon}|y(\epsilon) - z|^2 \geq J^\infty(z) \geq 0,$$

and this implies that $\{y(\epsilon) : \epsilon > 0\}$ is bounded and $y(\epsilon) \to z$ as $\epsilon \to 0$. Since $J^\infty(y(\epsilon)) \geq J^\infty(z)$, upper semicontinuity of J^∞ implies $J^\infty(y(\epsilon)) \to J^\infty(z)$. Now note that $J^\infty(y(\epsilon)) \leq J^\infty(y(\epsilon) + \rho(y(\epsilon))\xi)$. This follows either from Lemma 12.4 if the third coordinate of $y(\epsilon)$ is positive or from Lemma 12.6 (taking $\varphi(y) = \frac{1}{\epsilon}|y - z|^2$) if the third coordinate is zero. We have then

$$J^\infty(z) = \lim_{\epsilon \downarrow 0} J^\infty(y(\epsilon)) \leq \overline{\lim_{\epsilon \downarrow 0}} J^\infty(y(\epsilon) + \rho(y(\epsilon))\xi) \leq J^\infty(z + \rho(z)\xi). \qquad \square$$

For $w \in [0,\infty)^2$, we denote by $\zeta(w)$ the unique $z \in [0,\infty)^3$ for which $\omega(z) = w$ and $\gamma(z) = 0$. In terms of the function ρ constructed in Lemma 12.2, we have the formula

$$\zeta(\omega(z)) = z + \rho(z)\xi \quad \forall z \in [0,\infty)^3. \tag{12.60}$$

We set

$$\hat{J}(w) \triangleq J^\infty(\zeta(w)), \qquad w \in [0,\infty)^2. \tag{12.61}$$

It will be shown that $\hat{J}$ is a viscosity subsolution (defined below) of the partial differential equation

$$\alpha J - \hat{\mathcal{L}}J - h = C_0\delta_4 \quad \text{on} \quad (0,\infty)^2 \tag{12.62a}$$

and the Neumann boundary conditions

$$-J_1(0, w_2) = 0 \quad \forall w_2 \geq 0, \tag{12.62b}$$

$$-J_2(w_1, 0) = 0 \quad \forall w_1 \geq 0. \tag{12.62c}$$

Here, δ_4 and C_0 are the constants defined in Lemma 12.2(d), (e), $\hat{\mathcal{L}}$ is defined by (10.4), and $\hat{h}$ by (8.4). The value function $\hat{V}$ for the workload control problem is a classical solution of the related equations (10.2), (10.3). These facts will allow us to obtain an upper bound on $\hat{J}$ in terms of $\hat{V}$, and Proposition 12.1 will follow.

DEFINITION 12.8 (Crandall, Evans, and Lions (1984), Crandall and Lions (1984), Crandall, Ishii, and Lions (1992)). *We say that an upper semicontinuous function* $J : [0,\infty)^2 \to \mathcal{R}$ *is a* viscosity subsolution *of* (12.62a)–(12.62c) *if, whenever* $\bar{w} \in \arg\max_{[0,\infty)^2}(J-\varphi)$ *for some* $\varphi \in C^\infty(\mathcal{R}^2)$, *we have*

(a) *if* $\bar{w}_1 > 0$ *and* $\bar{w}_2 > 0$, *then*

$$\alpha J(\bar{w}) - \hat{\mathcal{L}}\varphi(\bar{w}) - \hat{h}(\bar{w}) \le C_0\delta_4;$$

(b) *if* $\bar{w}_1 = 0$ *and* $\bar{w}_2 > 0$, *then*

$$\min\{\alpha J(\bar{w}) - \hat{\mathcal{L}}\varphi(\bar{w}) - \hat{h}(\bar{w}) - C_0\delta_4, -\varphi_1(\bar{w})\} \le 0;$$

(c) *if* $\bar{w}_1 > 0$ *and* $\bar{w}_2 = 0$, *then*

$$\min\{\alpha J(\bar{w}) - \hat{\mathcal{L}}\varphi(\bar{w}) - \hat{h}(\bar{w}) - C_0\delta_4, -\varphi_2(\bar{w})\} \le 0;$$

(d) *if* $\bar{w}_1 = 0$ *and* $\bar{w}_2 = 0$, *then*

$$\min\{\alpha J(\bar{w}) - \hat{\mathcal{L}}\varphi(\bar{w}) - \hat{h}(\bar{w}) - C_0\delta_4, -\varphi_1(\bar{w}), -\varphi_2(\bar{w})\} \le 0.$$

Remark 12.9. A function J is a viscosity subsolution of (12.62a)–(12.62c) if and only if, for every C^∞ function $\varphi : \mathcal{R}^2 \to \mathcal{R}$ and every $\bar{w}$ which is a *strict maximum* of $J-\varphi$ over $[0,\infty)^2$, conditions (a)–(d) of Definition 12.8 hold. Indeed, if we have these conditions at strict maxima and if $\bar{w}$ maximizes $J-\varphi$, but perhaps not strictly, then $\bar{w}$ is a strict maximum of $J-\varphi_\delta$, where $\delta > 0$ and $\varphi_\delta(w) \triangleq \varphi(w) + \delta e^{-|w-w|^2}$. Writing conditions (a)–(d) for φ_δ and letting $\delta \downarrow 0$, we obtain these conditions for φ.

PROPOSITION 12.10. $\hat{J}$ *defined by* (12.61) *is a viscosity subsolution of* (12.62a)–(12.62c).

Let $\varphi : \mathcal{R}^2 \to \mathcal{R}$ be of class C^∞, and let $\bar{w}$ maximize $\hat{J}-\varphi$ over $[0,\infty)^2$. In light of Remark 12.9, we may assume that $\hat{J}-\varphi$ has a *strict* maximum over $[0,\infty)^2$ at $\bar{w}$.

For $\epsilon > 0$, put

$$\varphi^\epsilon(z) = \varphi(\omega(z)) + \frac{\epsilon}{2}(\gamma(z))^2, \qquad z \in (-1,\infty)^3.$$

We claim that $\bar{z} \triangleq \zeta(\bar{w})$ is a strict maximizer of $J^\infty - \varphi^\epsilon$ over $[0,\infty)^3$, i.e.,

$$J^\infty(z) - \varphi(\omega(z)) - \frac{\epsilon}{2}(\gamma(z))^2 < \hat{J}(\bar{w}) - \varphi(\bar{w}) \ \forall z \in [0,\infty)^3\backslash\{\bar{z}\}. \tag{12.63}$$

Consider $z \in [0,\infty)^3$, $z \neq \bar{z}$. If $\omega(z) = \bar{w}$, then $\gamma(z) \neq 0$ and Proposition 12.7 implies

$$J^\infty(z) - \varphi(\omega(z)) \le J^\infty(z + \rho(z)\xi) - \varphi(\omega(z)) = \hat{J}(\bar{w}) - \varphi(\bar{w});$$

inequality (12.63) follows. On the other hand, if $w(z) \neq \bar{w}$, then Proposition 12.7 implies

$$\begin{aligned} J^\infty(z) - \varphi(\omega(z)) &\le J^\infty(z + \rho(z)\xi) - \varphi(\omega(z)) \\ &= \hat{J}(\omega(z)) - \varphi(\omega(z)) \\ &< \hat{J}(\bar{w}) - \varphi(\bar{w}). \end{aligned}$$

Set

$$\Psi^{n,\epsilon}(z) \triangleq \sqrt{n}\nabla\varphi^\epsilon(z)\cdot(\xi - \xi^{(n)}) - \mathcal{A}^{(n)}\varphi^\epsilon(z), \qquad z \in (-1,\infty)^3,$$

where $\xi^{(n)}$ and $\mathcal{A}^{(n)}$ are given by (7.14) and (7.15). Let T be as in Step 3 of the proof of Proposition 11.1. For $z \in [0,\infty)^3$, set

$$f^{n,\epsilon}(z) \triangleq \int_0^{T(z)} \psi^{n,\epsilon}(z - r\xi^{(n)})dr, \qquad g^{n,\epsilon}(z) \triangleq \varphi^\epsilon(z) + \frac{1}{\sqrt{n}} f^{n,\epsilon}(z).$$

Note that

$$\nabla f^{n,\epsilon}(z)\cdot\xi^{(n)} = \psi^{n,\epsilon}(z) + O\left(\frac{1}{\sqrt{n}}\right), \qquad g^{n,\epsilon}(z) = \varphi^{\epsilon}(z) + O\left(\frac{1}{\sqrt{n}}\right),$$

where $O(\frac{1}{\sqrt{n}})$ is bounded in absolute value by $K(z)/\sqrt{n}$ and $K : [0,\infty)^3 \to [0,\infty)$ is locally bounded. Indeed, we may choose a linearly growing K because $f^{n,\epsilon}$ grows at most linearly.

Because $\bar{z}$ is a strict maximizer of $J^{\infty} - \varphi^{\epsilon}$ over $[0,\infty)^3$, we can find a compact set $G \subset \mathcal{R}^3$ containing $\bar{z}$ in its interior, a strictly increasing sequence of positive integers $\{k_n\}_{n=1}^{\infty}$, and a sequence $\{z^{k_n}\}_{n=1}^{\infty}$ such that $J^{(k_n)}(z^{k_n}) \to J^{\infty}(\bar{z})$ and each z^{k_n} maximizes $J^{(k_n)} - g^{k_n,\epsilon}$ over $G \wedge L^{(k_n)}$. To simplify typography, we assume that $k_n = n$, i.e.,

$$\lim_{n\to\infty} J^{(n)}(z^n) = J^{\infty}(\bar{z}), \tag{12.64}$$

$$J^{(n)}(z^n) - g^{n,\epsilon}(z^n) \geq J^{(n)}(z) - g^{n,\epsilon}(z) \quad \forall z \in G \cap L^{(n)}. \tag{12.65}$$

Directly from (3.1), this implies $\mathcal{L}^{n,1,U^n(z^n)}(J^{(n)} - g^{n,\epsilon})(z^n) \leq 0$, and Proposition 4.1 yields

$$\alpha J^{(n)}(z^n) - \mathcal{L}^{n,1,U^n(z^n)} g^{n,\epsilon}(z^n) \leq c\cdot z^n. \tag{12.66}$$

We now use expansion (7.13) to obtain

$$\begin{aligned}\mathcal{L}^{n,1,U^n(z^n)} g^{n,\epsilon}(z^n) &= \mathcal{L}g^{n,\epsilon}(z^n) + \theta^{(n)}\left[\nabla g^{n,\epsilon}(z^n)\cdot\xi^{(n)} + \frac{1}{\sqrt{n}}\mathcal{A}^{(n)} g^{n,\epsilon}(z^n)\right] \\ &\quad + \sum_{i=1}^{3} \mathcal{B}_i^{n,1,U^n(z^n)} g^{n,\epsilon}(z) + O\left(\frac{1}{\sqrt{n}}\right),\end{aligned} \tag{12.67}$$

where $\theta^{(n)} = \sqrt{n}[\frac{\lambda_1^{(n)}}{\mu_1^{(n)}} - U^n(z^n)]$ and $O(\frac{1}{\sqrt{n}})$ is bounded in absolute value by $K(z)/\sqrt{n}$ for some locally bounded function K (see Remark 7.1). Simplifying the right-hand side of (12.67) and using the inequalities $\theta^{(n)}\gamma(z^n) \leq 0$, $\nabla\gamma(z^n)\cdot\xi \geq 0$, we obtain

$$\begin{aligned}\mathcal{L}^{n,1,U^n(z^n)} g^{n,\epsilon}(z^n) &= \mathcal{L}\varphi^{\epsilon}(z^n) + \frac{\epsilon}{\sqrt{n}}\theta^{(n)}\gamma(z^n)\nabla\gamma(z^n)\cdot\xi \\ &\quad + \sum_{i=1}^{3} \mathcal{B}_1^{n,1,U^n(z^n)} g^{n,\epsilon}(z^n) + O\left(\frac{1}{\sqrt{n}}\right) \\ &\leq \mathcal{L}\varphi^{\epsilon}(z^n) + \sum_{i=1}^{3} \mathcal{B}_1^{n,1,U^n(z^n)} g^{n,\epsilon}(z^n) + O\left(\frac{1}{\sqrt{n}}\right)\end{aligned} \tag{12.68}$$

and

$$\mathcal{B}_1^{n,1,U^n(z^n)} g^{n,\epsilon}(z^n) = \sqrt{n}\mu_1^{(n)} U^n(z^n)\left[\varphi_1(\omega(z^n)) + \epsilon\gamma(z^n)\gamma_1(z^n) + O\left(\frac{1}{\sqrt{n}}\right)\right] 1_{\{z_1^n=0\}},$$

$$\mathcal{B}_2^{n,1,U^n(z^n)} g^{n,\epsilon}(z^n) = \sqrt{n}\mu_2^{(n)}(1 - U^n(z^n))\left[\varphi_2(\omega(z^n)) + \epsilon\gamma(z^n)\gamma_2(z^n) + O\left(\frac{1}{\sqrt{n}}\right)\right] 1_{\{z_2^n=0\}},$$

$$\mathcal{B}_3^{n,1,U^n(z^n)} g^{n,\epsilon}(z^n) = \sqrt{n}\mu_3^{(n)}\left[\varphi_2(\omega(z^n)) + \epsilon\gamma(z^n)\gamma_3(z^n) + O\left(\frac{1}{\sqrt{n}}\right)\right] 1_{\{z_3^n=0\}}.$$

Note that $\gamma(z^n) \geq 0$ whenever $z_2^n = 0$, which implies $(1 - U^{(n)}(z^n))1_{\{z_2^n=0\}} = 0$; hence $\mathcal{B}_2^{n,1,U^n(z^n)} g^{n,\epsilon}(z^n) = 0$.

Let us consider now the four cases (a)–(d) of Definition 12.8.

Case (a) ($\bar{w}_1 > 0, \bar{w}_2 > 0$). Because $\gamma(\bar{z}) = a(\bar{z}_1)a(\bar{z}_3) - b(\bar{z}_2) = 0$, we must have $\bar{z}_1 > 0, \bar{z}_2 > 0, \bar{z}_3 > 0$. Thus, for sufficiently large n,

$$\mathcal{B}_1^{n,1,U^n(z_n)} g^{n,\epsilon}(z^n) = \mathcal{B}_3^{n,1,U^n(z_n)} g^{n,\epsilon}(z^n) = 0,$$

and (12.68) implies

$$\mathcal{L}^{n,1,U^{(n)}(z^n)} g^{n,\epsilon}(z^n) \le \mathcal{L}\varphi^\epsilon(z^n) + O\left(\frac{1}{\sqrt{n}}\right). \tag{12.69}$$

From (12.66), we see that

$$\alpha J^{(n)}(z^n) - \mathcal{L}\varphi^\epsilon(z^n) \le c \cdot z^n + O\left(\frac{1}{\sqrt{n}}\right),$$

and letting first $n \to \infty$ and then $\epsilon \downarrow 0$, we obtain

$$\alpha J^\infty(\bar{z}) - \mathcal{L}\varphi(\omega(\bar{z})) \le c \cdot \bar{z}.$$

But $J^\infty(\bar{z}) = \hat{J}(\bar{w})$ and $\mathcal{L}\varphi(\omega(\bar{z})) = \hat{\mathcal{L}}\varphi(\bar{w})$, so (12.11) yields

$$\alpha \hat{J}(\bar{w}) - \hat{\mathcal{L}}\varphi(\bar{w}) \le \hat{h}(\bar{w}) + C_0\delta_4, \tag{12.70}$$

as required by Definition 12.8(a).

Case (b) ($\bar{w}_1 = 0, \bar{w}_2 > 0$). We have $\bar{z}_1 = \bar{z}_2 = 0, \bar{z}_3 > 0$. If $\varphi_1(\bar{w}) \ge 0$, the inequality in Definition 12.8(b) is satisfied and we are done, so we assume that $\varphi_1(\bar{w}) < 0$. But in this case

$$\lim_{n\to\infty}\left[\varphi_1(\omega(z^n)) + \epsilon\gamma(z^n)\gamma_1(z^n) + O\left(\frac{1}{\sqrt{n}}\right)\right] = \varphi_1(\bar{w}),$$

so we have $\mathcal{B}_1^{n,1,U^n(z^n)} g^{n,\epsilon}(z^n) \le 0$ for sufficiently large n. Since $\bar{z}_3 > 0$, we have $\mathcal{B}_3^{n,1,U^n(z^n)} g^{n,\epsilon}(z^n) = 0$ for sufficiently large n; (12.69) follows and leads to (12.70) as before.

Cases (c) and (d) are similar. □

PROPOSITION 12.11. *We have*

$$\hat{J}(w) \le \hat{V}(w) + \frac{C_0\delta_4}{\alpha} \quad \forall w \in [0,\infty)^2,$$

where δ_4 and C_0 are the constants defined in Lemma 12.2(d), (e).

Proof. For each $\epsilon > 0$, let $\hat{V}^\epsilon$ be the C^∞ function constructed in Step 1 of the proof of Proposition 11.1. For $\delta > 0$, define

$$\varphi^{\epsilon,\delta}(w) \triangleq \hat{V}^\epsilon(w) + \delta\eta(w), \qquad w \in [0,\infty)^2,$$

where $\eta(w) \triangleq w_1^2 + w_2^2 - w_1 - w_2$. Since $\hat{J}$ and $\hat{V}$ grow at most linearly, $\hat{J} - \varphi^{\epsilon,\delta}$ attains its maximum over $[0,\infty)^2$ at some point $\bar{w}^{\epsilon,\delta}$. Furthermore, the set $\{\bar{w}^{\epsilon,\delta}; 0 < \epsilon < 1\}$ is bounded for each fixed $\delta > 0$.

With $\delta > 0$ fixed, let $k^\delta > 0$ be such that $\bar{w}^{\epsilon,\delta} \in [0, k^\delta]^2$ for all $\epsilon \in (0,1)$. On the compact set $[0, k^\delta + 3]^2$, $\hat{V}_1$ and $\hat{V}_2$ are uniformly continuous. Consequently, there exists $\epsilon^\delta \in (0,1)$ such that for all $w, w' \in [0, k^\delta + 3]^2$, we have

$$|\hat{V}_i(w) - \hat{V}_i(w')| \le \frac{\delta}{2}, \qquad i = 1, 2,$$

whenever $|w - w'| < 4\epsilon^\delta$. In particular, if $\epsilon \le \epsilon^\delta$ and $\bar{w}_1^{\epsilon,\delta} = 0$, then because $\hat{V}_1(\bar{w}^{\epsilon,\delta}) = 0$ (see (10.2)), we have

$$\begin{aligned}\varphi_1^{\epsilon,\delta}(\bar{w}^{\epsilon,\delta}) &= \hat{V}_1^{\epsilon}(\bar{w}^{\epsilon,\delta}) - \delta \\ &= \int_{B_\epsilon(0)} \hat{V}_1(x + \bar{w}^{\epsilon,\delta} + 2\epsilon(1,1))\rho_\epsilon(x)dx - \delta \\ &\le -\frac{\delta}{2}.\end{aligned} \tag{12.71}$$

Similarly, if $\epsilon \le \epsilon^\delta$ and $\bar{w}_2^{\epsilon,\delta} = 0$, then $\varphi_2^{\epsilon,\delta}(\bar{w}^{\epsilon,\delta}) \le -\frac{\delta}{2}$.

We next show that

$$\alpha\hat{J}(\bar{w}^{\epsilon,\delta}) - \hat{\mathcal{L}}\varphi^{\epsilon,\delta}(\bar{w}^{\epsilon,\delta}) - \hat{h}(\bar{w}^{\epsilon,\delta}) \le C_0\delta_4 \quad \forall\delta > 0, \forall\epsilon \in (0, \epsilon^\delta). \tag{12.72}$$

For $\delta > 0$ and $0 < \epsilon < \epsilon^\delta$, if $\bar{w}_1^{\epsilon,\delta} > 0$ and $\bar{w}_2^{\epsilon,\delta} > 0$, inequality (12.72) follows immediately from Proposition 12.10 and (a) of Definition 12.8. If $\bar{w}_1^{\epsilon,\delta} = 0$ and $\bar{w}_2^{\epsilon,\delta} > 0$, then we use (12.71) and (b) of Definition 12.8. The case $\bar{w}_1^{\epsilon,\delta} > 0$, $\bar{w}_2^{\epsilon,\delta} = 0$ and the case $\bar{w}_1^{\epsilon,\delta} = 0$, $\bar{w}_2^{\epsilon,\delta} = 0$ are handled similarly.

From (11.4) we have

$$\begin{aligned}\alpha\varphi^{\epsilon,\delta}(\bar{w}^{\epsilon,\delta}) &= \alpha\hat{V}^{\epsilon}(\bar{w}^{\epsilon,\delta}) + \alpha\delta\eta(\bar{w}^{\epsilon,\delta}) \\ &\ge \hat{\mathcal{L}}\hat{V}^{\epsilon}(\bar{w}^{\epsilon,\delta}) + \hat{h}^{\epsilon}(\bar{w}^{\epsilon,\delta}) + \alpha\delta\eta(\bar{w}^{\epsilon,\delta}) - L_0\epsilon \\ &= \hat{\mathcal{L}}\varphi^{\epsilon,\delta}(\bar{w}^{\epsilon,\delta}) + \hat{h}^{\epsilon}(\bar{w}^{\epsilon,\delta}) + \delta[\alpha\eta(\bar{w}^{\epsilon,\delta}) - \hat{\mathcal{L}}\eta(\bar{w}^{\epsilon,\delta})] - L_0\epsilon.\end{aligned} \tag{12.73}$$

Combining this inequality with (12.72), and using the fact that $\bar{w}^{\epsilon,\delta}$ maximizes $\hat{J} - \varphi^{\epsilon,\delta}$, we see that

$$\begin{aligned}\alpha[\hat{J}(w) - \varphi^{\epsilon,\delta}(w)] &\le \alpha[\hat{J}(\bar{w}^{\epsilon,\delta}) - \varphi^{\epsilon,\delta}(\bar{w}^{\epsilon,\delta})] \\ &\le \alpha\bar{J}(\bar{w}^{\epsilon,\delta}) - \hat{\mathcal{L}}\varphi^{\epsilon,\delta}(\bar{w}^{\epsilon,\delta}) - \hat{h}^{\epsilon}(\bar{w}^{\epsilon,\delta}) \\ &\quad - \delta[\alpha\eta(\bar{w}^{\epsilon,\delta}) - \hat{\mathcal{L}}\eta(\bar{w}^{\epsilon,\delta})] + L_0\epsilon \\ &\le C_0\delta_4 + \hat{h}(\bar{w}^{\epsilon,\delta}) - \hat{h}^{\epsilon}(\bar{w}^{\epsilon,\delta}) \\ &\quad - \delta[\alpha\eta(\bar{w}^{\epsilon,\delta}) - \hat{\mathcal{L}}\eta(\bar{w}^{\epsilon,\delta})] + L_0\epsilon \\ &\le C_0\delta_4 + \hat{h}(\bar{w}^{\epsilon,\delta}) - \hat{h}^{\epsilon}(\bar{w}^{\epsilon,\delta}) + K\delta + L_0\epsilon \quad \forall\omega \in [0,\infty)^2,\end{aligned} \tag{12.74}$$

where

$$K \triangleq \max_{w\in[0,\infty)^2}\{-\alpha\eta(w) + \hat{\mathcal{L}}\eta(w)\} < \infty.$$

Letting first $\epsilon \downarrow 0$ and then $\delta \downarrow 0$ in (12.73), we conclude that

$$\alpha[\hat{J}(w) - \hat{V}(w)] \le C_0\delta_4,$$

which establishes the proposition. □

Proof of Proposition 12.1. Propositions 12.7 and 12.11 and (12.59), (12.60) imply that for any $z \in [0,\infty)^3$

$$J^{\#}(z) \le J^{\infty}(z) \le J^{\infty}(z + \rho(z)\xi) = \hat{J}(\omega(z)) \le \hat{V}(\omega(z)) + \frac{C_0\delta_4}{\alpha}. \tag{12.75}$$

But $C_0 = c_1\mu_1 - c_2\mu_2 + c_3\mu_3$ is constant, and $\delta_4 = \frac{1}{\mu_1 \wedge \mu_2} a^{-1}(\sqrt{\delta_2})$ can be made as small as we like by choice of the functions a and b introduced at the beginning of this section. □

COROLLARY 12.12. *Recalling that $J^{(n)} = J^{(n)}_{Y^n, U^n}$ denotes the cost of using the control law* (12.5a), (12.5b) *in the* nth *queueing network, and $J^{(n)}_*$ denotes the optimal cost in this network, we have*

$$\lim_{\delta \downarrow 0} \lim_{n \to \infty} \max_{\substack{\zeta \in L^{(n)} \\ |\zeta - z| \le \delta}} J^{(n)}(\delta) \le \lim_{\delta \downarrow 0} \lim_{n \to \infty} \min_{\substack{\zeta \in L^{(n)} \\ |\zeta - z| \le \delta}} J^{(n)}_*(\zeta) + \frac{C_0 \delta_4}{\alpha} \quad \forall z \in [0, \infty)^3. \tag{12.76}$$

Proof. Inequality (12.75) is simply

$$J^\infty(z) \le J_\#(z) + \frac{C_0 \delta_4}{\alpha}.$$

But Proposition 11.1 and 12.1 imply $J_\#(z) = J^\#(z) = \hat{V}(\omega(z))$ for all $z \in [0, \infty)^3$, and (12.76) follows from (12.74). □

REFERENCES

G. BARLES AND B. PERTHAME (1988), *Exit time problems in control and vanishing viscosity solutions of Hamilton-Jacobi-Bellman equations*, SIAM J. Control Optim., 26, pp. 1113–1148.

H. CHEN, P. YANG, AND D. YAO (1991), *Control and Scheduling in a Two-Station Queueing Network: Optimal Policies and Heuristics*, preprint.

K. L. CHUNG (1960), *Markov Chains with Stationary Transition Probabilities*, Springer-Verlag, New York.

M. G. CRANDALL AND P.-L. LIONS (1984), *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277, pp. 1–43.

M. G. CRANDALL, L. C. EVANS, AND P.-L. LIONS (1984), *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282, pp. 487–502.

M. G. CRANDALL, H. ISHII, AND P.-L. LIONS (1992), *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27, pp. 1–67.

J. DAI AND T. KURTZ (1995), *A multi-class station with Markovian feedback in heavy traffic*, Math. Oper. Res., 20, pp. 721–742.

W. H. FLEMING AND H. M. SONER (1993), *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York.

D. GILBARG AND N. TRUDINGER (1977), *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York.

J. M. HARRISON (1985), *Brownian Motion and Stochastic Flow Systems*, Wiley, New York.

——— (1988), *Brownian models of queueing networks with heterogeneous customer populations*, in Stochastic Differential Systems, Stochastic Control Theory and Applications, IMA Vol. Math. Appl. 10, P.-L. Lions and W. H. Fleming, eds., Springer-Verlag, New York, pp. 265–280.

J. M. HARRISON AND L. M. WEIN (1989), *Scheduling networks of queues: Heavy traffic analysis of a simple open network*, Queueing Systems Theory Appl., 5, pp. 265–280.

——— (1990), *Scheduling networks of queues: Heavy traffic analysis of a two-station closed network*, Oper. Res., 38, pp. 1052–1064.

D. IGLEHART AND W. WHITT (1970), *Multiple channel queues in heavy traffic* I, II, Adv. in Appl. Probab., 2, pp. 150–177, 355–364.

D. P. JOHNSON (1983), *Diffusion Approximations for Optimal Filtering of Jump Processes and Queueing Networks*, Ph.D. dissertation, University of Wisconsin, Madison, WI.

E. V. KRICHAGINA, S. X. C. LOU, S. SETHI, AND M. TAKSAR (1993), *Production control in a failure-prone manufacturing system: Diffusion approximation and asymptotic optimality*, Ann. Appl. Probab., 3, pp. 421–453.

E. V. KRICHAGINA, S. X. C. LOU, AND M. TAKSAR (1994), *Double band policy for stochastic manufacturing systems in heavy traffic*, Math. Oper. Res., 19, pp. 560–596.

H. J. KUSHNER AND F. L. MARTINS (1990), *Routing and singular control for queueing networks in heavy traffic*, SIAM J. Control Optim., 28, pp. 1209–1233.

——— (1991), *Limit Theorems for Pathwise Average per Unit Time Problems in Heavy Traffic*, preprint.

——— (1996), *Heavy traffic analysis of a controlled multi-class queueing network via weak convergence methods*, SIAM J. Control Optim., 34, pp. 1781–1797.

H. KUSHNER AND K. M. RAMACHANDRAN (1988), *Nearly optimal singular controls for wideband noise driven systems*, SIAM J. Control Optim., 26, pp. 569–591.

H. KUSHNER AND K. M. RAMACHANDRAN (1989), *Optimal and approximately optimal control problems for queues in heavy traffic*, SIAM J. Control Optim., 27, pp. 1293–1318.

N. C. LAWS (1992), *Resource pooling in queueing networks with dynamic routing*, Adv. in Appl. Probab., 24, pp. 699–726.

N. C. LAWS AND G. M. LOUTH (1990), *Dynamic scheduling of a four-station queueing network*, Probab. Engrg. Inform. Sci., 4, pp. 131–156.

W. P. PETERSON (1990), *A heavy traffic limit theorem for networks of queues with multiple customer types*, Math. Oper. Res., 16, pp. 90–118.

M. REIMAN (1984), *Open queueing networks in heavy traffic*, Math. Oper. Res., 9, pp. 441–458.

——— (1988), *A multiclass feedback queue in heavy traffic*, Adv. in Appl. Probab., 20, pp. 179–207.

L. WEIN (1990a), *Optimal control of a two-station Brownian network*, Math. Oper. Res., 15, pp. 215–242.

——— (1990b), *Scheduling of network queues: Heavy traffic analysis of a two-station network with controllable inputs*, Oper. Res., 38, pp. 1065–1078.

——— (1992), *Scheduling of network queues: Heavy traffic analysis of a multistation network with controllable inputs*, Oper. Res., 40, pp. S312–S334.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 6, pp. 2172–2179, November 1996

016

ON THE LAVRENTIEV PHENOMENON FOR OPTIMAL CONTROL PROBLEMS WITH SECOND-ORDER DYNAMICS*

CHIH-WEN CHENG† AND VICTOR J. MIZEL‡

Abstract. The present article examines control problems in one dimension for which there is an autonomous running cost and a specified terminal state. In this case, when the running cost involves only the control and the state, it is known that the infimal cost corresponding to any initial state is unaffected by the precise choice of L^p space ($1 \leq p < \infty$) which is specified for controls to be admissible. Here we show that the situation is different in the case of an autonomous running cost involving, in addition to the control, the state *and* its derivative. That is, despite the density of each space with higher exponent in those with lower exponent, the infimal cost will generally depend on the choice of p if sign constraints are present.

Key words. Lavrentiev gap, free zone, fully coercive running cost

AMS subject classifications. 49J05, 49J45

1. Introduction.

We are concerned with control problems of the autonomous form

$$\text{(P)} \qquad C[u] = \int_{-1}^{0} f(x, x', u)\, dt \quad \text{with } x'' = u,$$

over $A^p_{a,b} = \{u \in L^p(-1,0) | \, x(-1) = a, x'(-1) = b, x(0) = x'(0) = 0\}$, $1 \leq p < \infty$, where $f \geq 0$ is smooth and is convex in its third argument. Intuitively, one is evaluating the cost of parking a moving railroad car at the origin in unit time, but the availability of less artificial applications is not clear. (One possible exception is the control of second-order nonlinear materials with negative capillarity [CMM].) The question to be raised is whether—despite the density of L^{p_2} in L^{p_1} for each $p_1 < p_2$—the minimal cost $m_p := \inf_{A^p_{a,b}} C$ varies with the choice of exponent $p \in [1, \infty)$. If such an effect occurs, one says that Lavrentiev's phenomenon is present (cf., e.g., [La], [Ma], [Ce], [BM], [M1], [M2], [M3], [HM1], [HM2], [HM3], [L], [Da], [BuM]). It is known that in analogous autonomous control problems with *first-order* dynamics, namely,

$$C[u] = \int_{-1}^{0} f(x, u)\, dt, \quad \text{with} \quad x' = u, \quad \text{over}$$
$$A^p_a = \{u \in L^p(-1,0) \,|\, x(-1) = a, x(0) = 0\},$$

the minimizing u in A^p_a actually lies in L^∞, so such an outcome is impossible [CV1], [Da], [AAB]. In contrast, we present a class of examples in the second-order context for which such a phenomenon does take place on suitable subsets of $A^p_{a,b}$. This sheds light on a conjecture raised in [CV2] concerning higher-order integrands. It also provides an incentive for further study of the troubling issue raised in [BK], [NM], and [Li]: the invalidity in such examples of standard numerical methods for dealing with optimization problems.

*Received by the editors January 3, 1994; accepted for publication (in revised form) January 4, 1996. A version of this paper will be presented at the 35th IEEE Conference on Decision and Control, Kobe, Japan, December 11–13, 1996.

†Computer and Communication Research Laboratories, Industrial Technology Research Institute, Taiwan. The research of this author was supported in part by the Army Research Office and the NSF through the Center for Nonlinear Analysis.

‡Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213-3890. The research of this author was supported in part by NSF grants DMS-90-02562 and DMS-92-01221.

To illustrate the phenomenon, consider the following example:

$$C[u] = \int_{-1}^{0} (x')^6 (u)^8 \, dt, \quad \text{with } x'' = u,$$

over $A^p_{a,b} = \{u \in L^p(-1,0) \mid x(-1) = a > 0, x'(-1) = b < 0, x(0) = x'(0) = 0\}$. Now by Jensen's inequality and the chain rule for absolutely continuous functions (cf., e.g., [MM])

$$\begin{aligned} \int_s^0 (x'(t))^6 (u(t))^8 dt &= \int_s^0 (|x'(t)|^{3/4} x''(t))^8 dt \\ &= (4/7)^8 \int_s^0 (d/dt(|x'(t)|^{7/4}))^8 dt \\ &\geq (4/7)^8 |x'(s)|^{14}/|s|^7, \qquad -1 \leq s \leq 0. \end{aligned} \tag{1}$$

In particular, it follows that for all $u \in A^p_{a,b}$

$$C[u] \geq (4/7)^8 |b|^{14}. \tag{2}$$

Now consider the curve Γ given by $\Gamma = \{(x, y) | y = -x^{1/3}, x \geq 0\}$, and multiply the integrand by a weight factor $\theta(x, y)$, which vanishes along Γ and is positive everywhere else. Denoting the modified integrand by $f^\#$, we have

$$f^\#(x, x', u) = \theta(x, x')(x')^6 (u)^8 =: d(x, x')(u)^8.$$

Consider first the extreme case

$$\begin{aligned} \theta(x, y) &= 0 \quad \text{if } (x, y) \in \Gamma \\ &= 1 \quad \text{if } (x, y) \in [0, \infty) \times (-\infty, 0] \backslash \Gamma \\ &= \infty \quad \text{otherwise.} \end{aligned}$$

It is not hard to see why there is a Lavrentiev phenomenon for the cost function $C^\#$ corresponding to $f^\#$. Indeed consider the trajectory $x^\star$ given by

$$x_\star(t) = (-2t/3)^{3/2}, \qquad x'_\star(t) = -(-2t/3)^{1/2},$$

with corresponding boundary data $a = x_\star(-1)$, $b = x'_\star(-1)$. For this trajectory, $u_\star = x''_\star \in L^p(-1,0)$ for all $p \in [1,2)$, and $C^\#[u_\star] = 0$ ($d(x_\star(t), x'_\star(t)) = 0$ for all $t \in [-1,0]$). However, if we restrict attention to the smaller class $L^2(-1,0)$, then the optimizing control $u_\star$ is no longer admissible. It then follows from the previous estimate (1) that, in particular, we have the gap $(4/7)^8(2/3)^7$ for those controls $u \in L^2(-1,0)$ that coincide with $u_\star$ on an interval $[-1, s]$ for some $s < 0$ but are such that $\theta(x_\star(t), x'_\star(t)) \neq 0$ for all $t \in (s, 0)$. The full demonstration that there is a Lavrentiev gap for *all* L^2 controls for these and other a, b values will follow from the proof of Theorem A below.

Now the phenomenon is not restricted to cases in which the weight function d is non-smooth. Indeed, as follows from Theorem A stated below, the same basic result applies, for example, even to the polynomial weight function $d(x, y) = (x + y^3)^2$. More precisely, if we restrict attention to the state constrained subclasses

$$A^{p,+}_{a,b} = \{u \in L^p(-1,0) \mid x(-1) = a > 0, x'(-1) = b < 0, x(0) = x'(0) = 0 \text{ with } x(\cdot) \geq 0\},$$

then there will again exist for appropriate data a, b the relation

$$m_2^+ := \inf_{A_{a,b}^{2,+}} C^\# > 0 = m_1^+ = \min_{A_{a,b}^{1,+}} C^\#.$$

However, if one relaxes the sign constraint on x, then the *Lavrentiev gap* disappears:

$$m_p = \inf_{A_{a,b}^{p}} C^\# = 0 = m_1 = \min_{A_{a,b}^{1}} C^\# \quad \text{for all } p \in (1, \infty).$$

The initial class of problems that we discuss here retains the following features of the example analyzed above:

- There are zero-cost curves ("turnpikes") in the fourth quadrant of the (x, y) plane that correspond to trajectories arising from controls in $L^p(-1, 0)$ only for $1 \leq p < p_0$ for some $p_0 > 1$. The optimal trajectories for exponents $p < p_0$ follow these zero-cost curves, but such trajectories are not admissible when controls are restricted to $L^{p_0}(-1, 0)$.
- The weight function $d = d(x, y)$ has the following *homogeneity* property: for some $\gamma \in (1, 2)$ and $\alpha \geq 0$

$$d(x\lambda^\gamma, y\lambda^{\gamma-1}) = \lambda^\alpha d(x, y) \quad \text{for all } \lambda > 0.$$

This article is organized as follows. In §2 two propositions are presented that demonstrate that there is a certain region in the fourth quadrant of the (x, y) plane over which every trajectory in the subclass $A_{a,b}^{p_0,+}$ must cross, while such crossing can be avoided by trajectories in the classes $A_{a,b}^{p,+}$, $1 \leq p < p_0$. Then our main results for cost functions as above are presented. Finally, in §3, it is pointed out how perturbation of this special class of cost functions leads to *fully coercive* cost integrands that still exhibit the Lavrentiev phenomenon. Related results are presented in [CM].

2. Basic results. Our first result is a simple consequence of Hölder's inequality.

PROPOSITION 2.1. *Given $\gamma \in (1, 2)$, set $p_0 = 1/(2-\gamma)$. Suppose that $x \in W^{2,p_0}(-1, 0)$ satisfies $x(t_0) = 0$, $x'(t_0) = 0$. Then the functions $Y_0, Y_1 : [-1, t_0] \to \mathbb{R}$ defined by*

$$Y_0(t) = x(t)|t - t_0|^{-\gamma}, \qquad Y_1(t) = x'(t)|t - t_0|^{1-\gamma}$$

satisfy

$$Y_0(t) = o(1), \quad Y_1(t) = o(1) \quad \text{as } t \to t_{0^-}.$$

Our next result shows that with γ and p_0 as above, the phase vector $(x(t), x'(t))$ associated with any function in $A_{a,b}^{p_0^+}$ must occupy a plane sector $-c_2 \leq x^{(1-\gamma)/\gamma} y \leq -c_1$ for some nondegenerate time interval $[t_0, t_1]$. Note that Y_0, Y_1 above satisfy $Y_0'(t) = [Y_1(t) + \gamma Y_0(t)]/(t_0 - t)$.

PROPOSITION 2.2. *With γ and p_0 as above, suppose that $x \in A_{a,b}^{p_0,+}$. Put $t_0 = \min\{t \in [-1, 0] \mid x(t) = x'(t) = 0\}$, $Y_0(t) = x(t)|t - t_0|^{-\gamma}$, $Y_1(t) = x'(t)|t - t_0|^{1-\gamma}$, $Z(t) = x'(t)|x(t)|^{(1-\gamma)/\gamma}$.*

For any constants c and D satisfying $0 < c < D\gamma < a\gamma$ and any choice of c_1 and c_2 satisfying $0 < c_1 < c_2 < cD^{(1-\gamma)/\gamma}$ there exist t_2, $t_1 \in (-1, t_0)$ such that the following conditions hold: $Z(t_2) = -c_2$, $Z(t_1) = -c_1$, $0 \leq Y_0(t) \leq D$, $-c \leq Y_1(t)$, $-c_2 \leq Z(t) \leq -c_1$ for all $t \in [t_2, t_1]$.

Proof. Since $Y_0(-1) \geq a$ and, by Proposition 2.1, $Y_0(t) = o(1)$ near t_0, it follows by continuity that there is a $t \in [-1, t_0)$ such that $Y_0(t) = D$. Define $t_4 = \max\{t \in [-1, t_0) \mid Y_0(t) = D\}$, so $0 \leq Y_0(t) \leq D$ for all $t \in [t_4, t_0)$. Consequently

$Y_0'(t_4) \le 0$, which implies that $Y_1(t_4) \le -D\gamma$. Since by Proposition 2.1 $Y_1(t) = o(1)$ near t_0, it follows by continuity that there is a $t \in (t_4, t_0)$ such that $Y_1(t) = -c$. Defining $t_3 = \max\{t \in (t_4, t_0) \mid Y_1(t) = -c\}$, we obtain $-c \le Y_1(t)$ for all $t \in [t_3, t_0]$. Examine $G_2(t) = Y_1(t) + c_2|Y_0(t)|^{(\gamma-1)/\gamma} = [c_2 + Z(t)]x(t)^{(\gamma-1)/\gamma}(t_0 - t)^{1-\gamma}$. Note that $G_2(t_3) \le D^{(\gamma-1)/\gamma}[c_2 - cD^{(1-\gamma)/\gamma}] < 0$ by choice of c and c_2. If, for some $t \in [t_3, t_0)$, $Y_1(t) \ge 0$ or $Y_0(t) = 0$, then it follows by continuity that $G_2(s) = 0$ for some $s \in [t_3, t_0)$. ($x(t) \ge 0$ implies that also $Y_1(t) = 0$ in the latter case.) Otherwise, $Y_0(t) > 0$, $Y_1(t) < 0$ for all $t \in [t_3, t_0)$. We now show that in this case $\sup\{Z(t) \mid t \in [t_3, t_0)\} = 0$, so again $G_2(s) = 0$ for some $s \in [t_3, t_0)$. Otherwise, for some $k > 0$, $x'(t)|x(t)|^{(1-\gamma)/\gamma} \le -k$ for all $t \in [t_3, t_0)$.

Integrating both sides of this inequality over (t, t_0) gives $x(t) \ge (k/\gamma)^\gamma |t - t_0|^\gamma$, i.e., $Y_0(t) \ge (k/\gamma)^\gamma > 0$ for all $t \in [t_3, t_0)$, contradicting Proposition 2.1. A similar argument establishes that $G_1(t) = Y_1(t) + c_1|Y_0(t)|^{(\gamma-1)/\gamma}$ satisfies $G_1(s) = 0$ for some $s \in [t_3, t_0)$. On setting $t_2 = \max\{t \in [t_3, t_0) \mid G_2(t) = 0\}$, $t_1 = \min\{t \in [t_2, t_0) \mid G_1(t) = 0\}$, we obtain the desired conclusion.

We now state our main results.

THEOREM A. *Consider problem* (P) *with a cost integrand of the form*

$$f(x, y, u) = d(x, y)|u|^k,$$

where $k > 1$ *and* $d \in \mathrm{Cont}(\mathbb{R}^2)$ *satisfies the homogeneity condition*

(Hom) $\quad d(x\lambda^\gamma, y\lambda^{\gamma-1}) = \lambda^\alpha d(x, y)$ *for all* $\lambda > 0$ *for some* $\gamma \in (1, 2)$ *and* $\alpha \ge 0$.

Put $F = \{\beta > 0 \mid d(1, -\beta) = 0\}$, *and define the "free zone"* $\mathbb{F}$ *by*

$$\mathbb{F} = \{(x, y) \mid y = -\beta x^{(\gamma-1)/\gamma} \text{ with } \beta \in F\}.$$

Suppose that F *is nonempty with empty interior and that*

$(\star)$ $$\alpha - k(2 - \gamma) \le -1.$$

Then if the boundary data $a > 0$ *and* $b < 0$ *satisfy*

(#) $$(|b|/\beta)^{\gamma/(\gamma-1)} \le a \le |b| - (\gamma - 1)(|b|/\beta)^{\gamma/(\gamma-1)} \quad \text{for some } \beta \in F,$$

the Lavrentiev phenomenon holds for $A_{a,b}^{p_0,+}$, *where* $p_0 = 1/(2 - \gamma)$.

Remark. Note that (#) is nondegenerate, i.e., allows real values for a, if and only if $|b| \le \beta^\gamma/\gamma^{\gamma-1}$.

Proof. We begin by estimating $C[u]$ for controls $u \in L^{p_0}(-1, 0)$. By setting $\lambda = |y|^{1/(1-\gamma)}$ in (Hom) we can express f as follows:

$$f(x, y, u) = d(x|y|^{\gamma/(1-\gamma)}, -1)|y|^{\alpha/(\gamma-1)}|u|^k \quad \text{for } y < 0.$$

Now choose scalars c, D, c_1, and c_2 satisfying $0 < c < D\gamma < a\gamma$, $0 < c_1 < c_2 < cD^{(1-\gamma)/\gamma}$ such that $[-c_2, -c_1]$ contains no point of F. By Proposition 2.2, for the above choice of constants there exists $[t_2, t_1] \subset (-1, 0)$ such that

$$x'(t_2) = -c_2|x(t_2)|^{(\gamma-1)/\gamma}, \qquad x'(t_1) = -c_1|x(t_1)|^{(\gamma-1)/\gamma}$$

with $0 \le x(t) \le D$, $-c_2 \le x'(t)|x(t)|^{(1-\gamma)/\gamma} \le -c_1$ for all $t \in [t_2, t_1]$. Hence by continuity of d we conclude that there is a constant $\delta_1 = \delta_1(c_2, c_1) > 0$ such that

$$d(x(t)|x'(t))^{\gamma/(\gamma-1)}, -1) \ge \delta_1 > 0 \quad \text{for all } t \in [t_2, t_1].$$

Consequently, setting $\alpha/(\gamma-1)=\ell$ we have the following appraisal:

$$\begin{aligned} C[u] \geq \delta_1 \int_{t_2}^{t_1} |x'(t)|^{\ell}|u(t)|^k dt &= \delta_1 \int_{t_2}^{t_1} |x'(t)|^{\ell}|x''(t)|^k dt \\ &= \delta_1 \int_{x_1}^{x_2} |x'(t)|^{\ell+k-1}|dx'/dx|^k(t)dx \\ &= \Delta \int_{x_1}^{x_2} |d/dx|x'(t)|^{(\ell+2k-1)/k}|^k dx, \end{aligned}$$

where $\Delta=\delta_1(k/(\ell+2k-1))^k$, $x_1:=x(t_1)$, $x_2:=x(t_2)$, and we have employed the chain rule as well as the substitution rule for integrals [F, Thm. 3.2.6]. Putting $(\ell+2k-1)/k=r$ and applying Jensen's inequality and the chain rule we now obtain

$$\begin{aligned} C[u] &\geq \Delta \int_{x_1}^{x_2} |d/dx|x'(t)|^r|^k dx \\ &\geq \Delta \left|[|x'(t_2)|^r - |x'(t_1)|^r]/(x_2-x_1)\right|^k (x_2-x_1) \\ &= \Delta \left|[c_2^r x_2^{(\gamma-1)r/\gamma} - c_1^r x_1^{(\gamma-1)r/\gamma}]/(x_2-x_1)\right|^k (x_2-x_1) \\ &= \Delta c_2^{kr} x_2^{k(\gamma-1)r/\gamma+(1-k)} \left|1-(c_1/c_2)^r (x_1/x_2)^{(\gamma-1)r/\gamma}\right| A^{k-1}, \end{aligned}$$

where $A=[1-(c_1/c_2)^r(x_1/x_2)^{(\gamma-1)r/\gamma}]/(1-x_1/x_2) \geq 1-(c_1/c_2)^r$. Since $x_2 \in [0, D]$ and its exponent is nonpositive, we obtain

$$\begin{aligned} C[u] &\geq \Delta c_2^{rk} x_2^{(\alpha-k(2-\gamma)+1)/\gamma}|1-(c_1/c_2)^r|^k \\ &\geq \Delta c_2^{rk} D^{(\alpha-k(2-\gamma)+1)/\gamma}|1-(c_1/c_2)^r|^k =: \delta(D, c_1, c_2) > 0. \end{aligned}$$

Next we appraise $C[u]$ for controls $u \in L^1(-1,0)$. To do so we take β to be a point of F chosen to satisfy (#) and consider the specific control function $u^\star \notin L^{p_0}(-1,0)$ defined by

$$\begin{aligned} u^\star(t) &= 0 \quad \text{for all } t \in [-1, t^\star], \text{ where } t^\star+1=[a-(|b|/\beta)^{\gamma/(\gamma-1)}]/|b| \\ &= \omega\beta^\gamma(t_0-t)^{\gamma-2} \quad \text{for all } t \in [t^\star, t_0] \\ &= 0 \quad \text{for all } t \in [t_0, 0], \end{aligned}$$

where $\omega=(\gamma-1)\gamma^{1-\gamma}$ and $1+t_0=[a+(\gamma-1)(|b|/\beta)^{\gamma/(\gamma-1)}]/|b|$.

Simple integration gives

$$\begin{aligned} x^{\star\prime}(t) &= b \quad \text{for } t \in (-1, t^\star) \\ &= b+[\omega\beta^\gamma/(\gamma-1)][(t_0-t^\star)^{\gamma-1}-(t_0-t)^{\gamma-1}] \quad \text{for } t \in (t^\star, t_0) \\ &= b+[\omega\beta^\gamma/(\gamma-1)][(t_0-t^\star)^{\gamma-1}] \quad \text{for } t \in (t_0, 0). \end{aligned}$$

Our choice of t_0 and $t^\star$ ensures that $x^{\star\prime}(0-)=0$, which simplifies the expressions above and implies via a second integration that

$$\begin{aligned} x^\star(t) &= a+b(t+1) \quad \text{for } t \in [-1, t^\star] \\ &= (\beta/\gamma)^\gamma(t_0-t)^\gamma \quad \text{for } t \in (t^\star, t_0] \\ &= 0 \quad \text{for } t \in (t_0, 0]. \end{aligned}$$

Hence $C[u^\star] = 0$, since on the subset of $[-1, 0]$ where $u^\star \neq 0$, $x'(t) = -\beta(x(t))^{(\gamma-1)/\gamma}$ for all $t \in [t^\star, t_0]$, so $d(x^\star(t), x^{\star\prime}(t)) = 0$.

By the preceding calculations we conclude that, as claimed,

$$m_{p_0}^+ \geq \delta(D, c_1, c_2) > 0 = m_1^+.$$

Some analogue of the positivity restriction on trajectories in the above result is essential, as is shown below; cf., e.g., [J, Chap. 10].

THEOREM B. *Consider problem* (P) *with the same cost integrand as in the preceding result and with boundary data subject to* (#) *as before. Then the Lavrentiev phenomenon fails to take place over the spaces* $A_{a,b}^p$, $1 \leq p < \infty$. *That is,*

$$m_p = \inf_{A_{a,b}^p} C = 0 = m_1 = \inf_{A_{a,b}^1} C \quad \textit{for all } p \in (1, \infty).$$

Proof. We define the control $u^\star \in L^1(-1, 0) \backslash L^{p_0}(-1, 0)$ as before. Thus since $C[u^\star] = 0$, it remains to prove that $m_p = 0$ for $p \geq p_0$ as well. Recall that in the definition of $u^\star$ the time $t_0 \leq 0$ following which $u^\star$ is identically zero was determined by the boundary data according to $1 + t_0 = [a + (\gamma - 1)(|b|/\beta)^{\gamma/(\gamma-1)}]/|b|$. We separate our analysis into two cases, according to whether $t_0 < 0$ or $t_0 = 0$.

Case 1 ($t_0 < 0$). In this case, given any $\varepsilon > 0$, we construct a control $u_\varepsilon \in A_{a,b}^{p_0}$, which bifurcates from $u^\star$ and satisfies $C[u_\varepsilon] < \varepsilon$. Necessarily, this requires that the corresponding trajectory $(x_\varepsilon(t), x_\varepsilon'(t))$ avoid crossing the zone of the (x, y) plane defined by $-c_2 \leq y|x|^{(1-\gamma)/\gamma} \leq -c_1$, since otherwise by our previous calculations $C[u_\varepsilon] \geq \delta(D, c_1, c_2) > 0$.

Let $\rho > 0$ be taken sufficiently small. (Its value will be specified later.) On the interval $[t^\star, t_0]$ where $u^\star$ is nonzero, we select a point $(s_4, -\rho)$ on the graph $(t, x^{\star\prime}(t))$, $t \in [t^\star, t_0]$ such that $\sigma := t_0 - s_4 \in (0, |t_0|/4)$. (By a straightforward computation it is seen that $x^{\star\prime}(t) = -\beta[(\beta/\gamma)(t_0 - t)]^{\gamma-1}$ on this interval.) Subdivide the interval $[s_4, 0]$ into four subintervals $[s_4, s_3]$, $[s_3, s_2]$, $[s_2, -\sigma]$, $[-\sigma, 0]$ of lengths $|t_0|/2$, $|t_0|/4$, $|t_0|/4$, σ, respectively. Now we define u_ε by the requirement

$$\begin{aligned} u_\varepsilon(t) &= u^\star(t), \qquad t \in [-1, s_4] \\ &= 2\rho/|t_0|, \qquad t \in [s_4, s_3] \\ &= 4\tau/|t_0|, \qquad t \in [s_3, s_2] \\ &= -4\tau/|t_0|, \qquad t \in [s_2, -\sigma] \\ &= 0, \qquad t \in [-\sigma, 0], \end{aligned}$$

with $\tau \in (0, \rho)$ chosen so that $x_\varepsilon(s_4) = x^\star(s_4)$ and $x_\varepsilon(-\sigma) = x^\star(t_0) = 0$, i.e., so that $(\rho - \tau)t_0/4 = \int_{s_4}^{t_0} |x^{\star\prime}(t)|\, dt = x^\star(s_4)$. It is easily verified that $|u_\varepsilon(t)| \leq 4\rho/|t_0|$, $|x_\varepsilon'(t)| \leq \rho$, $|x_\varepsilon(t)| \leq \rho|t_0|/2$ for $t \in [s_4, 0]$. Thus by the continuity of d at $(0, 0)$, for ρ sufficiently small we obtain $d(x_\varepsilon(t), x_\varepsilon'(t)) < 1$ over $[s_4, 0]$, whereby

$$\begin{aligned} C[u_\varepsilon] &= \int_{s_4}^{-\sigma} d(x_\varepsilon(t), x_\varepsilon'(t))|u_\varepsilon(t)|^k dt \\ &\leq \int_{s_4}^{-\sigma} |u_\varepsilon(t)|^k dt \leq |4\rho/t_0|^k |t_0|. \end{aligned}$$

This value is $< \varepsilon$ for $\rho < (\varepsilon/4)^{1/k}(t_0/4)^{(k-1)/k}$, so the gap phenomenon does not occur. (Note that $u_\varepsilon \in L^\infty(-1, 0)$.)

Case 2 ($t_0 = 0$). In this case the preceding construction for u_ε as a direct perturbation of $u^\star$ is clearly no longer available. We still apply the above construction, but this time for each $\varepsilon > 0$ it is performed as a perturbation of a control $u^{\star\star} \in L^1(-1, 0)\backslash L^{p_0}(-1, 0)$ satisfying both $C[u^{\star\star}] < \varepsilon$ and the requirement that there be a $t_0' \in [-1, 0)$ such that $u^{\star\star}(t) = 0$ for all $t \in [t_0', 0]$. Examine the fourth quadrant of the (x, y) plane. The control $u^\star$ led to a curve proceeding from the initial point (a, b) in this quadrant to the origin by a path consisting of a horizontal segment which intersected the locus $\Gamma = \{(x, y) \mid y = -\beta x^{(\gamma-1)/\gamma}\}$ at a point $P^\star$, followed by that portion of Γ extending from the point of intersection to the origin. By the convexity of Γ, each tangent to this locus lies below the locus. Let $P \in \Gamma$ be a point to the left of $P^\star$, and let Q be the point of intersection of the tangent line at P with the horizontal segment $y = b$. It can be verified that the locus consisting of those points of the horizontal segment $y = b$ between (a, b) and Q, together with the points of the tangent line at P lying between Q and P and that portion of Γ lying between P and the origin, corresponds to a control $u^{\star\star}(t) = x^{\star\star\prime\prime}(t)$ for which the associated trajectory satisfies $x^{\star\star}(t_0') = 0$, $x^{\star\star\prime}(t_0') = 0$, with $t_0' < 0$ but arbitrarily near 0 as P approaches $P^\star$. Indeed $u^{\star\star}(t) = 0$ for all $t \in [t^{\star\prime}, 0]$, where $t^{\star\prime} - t^\star > 0$; $u^{\star\star}(t) = \rho e^{\theta(t-t^{\star\prime})}$ for all $t \in [t^{\star\prime}, t^{\star\star}]$ for appropriate $\rho > 0$, $\theta < 0$, and $t^{\star\star} > t^{\star\prime}$; $u^{\star\star}(t) = \text{ const } (t_0' - t)^{(\gamma-2)}$ for all $t \in [t^{\star\star}, t_0']$; and $u^{\star\star}(t) = 0$ for all $t \in [t_0', 0]$. Furthermore since $C[u^{\star\star}]$ achieves a positive contribution only for those t values associated with points of the tangent line between P and Q, it is clear that $C[u^{\star\star}]$ is smaller than any preassigned $\varepsilon > 0$, once P is chosen sufficiently near to $P^\star$. Therefore, by forming a control $u_\varepsilon \in L^\infty(-1, 0)$ bifurcating from $u^{\star\star}$ in the same manner as was done in Case 1 in relation to $u^\star$, we obtain $C[u^{\star\star}] < 2\varepsilon$. Hence the gap phenomenon does not occur.

3. Perturbations of f. In this section we point out how an additive perturbation of the type of running cost integrand appearing in §2 can lead to the Lavrentiev phenomenon, even for running cost integrands $f^\#$ satisfying the following "Tonelli-type" regularity and growth conditions:

$$f^\#_{uu}(x, y, u) > 0, \quad f^\#(x, y, u) \geq \varphi(u), \quad \text{with } \liminf_{|u|\to\infty} \varphi(u)/|u| = \infty.$$

COROLLARY. *Let $f = f(x, y, u)$ be an integrand satisfying the hypotheses of Theorem* A, *and let $x(-1) = a, x'(-1) = b$ denote boundary data satisfying the conditions of that theorem. Suppose that $e = e(x, y, u)$ is a nonnegative function satisfying $e_{uu} > 0$, $e(x, y, u) > \varphi(u)$, with φ as above. Let $u^\star$ denote the minimizer for C obtained in Theorem* A *for the given boundary data. Then if $e(x^\star(t), x^{\star\prime}(t), u^\star(t))$ is a function in $L^1(-1, 0)$, the Lavrentiev phenomenon $m^+_{p_0} > m^+_1$ also occurs for the cost C_ε associated with the integrand*

$$f_\varepsilon(x, y, u) = f(x, y, u) + \varepsilon e(x, y, u)$$

for $\varepsilon > 0$ taken sufficiently small.

This result is clear if ε is taken sufficiently small so that $C_\varepsilon[u^\star] < \delta(D, c_1, c_2)$ in the notation of Theorem A.

Example. Consider problem (P) with

$$C_\varepsilon[u] = \int_{-1}^{0} [(|x(t)|^{4/9} + x'(t))^{18}|u(t)|^k + \varepsilon|u(t)|^2]\,dt \quad \text{with } x'' = u$$

for data $x(0) = x'(0) = 0, x(-1) = a > 0, x'(-1) = b < 0$. By the corollary this can be treated as an additive perturbation for the cost functional

$$C[u] = \int_{-1}^{0} [(|x(t)|^{4/9} + x'(t))^{18}|u(t)|^k]\,dt.$$

The latter problem fits into Theorem A with $\gamma = 9/5$, $\alpha = 72/5$, $F = \{-1\}$, $k \geq 77$, and $p_0 = 5$, provided that a and b satisfy (#). The function $e(x^\star(t), x^{\star\prime}(t), u^\star(t)) \sim |t|^{-2/5}$ at $t = 0$ is integrable on $[-1, 0]$, so by the corollary for $\varepsilon > 0$ sufficiently small $C_\varepsilon[u]$ does exhibit the Lavrentiev phenomenon $m_5^+ = \inf_{A_{a,b}^{5,+}} > m_1^+ = \inf_{A_{a,b}^{1,+}}$.

Acknowledgments. We wish to thank an anonymous referee for several helpful suggestions improving the clarity of the presentation.

REFERENCES

[AAB] L. Ambrosio, O. Ascenzi, and G. Buttazzo, *Lipschitz regularity for minimizers of integral functionals with highly discontinuous integrands*, J. Math. Anal. Appl., 142 (1989), pp. 301–316.

[BK] J. M. Ball and G. Knowles, *A numerical method for detecting singular minimizers*, Numer. Math., 51 (1987), pp. 181–197.

[BM] J. M. Ball and V. J. Mizel, *One-dimensional variational problems whose minimizers do not satisfy the Euler-Lagrange equation*, Arch. Rational Mech. Anal., 90 (1985), pp. 325–388.

[BuM] G. Buttazzo and V. J. Mizel, *Interpretation of the Lavrentiev phenomenon by relaxation*, J. Funct. Anal., 110 (1992), pp. 434–460.

[Ce] L. Cesari, *Optimization Theory and Applications*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1983.

[CM] C.-W. Cheng and V. J. Mizel, *On the Lavrentiev phenomenon for autonomous second order integrands*, Arch. Rational Mech. Anal., 126 (1994), pp. 21–33.

[CMM] B. D. Coleman, M. Marcus, and V. J. Mizel, *On the thermodynamics of periodic phases*, Arch. Rational Mech. Anal., 117 (1992), pp. 321–347.

[CV1] F. H. Clarke and R. B. Vinter, *Regularity properties of solutions to the basic problem in the calculus of variations*, Trans. Am. Math. Soc., 292 (1985), pp. 73–98.

[CV2] ———, *A regularity theory for variational problems with higher derivatives*, Trans. Am. Math. Soc., 320 (1990), pp. 227–251.

[CWC] C.-W. Cheng, *The Lavrentiev Phenomenon and Its Applications in Nonlinear Elasticity*, Ph.D. dissertation, Department of Mathematics, Carnegie Mellon University, 1993.

[Da] A. M. Davie, *Singular minimizers in the calculus of variations in one dimension*, Arch. Rational Mech. Anal., 101 (1988), pp. 161–177.

[F] H. Federer, *Geometric Measure Theory*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1969.

[HM1] A. C. Heinricher and V. J. Mizel, *The Lavrentiev phenomenon for invariant variational problems*, Arch. Rational Mech. Anal., 102 (1988), pp. 57–93.

[HM2] ———, *A new example of the Lavrentiev phenomenon*, SIAM J. Control Optim., 26 (1988), pp. 1490–1503.

[HM3] ———, *A stochastic control problem with different value functions for singular and absolutely continuous control*, in Proc. 25th IEEE Conf. Decision and Control, Athens, Greece, 1986, pp. 134–139.

[J] V. Jurdjevic, *Geometric Control Theory*, Cambridge University Press, Cambridge, UK, 1996.

[K] I. A. K. Kupka, *The ubiquity of Fuller's phenomenon*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 313–350.

[La] M. Lavrentiev, *Sur quelques problèmes du calcul des variations*, Ann. Mat. Pura Appl., 41 (1926), pp. 107–124.

[Li] Z. Li, *Element removal method for singular minimizers in problems of hyperelasticity*, Math. Models Methods Appl. Sci., 5 (1995), pp. 387–399.

[L] P. D. Loewen, *On the Lavrentiev phenomenon*, Canad. Math. Bull., 30 (1987), pp. 102–108.

[Ma] M. Manià, *Sopra un esempio di Lavrentieff*, Boll. Un. Mat. Ital., 13 (1934), pp. 147–153.

[M1] V. J. Mizel, *The Lavrentiev phenomenon in both deterministic and stochastic optimization problems*, Proc. Internat. Workshop on Integral Functions in Calculus of Variations, Rend. Circ. Mat. Palermo (2), 15 (1987), pp. 111–130.

[M2] ———, *Developments in 1-dimensional calculus of variations affecting continuum mechanics*, in Proc. KIT Math. Workshop, Analysis and Geometry, Korea Inst. of Tech., 4 (1989), pp. 83–103.

[M3] ———, *Relevance of Lavrentiev's phenomenon in 1-dimension to continuum mechanics*, in Optimization and Nonlinear Analysis, A. Ioffe, M. Marcus, S. Reich, eds. Pitman Res. Notes in Math. 244, Longman, Essex, England, 1992, pp. 189–198.

[MM] M. Marcus and V. J. Mizel, *Absolute continuity on tracks and mappings of Sobolev spaces*, Arch. Rational Mech. Anal., 45 (1972), pp. 294–320.

[NM] P. V. Negron Marrero, *A numerical method for detecting singular minimizers of multidimensional problems in nonlinear elasticity*, Numer. Math., 58 (1990), pp. 135–144.